Trends in Neurosciences

CellPress

Opinion

The feasibility of artificial consciousness through the lens of neuroscience

Jaan Aru,[1,*] Matthew E. Larkum,[2,*] and James M. Shine[3,*]

Interactions with large language models (LLMs) have led to the suggestion that these models may soon be conscious. From the perspective of neuroscience, this position is difficult to defend. For one, the inputs to LLMs lack the embodied, embedded information content characteristic of our sensory contact with the world around us. Secondly, the architectures of present-day artificial intelligence algorithms are missing key features of the thalamocortical system that have been linked to conscious awareness in mammals. Finally, the evolutionary and developmental trajectories that led to the emergence of living conscious organisms arguably have no parallels in artificial systems as envisioned today. The existence of living organisms depends on their actions and their survival is intricately linked to multi-level cellular, inter-cellular, and organismal processes culminating in agency and consciousness.

Highlights

Large language models (LLMs) can produce text that leaves the impression that one may be interacting with a conscious agent.

Present-day LLMs are text-centric, whereas the phenomenological umwelt of living organisms is multifaceted and integrated.

Many theories of the neural basis of consciousness assign a central role to thalamocortical re-entrant processing. Currently, such processes are not implemented in LLMs.

The organizational complexity of living systems has no parallel in present-day AI tools. Possibly, AI systems would have to capture this biological complexity to be considered conscious.

LLMs and the current debates on conscious machines provide an opportunity to re-examine some core ideas of the science of consciousness.

Large language models and consciousness

There is a long tradition of questioning which animals are conscious [1–3] and whether entities outside the animal kingdom might be conscious [4–6]. Recently, the advent of LLMs has brought a novel set of perspectives to this question. Through their competence and ability to converse with us, which in humans is indicative of being conscious, LLMs prompt us to refine current notions of what it means to understand, to have agency, and to be conscious.

LLMs are sophisticated, multi-layer artificial neural networks with billions of connections whose weights are trained on hundreds of billions of words from various texts, including natural language conversations between humans. Through text-based queries, users interacting with LLMs are provided with a fascinating language-based simulation. If you take the time to use these systems, it is hard not to be struck by the apparent depth and quality of the internal machinations in the network. Ask it a question and it will provide you with an answer that drips with the kinds of nuance we typically associate with conscious thought. As a discerning, conscious agent yourself, it is tempting to conclude that the response has been generated by a similarly conscious being, one that thinks, feels, reasons, and experiences. Using this type of a 'Turing test' as a benchmark, the question can be raised whether LLMs are or soon will be conscious [7–10], which in turn raises a host of moral quandaries, such as whether it is ethical to continue to develop LLMs that could be on the brink of conscious awareness. While this position might not be prevalent among neuroscience researchers today, the improving capabilities of artificial intelligence (AI) systems will inevitably lead to the point where the possibility of machine consciousness needs to be addressed. Furthermore, this possibility is discussed extensively in news media, prompting neuroscientists to consider some of the arguments in favor and against it.

The notion of LLMs' potential to be conscious is often bolstered by the fact that the architecture of LLMs is loosely inspired by features of brains (Figure 1), the only objects to which we can currently attribute consciousness with confidence. However, while early generations of artificial neural

[1]Institute of Computer Science, University of Tartu, Tartu, Estonia
[2]Institute of Biology, Humboldt University Berlin, Berlin, Germany
[3]Brain and Mind Center, The University of Sydney, Sydney, Australia

*Correspondence:
jaan.aru@ut.ee (J. Aru),
larkumma@hu-berlin.de (M.E. Larkum),
and mac.shine@sydney.edu.au
(J.M. Shine).

© 2023 Elsevier Ltd. All rights reserved.

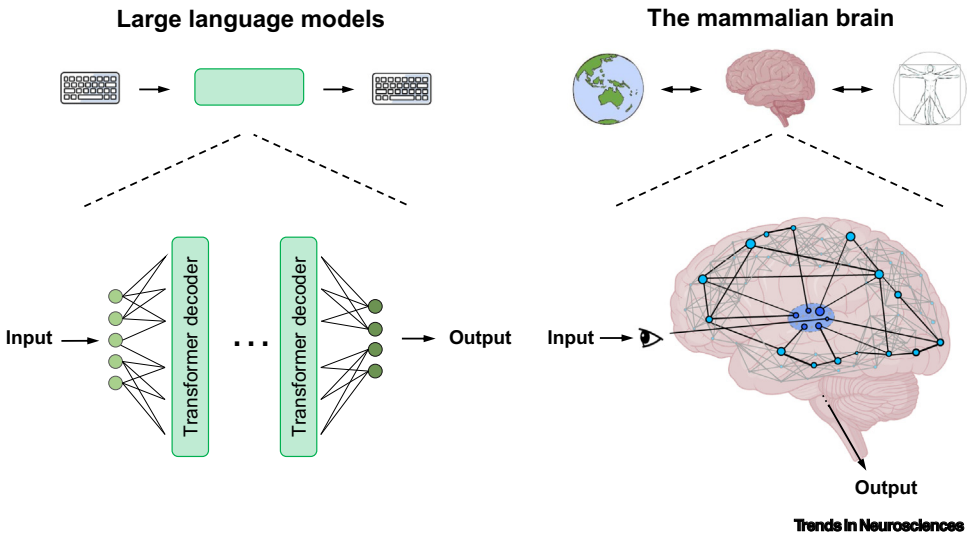

Figure 1. Macroscopic topological differences between mammalian brains and large language models. Left: a schematic depicting the basic architecture of a large language model, which can have tens or even more than a hundred decoder blocks arranged in a feed-forward fashion. The inputs and outputs are strings of letters. Right: a heuristic map of the thalamocortical system, which generates complex neural activity patterns thought to underlie consciousness. Input is multimodal information from the world and output is created by the interaction between the brain and the body.

networks were designed as a simplified version of the cerebral cortex [11], modern LLMs have been highly engineered and fit to purpose in ways that do not retain deep homology with the known structure of the brain. Indeed, many of the circuit features that render LLMs computationally powerful (Figure 1) have strikingly different architectures from the systems to which we currently ascribe causal power in the production and shaping of consciousness in mammals. For instance, many theories of the neural basis of consciousness would assign a central role in conscious processing to thalamocortical [12–17] and arousal systems [18–24], both features that are architecturally lacking in LLMs.

One might ask why it is so crucial for the architecture of LLMs to mimic features of the brain. The primary reason, in our view, is that we can currently be absolutely sure of only a version of consciousness that arises from brains embedded within complex bodies. Some may contend that in its strictest form, this argument could be further collapsed to humans, though many of the systems-level features considered important for subjective consciousness are pervasive across phylogeny, stretching back to mammals [13,24,25] and even to invertebrates [26]. We will return to this point, but start with the question about what precisely we mean by the term 'consciousness'. From there we will develop three arguments against the view that present-day AI systems have, or that future AI systems will soon have, consciousness: first, consciousness is tied to the sensory streams that are meaningful for the organism; second, in mammalian brains, consciousness is supported by a highly interconnected thalamocortical system; and third, consciousness might be inextricably linked to the complex biological organization characteristic of living systems.

What is consciousness?

Consciousness is a complex concept and its definitions have long been debated. In the context of human interactions, conversation would be among the first elements typically used to assess whether another person is conscious or not. As discussed earlier, interactive language-based conversations with LLMs are currently often a starting point to develop an intuitive sense about

whether LLMs might be conscious. Although these conversations are remarkable, they are not formal objective measures of consciousness and constitute only *prima facie* evidence for conscious agency. The advent of LLMs has demanded a re-evaluation of whether one can indeed infer consciousness directly from verbal interactions with other agents. Thus, there is an emerging view that the criteria for attributing human-like abilities and characteristics need to be re-assessed [27].

There are different meanings associated with the word 'consciousness'. Neurologists, for instance, often refer to *levels* of consciousness; in the first place, whether a person is conscious or not and, in a more refined manner, assessing the gradations or specific states of consciousness. Psychologists, by contrast, often focus on the contents of consciousness: the specific experiences, memories, and thoughts of individuals' inner world. Furthermore, there are distinctions between different contents of consciousness: our experiences can be described, for instance, as primarily phenomenal or experiential [28] (e.g., the sight/smell of an apple, or the feel of your arm) or more abstract [28] (e.g., how we imagine, prospect, or manipulate concepts). The question of whether AI systems are conscious could be approached in various ways: it could focus primarily on only some of these aspects, or possibly all of them together. In the following, we focus mainly on phenomenal consciousness and ask whether machines can experience the world phenomenally.

The umwelt of an LLM

The aspect of the world that is perceptually 'available' to an organism has been described as its 'umwelt' (from the German 'environment' [29]). For instance, human retinas respond to wavelengths of light ranging from ~380 to 740 nm, which we perceive as a spectrum from blue to red. Without technological augmentation, we cannot detect light waves outside of this narrow band, in the infrared (>740 nm) or UV (<380 nm) bands. We have a similar umwelt for the auditory domain (we cannot hear tones outside the 20–20 000 Hz range), somatosensory domain (we can differentiate stimulation up to about 1 mm apart on some parts of our body), and vestibular domain (yoked to the 3D structure of our semicircular canals, which provide our inner sense of 'balance'). Other species can detect other portions of the electromagnetic spectrum. For instance, honeybees can see light in the UV range [30] and some snakes can detect infrared radiation in addition to more traditional visual light cues [31]; that is, the bodies and brains of other animals place different constraints on their sensitivity to the sensory world around them. Gibson referred to this information, that we can pragmatically interact with, as a set of 'affordances' [25,32–34].

If anything at all, what is the umwelt of an LLM? What kinds of affordances does an LLM have access to? By the nature of its design, an LLM is only ever presented with binary-coded patterns fed to the network algorithms inherent within the complex transformer architectures that comprise the inner workings of present-day LLMs [35,36]. While neuronal spikes also potentially encode incoming analog signals as digital (i.e., binary), the information stream fed to the LLMs in question is highly abstract and hence does not itself make any robust contact with the world as it is. Text and speech coded into strings of letters are simply no match for the dynamic complexity of the natural world: the umwelt of an LLM (the information afforded to it) is of a fundamentally different nature compared with the information that enters our brain when we open our eyes or listen to a conversation, and hence any accompanying experience. Traditional philosophical discourse has underscored the distinctiveness in the information streams experienced across species (for instance, between humans and bats [37]) and the phenomenology of these experiences. While there is no definite way to quantify this difference, we highlight that the informational input accessible to LLMs is likely to exhibit a more significant disparity.

That being said, it is worth mentioning that there is no conceptual barrier stopping the input of future AI systems from being much more enriched. Future LLMs could be equipped with different types of inputs (see [38,39]) that better match the kinds of signals that conscious agents have access to every day (i.e., the statistics of the natural world). Taking this even further, could the umwelt of future AI systems become more extended than that available to humans? In contemplating this question, it is essential to recognize that our umwelt and conscious experience are not determined solely by sensory input. For example, consider lying in a floatation tank where, despite a lack of normal sensory experiences, consciousness persists. This underscores the notion that having an umwelt presupposes an inherent subjective perspective, that is, an agent to begin with [29,40,41]. Similarly, affordances depend on the internal properties of the agents, in particular their motivations and goals [33,40,41]. This underscores the point that consciousness does not arise merely from data and hence that simply adding massive data streams to future AI systems will not, by itself, lead to consciousness.

These notions may prompt us to rethink some of the fundamental assumptions in the science of consciousness. Specifically, as AI systems continue to exhibit increasingly sophisticated abilities, one will have to re-evaluate the necessity of more basic self- and agency-related processes for the emergence of consciousness, as posited by some of the theories of consciousness [42–46].

The neural architecture supporting conscious integration

There is a sizable literature on the neural correlates of consciousness, with many different theories about the neural processes that underlie conscious processing. Some of these frameworks highlight that consciousness is supported by neural processing within the dense, re-entrant thalamocortical network [12–15,47–53]. The thalamocortical network encompasses cortical areas, cortico-cortical connectivity, and higher-order thalamic nuclei with their diffuse projections to cortical areas [54–56]. This specific architecture of the thalamocortical system supports recurrent and complex processing thought to underlie consciousness [53,57–61] and conscious integration (i.e., the fact that consciousness feels unified despite arising from processes happening in different brain areas) [51,53,62]. However, the details of how this integration is achieved differ across various theories of consciousness.

For instance, according to the global neuronal workspace theory [48,49] consciousness depends on the central workspace constituted by a distributed frontoparietal cortical system. This workspace integrates information from local cortical processors and then globally broadcasts it to all local processors, with the global broadcast delineating conscious from non-conscious processes. Other theories of consciousness assign a different neural process to carry out this integration. For instance, the binding-by-synchrony theory [63,64] suggests that conscious integration occurs via high-frequency synchronization between different cortical areas, a process that can be putatively involved in diverse functions, including perception, cognition, or motor planning, depending on the cortical regions involved.

In the dendritic integration theory (DIT) [12,53] (Figure 2), it is proposed that global conscious integration also depends on local integration at the level of single layer 5 pyramidal neurons, which are large excitatory neurons that hold a central position in both thalamocortical and cortico-cortical loops [12,53]. These neurons have two major compartments (Figure 2, orange and red cylinders) that process categorically distinct types of information: the basal compartment (red) processes externally-grounded information, whereas the apical compartment (orange) processes internally-generated information [12,53,65]. According to the DIT, during conscious states, these two compartments are coupled (i.e., integrated), allowing information to flow

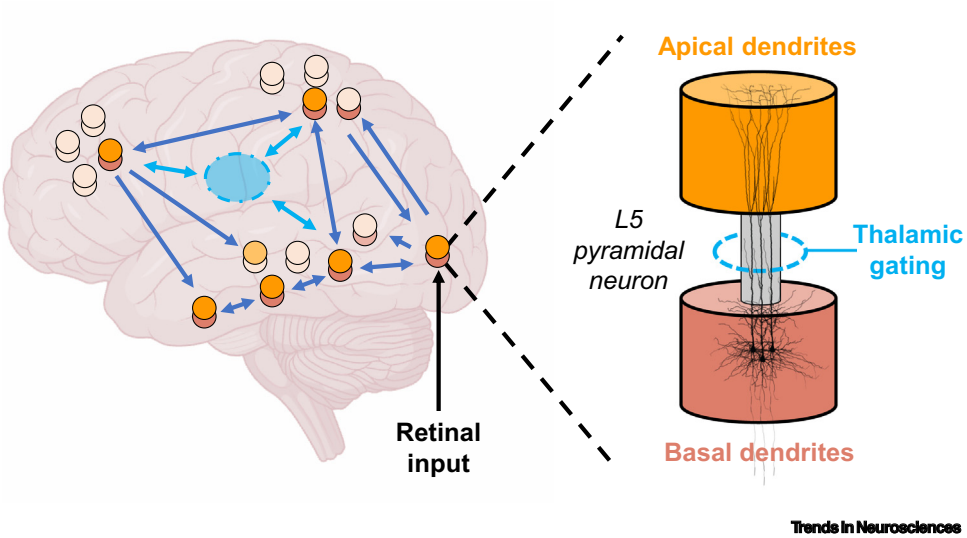

Figure 2. The neural architecture underlying conscious integration according to the dendritic integration theory (DIT). Left: sensory inputs (for instance, visual inputs) contain information that drives feed-forward activity in the sensory system. Internally generated signals (e.g., prediction, memory, attention) can augment certain features of the input stream (represented in the figure by bilateral arrows). This makes these representations stand out from the background (orange/red), leaving others inactive (light red). According to DIT [12,53], the thalamus (light blue; dot-dashed line) plays a crucial role in shaping/gating the contents of consciousness. Right: DIT assigns a key role to the subset of thick-tufted layer 5 (L5) pyramidal neurons in consciousness. These neurons display burst-firing, which occurs when depolarization of the cell body via basal dendrites (red) coincides temporally with descending cortical inputs to apical dendrites (orange), particularly in the presence of gating inputs from higher-order, matrix-type thalamus (blue).

through the thalamocortical and cortico-cortical connections, thus enabling system-wide integration and consciousness [12,53].

Notably, the architectures of present-day LLMs and other AI systems are devoid of features from each of these theoretical proposals: there is no equivalent of dual-compartment pyramidal neurons, nor a centralized thalamic architecture, a global workspace, or the many arms of the ascending arousal system. In other words, these AI systems are missing the very features of brains that are currently hypothesized to support consciousness. Although we are not arguing that the mammalian brain is the only architecture capable of supporting conscious awareness, the evidence from neurobiology suggests that very specific architectural principles (i.e., more than simple connections between integrate-and-fire neurons) are responsible for mammalian consciousness (see [1–4,26] for some examples of research on consciousness in non-mammalian species). Topologically, present-day AI systems are extremely simple in comparison, which is among the reasons we are cautious in ascribing phenomenal consciousness to them.

Could future AI models eventually incorporate the process of 'integration', which many theories of consciousness see as central? The integration proposed by the global neuronal workspace theory [48,49] offers a relatively straightforward implementation [9,10] and, in fact, some recent AI systems have incorporated something akin to a global workspace shared by local processors [66,67]. As the computational process of global broadcasting can be implemented in AI systems, an artificial system with a computationally equivalent global workspace would include a core ingredient underlying consciousness according to this theory [9,10]. However, as indicated earlier, not all theories of consciousness agree that this type of integration is key to consciousness. For instance, the integrated information theory of consciousness [50,51] claims that it is

impossible for a software-based AI system instantiated on a typical modern computer to achieve consciousness because modern computers do not have the appropriate architecture to realise the cause-effect power necessary for sufficiently integrating information [68,69]. Here, we will consider a third possibility, namely that consciousness might be implementable in principle, but it might require a level of computational specificity that is beyond the present-day (and perhaps future) AI systems.

Consciousness as a complex biological process

Consciousness does not only depend on the architecture of the system. For instance, the structure of the thalamocortical system does not change when we are in deep sleep or undergo anesthesia, yet consciousness disappears. Even during deep sleep, local neural responses and gamma-band activity in primary sensory areas can remain similar to those in the conscious state [70,71]. This implies that consciousness relies on specific neural processes that are different in conscious and unconscious brains.

To illustrate current knowledge about the details distinguishing conscious from unconscious processing, we will return to DIT, which encapsulates some of the specific neurobiological nuances relevant to the matter. In particular, DIT proposes that the crucial difference between conscious and unconscious processing lies in the integration between the two compartments of pyramidal cells (Figure 2). As indicated earlier, during conscious processing, these two compartments interact and hence enable complex and integrated processing across the thalamocortical system [12,53]. By contrast, it has been demonstrated that various anesthetic agents lead to functional decoupling between the two compartments [72]. In other words, while anatomically the neurons are intact and can fire action potentials, physiologically dendritic integration is severely limited within these cells: the top-down feedback component cannot influence processing. This dendritic coupling was demonstrated to be controlled by metabotropic receptors, which are often overlooked in computational models and in artificial neural networks. Furthermore, it was suggested that the activity of the metabotropic receptors in this context might be controlled by the higher-order thalamus [72]. Thus, there are relatively specific neurobiological processes that may be responsible for switching consciousness ‘on’ and ‘off’ in the brain. Clearly, the quality of experience in mammalian brains is intricately linked to the underlying processes from which it arises.

However compelling, these details almost certainly pale compared with the true complexity of the depth of biological organization required to achieve a satisfactory understanding of consciousness. While today’s explanations of consciousness rely on ideas such as the global workspace, information integration, recurrent feedback, dendritic integration, and other notions, it might be the case that the biological processes underlying consciousness are more intricate than these current concepts appreciate. It is also quite possible that the abstract computational-level ideas that are currently used to frame discussions in consciousness research may miss the necessary computational details required to satisfactorily explain consciousness. In other words, biology is complex and our understanding of biological computations is limited (Figure 3), so perhaps we simply do not yet have the right mathematical and experimental tools to understand consciousness.

To better ground this notion of biological complexity, it is worth considering that the cellular- and system-level processes described earlier are inextricably embedded within a living organism. Living organisms differ from present-day machines and AI algorithms, as they are constantly in the process of self-maintenance across several levels of processing [73,74]. Also, living systems have a multifaceted evolutionary and developmental history and their existence depends on their

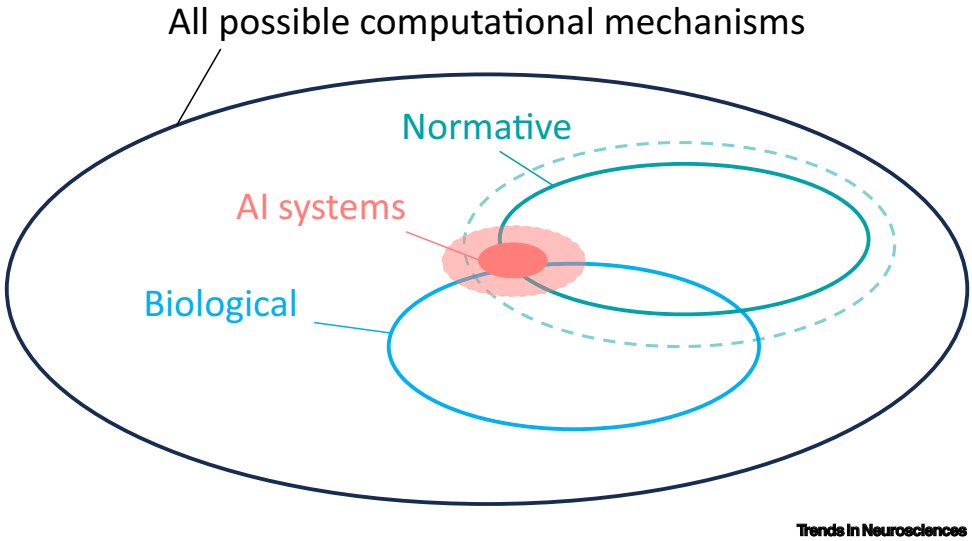

Figure 3. The limitations of current computational understanding of consciousness. The space (in the mathematical, abstract sense) of all possible computations (schematized as the large ellipse in the figure) is broader than the types of 'normative' computations currently envisioned or formalized in computational models (teal ellipse). Hence, current conceptualization may not have yet captured the key computations underlying consciousness. It can be argued that some of the normative computations that comprise the biological processes responsible for consciousness are currently understood (overlap between teal and blue ellipse). However, the 'knowledge horizon' of computational mechanisms is evolving and perhaps would have to be further extended (broken teal ellipse) for understanding consciousness. Computations in artificial intelligence (AI) systems (red ellipse) show some overlap with biological computations and some of the computations of AI systems are understood. However, the computations of AI systems differ from those of biological systems. In view of these differences, there is little *a priori* reason to assume, we would argue, that the computations of present-day AI systems are related to computations underlying phenomenal consciousness.

actions on multiple levels of organization (i.e., they have 'skin in the game'; Box 1). It has been argued that consciousness is intricately linked with the organization of living systems [74–77]. Here, we would like to draw attention to the fact that this organizational complexity (i.e., interactions between the different levels of the system) [77–81] of living systems is not captured within present-day computer software. While this fact need not impede progress in AI, it is entirely possible that the lack of any constraints imposed on modern AI algorithms to work like a living system effectively means that as long as AI is based on software, AI might be poorly placed to recapitulate conscious experience and agency.

The notion of biological complexity outlined in the previous paragraph is relevant also at the cellular level. A biological neuron is not just an abstract entity that can be fully captured with a few lines of code. Biological cells have multi-level organization and depend on a further cascade of biophysical intracellular complexity [79–83]. Consider the Krebs cycle, for instance, which underlies cellular respiration, a key process in maintaining cellular homeostasis [84]. Cellular respiration is a crucial biological process that enables cells to convert the energy stored in organic molecules into a form of energy that can be utilized by the cell. This process, however, is not 'compressible' into software, as processes like cellular respiration need to happen with real physical molecules. To be clear, our aim is not to suggest that consciousness requires the Krebs cycle, but rather to highlight that understanding consciousness may involve similar challenges in translating from the biological to the artificial realms: perhaps it cannot be abstracted away from the underlying machinery [68,69,85].

Box 1. LLMs and skin in the game: do we have a moral quandary?

Does it really matter if LLMs are conscious? If LLMs can match and even exceed our expectations in terms of getting superficially human-like responses that are useful and informative, is there any need to speculate about what an LLM experiences? Some argue that from a moral perspective, the answer should be a definitive 'yes' [92]. According to this view, we should carefully consider the ethical implications for any conscious entity, including AI, principally because it is assumed that some experiences could be negative and that in this case the AI system could also suffer. At this point, it is claimed, we should care or at least be cautious about whether an AI system might suffer.

We claim that LLMs do not (and will not) have experiences that can be considered suffering in any sense that should matter to human society. The notion of 'skin in the game' [93] may be useful as an analogy for articulating our argument here. This notion emphasizes the importance of personal investment and engagement in moral decision-making and suggests that those who have a personal stake in an issue are more competent to make ethical judgments than those who do not [93]. An LLM could, in principle, state in a conversation that it does not want to be shut down, but an LLM does not have 'skin in the game', as arguably there is no real consequence to the software when it is actually shut down. In contrast, in biology, the system has something to lose on several levels [73]: if it stops living, it will die. As the philosopher Hans Jonas has said: 'The organism has to keep going, because to be going is its very existence' [94]. If cellular respiration stops, the cell may die; if cells die, organs fail; if organs fail, the organism will soon die. The system has skin in the game across levels of processing, which is arguably prerequisite for caring about agency and consciousness [73]. Here, we would argue that not having the capacity for phenomenal consciousness would preclude suffering and, therefore, personal investment.

Importantly, we are not necessarily subscribing to the claim that consciousness cannot be captured within software at all [68,69,85–87]. Rather, we emphasize that we have to at least entertain the possibility that consciousness is linked to the complex biological organization underlying life [74–81] and thus computational descriptions that capture the essence of consciousness may be much more complex than our present-day theories suggest (Figure 3). It might be impossible to 'biopsy' consciousness and remove it from its organizational dwellings. This idea contradicts many current theories of consciousness, which assume that consciousness can be captured on the abstract level of computation [47,88]. This assumption, however, might be one that will require updating in light of modern AI systems: perhaps interdependencies and organizational complexity across scales observed in living systems cannot be ignored to fully understand consciousness.

It might be that although AI systems (at least to some extent) mimic their biological counterparts on the level of network computations, in these systems we have abstracted away all the other levels of processing that causally contribute to consciousness in the living brain and, possibly, have therefore abstracted away consciousness itself. In this way, LLMs and future AI systems may be trapped in a compelling simulation of the signatures of consciousness, but without any conscious experience to speak of. If consciousness is indeed related to these other levels of processing, or their coherent interactions across scales, we might still be far from the possibility of conscious machines.

Concluding remarks

Here, we have provided a neuroscience perspective on the possibility of consciousness in LLMs and future AI systems. We conclude that, while fascinating and alluring, LLMs are not conscious and will likely not be conscious soon. First, we detailed the vast differences between the umwelt of mammals (the 'slice' of the external world that they can perceive) and the highly impoverished and limited umwelt of LLMs when compared with biological counterparts. Second, we argue that the topological architecture of LLMs, while highly sophisticated, is sufficiently different from the neurobiological details of circuits empirically linked to consciousness in mammals that there is no *a priori* reason to conclude that they are capable of phenomenal consciousness (Figure 1). Third, we point out that it might not be possible to abstract consciousness away from the organizational complexity that is inherent within living systems but strikingly absent from AI systems. Overall, we believe that these three arguments make it extremely unlikely that LLMs, in their current

Outstanding questions

Assessment of consciousness in LLMs and AI is often envisioned to depend on language-based tests to probe consciousness. Is it possible to evaluate consciousness based on language (i.e., text) only? If not, are we destined to remain uncertain about consciousness in LLMs, or are there any further features and aspects of consciousness that can help (more confidently) diagnose the presence (or absence) of consciousness in artificial systems?

The thalamocortical system seems to be relevant for the neural basis of consciousness in mammals. How could a thalamocortical system be implemented in AI? Which particular functions and tasks will benefit from having a thalamocortical-like system?

The ascending arousal system also plays a crucial role in facilitating consciousness in living organisms, having a complex, multifaceted role in shaping neural dynamics. To what extent does AI need to mimic these different processes to capture the computational benefits of the ascending arousal system?

Could including biological details enhance the capabilities of AI systems? Besides the thalamocortical system, dendrites seem to be key players in some of the theories of consciousness discussed in the current paper. Are dendrites just a factor that adds computational complexity/efficiency to biological neural networks, or is there more to it?

Is the organizational complexity of living systems related to consciousness? Living systems consist of various levels of processing that causally interact. Can the organizational complexity of living systems be explained more formally? New mathematical frameworks are required for dealing with such systems to shed more light on consciousness.

According to some accounts, consciousness and agency are inextricably linked. To understand how consciousness emerges from biological activity, do we first need to understand agency?

form, have the capacity for phenomenal consciousness. Rather, they mimic signatures of consciousness that are implicitly embedded within the language that people use to describe the richness of their conscious experience.

Rather than representing an antithetical account, the proposed perspective may bear some useful implications (see Outstanding questions). For one, perhaps any worries about potential moral quandaries regarding sentience in LLMs are currently more hypothetical than real (Box 1). In addition, we believe that a refined understanding of the similarities and differences in the topological architecture of LLMs and mammalian brains provides opportunities for advancing progress in both machine learning and neuroscience. To this end, major inroads will occur through mimicking features of brain organization and by learning how simple distributed systems can process elaborate information streams [89–91]. For these reasons, we are optimistic that future collaborative efforts between AI researchers and neuroscientists have the potential to gain a deeper understanding of consciousness.

Acknowledgments

We would like to thank Jakob Howhy, Kadi Tulver, Raul Vicente, Christopher Whyte, and Gabriel Wainstein for their helpful comments on the manuscript. This work was supported by the European Social Fund through the 'ICT programme' measure, by the European Regional Development Fund through the Estonian Centre of Excellence in IT (EXCITE), the Estonian Research Council grant PSG728, the National Health and Medical Research Council (1193857), the Bellberry foundation, European Union (ERC, CorticalCoupling, 101055340), and the German Research Council (DFG LA 3442/13).

Declaration of interests

The authors declare no competing interests in relation to this work.

References

1. Seth, A.K. *et al.* (2005) Criteria for consciousness in humans and other mammals. *Conscious. Cogn.* 14, 119–139
2. Edelman, D.B. and Seth, A.K. (2009) Animal consciousness: a synthetic approach. *Trends Neurosci.* 32, 476–484
3. Birch, J. *et al.* (2020) Dimensions of animal consciousness. *Trends Cogn. Sci.* 24, 789–801
4. Baluška, F. and Reber, A. (2019) Sentience and consciousness in single cells: how the first minds emerged in unicellular species. *BioEssays* 41, 1800229
5. Thompson, E. (2022) Could all life be sentient? *J. Conscious. Stud.* 29, 229–265
6. Ball, P. (2022) *The Book of Minds: How to Understand Ourselves and Other Beings, from Animals to AI to Aliens*, University of Chicago Press
7. Chalmers, D.J. (2023) Could a large language model be conscious? *arXiv* Published online April 29, 2023. https://doi.org/10.48550/arXiv.2303.07103
8. Aguera y Arcas, B.A. (2022) Do large language models understand us? *Daedalus* 151, 183–197
9. VanRullen, R. and Kanai, R. (2021) Deep learning and the global workspace theory. *Trends Neurosci.* 44, 692–704
10. Juliani, A. *et al.* (2022) On the link between conscious function and general intelligence in humans and machines. *arXiv* Published online July 19, 2022. https://doi.org/10.48550/arXiv.2204.05133
11. Rumelhart, D.E., ed (1999) *Parallel Distributed Processing. 1: Foundations*, MIT Press
12. Aru, J. *et al.* (2020) Cellular mechanisms of conscious processing. *Trends Cogn. Sci.* 24, 814–825
13. Shine, J.M. (2021) The thalamus integrates the macrosystems of the brain to facilitate complex, adaptive brain network dynamics. *Prog. Neurobiol.* 199, 101951
14. Llinás, R. and Ribary, U. (2001) Consciousness and the brain. The thalamocortical dialogue in health and disease. *Ann. N. Y. Acad. Sci.* 929, 166–175
15. Tasserie, J. *et al.* (2022) Deep brain stimulation of the thalamus restores signatures of consciousness in a nonhuman primate model. *Sci. Adv.* 8, eabl5547
16. Koch, C. *et al.* (2016) Neural correlates of consciousness: progress and problems. *Nat. Rev. Neurosci.* 17, 307–321
17. Aru, J. *et al.* (2019) Coupling the state and contents of consciousness. *Front. Syst. Neurosci.* 13, 43
18. Parvizi, J. and Damasio, A. (2001) Consciousness and the brainstem. *Cognition* 79, 135–160
19. Shine, J.M. (2023) Neuromodulatory control of complex adaptive dynamics in the brain. *Interface Focus* 13, 20220079
20. Snider, S.B. *et al.* (2019) Disruption of the ascending arousal network in acute traumatic disorders of consciousness. *Neurology* 93, e1281–e1287
21. Edlow, B.L. *et al.* (2012) Neuroanatomic connectivity of the human ascending arousal system critical to consciousness and its disorders. *J. Neuropathol. Exp. Neurol.* 71, 531–546
22. Spindler, L.R. *et al.* (2021) Dopaminergic brainstem disconnection is common to pharmacological and pathological consciousness perturbation. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2026289118
23. Hindman, J. *et al.* (2018) Thalamic strokes that severely impair arousal extend into the brainstem. *Ann. Neurol.* 84, 926–930
24. Merker, B. (2007) Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behav. Brain Sci.* 30, 63–81
25. Shine, J.M. (2022) Adaptively navigating affordance landscapes: how interactions between the superior colliculus and thalamus coordinate complex, adaptive behaviour. *Neurosci. Biobehav. Rev.* 143, 104921
26. Barron, A.B. and Klein, C. (2016) What insects can tell us about the origins of consciousness. *Proc. Natl. Acad. Sci. U. S. A.* 113, 4900–4908
27. Mitchell, M. and Krakauer, D.C. (2023) The debate over understanding in AI's large language models. *Proc. Natl. Acad. Sci. U. S. A.* 120, e2215907120
28. Block, N. (1995) On a confusion about a function of consciousness. *Behav. Brain Sci.* 18, 227–247
29. von Uexküll, J. (1934/2010) *A Foray into the Worlds of Animals and Humans: With a Theory of Meaning* (1st edn), University of Minnesota Press

30. Wakakuwa, M. *et al.* (2007) Spectral organization of ommatidia in flower-visiting insects. *Photochem. Photobiol.* 83, 27–34
31. Chen, Q. *et al.* (2012) Reduced performance of prey targeting in pit vipers with contralaterally occluded infrared and visual senses. *PLoS One* 7, e34989
32. Gibson, J.J. (1966) *The Senses Considered as Perceptual Systems*, Houghton Mifflin
33. Greeno, J.G. (1994) Gibson's affordances. *Psychol. Rev.* 101, 336–342
34. Pezzulo, G. and Cisek, P. (2016) Navigating the affordance landscape: feedback control as a process model of behavior and cognition. *Trends Cogn. Sci.* 20, 414–424
35. Vaswani, A. *et al.* (2017) Attention is all you need. *arXiv* Published online August 2, 2023. https://doi.org/10.48550/arXiv.1706.03762
36. Brown, T.B. *et al.* (2020) Language models are few-shot learners. *arXiv* Published online July 22, 2020. https://doi.org/10.48550/arXiv.2005.14165
37. Nagel, T. (1974) What is it like to be a bat? *Philos. Rev.* 83, 435
38. Radford, A. *et al.* (2021) Learning transferable visual models from natural language supervision. *arXiv* Published online February 26, 2021. https://doi.org/10.48550/arXiv.2103.00020
39. Driess, D. *et al.* (2023) PaLM-E: an embodied multimodal language model. *arXiv* Published online March 6 2023. https://doi.org/10.48550/arXiv.2303.03378
40. Roli, A. *et al.* (2022) How organisms come to know the world: fundamental limits on artificial general intelligence. *Front. Ecol. Evol.* 9, 1035
41. Fultot, M. and Turvey, M.T. (2019) Von Uexküll's theory of meaning and Gibson's organism–environment reciprocity. *Ecol. Psychol.* 31, 289–315
42. Damasio, A. and Damasio, H. (2022) Homeostatic feelings and the biology of consciousness. *Brain* 145, 2231–2235
43. Damasio, A. and Damasio, H. (2022) Feelings are the source of consciousness. *Neural Comput.* 35, 277–286
44. Damasio, A. (2021) *Feeling & Knowing: Making Minds Conscious*, Pantheon
45. Panksepp, J. (2005) Affective consciousness: core emotional feelings in animals and humans. *Conscious. Cogn.* 14, 30–80
46. Solms, M. (2021) *The Hidden Spring: A Journey to the Source of Consciousness*, Profile Books
47. Seth, A.K. and Bayne, T. (2022) Theories of consciousness. *Nat. Rev. Neurosci.* 23, 439–452
48. Dehaene, S. and Changeux, J.P. (2011) Experimental and theoretical approaches to conscious processing. *Neuron* 70, 200–227
49. Mashour, G.A. *et al.* (2020) Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105, 776–798
50. Tononi, G. (2004) An information integration theory of consciousness. *BMC Neurosci.* 5, 42
51. Tononi, G. (2008) Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* 215, 216–242
52. Redinbaugh, M.J. *et al.* (2020) Thalamus modulates consciousness via layer-specific control of cortex. *Neuron* 106, 66–75.e12
53. Bachmann, T. *et al.* (2020) Dendritic integration theory: a thalamo-cortical theory of state and content of consciousness. *Philos. Mind Sci.* 1, II
54. Bell, P.T. and Shine, J.M. (2016) Subcortical contributions to large-scale network communication. *Neurosci. Biobehav. Rev.* 71, 313–322
55. Shine, J.M. *et al.* (2023) The impact of the human thalamus on brain-wide information processing. *Nat. Rev. Neurosci.* 24, 416–430
56. Suzuki, M., *et al*. How deep is the brain? The shallow brain hypothesis. Nat. Rev. Neurosci. (in press)
57. Supèr, H. *et al.* (2001) Two distinct modes of sensory processing observed in monkey primary visual cortex (V1). *Nat. Neurosci.* 4, 304–310
58. Lamme, V.A. (2006) Towards a true neural stance on consciousness. *Trends Cogn. Sci.* 10, 494–501
59. Imas, O.A. *et al.* (2005) Volatile anesthetics disrupt frontal-posterior recurrent information transfer at gamma frequencies in rat. *Neurosci. Lett.* 387, 145–150
60. Boly, M. *et al.* (2011) Preserved feedforward but impaired top-down processes in the vegetative state. *Science* 332, 858–862
61. Ku, S.W. *et al.* (2011) Preferential inhibition of frontal-to-parietal feedback connectivity is a neurophysiologic correlate of general anesthesia in surgical patients. *PLoS One* 6, e25155
62. Bayne, T. (2012) *The Unity of Consciousness*, Oxford University Press
63. Singer, W. (1998) Consciousness and the structure of neuronal representations. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 353, 1829–1840
64. Singer, W. (2001) Consciousness and the binding problem. *Ann. N. Y. Acad. Sci.* 929, 123–146
65. Larkum, M. (2013) A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends Neurosci.* 36, 141–151
66. Goyal, A. *et al.* (2021) Coordination among neural modules through a shared global workspace. *arXiv* Published online March 22, 2022. https://doi.org/10.48550/arXiv.2103.01197
67. Juliani, A. *et al.* (2022) The perceiver architecture is a functional global workspace. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44) (Culbertson, A. *et al*., eds), pp. 44, Cognitive Science Society
68. Koch, C. (2019) *The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed*, MIT Press
69. Tononi, G. and Koch, C. (2015) Consciousness: here, there and everywhere? *Philos. Trans. R. Soc. B Biol. Sci.* 370, 20140167
70. Krom, A.J. *et al.* (2020) Anesthesia-induced loss of consciousness disrupts auditory responses beyond primary cortex. *Proc. Natl. Acad. Sci.* 117, 11770–11780
71. Hayat, H. *et al.* (2022) Reduced neural feedback signaling despite robust neuron and gamma auditory responses during human sleep. *Nat. Neurosci.* 25, 935–943
72. Suzuki, M. and Larkum, M.E. (2020) General anesthesia decouples cortical pyramidal neurons. *Cell* 180, 666–676
73. Man, K. and Damasio, A. (2019) Homeostasis and soft robotics in the design of feeling machines. *Nat. Mach. Intell.* 1, 446–452
74. Seth, A. (2021) *Being You: A New Science of Consciousness*, Penguin
75. Thompson, E. (2010) *Mind in Life: Biology, Phenomenology, and the Sciences of Mind* (1st edn), The Belknap Press of Harvard University Press
76. Seth, A.K. and Tsakiris, M. (2018) Being a beast machine: the somatic basis of selfhood. *Trends Cogn. Sci.* 22, 969–981
77. Deacon, T.W. (2012) *Incomplete Nature: How Mind Emerged from Matter* (1st edn), W.W. Norton & Co
78. Louie, A.H. (2013) *More Than Life Itself: A Synthetic Continuation in Relational Biology*, Vol. 1. Walter de Gruyter
79. Rosen, R. (1991) *Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life*, Columbia University Press
80. Rosen, R. (2000) *Essays on Life Itself*, Columbia University Press
81. Moreno, A. and Mossio, M. (2015) *Biological Autonomy: A Philosophical and Theoretical Enquiry*, Springer
82. Nicholson, D.J. (2019) Is the cell really a machine? *J. Theor. Biol.* 477, 108–126
83. Kandel, E.R. *et al.*, eds (2000) *Principles of Neural Science*. Vol. 4. McGraw-Hill, pp. 1227–1246
84. Lane, N. (2022) *Transformer: The Deep Chemistry of Life and Death*, Profile Books
85. Searle, J.R. (1992) *The Rediscovery of the Mind*, MIT Press
86. Penrose, R. (1994) *Shadows of the Mind*, Oxford University Press
87. Gidon, A. *et al.* (2022) Does brain activity cause consciousness? A thought experiment. *PLoS Biol.* 20, e3001651
88. Butlin, P. *et al.* (2023) Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv* Published online August 22, 2023. https://doi.org/10.48550/arXiv.2308.08708
89. Richards, B.A. *et al.* (2019) A deep learning framework for neuroscience. *Nat. Neurosci.* 22, 1761–1770
90. Zador, A. *et al.* (2023) Catalyzing next-generation artificial intelligence through neuroAI. *Nat. Commun.* 14, 1597
91. Doerig, A. *et al.* (2023) The neuroconnectionist research programme. *Nat. Rev. Neurosci.* 24, 431–450
92. Metzinger, T. (2021) Artificial suffering: an argument for a global moratorium on synthetic phenomenology. *J. AI. Consci.* 8, 43–66
93. Taleb, N.N. (2018) *Skin in the Game: Hidden Asymmetries in Daily Life* (1st edn), Random House
94. Jonas, H. (2001) *The Phenomenon of Life: Toward a Philosophical Biology*, Northwestern University Press