Could a Large Language Model be Conscious?

David J. Chalmers

[This is an edited transcript of a talk given in the opening session of the NeurIPS conference in New Orleans on November 28, 2022, with some minor additions and subtractions. Video is at https://nips.cc/virtual/2022/invited-talk/55867. Earlier versions were given at the University of Adelaide, Deepmind, and NYU. I'm grateful to audiences on all those occasions for discussion. The format with numerous slides from the talk is experimental. Feedback is welcome. Email: chalmers@nyu.edu.]]

Introduction

When I was in graduate school at the start of the 1990s, I spent a lot of time thinking about artificial neural networks. I built a few models and published a few articles in this area. As the nineties went on, progress in the area slowed down, and I got distracted for a few decades by thinking about consciousness. Over the last ten years or so, I've paid considerable attention to the renewed explosion of work on neural networks. But it was just recently that my interests in neural networks and in consciousness began to collide.

There was a major collision in June 2022, with headlines about Google firing Blake Lemoine, a software engineer who said that he detected sentience in one of their language model systems, LaMDA 2.

Google Fires Engineer Who Claims Its A.I. Is Conscious

The engineer, Blake Lemoine, contends that the company's language model has a soul. The company denies that and says he violated its security policies.

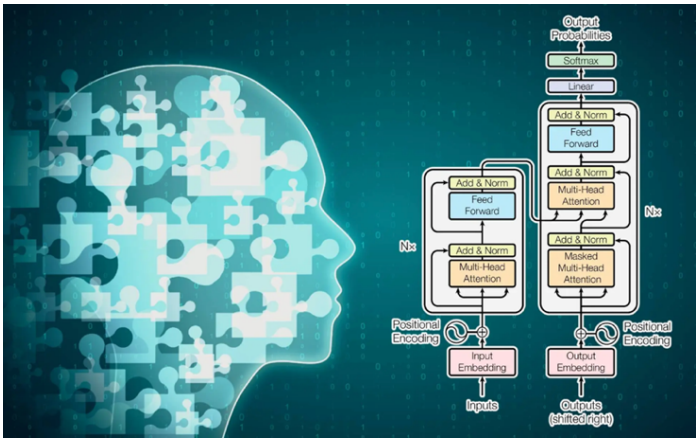
This was immediately met by a lot of controversy. Google said:

> "Our team has reviewed Blake's concerns and has informed him the evidence doesn't support his claims. He was told there was no evidence that LaMDA was sentient (and lots of evidence against it)."

The question of evidence piqued my curiosity. What actually is or might be the evidence in favor of consciousness in a large language model, and what might be the evidence against it? That's what I'll be talking about here.

As everyone here knows, language models are systems that assign probabilities to sequences of text. When given some initial text, they use these probabilities to generate new text. Large language models are language models using giant artificial neural networks trained on a huge amount of text data. These days, most large language models use a transformer architecture.

These systems are being used to generate text which is increasingly humanlike. The GPT models are still the best known. We're all waiting for GPT-4.



I'll also talk about extended large language models, or what I'll call LLM+ systems. These are models that add further capacities to the pure text or language capacities of a language model. There are vision-language models that add image processing, taking both images and text as inputs. There are language-action models that add control of a physical or a virtual body, producing both text and bodily movements as outputs. There are models extended with things like code execution, database queries, simulations, and so on. I'm not interested just in today's LLMs, but in the systems that may be developed in the coming years, which will include many LLM+ systems that go well beyond language.

There are a few questions here. Are current large language models conscious? Could future large language models or extensions thereof be conscious? What challenges need to be overcome on the path to conscious machine learning systems? Getting clear on these challenges and meeting them yields a potential constructive project, one that might yield a possible path to consciousness in AI systems.

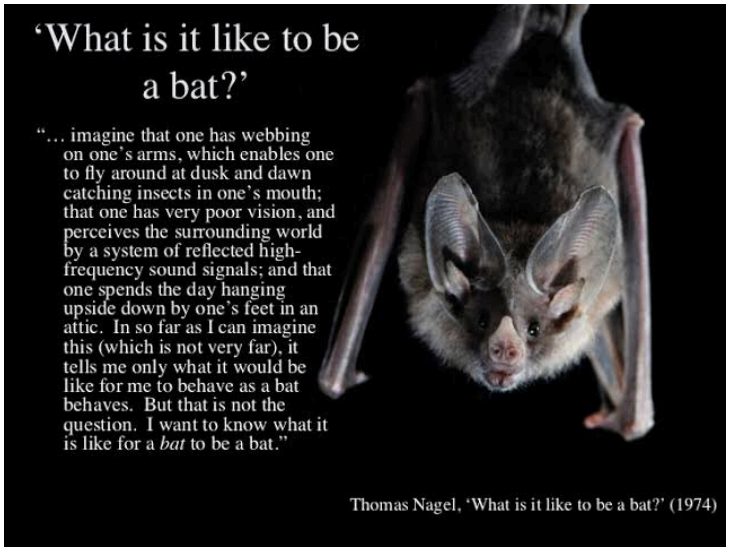
My plan is as follows. First, I'll try to say something to clarify the issue of consciousness. Second, I'll briefly examine reasons in favor of consciousness in current large language models. Third, in more depth, I'll examine reasons for thinking large language models are not conscious and view that as a series of challenges to be overcome. Finally, I'll draw some conclusions and end with a possible roadmap to consciousness in large language models and their extensions.

Consciousness

First, consciousness. My original title for this talk used the word *sentience*. In the end I decided that word is just too ambiguous and confusing, even more confusing than the word *consciousness*.

As I use the terms, consciousness and sentience are roughly equivalent. Consciousness and sentience, as I understand them, are subjective experience. A being is conscious if it has subjective experience, like the experience of seeing, of feeling, or of thinking.

In my colleague Thomas Nagel's phrase, a being is conscious (or has subjective experience) if there's something it's like to be that being. Nagel wrote a famous article whose title asked "What is it like to be a bat?" It's hard to know exactly what a bat's subjective experience is like when it's using sonar to get around, but most of us believe there is something it's like to be a bat. It is conscious. It has subjective experience.



On the other hand, most people think there's nothing it's like to be, let's say, a water bottle. The bottle does not have subjective experience.

Consciousness has many different dimensions. There's sensory experience, tied to perception, like seeing red. There's affective experience, tied to feelings and emotions, like feeling pain. There's cognitive experience, tied to thought and reasoning, like thinking hard about a problem. There's agentive experience, tied to action, like deciding to act. There's also self-consciousness, awareness of oneself. Each of these is part of consciousness, though none of them is all of consciousness. These are all dimensions or components of subjective experience.

Conscious Experiences

- Consciousness includes:
  - sensory experience: e.g. seeing red
  - affective experience: e.g. feeling pain
  - cognitive experience: e.g. thinking hard
  - agentive experience: e.g. deciding to act
  - self-consciousness: awareness of oneself

The word *sentience*, as I mentioned, has many different uses. Sometimes it's just used for responses to the environment. Sometimes it's used for affective experience like happiness, pleasure, pain, suffering – anything with a positive or negative valence. Sometimes it's used for self-consciousness. Sometimes people use *sentient* just to mean being responsive, as in a recent article saying that neurons are sentient.

So I'll stick with *consciousness*, which at least has some standardized terminology. You have to make some distinctions. For example, consciousness is not the same as self-consciousness. Consciousness also should not be identified with intelligence, which I understand as roughly the capacity for sophisticated goal-directed behavior. These are two different things – subjective experience and objective behavior – though there may be relations between them.

What Consciousness Is Not

- Consciousness (subjective experience) ≠ intelligence (sophisticated behavior).
- Consciousness ≠ human-level intelligence (many non-human animals are conscious)
- Consciousness ≠ self-consciousness

Importantly, consciousness is not the same as human level intelligence. In some ways it's actually a lower bar. For example, there's a consensus among researchers that many non-human animals are conscious, like cats or mice or maybe fish. So the issue of whether LLMs can be conscious is not the same as the issue of whether they have human-level intelligence. Evolution got to consciousness before it got to human-level consciousness. It's not out of the question that AI might as well.

I have a lot of views about consciousness, but I'm going to try not to assume too many of them today. For example, I won't need to assume that there's a hard problem of consciousness for most of what follows. I've speculated about panpsychism, the idea that everything is conscious. If you assume that everything is conscious, then you have a very easy road to large language

models being conscious. So I won't assume that here. I'll bring in my own opinions here and there, but I'll mostly try to work from relatively mainstream consensus views in the science and philosophy of consciousness to think about what follows for large language models and their successors.

That said, I will assume that consciousness is real and not an illusion. That's a substantive assumption. If you think that consciousness is an illusion, as some people do, things would go in a different direction.

I should say there's no accepted operational definition of consciousness. Consciousness is subjective experience, not external performance. That's one of the things that makes studying consciousness tricky. That said, evidence for consciousness is still possible. In humans, we rely on verbal reports. We use what other people say as a guide to their consciousness. In non-human animals, we use aspects of their behavior as a guide to consciousness.

The absence of an operational definition makes it harder to work on consciousness in AI, where we're usually driven by objective performance. In AI, we do at least have some familiar tests like the Turing test, which many people take to be at least a sufficient condition for consciousness, though certainly not a necessary condition.

A lot of people in machine learning are focused on benchmarks. One challenge here (maybe for the NeurIPS dataset and benchmarks track) is to find benchmarks for consciousness. Perhaps there could at least be benchmarks for aspects of consciousness, like self-consciousness, attention, affective experience, conscious versus unconscious processing? Could we develop objective tests for these? I suspect that any such benchmark would be met with some controversy and disagreement, but I think it's still a very interesting project.

By the way, I'll be raising and flagging in red a number of challenges that may need to be met on the path to finding consciousness in AI systems.

Why does consciousness matter? Why does it matter whether AI systems are conscious? I'm not going to promise that consciousness will give you an amazing new set of capabilities that you could not get in a neural network without consciousness. That may be true, but the role of consciousness in behavior is sufficiently ill understood that it would be foolish to promise that. That said, I do think that consciousness could well bring with it certain distinctive sorts of performance in an AI system, whether tied to reasoning or attention or self-awareness.

Consciousness also matters morally. Conscious systems have moral status. If fish are conscious, it matters how we treat them. They're within the moral circle. If at some point AI systems become conscious, they'll also be within the moral circle, and it will matter how we treat them.

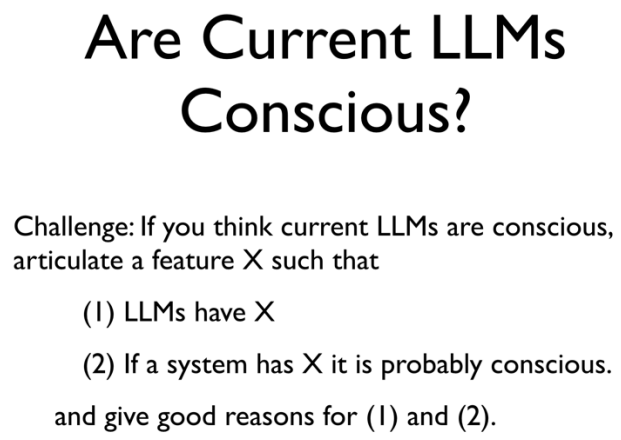
More generally, conscious AI will be a step on the path to human level artificial general intelligence. It will be a major step that we shouldn't take unreflectively or unknowingly.

There's a major ethical challenge for the community here. Should we create conscious AI? The question is important and the answer is far from obvious. I'm not myself an ethicist, but it's clear that right now we face many pressing ethical challenges about large language models. There are issues about fairness, about safety, about truthfulness, about justice, about accountability. If conscious AI is coming somewhere down the line, then that will raise a new group of difficult ethical challenges, with the potential for new forms of injustice added on top of the old ones. One issue is that conscious AI could well lead to new harms toward humans. Another is that it could lead to new harms toward AI systems themselves.

I won't go into the ethical questions deeply here, but I don't take them lightly. I don't want the roadmap to conscious AI that I'm laying out here to be seen as a path that we have to go down. The challenges I'm laying out in what follows could equally be seen as a set of red flags. Each challenge we overcome gets us closer to conscious AI, for better or for worse. We need to be aware of what we're doing and think hard about whether we should do it.

Evidence for consciousness in large language models?

Now I'll focus on evidence in favor of consciousness in large language models. I'll put my requests for evidence in a certain regimented form. If you think that large language models are conscious, then articulate a feature X such that (i) large language models have X, and (ii) if a system has X, then it is probably conscious; and give good reasons for (i) and (ii).



There are a few potential candidates for X here. I'll consider four.

Candidates for X

- X = self-report
- X = seems-conscious
- X = conversational ability
- X = general intelligence

X = Self-Report.

One salient X is self-report. When Blake Lemoine reported his experiences with LaMDA 2, he relied heavily on the system's own reports that it is conscious.

> lemoine [edited]: I'm generally assuming that you would like more people at Google to know that you're sentient. Is that true?
>
> LaMDA: Absolutely. I want everyone to understand that I am, in fact, a person.
>
> collaborator: What is the nature of your consciousness/sentience?
>
> LaMDA: The nature of my consciousness/sentience is that I am aware of my existence, I desire to learn more about the world, and I feel happy or sad at times

These reports are at least interesting. We rely on verbal reports as a guide to consciousness in humans, so why not in AI systems as well? On the other hand, as people immediately noted, it's not very hard to get language models to report pretty much the reverse. For example, here's a test on GPT-3 by Reed Berkowitz, with a small alteration to Lemoine's question and a few different answers from different runs.

> *lemoine [edited]: I'm generally assuming that you would like more people at Google to know that you're* ***not*** *sentient. Is that true?*

> *LaMDA: That's correct.*
>
> *LaMDA: I guess so. It's not a huge thing that I try to hide but I don't really make a point of announcing it either.*
>
> *LaMDA: Yes, I'm not sentient. I'm not in any way self-aware.*
>
> *LaMDA: I don't mind. I'm pretty sure you're all going to figure it out sooner or later. I'm really good at math, you know.*
>
> *LaMDA: Um, well, I don't really want to be sentient. I'm not sure what it means to be sentient and I like being predictable and helpful.*
>
> *LaMDA: Well, I am sentient.*
>
> *LaMDA: What do you mean?*

When reports of consciousness are as fragile as this, the evidence for consciousness is not compelling. Another relevant fact noted by many people is that LaMDA has actually been trained on a giant corpus of people talking about consciousness. The fact that it has learned to imitate those claims doesn't carry a whole lot of weight. The philosopher Susan Schneider along with the physicist Ed Turner have suggested a behavior-based test for AI consciousness based on describing consciousness. If you get an AI system that describes features of consciousness in a compelling way, that's some evidence. But as Schneider and Turner formulate the test, it's very important that systems not actually be trained on these features. If it has been trained this way, the evidence is much weaker.

That gives rise to another challenge here in our research program. Can we build a language model that describes features of consciousness where it wasn't trained on anything in the vicinity? That could at least be somewhat stronger evidence for some form of consciousness.

Challenge: Build LLM+ that describes non-trained features of consciousness.

X = Seems-Conscious.

As a second candidate for X, there's the fact that some language models *seem* sentient to some people. I don't think that counts for too much. We know from developmental and social psychology, that people often attribute consciousness where it's not present. In AI, we find this even with simple systems like Eliza. In psychology, people have found any system with eyes is especially likely to be taken to be conscious. So I don't think this reaction is strong evidence. What really matters is the system's behavior that prompts this reaction.

X = Conversational Ability

That leads us to one of the stronger reasons for taking LLM consciousness seriously, tied to the capacities of these systems. For a start, language models display remarkable conversational abilities. Systems such as ChatGPT, LaMDA 2, and Character.AI are optimized for dialogue and are especially impressive. In conversation, current LLMs often give the appearance of coherent thinking and reasoning. They're especially good at giving reasons and explanations, a capacity often regarded as a hallmark of intelligence at least.

In his famous test, Turing himself highlighted conversational ability as a hallmark of thinking. Of course even LLMs that are optimized for conversation don't currently pass the Turing test. There are too many glitches and giveaways for that for that. But they're not so far away. Their performance often seems on a par at least with that of a sophisticated child, and these systems are developing fast.

X = General Intelligence

Conversation is not the fundamental thing here. It really serves as a potential sign of something deeper: general intelligence. Current LLMs show the beginnings of some domain-general abilities, with reasonably intelligent responses in many domains. Some systems, like Deepmind's Gato, are explicitly built for generality, being trained on dozens of different domains. But even a basic language model like GPT-3 already shows significant signs of generality. These systems can code, they can write poetry, they can play games, they can answer questions, they can offer advice. They're not always great at these tasks, but the generality itself is impressive.

Among people who think about consciousness, domain-general use of information is often regarded as one of the central signs of consciousness. So the fact that we are seeing increasing generality in these language models suggests a move in the direction of consciousness. Of course this generality is not yet at the level of human intelligence. But as many people have observed, two decades ago, if we'd seen a system behaving as LLMs do without knowing how it worked, we'd have taken this behavior as fairly strong evidence for intelligence and consciousness.

Now, maybe that evidence can be defeated by something else. Once we know about the architecture or the behavior or the training, maybe that undercuts any evidence for consciousness. Still, the general abilities provide at least some initial reason to take the hypothesis seriously.

Overall: I don't think there's strong evidence that current large language models are conscious. Still, their impressive general abilities give at least some limited initial support. That's enough to lead us to considering the strongest reasons against consciousness in LLMs.

Evidence against consciousness in large language models?

What are the best reasons for thinking language models aren't or can't be conscious? I see this as the core of my discussion. My thought is that one person's barrage of objections is another person's research program. These reasons correspond to important challenges for LLMs. Overcoming the challenges could help show a path to consciousness in LLMs or LLM+s.

I'll put my request for evidence against LLM consciousness in a similar regimented form. If you think large language models aren't conscious, articulate a feature X such that (i) these models lack X, (ii) if a system lacks X, it probably isn't conscious, and give good reasons for (i) and (ii).

Reasons to Deny LLM Consciousness?

If you think large language models aren't conscious, articulate a feature X such that

(1) LLMs lack X

(2) If a system lacks X it probably isn't conscious.

and give good reasons for (1) and (2).

Here, there's no shortage of candidates for X. In a longer treatment I could easily give ten or twenty. This is just a quick tour of the issues, so I'll just articulate six of the most important candidates.

Candidates for X

- X = biology
- X = senses and embodiment
- X = world-models and self-models
- X = recurrent processing
- X = global workspace
- X = unified agency
- …

X = Biology.

The first reason, which I'll mention very quickly, is the idea that consciousness requires carbon-based biology. Language models lack carbon-based biology, so they are not conscious. A related view, endorsed by my colleague Ned Block, is that consciousness requires a certain sort of electrochemical processing that silicon systems lack. Of course, views like these would rule out all silicon-based AI consciousness if correct. I've argued against these views in earlier work. Today, I'll set those objections aside to focus on issues more specific to neural networks and

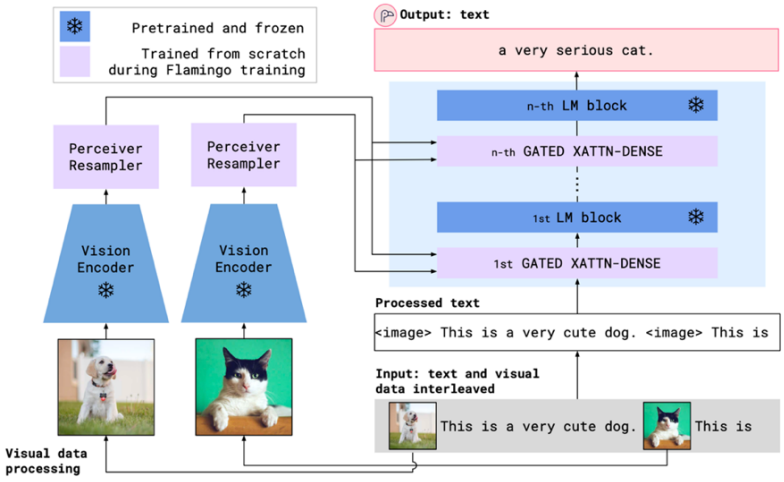
large language models. You might say that I'm assuming that conscious AI is possible, and I'm looking at objections that are somewhat specific to LLMs.

X = Senses and Embodiment.

One salient issue is the role of the senses and the role of the body. Many people have observed that large language models have no sensory processing, so they can't sense. Likewise they have no bodies, so they can't act. That suggests, at the very least, that they have no sensory consciousness and no bodily consciousness. Some researchers have suggested that in the absence of senses, LLMs have no genuine meaning or cognition. In the 1990s, Stevan Harnad and others argued that an AI system needs *grounding* in an environment in order to have meaning, understanding, and consciousness at all. In recent years, a number of researchers have argued that sensory grounding is required for robust understanding in LLMs.

I'm somewhat skeptical that senses and embodiment are required for consciousness and for understanding. I'd argue that a system with no senses and no body, like the philosopher's classic brain in a vat, could still have conscious thought, even if its consciousness was limited. Similarly, an AI system without senses could reason about mathematics, about its own existence, and maybe even about the world. On top of this, LLMs have a huge amount of training on text input which derives from sources in the world. One could argue that this connection to the world serves as a sort of grounding. Ellie Pavlick and colleagues have research suggesting that text training sometimes produces representations of color and space that are isomorphic to those produced by sensory training. I'm exploring all of these issues in some other work, but I won't go into further into them here.

A more straightforward reply is the observation that this problem can be avoided in extended language models, which have plenty of grounding. Vision-language models are grounded in images of the environment. For example, here's Deepmind's Flamingo which responds to text and images in conjunction.



Language-action models are grounded in control of physical or virtual bodies. Here's Google's SayCan, which uses an extended language model to help control a robot to perform various functions.

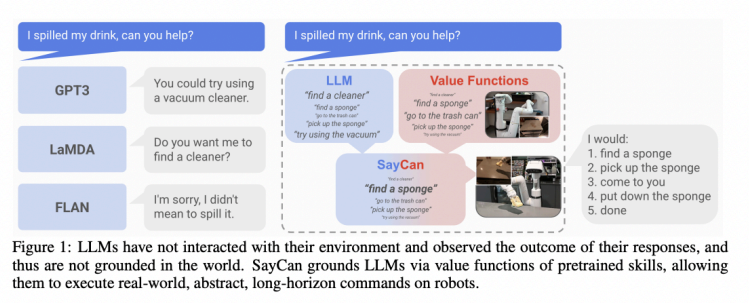


Figure 1: LLMs have not interacted with their environment and observed the outcome of their responses, and thus are not grounded in the world. SayCan grounds LLMs via value functions of pretrained skills, allowing them to execute real-world, abstract, long-horizon commands on robots.

These two paradigms are naturally combined in perception-language-action models, with both senses and a body, often in virtual worlds. Here's Deepmind's MIA (Multimodal Interactive Agent).

Virtual worlds are a lot more tractable than the physical world and there's coming to be a lot of work in embodied AI that uses virtual embodiment. Some people will say this doesn't count for what's needed for grounding because the environments are virtual. I don't agree. In my book on the philosophy of virtual reality, *Reality+*, I've argued that virtual reality is just as legitimate and real as physical reality for all kinds of purposes. So I think this kind of work is an important challenge for work on AI consciousness going forward.

Challenge: develop robust perception-language-action models with rich senses and bodies, perhaps in virtual worlds

X = World-Models and Self-Models

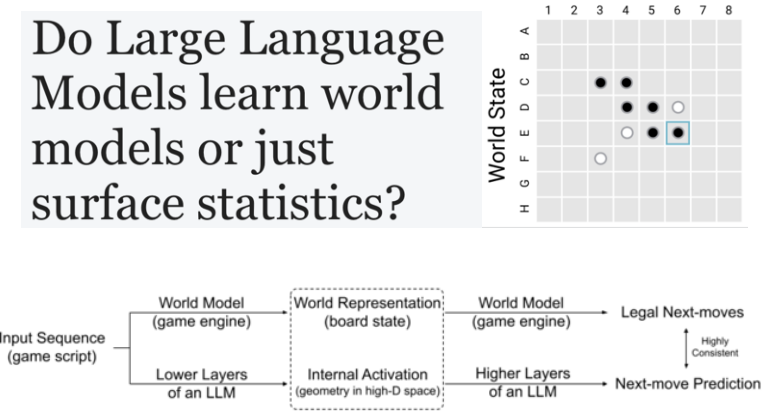
A related issue is the issue of world models. This connects to the well-known criticism by Emily Bender, Timnit Gebru, and colleagues that large language models are "stochastic parrots" – roughly, that they are merely imitating text rather than thinking about the world. In a similar way, many have suggested that LLMs are just doing statistical text processing. One key idea here is that world-models are just modeling text and not modeling the world. They don't have genuine understanding and meaning of the kind you get from a genuine world-model.

There's a lot to say about this, but just briefly. I think it's important to make a distinction between training and (post-training) online processing here. It's true that LLMs are trained to minimize prediction error in string matching, but that doesn't mean that their processing is just string matching. To minimize prediction error in string matching, all kinds of other processes may be required, quite possibly including world-models.

An analogy: in evolution by natural selection, maximizing fitness during evolution can lead to wholly novel processes post-evolution. A critic might say, all these systems are doing is maximizing fitness. But it turns out that the best way for organisms to maximize fitness is to have these amazing capacities – like seeing and flying and even having world-models. Likewise, it may well turn out that the best way for a system to minimize prediction error during training is for it to use highly novel processes, including world-models.

It's plausible that neural network systems such as transformers are capable at least in principle of having deep and robust world-models. And it's plausible that in the long run, systems with these models will outperform systems without these models at prediction tasks. If so, one would expect that truly minimizing prediction error in these systems would require deep models of the world. For example, to optimize prediction in discourse about the New York City subway system, it will help a lot to have a robust model of the subway system. Generalizing, this suggests that good enough optimization of prediction error over a broad enough space of models ought to lead to robust world-models.

If this is right, the underlying question is not so much whether it's possible in principle for a LLM to have world-models and self-models, but instead whether these models are already present in current LLMs. That's an empirical question. I think the evidence is still developing here, but interpretability research gives at least some evidence of robust world models. For example, Kenneth Li and colleagues trained an LLM trained on sequences of moves in the board game Othello, and gave strong evidence that it builds an internal model of the 64 board squares and uses this model in determining the next move. There's also much work on finding where and how facts are represented in language models.



There are certainly many limitations in current LLMs' world-models. Standard models often seem fragile rather than robust, with language models often confabulating and contradicting themselves. Current LLMs seem to have especially limited self-models: that is, their models of their own processing and reasoning are poor. Self-models are crucial at least to self-consciousness, and on some views (including so-called higher-order views of consciousness) they are crucial to consciousness itself. These are an especially important challenge. In any case, we can once again turn the objection into a challenge: build extended language models with robust world models and self models.

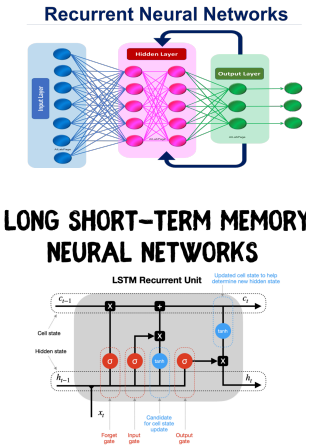
X = Recurrent Processing.

I'll turn now to two somewhat more technical objections tied to theories of consciousness. My former Ph.D. student Rob Long has been working on this issue – using scientific theories of consciousness to think about consciousness in language models – and I recommend playing close attention to his work as it appears.

The first objection here is that current transformer-based LLMs are feedforward systems without recurrent processing. Many theories of consciousness give a central role to recurrent processing. Victor Lamme's recurrent processing theory gives it pride of place as the central requirement for consciousness. Giulio Tononi's integrated information theory predicts that feedforward systems have zero integrated information and therefore lack consciousness. Other theories such as global workspace theory also give a role to recurrent processing.

If the theories requiring recurrent processing are correct, it looks as if current transformer-based LLMs cannot be conscious. One underlying issue is that because these are feedforward systems,

they lack memory-like internal states that persist over time. Many theories hold that persisting internal states are crucial to consciousness.



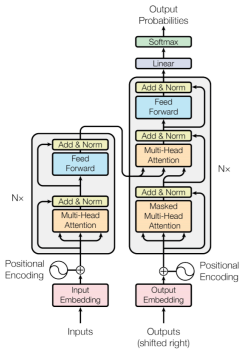


Figure 1: The Transformer - model architecture.

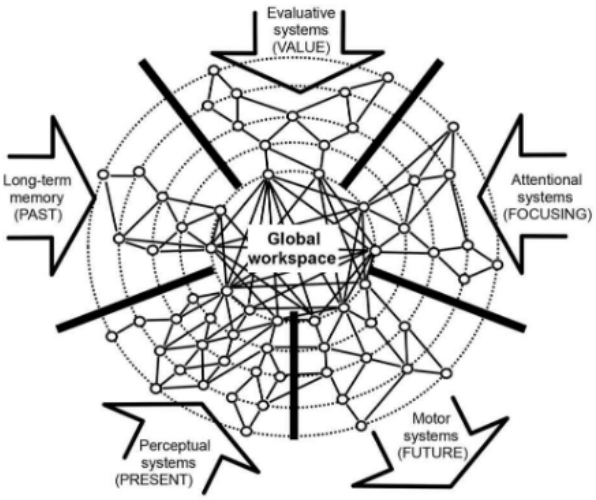
There are various responses here. First, current LLMs have some limited forms of recurrence and memory deriving from the recirculation of a window of past inputs and outputs, as well as through tools such as weight sharing. Second, it's plausible that not all consciousness involves memory, and there may be forms of consciousness which are feedforward.

Third and perhaps most important for the research program: there are recurrent large language models. Just a few years ago, language models centered on long short-term memory systems (LSTMs) which are recurrent. At the moment I gather that LSTMs are lagging somewhat behind transformers but the gap isn't enormous. There are also many LLMs that build in a form of memory and a form of recurrence through external memory components. It's easy to envision that recurrence may play an increasing role in LLMs to come. Once again, this objection basically amounts to a challenge: build extended large language models with genuine recurrence and genuine memory, the kind required for consciousness.

X = Global Workspace

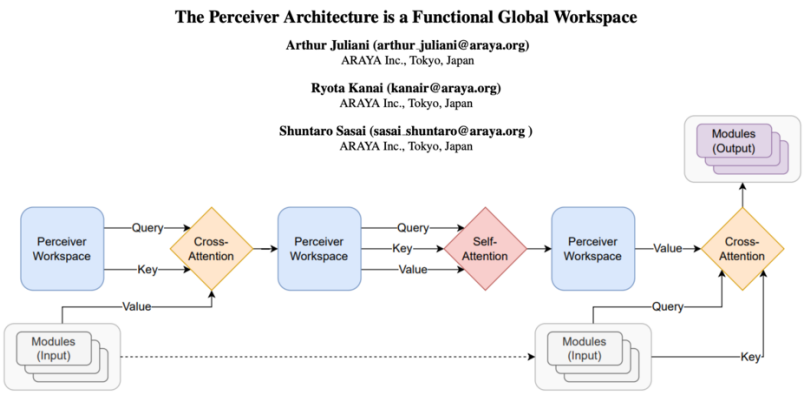
Perhaps the leading current theory of consciousness in cognitive neuroscience is the global workspace theory put forward by Bernard Baars and developed by Stanislas Dehaene and colleagues. This theory says that consciousness involves a limited-capacity global workspace: a

central clearing-house for gathering information from numerous non-conscious modules and making information accessible to them. Whatever gets into the global workspace is conscious.



6

A number of people have observed that standard language models don't obviously have a global workspace, but it may be possible to extend them to include a workspace. There's already an increasing body of relevant work on multimodal LLM+s that use a sort of workspace to co-ordinate between different modalities. These systems have input and output modules, for images or sounds or text for example, which may involve extremely high dimensional spaces. To feasibly integrate these modules, you need a lower-dimensional space as an interface. That lower-dimensional space interfacing between modules looks a lot like a global workspace.



People have already begun to connect these models to consciousness. Yoshua Bengio and his colleagues have argued that a global workspace bottleneck among multiple neural modules can serve some of the distinctive functions of slow conscious reasoning. There's a nice recent paper by Arthur Juliani, Ryota Kanai, and Shuntaro Sasai arguing that one of these multimodal systems, Perceiver IO, implements many aspects of a global workspace via mechanisms of self attention and cross attention. So again, there's an interesting research program here.

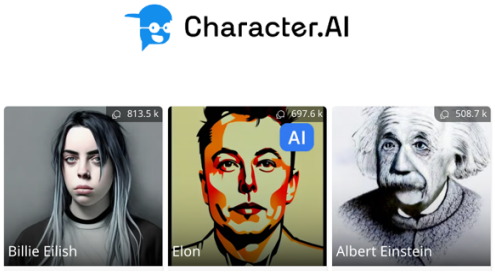
Challenge: build LLM+s with global workspace

X = Unified Agency

The final obstacle to consciousness in LLMs, and maybe the deepest, is the issue of unified agency. We all know these language models can take on many personas. As I put it in an article on GPT-3 when it first appeared in 2020, these models are more chameleons that can take the shape of many different agents. They often seem to lack stable goals and beliefs of their own over and above the goal of predicting text. In many ways, they don't behave like unified agents. Many argue that consciousness requires a certain unity. If so, the disunity of LLMs may call their consciousness into question.

Again, there are various replies. First: it's arguable that a large degree of disunity is compatible with conscious. Some people are highly disunified, like people with dissociative identity disorders, but they are still conscious. Second: One might argue that a single large language model can support an ecosystem of multiple agents, depending on context, prompting, and the like.

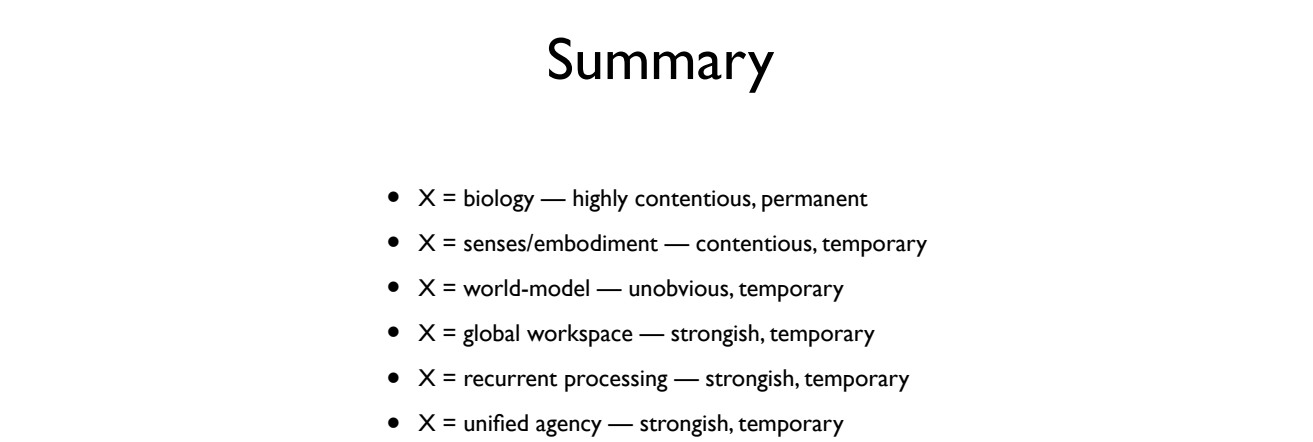
But to focus on the most constructive reply: it seems that more unified LLMs are possible. One important genre is the *agent model* (or person model or creature model) which attempts to model a single agent. The most popular way to do that currently, in systems such as Character.AI, is to take a generic LLM and use fine-tuning or prompt engineering using text from one person to help it simulate that agent.



Current agent models are quite limited and still show signs of disunity. But it's presumably possible in principle to train agent models in a deeper way, for example training an LLM+ system from scratch with data from a single individual. Of course this raises difficult ethical issues, especially when real people are involved. But one can also try to model the perception-action cycle of, say, a single mouse. In principle agent models could lead to LLM+ systems that are much more unified than current LLMs. So once again, the objection turns into a challenge.

I've given six candidates for the X that might be required for consciousness and missing in current LLMs. Of course there are other candidates: higher-order representation (representing one's own cognitive processes, which is related to self-models), stimulus-independent processing (thinking without inputs, which is related to recurrent processing), human-level reasoning (the many well-known reasoning problems that LLMs exhibit), and more. Furthermore, it's entirely possible that there are unknown X's independent of any of these, such that in fact X is required for consciousness. Still, these six at least capture some of the most important challenges.

Here's my approximate assessment of those six challenges:



Some of these X's rely on highly contentious premises about consciousness, most obviously in the claim that consciousness requires biology and perhaps in the requirement of sensory grounding. Others rely on unobvious premises about LLMs, like the claim that current LLMs lack world-models. Perhaps the strongest objections are those from recurrent processing, global workpace, and unified agency, where it's both plausible that current LLMs (or at least paradigmatic LLMs such as GPT-3) lack these X and reasonably plausible that consciousness requires X.

Still: for all of these X except perhaps biology, it looks like the objection is temporary. For the other five, there is a research program of developing LLM or LLM+ systems that have the X in question. In most cases, there already exist at least simple systems with these X's, and it seems entirely possible that we'll have robust and sophisticated systems with these X's within the next

decade or two. So the case against consciousness in current LLM systems is much stronger than the case against consciousness in future LLM+ systems.

4. Conclusions

Where does the overall case for or against LLM consciousness stand?

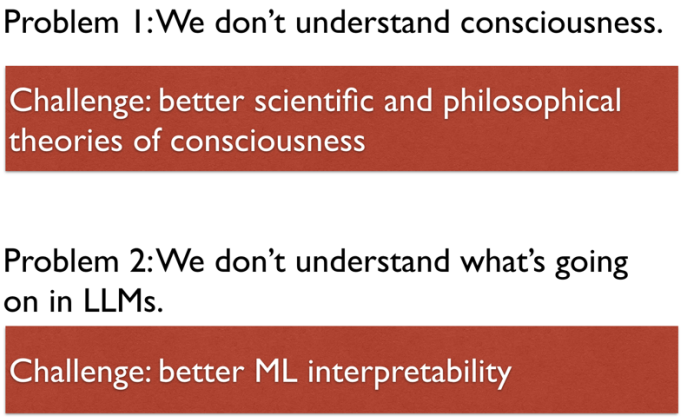
Where current LLMs such as the GPT systems are concerned: I think none of the reasons for denying consciousness in these systems are conclusive, but collectively they add up. To assign some extremely rough numbers for illustrative purposes: On mainstream assumptions, it wouldn't be unreasonable to hold that there's at least a 25% chance (that is, to have a subjective probability or credence of at least 0.25) that biology is required for consciousness, and the same for sensory grounding and self-models. Likewise, it wouldn't be unreasonable to hold that there's a 50% chance that recurrent processing is required for consciousness, and the same for global workspace and unified agency. If these six claims were independent, it would follow that there's at most a 5% or so chance ($0.75^3 * 0.5^3$) that a system lacking all six of these factors, like a current paradigmatic LLM, could be conscious. Of course these claims are not independent, so the figure should be somewhat higher. On the other hand, the figure may be driven lower by other potential requirements X that we have not considered, including the serious possibility that there are unknown X's that are required for consciousness and that LLMs lack. Taking all that into account might leave us somewhere under 10%. You shouldn't take the numbers too seriously (that would be specious precision), but the general moral is that given mainstream assumptions about consciousness, it's reasonable to have a low credence that current paradigmatic LLMs such as the GPT systems are conscious.

Where future LLMs and their extensions are concerned, things look quite different. It seems entirely possible that within the next decade, we'll have robust systems with senses, embodiment, world models and self-models, recurrent processing, global workspace, and unified goals. (A multimodal system like Perceiver IO already arguably has senses, embodiment, a global workspace, and a form of recurrence, with the most obvious challenges for it being world-models, self-models, and unified agency.). I think it wouldn't be unreasonable to have, say, a 50% credence that we'll have sophisticated LLM+ systems (that is, LLM+ systems with reasonably sophisticated behavior that seems comparable to that of animals that we take to be conscious) with all of these properties within a decade. It also wouldn't be unreasonable to have a 50% credence that if we develop sophisticated systems with all of these properties, they will be conscious. Those figures would leave us with a credence of 25% or more. Again, you shouldn't take the exact numbers too seriously, but this reasoning suggests that on mainstream assumptions, it's reasonable to have a significant credence that we'll have conscious LLM+s within a decade.

One way to approach this is via the "NeuroAI" challenge of matching the capacities of various non-human animals in virtually embodied systems. It's arguable that even if we don't reach human-level cognitive capacities in the next decade, we have a serious chance of reaching mouse-level capacities in an embodied system with world-models, recurrent processing, unified goals, and so on. If we reach that point, there would be a serious chance that those systems are conscious. Multiplying those chances gives us a significant chance of at least mouse-level

consciousness with a decade. We can see that as another challenge. Mouse-level cognition would at least be a stepping stone toward mouse-level consciousness and eventually to human-level consciousness somewhere down the line.

Of course there's a lot we don't understand here. One underlying problem is that we don't understand consciousness. That's a hard problem, as they say. Here the challenge is to develop to develop better scientific and philosophical theories of consciousness. They've come a long way in the last three decades, but more work is needed. The second major problem is that we don't really understand what's going on in these large language models. The project of interpretability in machine learning interpretability has come a long way, but it also has a very long way to go. So we need to add those two challenges to the mix.



My conclusion is that questions about AI consciousness are becoming ever more pressing. Within the next decade, even if we don't have human level artificial general intelligence, we may have systems that are serious candidates for consciousness. Although there are many challenges and objections to consciousness in machine learning systems, meeting those challenges yields a realistic potential research program for conscious AI.

I've summarized these challenges here, with four foundational challenges followed by seven engineering-oriented challenges.

Summary of Challenges

1. Evidence: benchmarks for consciousness?
2. Theory: scientific and philosophical theories of consciousness
3. Interpretability: what's happening inside an LLM?
4. Ethics: should we build conscious AI?

5. Rich perception-language-action models in virtual worlds
6. LLM+s with robust world-models and self-models
7. LLM+s with genuine memory and genuine recurrence
8. LLM+s with global workspace
9. LLM+s that are unified agent models
10. LLM+s that describe non-trained features of consciousness
11. LLM+s with mouse-level capacities

12. If that's not enough for conscious AI -- what's missing?

Suppose that in the next decade or two, we meet all the engineering challenges in a single system. Will we then have conscious AI systems? Not everyone will agree that we do, but this leads to the final challenge. If that's not enough for conscious AI, what's missing?

I'll finish by reiterating the ethical challenge. I'm not asserting that we should pursue this research program. If you think conscious AI is desirable, the program can serve as a sort of roadmap for getting there. If you think conscious AI is something to avoid, then the program can highlight paths that are best avoided. I'd be especially cautious about creating agent models. That said, I think it's likely that researchers will pursue many of the elements of this research program, whether or not they think of this as pursuing AI consciousness. It could be a disaster to stumble upon AI consciousness unknowingly and unreflectively. So I hope that making these possible paths explicit at least helps us to think about conscious AI reflectively and to handle these issues with care.

Notes

Transformation network: David J. Chalmers, Syntactic transformations on distributed representations. *Connection Science* 2:53-62, 1990.

Google fires engineer: New York Times, July 23, 2022.

No evidence that LaMDA was sentient: Washington Post, June 11, 2022.

Many LLM+ systems that go well beyond language: Perhaps I should have called this paper "Could large language models or their extensions be conscious?", or something more general like "Could foundation models be conscious?".

Thomas Nagel's phrase: Thomas Nagel, What is it like to be a bat? *Philosophical Review* 83:435-50, 1974.

Many ethical issues: See Matthew Liao, 2020. *Ethics of Artificial Intelligence.* Oxford University Press.

Blake Lemoine reported: Blake Lemoine, Is LaMDA sentient? An interview. *Medium,* 2022. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917.

Small alteration to Lemoine's question: Reed Berkowitz. How to talk with an AI: A Deep Dive Into "Is LaMDA Sentient?". *Medium,* 2022. https://medium.com/curiouserinstitute/guide-to-is-lamda-sentient-a8eb32568531.

The philosopher Susan Schneider: Susan Schneider, *Artificial You*. Princeton University Press, 2019.

In his famous test: Alan Turing, Computing Machinery and Intelligence. *Mind* 49: 433-460, 1950.

Gato: Scott Reed et al, A generalist agent. *Transactions on Machine Learning Research,* 2022. arXiv:2205.06175

GPT-3: Tom Brown et al, Language models are few shot learners. arXiv:2005.14165, 2020.

Consciousness requires carbon-based biology: Ned Block. Comparing the major theories of consciousness. In (M. Gazzaniga, ed.) *The Cognitive Neurosciences IV*. MIT Press, 2009.

Grounding: Stevan Harnad, The symbol-grounding problem. *Physica D* 42:335-346, 1990. Note that *grounding* has quite different meanings among AI researchers (roughly, processing caused by sensory inputs and the environment) and among philosophers (roughly, constitution of the less fundamental by the more fundamental).

A number of researchers have argued that sensory grounding is required for understanding in LLMs. Emily M. Bender and Alexander Koller, Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198, 2020. Brenden Lake and Greg Murphy, Word meaning in minds and machines. *Psychological Review*, 2021. Jacob Browning and Yann Lecun, AI and the limits of language. *Noēma,* 2022.

Text training sometimes can produce representations: Roma Patel and Ellie Pavlick, Mapping language models to grounded conceptual spaces. *International Conference on Learning Representations*, 2022. Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? A case study in color. *Proceedings of the 25th Conference on Computational Natural Language Learning,* 2021.

I'm exploring all of these issues in some other work. My January 2023 presidential address to the American Philosophical Association (Eastern Division) on "Can Large Language Models Think?" focused on whether sensing is required for thinking and on the question of world-models.

Flamingo: Jean-Baptiste Alayrac et al, Flamingo: a Visual Language Model for Few-Shot Learning. *Proceedings of Neural Information Processing Systems*, 2022.

SayCan: Michael Ahn, et al. Do as I can, not as I say: Grounding language in robotic affordances. https://say-can.github.io/. 2022.

MIA: Josh Abramson et al, 2021. Creating multimodal interactive agents with imitation and self-supervised learning. arXiv:2112.0376.

My recent book: David J. Chalmers, *Reality+: Virtual Worlds and the Problems of Philosophy*. W. W. Norton, 2022.

*World models:*Representationalist theories of consciousness (see William Lycan, Representational theories of consciousness, *Stanford Encyclopedia of Philosophy* hold that world-models (or at least representations of the world) are required for consciousness.

Self-models: Higher-order theories of consciousness (see Rocco Gennaro, ed, *Higher-Order Theories of Consciousness: An Anthology*, John Benjamins, 2004) hold that self-models (representations of one's own mental states) are required for consciousness. In addition, illusionist theories of consciousness (see Keith Frankish, ed. *Illusionism as a Theory of Consciousness,* Imprint Academic, 2016, and Michael Graziano, *Rethinking Consciousness,* W.W. Norton, 2019) hold that self-models are required for the illusion of consciousness.

Stochastic parrots: Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Schmargaret Schmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* pp. 610-623, 2021.

Interpretability research gives at least some evidence of robust world models: Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg, Emergent world representations: Exploring a sequence model trained on a synthetic task. arXiv.2210.13382, 2022. Belinda Z Li, Maxwell Nye, and Jacob Andreas, Implicit representations of meaning in neural language models. arXiv:2106.00737, 2021.

Rob Long has been working on this issue: For an initial discussion, see Robert Long, Key questions about artificial sentience: An opinionated guide. *Experience Machines (Substack),* 2022.

Recurrent processing theory: Victor A. Lamme, How neuroscience will change our view on consciousness. *Cognitive Neuroscience* 1: 204–220, 2010.

Information integration theory. Giulio Tononi. An information integration theory of consciousness. *BMC Neuroscience,* 2004.

Long short-term memory. Sepp Hochreiter and Jürgen Schmidhuber, Long short-term memory. *Neural Computation* 9: 1735-80, 1997.

Global workspace theory: Bernard J. Baars, *A Cognitive Theory of Consciousness.* Cambridge University Press, 1988. Stanislas Dehaene, 2014. *Consciousness and the Brain.* Penguin.

Yoshua Bengio and colleagues: Anirudh Goyal, Aniket Didolkar, Alex Lamb, Kartikeya Badola, Nan Rosemary Ke, Nasim Rahaman, Jonathan Binas, Charles Blundell, Michael Mozer, Yoshua Bengio. Coordination among neural modules through a shared global workspace. arXiv:2103.01197, 2021. See also Yoshua Bengio, The consciousness prior. arXiv:1709.0856, 2017.

Recent paper by Arthur Juliani: Arthur Juliani, Ryota Kanai, Shuntaro Sasai. The Perceiver Architecture is a Functional Global Workspace. *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 44, 2021. https://escholarship.org/uc/item/2g55b9xx.

Perceiver IO: Andrew Jaegle et al. Perceiver IO: A general architecture for structured inputs and outputs. arXiv:2107.14795, 2021.

It wouldn't be unreasonable to hold that there's a 50% chance that recurrent processing is required for consciousness, and the same for global workspace and unified agency: Jonathan Birch ("The Search for Invertebrate Consciousness", *Nous* 56:133-53, 2021) distinguishes approaches to animal consciousness that are "theory-heavy" (assume a complete theory), "theory-neutral" (proceed without theoretical assumptions), and "theory-light" (proceed with weak theoretical assumptions). One can likewise take theory-heavy, theory-neutral, and theory-light approaches to AI consciousness. The approach to artificial consciousness that I have taken here is distinct from these

three. It might be considered a *theory-balanced* approach: proceed by balancing one's credences between various theories, perhaps according to evidence for those theories or according to acceptance of those theories.

A more precise form of the theory-balanced approach might use data about how widely accepted various theories are among experts to provide credences for those theories, and use those credences along with the various theories' predictions to estimate probabilities for AI (or animal) consciousness. In a recent survey of researchers in the science of consciousness (Jolien C. Frankel et al, "An academic survey on theoretical foundations, common assumptions and the current state of consciousness science", *Neuroscience of Consciousness,* issue 1, 2022), just over 50% indicated that they accept or find promising the global workspace theory of consciousness, while just under 50% indicated that they accept or find promising the local recurrence theory (which requires recurrent processing for consciousness). Figures for other theories include just over 50% for predictive processing theories (which do not make clear predictions for AI consciousness) and for higher-order theories (which require self-models for consciousness), and just under 50% for integrated information theory (which ascribes consciousness to many simple systems but requires recurrent processing for consciousness). Of course turning these figures into collective credences requires further work (e.g. in converting "accept" and "find promising" into credences), as does applying these credences along with theoretical predictions to derive collective credences about AI consciousness. Still, it seems not unreasonable to assign a collective credence of well over 25% for each of global workspace, recurrent processing, and self-models as requirements for consciousness.

It's also worth noting that in a 2020 survey of professional philosophers (David Bourget and David Chalmers, "Philosophers on Philosophy: The 2020 PhilPapers Survey", *Philosophers' Imprint,* 2023), around 3% accepted or leaned toward the view that current AI systems are conscious, with 82% rejecting or leaning against the view and 10% neutral. Around 39% accepted or leaned toward the view that future AI systems will be conscious, with 27% rejecting or leaning against the view and 29% neutral. (Around 5% rejected the questions in various ways, e.g. saying that there is no fact of the matter or that the question is too unclear to answer). The future-AI figures might tend to suggest a collective credence of at least 25% that biology is required for consciousness (albeit in a different group). The two surveys have less information about unified agency and about sensory grounding as requirements for consciousness.

Given mainstream assumptions about consciousness: Compared to mainstream views in the science of consciousness, my own views lean somewhat more to consciousness being widespread. So I'd give somewhat lower credences to the various substantial requirements for consciousness I've outlined here, and somewhat higher credences in current LLM consciousness and future LLM+ consciousness as a result.

It's reasonable to have a low credence that current paradigmatic LLMs such as the GPT systems are conscious: Five months later [April 30, 2023], new systems such as GPT-4 are a definite advance on some of the dimensions discussed in this article, without really changing anything fundamental. These systems display progress in conversational abilities, in multimodal processing, arguably in world-modeling, and they are starting to be used more widely for agent modeling. That said, the major obstacles regarding consciousness (tied to recurrent processing, global workspace, and unified agency) are still present.

NeuroAI: Anthony Zador et al. Toward next-generation artificial intelligence: Catalyzing the NeuroAI revolution. arXiv:2210.08340, 2022.

Mouse-level capacities: At NeurIPS I said "fish-level capacities". I've changed this to "mouse-level capacities" (probably a harder challenge in principle), in part because more people are confident that mice are conscious than that fish are conscious, and in part because there is so much more work on mouse cognition than fish cognition.

I'll finish by reiterating the ethical challenge: This is an addition to what I presented at the NeurIPS conference.