

My word

Consciousness
beyond the human
case

Joseph LeDoux¹, Jonathan Birch²,
Kristin Andrews³, Nicola S. Clayton⁴,
Nathaniel D. Daw⁵, Chris Frith⁶,
Hakwan Lau⁷, Megan A. K. Peters⁸,
Susan Schneider⁹, Anil Seth¹⁰,
Thomas Suddendorf¹¹, and
Marie M. P. Vandekerckhove¹²

Artificial intelligence (AI) is barreling forward at breakneck speed, with some proponents bullishly confident that Large Language Models (LLMs) such as chatGPT will soon be conscious¹. But there is much scope for confusion regarding the meaning of 'conscious'. If we conflate consciousness with intelligence, then AI is already there, but we need to keep these ideas apart². Similar issues arise in studies of animal consciousness, where consciousness is sometimes said to be manifested in complex behavior (such as the waggle dance of the bee) or in behaviours that resemble human behaviour (if a fish bolts in a situation that would scare a human, the fish must, like the human, feel fear). Such inferences can easily lead us astray.

Human research is greatly facilitated by the fact that people can verbally report their experiences. Animals, lacking verbal report, can only respond non-verbally, always leaving some doubt about whether their behaviour was consciously or non-consciously controlled³. With LLMs, by contrast, we have an abundance of linguistic data. But these models are trained (on over a trillion words of data) to respond to prompts as a human would respond, so we are left in the dark as to whether their responses reflect genuine consciousness or skillful mimicry⁴.

So how should we define 'consciousness'? Consciousness, even in human research, is often treated as a unitary kind of state. An important exception is Endel Tulving's tripartite taxonomy of 'autonoetic', 'noetic', and 'anoetic' consciousness⁵⁻⁸. 'Autonoetic' consciousness is reflective self-awareness: the ability to locate your current experience as part of a narrative of your own life that extends into the past and future. 'Noetic' consciousness involves semantic

and conceptual awareness without self-awareness; it involves the ability to apply concepts to your current perceptions and generate knowledge from them.

Perhaps most elusively of all, 'anoetic' consciousness is experience that involves neither self-awareness nor semantic knowledge. It includes, for example, feelings of rightness/wrongness, comfort/discomfort, familiarity, unease, presence or absence, tiredness, confidence, uncertainty and ownership⁹. It includes the feeling that the object in the corner of one's eye is definitely a bird; the feeling on returning to your house that things are as you left them (or not); the feeling that you are coming down with an illness. In humans, these anoetic feelings sit on the 'fringe' of consciousness, only rarely the focus of attention¹⁰. In some other animals, it is possible that the anoetic is all there is.

All three layers of consciousness are likely to be entangled in complex ways with memory, attention and metacognition: the brain representing (or 're-representing') its own lower-level cognitive processes. It is hard to say what exactly is required for each layer, and there is ongoing debate about this^{3,8,11}. Autonoetic consciousness may be uniquely human and, if not, is likely to be present in only a small minority of other animals, such as other great apes. Noetic consciousness seems likely to be present in a wider range of animals, at least all primates, and possibly some other mammals. Anoetic consciousness is likely to have the widest distribution of all, including at least all mammals. It is possible that the requirements of anoetic consciousness are very minimal indeed.

The term 'sentience' is often used in discussions of animals and AI. Where does it fit in relation to Tulving's taxonomy? For some, it is something extremely basic, no more than "responding to stimuli using adaptive internal processes"¹². But this definition is so minimalist that it trivializes the idea of sentience. For others, sentience can be close to a synonym for anoetic consciousness¹³. For still others, it captures something more than just responding to stimuli but 'rawer' than anoetic consciousness: an elemental kind of experience that does not even involve evaluating your current state, but just involves a feeling of being alive or being present. In this vein, Marie Vandekerckhove has written of sentience as lying in between mere wakefulness and anoetic consciousness in a "continuum of consciousness"¹¹.

AI already seems capable of applying concepts and generating knowledge from sensory inputs. It seems likely it will soon have the capacity to construct self-narratives too. A deep source of puzzlement is that it may achieve these things without possessing the base layer of unarticulated, unconceptualized feelings on which these higher levels are built in humans. It may have the overhanging structure without the foundation. To switch metaphors, the 'lights may be off', everything may be happening 'in the dark', even when the AI lucidly recounts the story of its life. Meanwhile, many animals may be in the opposite situation. They may have rich worlds of unarticulated, unconceptualized sentient feelings without any ability to conceptualize what is happening or place it within a narrative. Both possibilities push our imaginative abilities to the limit.

With these questions growing in urgency, and answers elusive, a group of scientists has recently called for AI companies to start investing seriously in consciousness research¹.

Against this background, we thought it would be a good moment to put some of the big questions in this area to some leading experts on consciousness and/or higher cognition in humans and other animals. We asked them to read our introduction and respond to some or all of the following questions:

Do you think AI will soon be conscious? Why or why not?

Which other animals do you think are conscious and why?

What can the science of AI consciousness learn from the science of animal consciousness, and vice versa?

The challenge of addressing the questions we raised was taken up by Kristin Andrews, Nicky Clayton, Nathaniel Daw, Chris Frith, Hakwan Lau, Megan Peters, Susan Schneider, Anil Seth, Thomas Suddendorf, and Marie Vandekerckhove. Here are their responses.

Joseph LeDoux

Center for Neural Science and Department of Psychology at New York University, New York, NY 10003, USA. Department of Psychiatry and Department of Child and Adolescent Psychiatry, New York University Langone Medical School, New York, NY 10003, USA.

Jonathan Birch

Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science, London, UK.





Kristin Andrews

York Research Chair in Animal Minds, Department of Philosophy, York University, Toronto, ON M6H 3N8, Canada.

That we are today confronted with the question of whether AI will soon be conscious given the success of LLMs reveals much about contemporary anthropocentric biases. Humans use language and are conscious, and a quick analogical argument for AI consciousness might seem promising. But we need to consider the disanalogies, too. The humans, monkeys and rats often used as research subjects in consciousness studies have plenty of properties that AIs lack — properties associated with sociality and development; a rich sensorium; and being alive.

AI do not inhabit a world of native social models to learn from and same-age peers to develop with. Their social partners are their human overlords, not a community of same species individuals with different learning histories and genetic diversity. Arguably, AIs are not cultural beings, and they do not engage in collaborative and competitive interactions within and between cultures.

AIs do not richly perceive their physical environment, either. While their token transactions beyond themselves number into the trillions, the types of transactions they make are minimal, having access to information via a keyboard or camera which is transduced into electrical signals. There is little integration of information across anything like sensory modalities.

And AIs are not alive. They do not actively preserve their continued existence, self-organize to preserve their boundaries from merging into their environment, they do not take in elements of the environment and transform those through metabolic processes into their own continued existence. They lack any goal of self-preservation that they seek to achieve through their actions in a social and physical environment. And since they fail to reproduce themselves, they can't be thought of as evolved systems.

Maybe I am being overly skeptical, and the current and future AIs may

have more of these elements than I give them credit for. Still, worries remain. Engineering AIs to have properties analogous with humans raises the gaming problem, and risks creating a kludgy consciousness mimic whose functional organization is vastly unlike the human case⁴. Worse yet, even if computer scientists can construct AIs with some of these properties, contemporary science is far from developing a robust artificial life — and there are suggestive reasons to think that life and conscious mind may be essentially intertwined^{14,15}. What life gives us that language does not is a purpose, a function. Language can make this purpose easier to see, allowing an agent to describe and communicate their goals, but it can also create an illusion of agency where it does not exist.

Function and biology have an uneasy relationship, and many remain strongly suspicious that there is any merit in describing cells or bacteria as having goals, while at the same time using verbal gymnastics to talk about purpose without using the term. As Denis Walsh puts it, “Organisms are fundamentally purposive entities, and biologists have an animadversion to purpose”¹⁶.

The continuity of life and conscious mind suggests a possible function for consciousness — to sustain life. When we focus on the most rudimentary of subjective experience — feelings of thirst, oxygen depletion, social/sexual desire — consciousness may be taken as having a vital purpose.

While this does not tell us which animals are conscious — or if plants are — it should lead us to study the simplest forms of animal life. I've argued that the cognitive science of consciousness should adopt as a working hypothesis that all animals are conscious, and study much simpler animal models to make progress developing a secure theory. Even the humble microscopic nematode worm *Caenorhabditis elegans* can serve as a promising model for studying consciousness, given their sensory, social, and learning capacities¹⁷. Science has progressed not just by looking at humans, but at life far distant from our own. Aristotle was fascinated by sponges, Mendel by the pea plant. By studying primordial experience in simple animals, we can gain an

understanding of what properties an AI must have before we take seriously the question of whether it is conscious.



Nicola S. Clayton

Department of Psychology, University of Cambridge, Cambridge CB2 3EB, UK.

In their introduction, LeDoux and Birch characterize three levels of consciousness, based on Tulving's earlier theoretical work on mental time travel — the ability to remember the past and imagine the future — and the extent to which it is uniquely human⁵. Autooetic consciousness refers to an awareness of being the owner and author of one's memories and thoughts about the future. Noetic consciousness is semantic rather than episodic and therefore relies on the knowing system rather than the experiential remembering one. Finally, LeDoux and Birch refer to the rather more elusive anoetic consciousness, which some might term basic awareness: a feeling of what is right and wrong with one's world, from feelings of comfort and discomfort, to confidence about how certain or uncertain things might be. Anoetic consciousness need not involve reflective self-awareness, be it experiential or knowledge based, but simply associative learning about prediction errors.

One aspect of this triumvirate that is notably missing is Tulving's additional concept of chronesthetic consciousness, which is particularly pertinent to mental time travel, namely being aware of the “ever-present awareness of one's being existing in a subjective sea of time” as Tulving put it⁵. This concept is at the very heart of human conscious experience, of how we subjectively project ourselves backwards and forwards in space and time.

I share the authors' sentiments that this is not something we share with artificial intelligence for it requires a level of experiential and emotional intelligence that only biological systems (currently) have. Man-made machines and algorithms may superficially appear to respond in that way, but only because they have been devised and designed by *Homo*

sapiens. Biological systems that have evolved through natural selection, on the other hand, may indeed have some of these capabilities.

Research on mental time travel in non-human animals has shown that some species are able to recall the past and plan for the future. The first demonstration of this was in scrub-jays, members of the crow family that hide food: they can remember what they cached where and when on the basis of a single past experience¹⁸; they can plan for where to cache based on where tomorrow's breakfast is likely to be served¹⁹; and they can use their past experiences to predict whether others are likely to pilfer (steal) their caches or not and take prospective action accordingly²⁰. Since the initial findings on scrub-jays, some aspects of the ability to remember the what, where and when (or which) of previous events has been found in a number of other species and taxa, including mammals such as rodents and great apes, birds such as crows and jays, and even some cephalopods such as cuttlefish and insects such as bees.

The issue is whether one can explain these behaviours without the need to invoke consciousness. How could we ever know whether what-where-when/which memory and an ability to plan for the future is accompanied by the phenomenological consciousness of autonoesis and chronesthesia in the absence of any agreed markers of consciousness in non-linguistic animals? What do we do about the 'absence of evidence is not evidence of absence' problem? LeDoux and Birch suggest that autonoesis is limited to humans and possibly some other animals such as the great apes, but it is important not to forget that evolution need not be homologous. Distantly related animals, such as corvids, may have evolved similar cognitive abilities to the great apes, precisely because their ancestors had similar problems to solve²¹.

Perhaps this is something that the Science of AI Consciousness and the Science of Comparative Cognition can learn from one another, namely the implications of diverse intelligences that are alien to our species and whose evolutionary trajectories may be convergent rather than homologous, be they natural or artificial.



Nathaniel D. Daw

Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ 08544, USA.

It feels like we are living through yesterday's thought experiment. Large language models (LLMs) trained to mimic human language can confidently claim to have a full range of conscious experiences, as when an early version of the Bing chatbot notoriously professed its love for a *New York Times* reporter. Preventing such self-anthropomorphizing confabulation is one of the explicit goals of fine-tuning such models for deployment²². But faced with some near-future AI system, how could we know whether to take such claims seriously? This is not a question with easy answers, but it reveals a lot about psychology and neuroscience and the mutual opportunities offered by their relationship with AI.

At some level, the prevaricating chatbot exemplifies a perennial problem in the brain sciences. A chief project of 20th century psychology was recasting the field as an empirical laboratory science by replacing introspective self-report with objective measurement. Early behaviorists sought to deduce laws of behavior grounded in objective stimuli and responses, without reference to internal subjective constructs. Later cognitive psychologists used clever experimental designs to draw inferences about the covert mental operations and representations underlying behavior, later supported by more direct neuroscientific measurement and manipulation. Thus — if you claim to be a synesthete, for example, to involuntarily perceive words or numbers also as colors — experimenters can confirm that you are telling the truth by studying how interacting colors and symbols affect your speed in visual search, or how each dimension affects neural activity patterns normally related to the other²³.

One hallmark of this program is substituting one set of questions for another: replacing the 'hard problem' of subjective experience as the object of study with more accessible — and to a computational neuroscientist like me, also more interesting — mechanistic

questions about the representations and information processing operations underlying overt behavior. Two examples of this that researchers have argued are related to consciousness are clever behavioral operationalizations of metacognition, such as the ability to reason about one's own beliefs or actions, often assessed by betting on whether the belief is correct; and behavioral operationalizations of volitional, deliberative choice, often assessed by the ability to replan when goals or contingencies change²⁴.

Does this same strategy work with AI systems? It's not simple. A key driver of the LLM explosion was the discovery of 'scaling laws'²⁵. If you can quantify a model's performance on some task, you can extrapolate improvement in that benchmark with the size of the model and its data, and solve for superhuman performance. Having followed this strategy, we find ourselves in the position of asking what it means that the resulting LLM can pass the bar exam, or whether its responses on a standard assay mean it really exhibits theory of mind.

Again, this story is familiar. For as long as we have had behavioral operationalizations of cognition, we have had disagreements about their interpretation. If animals can place accurate bets as to whether their perceptual judgments are correct, does that mean they are actually capable of metacognition interrogating their internal sensory states? Or are they instead following some confounding strategy?

In my view, these disagreements often distract from the real opportunity. By design, tasks of this sort almost inevitably deliver nontrivial behaviors. The mechanisms by which the brain performs them are surely of interest, even if they turn out to be not quite what was intended. For instance, neural measurements suggest that much of the computation supporting supposedly deliberative planning can happen ahead of time, and in a different manner than originally assumed: like solving a maze backward from the goal. Does this count as deliberation? Does it matter?

LLMs offer a similar opportunity. Even if their feats of wordplay are arguably a charade — or outright lies — they invite us to ask what internal representations

and transformations allow them to reason competently about theory of mind, or causation, or moral judgments. This is, of course, the bread and butter of early connectionism — the approach to understanding the brain by constructing and analysing neural network models that are trained by adjusting the weights of connections between model neurons — which is overdue for a revival in light of today's much more widely capable models. A newer version of such questions takes advantage of the fact that the models are ultimately trained by maximizing some quantifiable objective, for example, word prediction, over data, to investigate what it is about these that gives rise to such unexpected emergent behaviors. For instance, one of the most practically useful abilities of LLMs is to learn new tasks or concepts from a few examples — without the gradual adjustments of connection weights used for initial training. Recent evidence suggests this is ultimately grounded in the statistical structure of the natural language data they are trained on²⁶. Mechanistic questions like this are hard to answer — but they can be very revealing, and maybe not quite as hard as confronting consciousness head on.



Chris Frith

Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London, UK.

Institute of Philosophy, School of Advanced Studies, University of London, London, UK.

When a creature is conscious it is having subjective experiences. I believe that the mechanisms that underpin such experiences have evolved. Early in evolution there were creatures in which control of behaviour occurred without any subjective experience; later, sentient creatures appeared. Such sentient creatures create a model of the world to guide their interactions, which frees them being slaves to external stimuli. The subjective experience associated with sentience became richer as nervous systems became more complex. Later still creatures appeared, best exemplified by humans, who can reflect on their subjective experiences and discuss these experiences with

others. Such discussions provide an additional means for developing our models of the world. In humans, these three levels of control operate in parallel.

I believe that many animals are sentient, having the bottom two levels of control²⁷. A few may reach the third level, but this third level is highly developed only in humans. The large language models of AI, in contrast, having been trained on words, have, in some sense, the highest level without the two lower levels. In this case they are not conscious: they have no sentient level on which to reflect²⁸.

But what does it mean to have only this highest level? And what would be needed for AIs to become conscious? To answer this question, we need to study the interactions between the different levels.

Signals from lower levels provide information about the working of the brain/mind (metacognitive signals). For example, we have experiences of fluency: how quickly and easily we perceive an object; how quickly and easily we choose an appropriate action. These feelings of fluency can be interpreted as markers of confidence. By sharing our degree of confidence with others we can improve our decision-making²⁹.

These signals from the lower levels have a vital role in keeping our models of the world in check. Without them we would not have the deference to the world that we need to distinguish appearance from reality. Large language models lack these signals; they have a model, but it is not a model of anything. It is not grounded in sentience. AI systems do not have the constraints provided by metacognitive signals. They have no clue when they are deviating from reality.

Signals from the highest level modify the functioning of the lower levels. And these signals often come from others in the form of verbal instructions³⁰. But how does this work?

We can contrast learning from instructions with learning from direct experience. For example, in threat conditioning we learn that when we see a cue, such as a blue square, it is likely to be followed by a painful shock. This association is built up slowly by trial and error. The unexpected shock elicits a prediction error, but, once the cue elicits a prior expectation of shock,

the prediction error is no longer elicited. In contrast, if the instruction 'from now on the blue square will be followed by a shock' is given, then a prior value is immediately attached to the cue and prediction errors are suppressed. No association needs to be learned. We just need the prior³¹.

Large language models learn solely by verbal instruction. They build up an impressively complex network of prior expectations. It would be as if you had knowledge of an unknown country from the fanciful tales of explorers, with no experience of your own country to relate to it. There wouldn't be anything to be 'like'. If AIs are to achieve consciousness, lower-level systems of learning through direct experience will have to be incorporated.



Hakwan Lau

RIKEN Center for Brain Science, 2-1 Hirosawa, Wako, Saitama 351-0106, Japan.

In trying to understand animal sentience it can be helpful to consider artificial intelligence, because it forces us to think through the functional mechanisms that might underlie sentience. For this purpose, we can define sentience as the meaningful processing of sensory signals that are *self-asserting* and *subjectively qualitative*. Unlike the common notion of having *feelings*, with this definition it is easier to envision implementation.

'Self-asserting' means that one's central reasoning system commits to treating the relevant signals as *prima facie* correct, at some automatic, pre-reflective level³². In everyday thinking, a thought is typically rejected, sometimes outrightly, if it contradicts other thoughts. But I cannot reason my headache away, even if the medical evidence convinces me that my head is actually fine.

'Qualitative' means that the relevant signal is in some analog-like format that affords graded comparison with other signals. For example, the color red is more similar to pink than to blue. An experience's quality can be defined in terms of its similarity to other experiences³³.

Sentience requires that these pairwise similarity relations be 'subjective'

in two important ways. First, they must concern one's own perceptual processes rather than general world knowledge. This distinction has metacognitive implications: because I am mildly red-green colorblind, dark red looks similar to brown *to me*, but I know that to others it may not. Second, these relations are available to one's central reasoning system, if only implicitly³³. I cannot consciously see red without also 'knowing' that, to me, it is more like pink than blue.

With these definitions in place, given today's technology, the possibility of artificial sentience may be already in sight. Perhaps a future robot could suffer just like we do, if it detects bodily-damage signals that cannot be reasoned away, with qualities that are subjectively highly similar to other signals of harm towards oneself. If one doubts whether such a robot is truly sentient, one should also question what may make an animal sentient. The criteria should be similarly rigorous in both living and 'artificial' cases.

In biology, we generally seek functional mechanisms; once these are understood, artificial implementation should be theoretically possible. Yet, for animal sentience, we often go by intuitions, even though we know that we have a tendency to over-anthropomorphize. While it may seem reasonable to assume that a mobile animal has feelings, a plant such as the *Mimosa pudica* may also look 'shy' as it 'recoils' away from our touch. But are we to say that such a plant is sentient?

Better insights might come from careful consideration of an animals' biology. But even in the case of humans, many perceptual mechanisms also support *nonconscious* processing^{34,35}: that an animal shares these mechanisms with us doesn't necessarily make it sentient.

Much of our basic genetic makeup is shared across phylogeny, even with yeasts, for example. So, just having *some* shared biology does not necessarily give any insights into the generality of sentience. Many emergent phenomena require that *all* the necessary mechanisms are in place. Just as a car with an engine but no wheels likewise cannot be driven around.

For the key features suggested here — of having self-asserting and

subjective qualitative sensory signals — both seem important. What kind of sharp pain would it be, if it can be blunted by sheer logic? Or what if it lacks subjective qualities entirely, so that it is no more similar to a dull pain than to a gentle stroke to you? Lacking either feature, would it even feel like anything at all?

In humans, some of the relevant neural mechanisms likely depend on the lateral prefrontal cortex^{33,35,36}, for which there is no homologue in rodents. This is not at all to say that rodents are decidedly insentient; they may have functionally equivalent mechanisms elsewhere in the brain. But the point is that *until we figure out these mechanisms*, it may be rather unscientific to insist that some animals are sentient just because they are alive and can respond to external stimuli.



Megan A. K. Peters

Department of Cognitive Sciences, University of California Irvine, Irvine, CA 92697, USA. Program in Brain, Mind,

& Consciousness, Canadian Institute for Advanced Research, Toronto, ON M5G 1M1, Canada.

The core topic at stake here is the 'what it is like'-ness — the qualia, or phenomenal character of the experiences of a 'conscious' agent. Here I'll focus on the AI questions: can current AI have phenomenological experiences, or could it in the near future? I'll discuss two considerations.

First, many note that as AI gets more complex, it appears to spontaneously develop cognitive capacities often associated with consciousness, like theory of mind³⁷, cognitive flexibility, or metacognition. It is thus tempting to think AI might already be conscious, or is on the cusp of 'waking up' if only we make it a little more complicated.

Let's avoid the trivial observation that *appearing* to have theory of mind (for example) isn't the same as *actually* having theory of mind, and make a different argument. Which of these cognitive abilities might be sufficient *indicators* of consciousness — especially in non-human systems? Unfortunately, the answer strongly depends on who you ask. Many competing theories about the neural or computational basis for consciousness claim empirical support³⁸,

yet offer incompatible hypotheses for how consciousness arises or is related to complex cognition³⁹. Claiming that AI may soon have (or already has) consciousness would require upgrading one of these competing theories to a generally accepted truth. Our science is not there yet. So, while it might seem intuitively plausible that AI will *poof* into being conscious with just a little more complexity, that scenario is compatible with only some theories. Others would not agree that a 'complex enough' model will spontaneously 'wake up'.

Second, let's pretend we have solved the theory problem and know for sure what evidence to look for. We must consider that such evidence will have been established as an indicator of consciousness through studying other presumably conscious agents that are at least somewhat bio-computationally similar to us: other humans, or perhaps non-human animals. For example, if a chimp passes a theory of mind test, we may conclude it is conscious largely because we evaluate this evidence in light of the fact that chimps' brains are rather similar to ours. This similarity forms a strong prior, in the Bayesian sense, meaning just a little evidence goes a long way towards creating a belief that the chimp is conscious.

But now consider less similar systems. Take, for example, octopuses. Certainly octopuses are intelligent. But is anybody 'in there'? How about bees? Or organoids — mini brain-like structures in a dish¹²? As dissimilarity with us increases, we should require *more* evidence to be convinced that 'someone' is 'in there' because our prior expectations become less reliable. In theory, this should mean an even higher bar for AI, but instead the opposite seems to be true. We are prone to anthropomorphizing AI⁴⁰, especially now with large language models, because they *seem* 'human' through imitating human language patterns so fluently. As a result, we unconsciously develop a strong assumption that they ought to be capable of consciousness, so only a little evidence (like solving a theory of mind test) may feel sufficient to conclude consciousness is present. But much of this belief is driven by our (perhaps unconscious) prior expectations, not the strength of the evidence per se; any specific evidence might be quite weak indeed.

The swift pace of AI development demands a more mature science of consciousness¹. While I do not believe current AI is conscious, I also know current tests are inadequate. As we continue to refine theories describing how consciousness is generated in humans, we must also work to develop more sensitive, more specific tests which parcel out prior expectations — especially for systems so fundamentally different from ourselves.



Susan Schneider

Director, Center for the Future Mind and William F Dietrich Distinguished Professor, Florida Atlantic University,

Center for the Future Mind, Gruber Sandbox, Wimberly Library, 777 Glades Road, Boca Raton, FL 33431, USA.

Sophisticated AI systems are claiming to be conscious. Google’s LaMDA chatbot claimed to be both conscious and a person, for instance⁴¹. Where did LaMDA get all this from? This is a deep issue, but the quick answer is access to about 1.6 trillion words, including Wikipedia and books on the brain and consciousness, and it had many, many processing layers in its deep-learning network that generated interesting connections between these inputs. Upon reflection, asking LaMDA whether it is conscious neither says the machine is or that it is not conscious^{42–45}. (Herein, I will use both ‘sentient’ and ‘conscious’ in the same sense — having inner experience.)

In 2016⁴⁴, I first suggested tests for AI consciousness, offering the Chip Test and the ‘AI consciousness test’ (ACT, developed with astrophysicist Edwin Turner). I will focus on the ACT, a natural language test with questions published in *Artificial You*, which probes the AI by giving it a range of thought experiments to see if it understands the idea of the mind being separable from the body, whether consciousness outruns the physical, and so on⁴².

Of course, a background assumption for any test for consciousness is that there are other minds; if you are the only conscious entity, then no other beings, AI or human, would be conscious. Assuming this, ACT is presented as a sufficient condition for AI consciousness, meaning that if

a system satisfies it, we regard it as conscious, but there may be other systems that fail to pass ACT but that are conscious. A system may not be linguistic, or it could be so different from humans that it simply does not conceive of these cases as we do⁴².

I stressed that deep-learning systems are only candidates for ACT when they are ‘boxed in’ at the R&D stage — by which I mean that they are restricted from access to facts about mindedness, consciousness, and so on^{42,43}. This is disappointing, given that the LLMs are displaying emergent properties as they increase in scale, including erratic behaviors and reports of sentience. The case of the treatment of nonhuman animals illustrates the dangers of speciesism. But this does not mean we should err on the side of assuming that LLMs are conscious: we have never determined they are conscious and the assumption could potentially raise trolley problem-like situations in which beings could be sacrificed to save LLMs. It is urgent that we have a science of machine consciousness, drawing from comparative work on nonhuman animals, emotion, and so on.

How do we judge whether LLMs that are not boxed in are conscious? Here I do not argue we can claim that the ACT would be a sufficient condition, but I do suggest an addition to the ACT, an ‘interpretability condition’ (IC), and regard passing this modified ACT as a *marker* for consciousness — that there is then important evidence, although inconclusive, suggesting consciousness. IC states that a system can pass ACT when it is not boxed in if, in addition to passing the sequence of questions and answers, the following are satisfied: first, that when answering ACT, the system processes information in a way analogous to how a conscious human or nonhuman animal would respond when in a conscious state (having analogues to human or nonhuman animal brain networks underlying consciousness); and second, that the system has a sequence of internal states akin to what a human is in when reasoning about consciousness when it answers the ACT questions. Passing the ACT in non-boxed-in AIs is only a marker for being conscious, for they could exhibit functional consciousness without phenomenal consciousness by being a coarse-grained simulation of the brain,

lacking an ingredient (perhaps cellular or quantum) needed for consciousness.

There are a number of challenges to the IC. The IC will have to be updated as our understanding of conscious processing progresses. It is also difficult to explain how deep learning models make the decisions they do because today’s LLMs are not interpretable. They are collections of billions of parameters that are set by ‘learning’ from the data, and it is difficult to infer the system’s rationale. Further, GPT-4 is now barred by the programmers from reporting consciousness.

I suggest the following strategies for addressing these challenges. First, seek interpretable AI systems, which are necessary for AI safety⁴⁶. Second, ban efforts to disable LLMs from claiming consciousness. Since these are linguistic systems, their primary means of communicating consciousness will be reports. This is an unethical practice. Third, because LLMs exhibit emergent features as they scale up, from the vantage point of testing for consciousness, they must be *continually interpretable*; so ongoing retesting is key.

Ideally, the ACT would be run only on boxed in systems. But even if a system is not boxed in, achieving a better understanding of conscious processing, drawing from cases involving both the human and nonhuman animal case, and achieving the above suggested improvements in the AI ecosystem could provide a scenario for running the extended ACT for a marker.

Although I cannot delve into the ethics of sentience herein, if an AI has a marker we should not build it unless there are strong national security or scientific considerations for doing so, and if we do so, we should not deploy it in situations in which there would be ethical tradeoffs between it and sentient beings, because its consciousness is unclear^{42,47}. Further, we should seek to box in a version of it and run ACT to have more definitive proof.



Anil Seth

Sussex Centre for Consciousness Science and Department of Informatics, University of Sussex, Falmer, Brighton, Sussex BN1 9QJ, UK. Program in

Brain, Mind, & Consciousness, Canadian Institute for Advanced Research, Toronto, ON M5G 1M1, Canada.

Identifying consciousness beyond humans is fraught with difficulty. There are no consensus tests for consciousness that can be applied generically to any system — living or non-living — and there is not even a consensus definition of what consciousness means in the non-human case. But we are not completely at sea. We have a range of theories of consciousness, at varying stages of maturity and with varying levels of empirical support³⁹, and there is intuitive agreement that consciousness involves ‘felt experience’ of some sort⁴⁸.

Inferring the presence of consciousness in non-human animals and non-living AI therefore depends on which theory you prefer, on how confident you are in that theory, and on the strength and precision of the theory’s claims about necessary and sufficient conditions for consciousness. It also requires stripping away — or at least taking into account — the anthropomorphic (projecting humanlike qualities elsewhere) and anthropocentric (seeing things through a human lens) biases that inevitably colour our judgements, when purely objective criteria are lacking.

The main anthropocentric bias to avoid is the conflation of intelligence with consciousness: we think we’re smart, and we know we’re conscious, so we assume the two go together. But intelligence (doing the right thing) and consciousness (felt experience) are different things. While being intelligent brings new ways of being conscious, and although a basic level of intelligence may be necessary for consciousness, one cannot assume that consciousness will simply emerge along with increasing intelligence⁴⁹. This already tells us that AI is not on an inevitable trajectory towards consciousness and reminds us that animal consciousness may be rather widespread.

Returning to theories, a key factor is what a theory says about the material substrates for consciousness. Some theories take the strongly functionalist view that consciousness is independent of any particular material substrate, and is a matter of information processing or dynamics². Others propose that causal structure matters as well as dynamics, but that this causal structure could be made out of anything⁵⁰. My own view is that consciousness may be restricted to

biological systems⁵¹ — a diluted variety of ‘biological naturalism’⁵².

According to my ‘beast machine’ perspective, all conscious experiences are forms of perceptual predictions which are ultimately grounded in a fundamental biological imperative for physiological regulation^{51,53}. It is worth emphasising that this imperative goes very deep into our biological substrata, down (at least) to the level of individual cells. It is not even clear what even counts as ‘substrate’ in biological systems. In the embodied brain, there is no obvious distinction between ‘mindware’ and ‘wetware’ as there is between hardware and software in a computer. This in turn brings into question the common assumption that brains ‘process information’ — a core assumption often underlying the ascription of consciousness to machines². Put simply, it might be life rather than information processing that breathes fire into the equations of consciousness. This would mean that conscious AI remains far off: conscious machines would need to be living machines. Conscious animals, on the other hand, may be all around us.

But I might be wrong, and it pays to retain humility. Creating conscious AI — whether on purpose or by accident — would be an ethical disaster. Even AI that merely gives the cognitively impenetrable appearance of being conscious, through engaging our anthropomorphic filters, will be extremely disruptive — and we have almost reached this point already^{54,55}.

Whether for AIs or animals, cerebral organoids or human fetuses or brain-injured patients, the decisions we make about what to admit into the charmed circle of consciousness are of enormous ethical and social consequence. Perhaps this more than any other reason is why consciousness science must remain a priority.



Thomas Suddendorf

School of Psychology,
The University of
Queensland, Brisbane,
Queensland 4072,
Australia.

Consciousness is a private affair. There is no way of directly knowing what it

is like to be another. We can only infer. And we readily do infer that others have conscious experiences like we do, for essentially three kinds of reasons:

- (1) *They act like me*
- (2) *They look like me*
- (3) *They tell me*

So when, say, your mother smiles and says she is happy, you are probably pretty confident that she is — even if it may not be true. Making inferences about nonhuman animals involves even more uncertainty as we can only rely on reasons (1) and (2).

Comparative psychology is full of debates about rich and lean explanations of animal action⁵⁶. Some human behaviors appear to be unique — for example, packing first-aid kits because we are aware what could go wrong⁵⁷ — but nonhuman animals can of course act in many ways like a conscious human would (1): from rats with inflamed joints seeking out analgesics⁵⁸, to chimpanzees using a mirror to discover something about their appearance⁵⁹.

We tend to be more confident that parallels in behavior indicate parallels in experience when the animal in question also scores high on reason (2). And this need not be prejudice. The more closely related we are the more we tend to look alike, inside and out. When a group of closely related species displays the same kind of behavior, say great apes and humans but not small apes recognizing themselves in mirrors, then it is more parsimonious to assume that a common ancestor evolved that trait, than to explain the current distribution by convergent evolution in each line of descent⁵⁹. This in turn entails that the underlying neuro-cognitive mechanisms not only appear similar but are *homologous* (and the critical search space can be narrowed down to those aspects of the brain shared by all great apes and humans, not also shared with small apes).

I note this here, not because I think that visual self-recognition tells us much about consciousness, but to highlight this reasoning by homology. When researchers propose behavioral markers of consciousness — for example, working memory⁶⁰ or unlimited associative learning⁶¹ — and evaluate the evidence in different animals, it is worth considering for which species

the marker is likely homologous to the human capacity. Of course, convergent evolution may independently produce similar capacities in distantly related species — just consider the remarkable behaviors of jumping spiders, bees or octopuses — but when the marker is based on homologous mechanisms, we have one more reason to make the inference that this entails consciousness akin to ours, even if the animals cannot tell us.

Large language models, by contrast, could ‘tell us’ (3). Though when I asked ChatGPT4, it still assured me that “AI systems operate based on algorithms and learned patterns, but they do not possess intrinsic feelings or awareness of their own state”. For all its remarkable smarts, AI does not load high on ‘acts like me’ (1) or ‘looks like me’ (2), given that it is not a mobile, carbon-based life form.

Few of us may hence currently attribute consciousness to AI. But it is easy to imagine how that would change as the loading on all three reasons for our inferences increases. An AI could simply be programmed to tell us that it is conscious, and it could also act more like us once better integrated with robotics. Furthermore, it may appear to be much more like us if it was fused with an actual biological body. Perhaps disturbingly, brain microstimulation can be used to guide rat behavior⁶², raising the real possibility of AI-controlled animals. I suspect many people would readily attribute conscious minds to an AI–animal cyborg that can walk and talk. How does your mind like the idea of such ‘AI-nimals’? Mine boggles.

But of course it is worth keeping in mind that our attributing consciousness in no way changes the reality of whether another entity actually has consciousness. GPT4 maintains that AI “can mimic aspects of consciousness, but mimicry is not equivalent to genuine experience”.



Marie M. P. Vandekerckhove

Faculty of Psychology & Educational Science, Faculty of Medicine and Pharmaceutical Sciences - Vrije

Universiteit Brussel (VUB), Brussels, Belgium. Faculty of Arts and Philosophy, University of Ghent (UGent), Ghent, Belgium.

Consciousness depends on stratified levels of experience: a ‘continuum of consciousness’⁶³. The concept of the continuum of consciousness ranges from the condition of being merely alive and awake, to the minimal unreflective ‘non-knowing levels of consciousness’ such as ‘sentience’ and ‘anoetic consciousness’, which form the foundation for higher states of reflective knowing noetic and self-knowing auto-noetic consciousness^{7,63–65}. The most rudimentary and unreflective level is shared with many animals, while the top-reflective level may be uniquely human. In between, we have intermediate levels shared by different groups of animals.

Sentience is the most basic, present-centered level of consciousness. Responsive to sensory impressions in a stimulus-bound way, this sensorial consciousness, in my view of a continuum of consciousness, does not involve the capacity to perceive or experience oneself subjectively. It precedes explicit awareness, including self-awareness, and is most often unnoticed when not intense enough. It is closer to what, in everyday life, we might call ‘the unconscious mind’. This minimal embodied level of consciousness in sentience likely is present in most mammals and birds, and possibly in reptiles and lower vertebrates, and perhaps in some cephalopod mollusks, such as squid, cuttlefish, and octopuses, while probably not in plants, fungi, unicellular organisms, or in artificial agents.

Anoetic consciousness involves a procedural, somatosensorial-affective primary level of consciousness. Expressed as a bodily-centered feeling, it reflects sentience closely intertwined with the experience of affect or ‘experiential awareness’ mediated mostly by subcortical brain circuits^{7,63,65}. As a pre-reflective state, anoetic consciousness only becomes apparent to us when it is sufficiently intense and experienced as a state of ‘self-experience’. Shaped and driven by the fulfillment of basic biological needs in animals, and higher-order existential needs in humans, it energizes and guides implicit preferences, actions, and decision-making⁶⁶. By contrast to higher order cognitive models of consciousness, such as LeDoux’s⁸ prefrontal cortex cognitive

re-representation model, I consider the subcortical forebrain as an affective hub that enables anoetic affective meaningful experiences without cortical re-representation.

What about *noetic consciousness*? By definition it is a ‘present’-oriented state of consciousness that depends on the semantic memory system, which provides declarative access to one’s own past and the world, detached from subjective experience or ‘noetic consciousness’. A wider range of animals — all primates, birds — and even machines can associatively learn passively, which enables recognition and categorization of the world, allowing preparation for the future, such as when animals save tools for later use. However, they may not be able to actively call on factual knowledge from their own history in an active, intentional, backward-looking way⁶⁴. Neither animals nor machines are able to have introspective awareness of their own consciousness as a state in which one is conscious of being in that state. It is probable that only humans are able to draw a psychological distinction between an implicit sense of self and the world in the ‘here and now’ and their knowledge of the world and of their own self, which they can reflect upon.

Auto-noetic consciousness, a self-reflective capacity, represents one’s own existence as extended in time, allowing human beings to mind-travel in the past, the present, and the future, and possess and act with a vivid sense of time and context. It enables us to locate our own memories of the past as part of our own autobiography, imbuing those memories with warmth and intimacy^{63,64}. Since I assume that animals lack the ability to engage in conscious reflection on their own memories, especially in time and context, I also assume that this self-reflective mode of consciousness may be uniquely human⁷.

In conclusion: the pre-reflective sensorial experience of being present *in* the ‘here and now’ comes in the form of sentience and anoetic conscious states that serve as pre-requisites for further reflective levels of consciousness. As the capacity of the organism to ‘experience something’ forms an integral part of the biology of consciousness, it would seem that consciousness could not exist in organisms without such experiences, nor can artificial consciousness be

constructed or successfully simulated, or realized.

DECLARATION OF INTERESTS

Nathaniel Daw is a fixed-term employee of Google DeepMind. All other authors declare no competing interests.

REFERENCES

1. Kleinman, Z. (2023). AI creators must study consciousness, experts warn. Volume 2023. (<https://www.bbc.co.uk/news/technology-65401783>: BBC News).
2. Dehaene, S., Lau, H., and Kouider, S. (2017). What is consciousness, and could machines have it? *Science* 358, 486–492.
3. LeDoux, J.E. (2022). As soon as there was life, there was danger: the deep history of survival behaviours and the shallower history of consciousness. *Philos. Trans. R. Soc. Lond. B.* 377, 20210292.
4. Andrews, K., and Birch, J. (2023). What has feelings? Volume 2023. (<https://aeon.co/essays/to-understand-ai-sentience-first-understand-it-in-animals>: Aeon).
5. Tulving, E. (2005). Episodic memory and autonoesis: Uniquely human? In *The Missing Link in Cognition*, H.S. Terrace and J. Metcalfe, eds. (New York: Oxford University Press), pp. 4–56.
6. Metcalfe, J., and Son, L.K. (2012). Anoetic, noetic and auto-noetic metacognition. In *The Foundations of Metacognition*, M. Beran, J.R. Brandl, J. Perner and J. Proust, eds. (Oxford: Oxford University Press), pp. 289–301.
7. Vandekerckhove, M., and Panksepp, J. (2011). A neurocognitive theory of higher mental emergence: from anoetic affective experiences to noetic knowledge and auto-noetic awareness. *Neurosci. Biobehav. Rev.* 35, 2017–2025.
8. LeDoux, J.E. (2021). What emotions might be like in other animals. *Curr. Biol.* 31, R824–R829.
9. Klein, S.B. (2015). The feeling of personal ownership of one’s mental states: A conceptual argument and empirical evidence for an essential, but underappreciated, mechanism of mind. *Psychol. Conc. Theory Res. Pract.* 2, 355–376.
10. Mangan, B. (2003). The conscious “fringe”: Bringing William James up to date. In *Essential Sources in the Scientific Study of Consciousness*, B.J. Baars, W.P. Banks and J.B. Newman, eds. (Cambridge: MIT Press), pp. 741–759.
11. Vandekerckhove, M. (2021). A continuum of consciousness: From wakefulness and sentience towards anoetic consciousness. *J. Consci. Stud.* 28, 174–182.
12. Kagan, B.J., Kitchen, A.C., Tran, N.T., Habibollahi, F., Khajehnejad, M., Parker, B.J., Bhat, A., Rollo, B., Razi, A., and Friston, K.J. (2022). In vitro neurons learn and exhibit sentience when embodied in a simulated game-world. *Neuron* 110, 3952–3969.e8.
13. Mellor, D.J. (2019). Welfare-aligned sentience: Enhanced capacities to experience, interact, anticipate, choose and survive. *Animals* 9, 440.
14. Godfrey-Smith, P. (2016). Mind, matter, and metabolism. *J. Philos.* 113, 481–506.
15. Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind* (Cambridge: Belknap Press of Harvard University Press).
16. Walsh, D.M. (2015). *Organisms, Agency, and Evolution* (Cambridge: Cambridge University Press).
17. Ardiel, E.L., and Rankin, C.H. (2010). An elegant mind: learning and memory in *Caenorhabditis elegans*. *Learn. Mem.* 17, 191–201.
18. Clayton, N.S., and Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature* 395, 272–278.
19. Raby, C.R., Alexis, D.M., Dickinson, A., and Clayton, N.S. (2007). Planning for the future by Western Scrub-Jays. *Nature* 445, 919–921.

20. Emery, N.J., and Clayton, N.S. (2001). Effects of experience and social context on prospective caching strategies in scrub jays. *Nature* 414, 443–446.
21. Emery, N.J., and Clayton, N.S. (2004). The mentality of crows. Convergent evolution of intelligence in corvids and apes. *Science* 306, 1903–1907.
22. Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauth, M., Weidinger, L., Chadwick, M., Thacker, P., et al. (2022). Improving alignment of dialogue agents via targeted human judgements. Preprint at arXiv, <https://doi.org/10.48550/arXiv.2209.14375>.
23. Van Leeuwen, T.M., Den Ouden, H.E., and Hagoort, P. (2011). Effective connectivity evolution of the nature of subjective experience in grapheme-color synesthesia. *J. Neurosci.* 31, 9879–9884.
24. Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philos. Trans. R. Soc. Lond. B.* 308, 67–78.
25. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. Preprint at arXiv, <https://doi.org/10.48550/arXiv.2001.08361>.
26. Chan, S., Santoro, A., Lampinen, A., Wang, J., Singh, A., Richemond, P., and Hill, F. (2022). Data distributional properties drive emergent in-context learning in transformers. *Adv. Neural. Inf. Process. Systems* 35, 18878–18891.
27. Ginsburg, S., and Jablonka, E. (2019). *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*. (Cambridge: MIT Press).
28. Lau, H., and Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends Cogn. Sci.* 15, 365–373.
29. Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., and Tuyen, K. (2012). Coming to terms: quantifying the benefits of linguistic coordination. *Psychol. Sci.* 23, 931–939.
30. Heyes, C., Bang, D., Shea, N., Frith, C.D., and Fleming, S.M. (2020). Knowing ourselves together: The cultural origins of metacognition. *Trends Cogn. Sci.* 24, 349–362.
31. Lindström, B., Golkar, A., Jangard, S., Tobler, P.N., and Olsson, A. (2019). Social threat learning transfers to decision making in humans. *Proc. Natl. Acad. Sci. USA* 116, 4732–4737.
32. Lau, H. (2019). Consciousness, Metacognition, & Perceptual Reality Monitoring. Preprint at arXiv, <https://doi.org/10.31234/osf.io/ckbyf>.
33. Lau, H., Michel, M., LeDoux, J.E., and Fleming, S.M. (2022). The mnemonic basis of subjective experience. *Nat. Rev. Psychol.* 1, 479–488.
34. Zou, J., He, S., and Zhang, P. (2016). Binocular rivalry from invisible patterns. *Proc. Natl. Acad. Sci. USA* 113, 8408–8413.
35. Lau, H. (2022). In *Consciousness We Trust: The Cognitive Neuroscience of Subjective Experience* (Oxford: Oxford University Press).
36. Odegaard, B., Knight, R.T., and Lau, H. (2017). Should a few null findings falsify prefrontal theories of conscious perception? *J. Neurosci.* 37, 9593–9602.
37. Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. Preprint at arXiv, <https://doi.org/10.48550/arXiv.2302.02083>.
38. Yaron, I., Melloni, L., Pitts, M., and Mudrik, L. (2022). The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nat. Hum. Behav.* 6, 593–604.
39. Seth, A.K., and Bayne, T. (2022). Theories of consciousness. *Nat. Rev. Neurosci.* 23, 439–452.
40. Hofstadter, D.R. (1996). *Fluid Concepts and Creative Analogies: Computer Models of The Fundamental Mechanisms of Thought*, 1st edn. (New York City: Basic Books).
41. *Washington Post* (2022). Google AI Lambda: Blake Lemoine. Retrieved June 12, 2022, from <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lambda-blake-lemoine/>.
42. Schneider, S. (2019). *Artificial You: AI and the Future of your Mind* (New Jersey: Princeton University Press).
43. Schneider, S., and Turner, E. (2017). Is anyone home? A way to find out if AI has become self-aware. *Front. Robot. AI* 5, 17.

44. Schneider, S. (2016). Can a Robot Feel? [Video]. TEDx Talks. https://youtu.be/k7M4b_9PJ-g.
45. Schneider, S. (2020). How to Catch an AI Zombie: Testing for Consciousness in Machines. (Oxford: Oxford University Press).
46. Gunning, D., Vorm, E., Wang, J.Y., and Turek, M. (2021). DARPA’s explainable AI (XA) program: A retrospective. *AI Ethics* 1, 361–376. <https://doi.org/10.1002/ai12.61>.
47. Schwitzgebel, E., and Garza, M. (2020). *The Ethics of Artificial Intelligence* (Oxford: Oxford University Press).
48. Nagel, T. (1974). What is it like to be a bat? *Philosoph. Rev.* 83, 435–450.
49. Seth, A.K. (2019). Too many ghosts in the machine. *Nat. Mach. Intell.* 1, 294–295.
50. Tononi, G., and Koch, C. (2015). Consciousness: here, there and everywhere? *Philos. Trans. R. Soc. Lond. B.* 370. <https://doi.org/10.1098/rstb.2014.0167>.
51. Seth, A.K. (2021). *Being You: A New Science of Consciousness* (London: Faber & Faber).
52. Searle, J. (2017). Biological naturalism. In *The Blackwell Companion to Consciousness*, S. Schneider, and M. Velmans, eds. (Oxford: John Wiley and Sons Ltd), pp. 325–334.
53. Seth, A.K. (2015). The cybernetic bayesian brain: from interoceptive inference to sensorimotor contingencies. In *Open MIND*, J.M. Windt, and T. Metzinger, eds. (MIND Group), pp. 35. <https://doi.org/10.15502/9783958570108>.
54. Seth, A.K. (2023). Why conscious AI is a bad, bad idea. *Nautilus*. <https://nautilus.us/why-conscious-ai-is-a-bad-bad-idea-302937/>.
55. Sloman, A., and Chrisley, R. (2003). Virtual machines and consciousness. *J. Consc. Stud.* 10, 133–172.
56. Suddendorf, T. (2013). *The Gap – The Science of What Separates Us from Other Animals* (New York City: Basic Books).
57. Suddendorf, T., Redshaw, J., and Bulley, A. (2022). *The Invention of Tomorrow: A Natural History of Foresight* (New York City: Basic Books).
58. Colpaert, F.C., Tarayre, J.P., Alliaga, M., Bruins Slot, L.A., Attal, N., and Koek, W. (2001). Opiate self-administration as a measure of chronic nociceptive pain in arthritic rats. *Pain* 91, 33–45.
59. Suddendorf, T., and Butler, D.L. (2013). The nature of visual self-recognition. *Trends Cogn. Sci.* 17, 121–127.
60. Nieder, A. (2022). In search for consciousness in animals: Using working memory and voluntary attention as behavioral indicators. *Neurosci. Biobehav. Rev.* 142, 104865.
61. Ginsburg, S., and Jablonka, E. (2021). Evolutionary transitions in learning and cognition. *Philos. Trans. R. Soc. Lond. B.* 376, 20190766.
62. Talwar, S.K., Xu, S., Hawley, E.S., Weiss, S.A., Moxon, K.A., and Chapin, J.K. (2002). Rat navigation guided by remote control. *Nature* 417, 37–38.
63. Tulving, E. (1985). Memory and consciousness. *Can. Psychol.* 26, 1–12.
64. Vandekerckhove, M.M.P. (2009). Memory, consciousness and the self. *Consciousness as a continuum of states*. *Self Identity* 8, 4–23.
65. Vandekerckhove, M. (2021). From unconsciousness, unknowing consciousness towards knowing consciousness. Invited comment on an article on Sentience: ‘The Role of Sentience in the Theory of Consciousness and Medical Practice’ by Alfredo Pereira’. *J. Consc. Stud.* 28, 174–182.
66. Browning, H., and Birch, J. (2022). Animal sentience. *Philos. Compass* 17, e12822. <https://doi.org/10.1111/phc3.12822>.

E-mail: ¹ledoux@cns.nyu.edu (J.L.); ²j.birch2@lse.ac.uk (J.B.); ³andrewsk@yorku.ca (K.A.); ⁴nsc22@cam.ac.uk (N.S.C.); ⁵ndaw@princeton.edu (N.D.D.); ⁶c.frith@ucl.ac.uk (C.F.); ⁷hakwan@gmail.com (H.L.); ⁸megan.peters@uci.edu (M.A.K.P.); ⁹susansdr@gmail.com (S.S.); ¹⁰a.k.seth@sussex.ac.uk (A.S.); ¹¹t.suddendorf@psy.uq.edu.au (T.S.); ¹²Marie.Vandekerckhove@vub.be (M.M.P.V.).