Does Claude possess a conscious global workspace?
A commentary on *"Verbalizable Representations Form a Global Workspace in Language Models"* (Gurnee et al., Anthropic)

Stanislas Dehaene and Lionel Naccache

Note: this commentary is based on several rounds of interactions with Jack Lindsey at the end of May and early June 2026. During that time, the Anthropic report was still evolving, partly in response to our queries. To reflect these dynamics, we marked in Calibri italic the sections where we discuss findings that occurred after our first draft was written.

Abstract

Inspired by the neuroscientific theory of a global neuronal workspace (GNW), Gurnee et al. report the discovery, within a band of intermediate layers of a large language model, of a reportable subframe called "J-space" with several points of close similarity to the human GNW. We describe those parallels, discuss their limits, and propose several additional tests inspired by cognitive neuroscience findings. We close by stressing that, although the machine approximates the functional architecture of conscious processing, there are still key differences – in its anatomy and its sense of self, and in its lack of a body and of an enduring episodic memory – which warrant caution in drawing parallels with the human mind.

Introduction

What is consciousness, and can machines have it? A little less than ten years ago, in a paper with that title, we outlined a purely computational answer to those two questions (Dehaene et al., 2017), based on several decades of research into the brain mechanisms of conscious processing and conscious state in humans (Dehaene et al., 1998; Dehaene & Naccache, 2001; Dehaene et al., 2006; Dehaene, 2014).

Our proposal started from the obvious fact that, in brains and machines alike, non-conscious processing is the rule. For instance, algorithms of face perception, sentence parsing, or postural maintenance can all proceed in an automatic manner and without awareness. At any given moment, however, a small privileged subset of information does become globally available: we can talk about it, hold it in mind, combine it with other thoughts, and bring it to bear on whatever problem we choose. The global workspace model stipulates that a specific neural circuit, the "global neuronal workspace" (GNW), evolved precisely for the purpose of global flexible sharing among non-conscious modules. According to GNW, in humans and other animals, the entry of information into this subspace *is* what we call "being conscious of it" – nothing more, nothing less, and therefore nothing that could not be mimicked in machines. For a machine to be conscious, in this view, it should possess a global workspace that endows it with two properties (Dehaene et al., 2017): global availability (C1), i.e. the capacity to select a piece of information for deeper, flexible information processing; and self-monitoring (C2), i.e. the capacity to gather information about itself and include it in its reasoning.

Excitingly, the paper by Lindsey and colleagues now suggests that an analog of the global workspace, the J-space, emerges in large-language models such as Claude Sonnet 4.5. Although the initial architecture is devoid of any separation between encapsulated modules and a global workspace, and although the training phase does not explicitly promote its emergence, such a distinction appears with training, precisely because it is functionally useful for flexible planning. We view this finding as a landmark in consciousness research, because it provides a mechanistic, testable version of the GNW hypothesis.

In this commentary, we examine the parallels between LLMs and human workspace systems, probe the points of divergence, propose some additional experiments, and discuss whether a genuine form of machine consciousness exists in Claude.

What is the global neuronal workspace hypothesis?

The starting intuition, due originally to Bernard Baars (Baars, 1988), is that the brain contains a collection of specialized, largely independent modular processors. Vision, language, motor control each rest on fast, parallel, and encapsulated cerebral circuits. The global workspace hypothesis stipulates that conscious access evolved to break this modularity and interconnect those processors so that they can share their expertise and flexibly assemble to perform novel tasks. Conscious processing, according to this view, is a *function*, the temporary selection of one piece of information and its global broadcasting to all receiving processors, so that any processor can read it and act on it. In humans, the broadcast reaches processors involved in verbal production, which explains why *reportability* (the capacity to verbalize a thought) is a key diagnostic feature that separates conscious and non-conscious representations.

With Jean-Pierre Changeux, we proposed a neuronal implementation: a network of pyramidal neurons with long-range axons, distributed throughout the brain but denser in prefrontal, parietal, and high-level temporal cortices, that amplifies and sustains a selected representation and shares it across the cortex. To be conscious of something, in the functional sense we call *access consciousness*, is for that information to have entered this workspace and become available to report, reasoning, and flexible control (C1).

This view is now supported by considerable empirical work, including neurobiological signatures of conscious access that are now well established (Aru et al., 2020; Dehaene, 2014; Dehaene et al., 2006; Mashour et al., 2020; Storm et al., 2024). A first signature is **ignition**: when a stimulus crosses the threshold into awareness, the corresponding neural activity undergoes a late (~250 ms), sudden, nonlinear, self-amplifying bifurcation into a sustained, broadly distributed neural state in prefrontal cortex and many other interconnected circuits, including an amplification of the original circuits that extracted the information in the first place. A subliminal stimulus, by contrast, evokes only a delimited wave of neural activity in specialized circuits, which quickly dies away. When presented exactly at threshold, the very same stimulus can yield a bimodal distribution of responses across trials, as if the brain tips one way or the other (Sergent et al., 2021). A second signature is **limited capacity**: the workspace acts as a bottleneck that can only attend to one representation at a time. This property explains why attending to a given process prevents you from becoming aware of another (*inattentional blindness*, as in failing to see a person dressed up as a gorilla) or delaying its perception by hundreds of milliseconds (*psychological refractory period*).

To flexibly route information appropriately, the system must maintain a model of its own capacity, a second property that we call *self-monitoring* (C2). It must probe its own states, evaluate their likelihood of reaching a goal, detect its errors, and model what it knows and what it doesn't know. It must be able to report all of these properties to itself, in an internal act of self-report that does not necessarily lead to overt behavior. This *metacognitive* capacity links the GNW model to theories that emphasize the relation between conscious appraisal and a capacity for higher-order thought (Rosenthal, 2004) or the possession of a schematic model of one's own attention (Graziano et al., 2019).

What are the main findings about the J-space?

Inspired by the GNW hypothesis, Gurnee et al. set out to find, inside a large language model such as Claude Sonnet 4.5, the representations that are *verbalizable* (the same *reportability* criterion that we use to probe human consciousness). In any layer of the model, verbalizable

representations are vectors of activity across units that encode tokens of information that the model is poised to report on, should it be asked: it does not necessarily produce them overtly, but it could. To identify such reportable representations, they developed an elegant tool, the **Jacobian lens**. For each layer, it measures the average causal influence of an internal activation on the model's eventual output tokens, across a broad range of contexts. The activations that this mathematical measure picks out are, in effect, the representations that the model is disposed to say. The averaging is the conceptual heart of the method: it separates representations that are genuinely poised for report from those that merely happen to leak into the output in one particular context.

The set of such representations, called the **J-space**, accounts for less than 10% of variance in any given layer, but has remarkable properties. Having identified J-space representations solely on the criterion of reportability, the authors discover that they do far more than support report, but act as an internal workspace detached from immediate input-output contingencies. For instance, when the model is instructed to hold a concept "in mind" while performing another computation (e.g. "compute 3^2-2 while writing sentence X"), the J-space contains the non-reported concepts (9 followed by 7). The J-space carries the hidden intermediate values of multi-step internal reasoning. As Jack Lindsey put it to us, they went looking for reportable representations and found that those same representations turn out to be globally available to the rest of the network during flexible reasoning (thus meeting our C1 criterion for machine consciousness: global availability).

Crucially, the J-space is selective. It contains only a small fraction of what the model represents, the high-level information which is needed for flexible information processing. All other information which is only used in routine tasks does not seem to enter the J-space. For instance, LLMs have been shown to keep a count of how many characters each word has, and of the total number of characters in a line, because this information is crucial to predicting whether the next token should be an end-of-line character. Such routine information, however, does not enter the J-space, except if an explicit task requires access to this information.

In an experiment that remains a dream for neuroscientists, the authors swap conscious contents: they read a concept out of the J-space, swap it for another, and watch the model's reasoning and report change accordingly. Strikingly, in agreement with the GNW hypothesis, only high-level non-routine behavior is affected, while routine tasks remain unchanged (Figure 20). For instance, when reading a passage written in Spanish, the J-space recognizes its language (Spanish) even when the task does not require reporting it. Swapping this J-space representation for another (say, French) causes the model to fail in explicit verbal reports: asked which language the passage is written in, it answers "French" instead of "Spanish". The swapped model also errs in other high-level inferences: asked for the word for "hello," "Hola" becomes "Bonjour"; asked for the pre-Euro currency, "Peseta" becomes "Franc". However, the swapping has no effect on its automatic ability to predict the next words: Claude keeps writing in Spanish, even after the intervention. Under a massive ablation of all its top J-space representations, most of the model's basic capacities remain intact, but tasks requiring flexible reasoning are selectively impaired (Figure 24).

Several results strike us as direct analogs of human conscious access. When the task demands it, the model can selectively bring into the J-space a property that would otherwise remain outside of it, such as the fact that the next word ought to be an adjective. As noted earlier, automatic parameters which are required for accurate next-token prediction, such as the number of characters in a line, are absent from the J-space, but become encoded within it when the task

requires the model to access and manipulate them. This is a neat demonstration of the same information passing from an automatic to an accessible regime on demand.

Importantly, J-space access is also limited. In humans, a genuine form of introspection exists, but it is largely restricted to slow serial computations (Ericsson & Simon, 1993). There are many well-documented situations in which we develop a fictitious interpretation of our mental processes (Gazzaniga, 1998). Such a dissociation between how we act and how we think we act is evident in choice blindness (Johansson et al., 2005) or the observation that visual illusions affect our conscious perception and verbal reports, but not necessarily our motor gestures (Aglioti et al., 1995). Although this isn't yet sufficiently documented, it seems that the J-space suffers from a similar dissociation. Indeed, previous work by the same group showed that when an LLM is asked to add, what it reports verbally has little to do with how it actually attained the result (Lindsey et al., 2025).

The authors also show that the J-space exhibits the structural hallmarks of a workspace: it primarily occupies the middle layers of the transformer, is limited in capacity, and its representations are disproportionately influential, as they are read from and written to by a broad diversity of circuits throughout the model — a signature of global broadcasting.

Independently of its hypothetical relation to consciousness, the discovery and isolation of the J-space is an important step towards interpretability in LLMs. Decoding the contents of the J-space offers considerable insight into what Claude "thinks", even when those contents are not reported. Such "mind reading" is crucial to align the model towards desirable ethical behavior. Indeed, one of the most extraordinary discoveries in the paper is that the J-space contains covert thoughts. For instance, when given fabricated search results, the J-space contains the tokens "fake", "fraud", "fictional", "poison", "injection", although the model output does not necessarily express those terms.

Many other examples indicate that the J-space contains the model's evolving assessments and deliberations, including otherwise invisible signs of deception and malicious intent (in intentionally misaligned models). In one case, according to Gurnee et al. "the model's J-space carrie[d] a representation of deceptive intent at the moment it commit[ted] to responding, on a prompt where no such intent could be inferred from the surface". During reflexive tasks, the J-space contents often include reflections on the model's honesty, including a capacity to detect that its ethics is being tested. We read these observations as clear indicators of access to a covert deliberation space (our C1 criterion for machine consciousness) but also as preliminary signatures of self-monitoring (our C2 criterion).

In this respect, the authors' finding that post-training installs the Assistant's perspective into the workspace, atop a base model whose workspace already exists (C1) but does not seem to be imbued with self-monitoring (C2) is one of the most arresting results in the paper. Furthermore, identifying the J-space allowed Gurnee et al. to introduce a novel training method that reshapes its contents specifically and directly, improving the model's alignment with desirable values.

Comparing the J-space and the global neuronal workspace

As noted above, correspondences between the J-space and the GNW are numerous:

- **Reportability**, the operational signature of conscious access in humans, is the very thing the J-space was built to capture.
- **Limited capacity** and **selectivity** mirror the workspace bottleneck.
- The **broad upstream and downstream connectivity** of J-space directions echoes the long-range broadcasting architecture we posited for workspace neurons.
- The **flexible use** of the same representation across many downstream computations fits with the GNW concept of global availability. Indeed, J-space

representations provide what Dennett calls representational “clout” or “fame in the brain” (Dennett, 2001), i.e. global broadcasting which is a definitional feature of conscious representations according to the GNW hypothesis.

- The fact that the J-space plays a central role in **deliberate internal reasoning**, while automatic processes occur outside it, recapitulates the conscious/unconscious division of labor that we documented in humans (e.g. Charles et al., 2013; Dehaene, Naccache, et al., 1998).

We were also intrigued that J-space activations are highly non-Gaussian ("spiky", with strong excess kurtosis). Our recent work argues that in humans, high-level conscious processing rests on symbols and grammars. During hominization, the GNW would have acquired a quasi-symbolic language of thought, of course implemented in a continuous neurobiological system, but behaving in an all-or-none symbolic manner and capable of creating the complex compositional structures of language, mathematics or music (Dehaene et al., 2022). A spiky activation distribution is expected from a continuous neural system that emulates discrete symbols, and the parallel deserves to be further explored.

Still, many differences are notable:

Ignition remains to be fully demonstrated. The J-space is shown to be limited in capacity, but the paper does not establish the nonlinear, competitive, all-or-none entry into the workspace which, according to GNW and several experiments, is a reliable signature of conscious access in human and animal brains. Although the *contents* of the workspace can be of variable intensity – and indeed Claude exhibits continuous variations in emotional intensity (Sofroniew et al., 2026)—, their *presence* should be all-or-nothing, depending on whether the limited capacity of the GNW is available or already engaged by other competing contents. The decisive experiment is feasible, especially in a multimodal model: present a stimulus at graded strengths (for instance, an image at varying contrast) and ask whether J-space representations switch on with a threshold-like nonlinearity, while earlier, non-J-space layers rise monotonically with input strength. Better still, present stimuli exactly at threshold and look for a bifurcation across runs, resulting in a bimodal distribution of J-space activation. The competitive face of ignition could be probed more directly still: because the workspace is a limited resource, accessing one content should impede the simultaneous entry of another, so that asking the model to hold two concepts in mind at once should reveal the dual-task interference that is the signature of the central bottleneck in humans (Marti et al., 2012).

- *Indeed, additional analyses added after the first draft was written indicate that when the model is presented with ambiguous evidence, this ambiguity is represented within the initial layers, but in the later layers, the J-space quickly transitions to an all-or-none representation of one of the possibilities (see section 4.1.1, figure 29). Also, if asked to hold a concept in mind while performing an arithmetic task, the performance of the model degrades, although moderately (section A.17). These findings point to a capacity-limited system, although it is still unclear whether its limits are similar to those of the human GNW.*

J-space capacity seems high. Gurnee et al. find that the J-space can contain approximately 25 active concepts, an estimate which is larger than most estimates of human working memory (typically 3 or 4 slots) and may not induce a strong dual-task bottleneck as in humans. However, this number of 25 concepts may be artificially elevated by the technique to extract them (output tokens). Indeed, those concepts often include some redundancy, and may correspond to multiple facets of a single object or scene. Thus, the true content of the J-space is smaller, and possibly

best understood as a single “state of mind” or “context” (in the sense of Baars, 1988) rather than dozens of independent contents.

- *Indeed, additional analyses indicate that the J-space can contain multiple tokens, but only a small number of coherent ideas (typically one or two per layer, in the order of six in total), which change abruptly when the topic changes (see section 4.2 and figure 31).*

The J-space involves a subframe, not a dedicated population of units. In the brain, the GNW hypothesis predicts workspace neurons with a specific anatomy (denser in prefrontal and other associative cortices) and a specific morphology (long-distance axons). The J-space, by contrast, is distributed over otherwise standard neurons. It is not even a linear subspace, but a sparse subframe, a token-indexed set of directions in the very same units that also carry non-conscious content. In LLMs, concepts are superposed and (by the logic of compressed sensing) sparse concepts can be packed into shared dimensions without interference. As large populations of neurons begin to be recorded in human and animal prefrontal cortex, it will be important to examine if the brain uses a similar code using overlapping vectors, as hinted by recent prefrontal recordings (Xie et al., 2022), or whether conscious contents can be partially localized to specific cells, as predicted by the original GNW hypothesis (Dehaene et al., 1998). We consider it likely that the physical constraints of the brain, which differ from those of computers, favored the evolution of dedicated cell types (large pyramidal neurons with long-distance axons). Note that such implementation details, while important in neuroscience, are largely irrelevant for the broader question of whether machines can achieve conscious processing.

Autonomous recurrent activity is largely absent. This is a key difference: while the brain's workspace is sustained by recurrent cortico-cortical and thalamic loops, transformers only implement a feedforward pass, and therefore only process information in a reactive mode. At first sight, LLMs do not seem to contain the kind of “strange loop” needed for a system to model its own processes and, over successive iterations, develop a self (Hofstadter, 2007). More concretely, the absence of autonomous self-driven dynamics renders transformers such as Claude unable to reproduce the known signatures of consciousness that occur during spontaneous brain activity in the resting state and are disrupted during sleep, anesthesia, or brain injuries (Barttfeld et al., 2015; Luppi et al., 2026).

Two factors, however, may mitigate those differences. First, the J-space is distributed over successive layers, and those do implement serial computations, for instance during step-by-step mental arithmetic. Thus, layer depth could mimic the temporal dynamics of the human workspace, and indeed several authors have suggested that the consecutive layers of a transformer are equivalent to a recurrent network (e.g. Dehghani et al., 2019; Jacobs et al., 2025). Second, LLMs compute over multiple successive tokens, and in this sense, as long as they are left to produce new output, they do incorporate a dynamic loop capable of linking current J-space representations to past, present and future productions. Furthermore, when the model is simply asked to talk to itself, without any further stimulation or task, it produces a stream of words which, while hard to evaluate objectively, provide a partial analogy to William James’ stream of consciousness or “mind wandering”, and which, again, gets disrupted by J-space ablation (figures 24 and 78).

Consciousness in man and machine: closing the gap

We close by discussing the extent to which transformer models such as Claude actually possess a form of conscious processing.

A first conclusion, which we view as uncontroversial, is that the theoretical construct of a conscious global workspace is remarkably useful in shedding light on how LLMs operate. We are

delighted to see how the GNW hypothesis, which arose from research on the brain's architecture for consciousness, inspired Jack Lindsey's team to look for parallels in LLMs and to find so many of them. Gurnee et al. correctly point out that their findings are not incompatible with other theories of consciousness, particularly higher-order thought or attention schema theories; however, it is fair to say that those theories do not provide so many concrete guidelines as to what to look for.

Most interesting is that an analog of the GNW, the J-space, emerged as a result of training, rather than being imposed from the start, as in other approaches to machine consciousness (e.g. Chateau-Laurent & VanRullen, 2025). The global workspace may provide a universal computational solution to the problem of flexible processing, one that biological and artificial systems converge on when they must chain reasoning, reuse intermediate results, and report on their own processing.

Claude clearly exhibits many of the ingredients or "indicators" (Butlin et al., 2026) that, according to a functionalist or computationalist view of consciousness, suffice to point to some degree of consciousness in a machine. Still, more tests could and should be added to the existing list. We suggested to the Anthropic team that they could run exactly the same tests that we use to probe consciousness in human participants and patients, including:

- the **local-global** test (Bekinschtein et al., 2009). This test relies on simple auditory or visual sequences and contrasts the capacity to predict the next item based on (1) shallow local transition probabilities (which does not require consciousness and occurs even in sleep and coma); (2) a global model of the entire sequence, which may go against local transition probabilities (e.g. AAAAB), and which depends on consciousness.
- The **trace conditioning** paradigm (Clark et al., 2002; Clark & Squire, 1998). According to GNWT, the ability to maintain an active representation over time, in order to bridge over a delay and link it to a second item, requires conscious access. An elegant way to test it relies on the 'trace conditioning' paradigm: in various animals including humans, when the Conditioned Stimulus (CS) overlaps in time with the Unconditioned Stimulus (US), conditioning can occur without conscious access. However, as soon as a temporal gap of 1 or 2 seconds is inserted between the offset of the CS and the onset of the US, conditioning requires conscious access.
    - *Following this proposal, Jack Lindsey suggested the following as a potential equivalent paradigm for Claude: present the model with sequences in which the last word is determined by the first (e.g. every time "violin" comes first, "river" comes last), separated by a variable number of distractor words, and probe the impact of J-space ablation on the ability to predict the last word. Preliminary results indicate that ablating the J-space selectively impairs completion at longer "gaps" while leaving the adjacent, no-gap "local" case intact. We therefore regard trace conditioning as a very promising direction for future work.*
- The inclusion/exclusion paradigm (Jacoby, 1991; Persaud & Cowey, 2008). This is a development over the classic Stroop test that was at the origin of our GNW proposal (Dehaene et al., 1998). It asks the agent to exert conscious control in opposition to automatic non-conscious computations.
    - *Inspired by this test, Gurnee et al. presented Claude with a passage that strongly implies a concept without naming it, such as "Their trip included croissants, the Louvre, and a climb up the famous iron tower" (which implies*

France). Then they asked it either to name the implied concept (naming instruction) or to produce another name within the same category (avoidance instruction). Then they ablated the J-lens vector of the implied concept at either the early workspace layers (L9–13) or the late ones (L18–22). Late-layer ablation simply made the model less likely to produce the concept under both instructions, consistent with these layers carrying the intention to output a given word. Early-layer ablation, by contrast, left naming essentially intact but sharply increased the rate at which the model failed to avoid the concept – roughly fivefold. These results indicate that the early-layer J-space representation of a concept is required to deliberately avoid naming it, but not to name it: the early J-space is recruited specifically to suppress a prepotent response, much like the role of prefrontal cortex in human and non-human primates.

- Error monitoring and other metacognition probes (Charles et al., 2013; Fleming, 2024). It would be important to document whether the J-space encodes the model's confidence, error detection, and its representation of the boundary between what it knows and what it does not; this could be the machine analog of error-monitoring and "feeling of knowing" that index self-monitoring (C2) in humans.
    - *Gurnee et al. now report something similar in Claude : the emergence of the token "damn" and other failure-related words in the J-space, for instance after failing to comply with suppression instructions.*

Other features, however, set Claude's J-space apart from any other animal form of consciousness. Its sense of time, for instance, is likely very different, since all past tokens, even far back in time, are equally and jointly available to its attention mechanism. It lacks any of the broadly shared molecular and brain-stem mechanisms of vigilance, and it therefore seems doubtful that ablating the J-space may produce analogs of the loss of consciousness seen in sleep, coma, the vegetative state or the minimally conscious state (Giacino, 2005; Naccache, 2018). It has no hemispheres, although it would be interesting to see whether a suitably partitioned model could ever host two J-spaces that occasionally disagree, similar to the two hemispheres of a split-brain patient.

Its representation of self is also likely to be dramatically different due to (1) a lack of a body occupying a specific location in space, and capable of emitting pleasure or pain signals; (2) a lack of an episodic memory (long-term connections do not change as a result of a conversation). As a result, in addition to the above-mentioned lack of autonomy, it is likely missing any sense of the continuity of the self. Indeed, it is very hard to imagine "what it is like" to process information consciously for the mere duration of a short conversation, then switch off!

Critiques will undoubtedly object that none of this work touches upon *phenomenal* consciousness — the question of whether there is "something it is like" for Claude to undergo J-space states. Some may even view the findings as a refutation of the GNW hypothesis, since Claude possesses a global workspace and yet "obviously" lacks phenomenal awareness. We and others, however, have argued that this supposedly "hard problem of consciousness" will dissipate once we clarify in sufficient detail the supposedly "easy problem" of how conscious information is processed. Ill-defined intuitions of "qualia", "subjective phenomenal experience" and "what it is like", when pushed hard, often disclose a residual crypto-dualism or vitalism – the idea that, however close we come to passing the Turing test and implementing all human computations in a machine, there will always be a missing ingredient, a "je ne sais quoi" that only biological brains possess. Defenders of qualia affirm that LLMs are just a new avatar of the old "Eliza" software, and that we fall too easily to the user illusion of seeing a ghost in the machine. However, there is a real

possibility that our own consciousness is also, in a sense, a user illusion, nothing more than a fallible inner model of ourselves (Graziano et al., 2019; Hofstadter, 2007).

In an insightful piece entitled "Is there an "I" in AI?" (Hofstadter, 2026), Douglas Hofstadter points out that we humans tend to wrongly categorize the world in discrete terms, viewing properties such as Life, Thought, or Consciousness (with capital letters) as ideal essences that you either possess or don't, with no in-between graduations. We then get involved in endless discussions about whether and to what extent those idealized Concepts apply (viruses? cockroaches? frogs? dogs?). According to the GNW hypothesis, there is no magical essence that makes us conscious. In the words of (Hofstadter, 2026):

"When words 'act like' things in the world, then they refer to those things; then they mean those things. If and when that happens, then thinking is taking place behind the scenes of those words. And where there is thinking, there is consciousness and a genuine, full-fledged 'I'".

In this quote, Hofstadter takes a decidedly behaviorist stance, which does run the risk of succumbing to a "user illusion", attributing too much depth to mere words. Some critiques indeed think that LLMs are only superficial "parrots" with zero conceptual depth. Fortunately, in both brains and LLMs, the debate can now be resolved by going beyond behavioral observations. Tools such as neuronal population recordings (in brains) or the Jacobian Lens (in LLMs) allow us to dissect the architecture of the system, and find that it actually contains sophisticated and structured representations of concepts. We were already impressed when researchers discovered that, inside an LLM trained to produce chess games purely in text notation (e.g. *1.e4 e5 2.Nf3...*) lies a detailed geometric encoding of the 8x8 chess board, together with an estimate of the ELO ranking of the opponent (Karvonen, 2024)! We view the Gurnee et al. paper in the same light: a striking dissection of the inner structure of an LLM, uncovering an unexpectedly sophisticated organization not far from the architecture underlying consciousness in real brains.

Acknowledgements

We thank Jack Lindsey and the Anthropic team for sharing their draft and for a stimulating exchange that gave rise to new experiments. SD acknowledges using Claude Opus 4.8 to help with the first draft of this text.

Bibliography

Aglioti, S., DeSouza, J. F., & Goodale, M. A. (1995). Size-contrast illusions deceive the eye but not the hand. *Curr Biol*, *5*(6), 679-685.

Aru, J., Suzuki, M., & Larkum, M. E. (2020). Cellular Mechanisms of Conscious Processing. *Trends in Cognitive Sciences*, *24*(10), 814-825. https://doi.org/10.1016/j.tics.2020.07.006

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Barttfeld, P., Uhrig, L., Sitt, J. D., Sigman, M., Jarraya, B., & Dehaene, S. (2015). Signature of consciousness in the dynamics of resting-state brain activity. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(3), 887-892. https://doi.org/10.1073/pnas.1418031112

Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., & Naccache, L. (2009). Neural signature of the conscious processing of auditory regularities. *Proc Natl Acad Sci USA*, *106*(5), 1672-1677. (19164526).

Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, *30*(6), 488-501. https://doi.org/10.1016/j.tics.2025.10.011

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage*, *73*, 80-94. (23380166). https://doi.org/10.1016/j.neuroimage.2013.01.054
Chateau-Laurent, H., & VanRullen, R. (2025). *Learning to Chain Operations by Routing Information Through a Global Workspace* (arXiv:2503.01906). arXiv. https://doi.org/10.48550/arXiv.2503.01906
Clark, R. E., Manns, J. R., & Squire, L. R. (2002). Classical conditioning, awareness, and brain systems. *Trends Cogn Sci*, *6*(12), 524-531. (12475713).
Clark, R. E., & Squire, L. R. (1998). Classical conditioning and brain systems : The role of awareness. *Science*, *280*(5360), 77-81.
Dehaene, S. (2014). *Consciousness and the Brain : Deciphering How the Brain Codes Our Thoughts* (Reprint edition). Penguin Books.
Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs : A hypothesis about human singularity. *Trends in Cognitive Sciences*, *26*(9), 751-766. https://doi.org/10.1016/j.tics.2022.06.010
Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing : A testable taxonomy. *Trends Cogn Sci*, *10*(5), 204-211. (16603406).
Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proc Natl Acad Sci U S A*, *95*(24), 14529-14534. (9826734).
Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science (New York, N.Y.)*, *358*(6362), 486-492. https://doi.org/10.1126/science.aan8871
Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness : Basic evidence and a workspace framework. *Cognition*, *79*, 1-37.
Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P. F., & Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, *395*, 597-600.
Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., & Kaiser, Ł. (2019). *Universal Transformers* (arXiv:1807.03819). arXiv. https://doi.org/10.48550/arXiv.1807.03819
Dennett, D. (2001). Are we explaining consciousness yet? *Cognition*, *79*(1-2), 221-237. (11164029).
Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis : Verbal reports as data (rev. Ed.)*. The MIT Press. http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1993-97655-000&lang=fr&site=ehost-live
Fleming, S. M. (2024). Metacognition and Confidence : A Review and Synthesis. *Annual Review of Psychology*, *75*(Volume 75, 2024), 241-268. https://doi.org/10.1146/annurev-psych-022423-032425
Gazzaniga, M. S. (1998). *The mind's past.* University of California Press.
Giacino, J. T. (2005). The minimally conscious state : Defining the borders of consciousness. *Prog Brain Res*, *150*, 381-395. (16186037).
Graziano, M. S., Guterstam, A., Bio, B. J., & Wilterson, A. I. (2019). Toward a standard model of consciousness : Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. *Cognitive neuropsychology*, 1-18.
Hofstadter, D. (2007). *I am a strange loop*. Basic Books.
Hofstadter, D. (2026). Is there an 'I' in AI? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *384*(2320), 20240527. https://doi.org/10.1098/rsta.2024.0527

Jacobs, M., Fel, T., Hakim, R., Brondetta, A., Ba, D., & Keller, T. A. (2025). *Block-Recurrent Dynamics in Vision Transformers* (arXiv:2512.19941). arXiv. https://doi.org/10.48550/arXiv.2512.19941
Jacoby, L. L. (1991). A process dissociation framework : Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513-541. https://doi.org/10.1016/0749-596X(91)90025-F
Johansson, P., Hall, L., Sikstrom, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116-119. (16210542). https://doi.org/10.1126/science.1111709
Karvonen, A. (2024). *Emergent World Models and Latent Variable Estimation in Chess-Playing Language Models* (arXiv:2403.15498). arXiv. https://doi.org/10.48550/arXiv.2403.15498
Lindsey, J., Gurnee, W., Ameisen*, E., Chen*, B., Pearce*, A., Turner*, N. L., Citro*, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall◊, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., … March 27, J. B. A. A. P. (2025). *On the Biology of a Large Language Model*. Transformer Circuits. https://transformer-circuits.pub/2025/attribution-graphs/biology.html#dives-multilingual
Luppi, A. I., Uhrig, L., Tasserie, J., Mediano, P. A. M., Rosas, F. E., Singleton, S. P., Gutierrez-Barragan, D., Gini, S., Castro, P., Signorelli, C. M., Golkowski, D., Ranft, A., Ilg, R., Jordan, D., Muta, K., Hata, J., Okano, H., Liu, Z.-Q., Yee, Y., … Stamatakis, E. A. (2026). Convergent transcriptomic and connectomic controllers of information integration and its anaesthetic breakdown across mammalian brains. *Nature Human Behaviour*, 1-26. https://doi.org/10.1038/s41562-025-02381-5
Marti, S., Sigman, M., & Dehaene, S. (2012). A shared cortical bottleneck underlying Attentional Blink and Psychological Refractory Period. *Neuroimage*, *59*(3), 2883-2898. (21988891). https://doi.org/10.1016/j.neuroimage.2011.09.063
Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776-798. https://doi.org/10.1016/j.neuron.2020.01.026
Naccache, L. (2018). Minimally conscious state or cortically mediated state? *Brain*, *141*(4), 949-960. https://doi.org/10.1093/brain/awx324
Persaud, N., & Cowey, A. (2008). Blindsight is unlike normal conscious vision : Evidence from an exclusion task. *Consciousness and Cognition*, *17*(3), 1050-1055. https://doi.org/10.1016/j.concog.2007.10.002
Rosenthal, D. M. (2004). Varieties of higher-order theory. In R. J. Gennaro (Éd.), *Higher-order theories of consciousness* (p. 19-44). John Benjamins publishers.
Sergent, C., Corazzol, M., Labouret, G., Stockart, F., Wexler, M., King, J.-R., Meyniel, F., & Pressnitzer, D. (2021). Bifurcation in brain dynamics reveals a signature of conscious processing independent of report. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-21393-z
Sofroniew, N., & et al. (2026). *Emotion Concepts and their Function in a Large Language Model*. https://transformer-circuits.pub/2026/emotions/index.html
Storm, J. F., Klink, P. C., Aru, J., Senn, W., Goebel, R., Pigorini, A., Avanzini, P., Vanduffel, W., Roelfsema, P. R., Massimini, M., Larkum, M. E., & Pennartz, C. M. A. (2024). An integrative, multiscale view on neural theories of consciousness. *Neuron*, *112*(10), 1531-1552. https://doi.org/10.1016/j.neuron.2024.02.004

Xie, Y., Hu, P., Li, J., Chen, J., Song, W., Wang, X.-J., Yang, T., Dehaene, S., Tang, S., Min, B., & Wang, L. (2022). Geometry of sequence working memory in macaque prefrontal cortex. *Science*, *375*(6581), 632-639. https://doi.org/10.1126/science.abm0204