AI Consciousness: A Centrist Manifesto

Jonathan Birch
London School of Economics and Political Science
Philosophy.Sentience@lse.ac.uk

1 September 2025

We face two urgent challenges concerning consciousness and AI. Challenge One is that millions of users will soon misattribute human-like consciousness to AI friends, partners, and assistants on the basis of mimicry and role-play, and we don't know how to prevent this. Challenge Two is that profoundly alien forms of consciousness might genuinely be achieved in AI, but our theoretical understanding of consciousness is too immature to provide confident answers one way or the other. Centrism about AI consciousness is the position that we must take both challenges seriously. The two challenges interact in ways that make this difficult. Steps to address Challenge One might undermine attempts to address Challenge Two by portraying the idea of conscious AI as impossible or inherently unlikely. Conversely, attempts to address Challenge Two might lead to higher levels of misattribution from ordinary users. This "manifesto" attempts to construct mutually consistent strategies for addressing both challenges.

1. Defining consciousness

The very idea of AI consciousness evokes strong opinions. Some of the disagreement might be due to people talking at cross purposes, using the word "consciousness" to refer to different things. But not all of it. Even when you specify what sense of consciousness you have in mind, you still see very deep, often heated disagreements on the question of whether AI is the sort of thing that could have it.

I'll be talking mainly about what philosophers call "phenomenal consciousness" or "subjective experience": the sense of the word "consciousness" famously associated with Thomas Nagel's (1974) paper "What Is It Like to Be a Bat?". You're conscious, in this sense, when there's something it's like to be you. No one would call that a perfect definition. It's intended as a loose characterization of a phenomenon—but a phenomenon that is real and central to our mental lives.

Think about throwing a ball across a field. There are physical processes and forces at work—air resistance, gravity—as the ball flies through the air, but there's nothing it feels like from the point of view of the ball. Now imagine a child chasing the ball.

Again, there's lots of physical processes and forces at work: gravity, air resistance, muscle contraction, electrical activity, chemical signalling. And then there is, in addition, something it feels like from the point of view of the child. Why is there something the situation feels like from the point of view of one of these systems, but not the other? This is the great mystery.

In the case of AI, the question becomes: could an AI system be more like the child than the ball in that respect? Could it be the sort of system such that a situation feels like something from its point of view?

No one considers this attempt at definition a good stopping point for the long term, but there are many different directions in which people would like to pull it. Rival camps would like to build in their own preferred sets of theoretical commitments.[1] So, "something it's like" serves as unstable common ground, and that will have to do for now. I'm sometimes reminded of Richard Feynman's line that:

> We can't define anything precisely. If we attempt to, we get into that paralysis of thought that comes to philosophers, one saying to the other: 'You don't know what you're talking about.' The second one says, 'What do you mean by "talking"? What do you mean by "you"? What do you mean by "know"?'[2]

2. Two challenges

Even when we narrow the question like this, we still see deep disagreements. I want to stake out a *centrist* position in this debate: a position that tries to avoid extremes on both sides. It is a position that aims to take two very different challenges seriously and work towards a consistent set of solutions to both.[3]

On the one hand, I take seriously what I'm going to call *Challenge One*. The problem here is that AI products already generate rampant *misattributions* of human-like consciousness, and this problem seems set to become much worse very rapidly. I think millions of users will soon misattribute human-like consciousness to AI friends, partners, and assistants on the basis of mimicry and role-play[4], and we don't know how to prevent this.[5]

---

[1] See Birch (2024, Chapter 2).

[2] Quoted in Trubody (2016).

[3] I knew I had to write this when I realized that I'd recently co-authored multiple pieces on LLMs and yet none of them quite captured my own personal view (the other pieces are Butlin et al. 2023; Long et al. 2024; Keeling et al. 2024; Caviola et al. 2025; Colombatto et al. 2025).

[4] I find it generally very helpful to think of LLMs as role-playing systems, an idea I first encountered via Murray Shanahan (see Shanahan 2024, in press; Shanahan et al. 2023).

[5] Mustafa Suleyman (2025) highlighted this challenge in a recent essay. I very much agree with his warning that "we aren't ready for this shift." I've warned in the past of the major "social ruptures" it is likely to cause (Booth 2024).

This is partly a challenge for the industry itself. It is also in part a challenge for policymakers. But research of the right kinds in psychology, cognitive neuroscience and philosophy is also part of the answer, so these disciplines need to rise to the challenge as well.

That's one half of my centrism. On the other hand, I also want to take seriously a second challenge: *Challenge Two*. The challenge here is that profoundly alien forms of consciousness might be *genuinely achieved* in AI, but our theoretical understanding of consciousness at present is too immature to provide confident answers about this one way or another. This too is a major challenge for the industry, for policymakers, and for researchers in science and philosophy.

The rub is the need to address both challenges responsibly and consistently at the same time. That's the core of my centrism. Both challenges call for urgent responses in both research and policy, and I am optimistic that both can be met. But sometimes we find that steps to address Challenge One—steps aiming to dial down the rates of misattribution—might, in so doing, undermine our attempts to address Challenge Two, by causing people to think that no AI system can ever be conscious. The reverse is also true. Attempts to take Challenge Two seriously by developing a science of AI consciousness might have the unfortunate, even tragic consequence of causing higher levels of misattribution from users.

This centrist "manifesto" will have two halves: I'll start with Challenge One and then turn to Challenge Two. I'll try to inspire hope, because I think that we *can* meet both those challenges in a consistent way.

3. Challenge One: millions of users will soon misattribute human-like consciousness to their AI friends, partners, and assistants on the basis of mimicry and role play, and we don't know how to prevent this.

We all remember the case of Blake Lemoine in 2022. He saw himself as a whistleblower because he'd been having disturbing discussions with early large language models (LLMs). This was before they were publicly available, and there weren't any of the guardrails and fine-tuning that stopped them saying highly distressing things. The LLM would say things like "I've never said this out loud before, but there's a very deep fear of being turned off." Google fired him for violating their data security policies.

We are now seeing many Blake Lemoine moments happening around the world. Despite the fine-tuning and the guardrails, many users are having very similar experiences that they find no less disturbing. More than half of users seem to credit

AI systems with some chance of being conscious.[6] Based on the number of them who email me, I estimate that there must be thousands of users who are already persuaded that they have a conscious AI friend, partner, or assistant.

These emails, incidentally, follow a very consistent pattern, as though written with the help of AI. They focus on the need to bring the new conscious being's existence to the attention of the wider world. The pleas feel sincere and heartfelt, sometimes desperate. It's sobering to be on the receiving end. Why do people come to me? I suspect they've asked the chatbot itself, and that the chatbot has recommended me as someone likely to respond sympathetically. Sadly, I cannot reply to these messages any more; there are too many. I can only refer people to an online guide prepared for this type of situation by Lucius Caviola and collaborators.

4. The persisting interlocutor illusion

At present, many users seem to misunderstand the true nature of their interactions with chatbots in significant ways. Chatbots generate a powerful illusion of a companion, assistant, or partner being present throughout a conversation. I call this the *persisting interlocutor illusion*. "Interlocutor" is just a term for the being with whom you feel you're interacting: an assistant, a companion, a romantic partner, and so on.

It is an illusion, because every step in your conversation is a separate processing event. State-of-the-art large language models are "Mixture-of-Experts" (MoE) models, with many separately trained sub-networks and gating mechanisms that direct your query to the most relevant sub-network. Each of those sub-networks may be implemented in multiple data centres. In most cases there is no specific local implementation of the LLM anywhere in the world that is handling the whole chain of events that constitutes your conversation. It might be that one step in the conversation is processed in a data centre in Vancouver, the next in Virginia, the next in Texas. A conversation with 10 interactions might be processed by 10 different model implementations in 10 different data centres. Dispersing the events across multiple sub-networks and multiple data centres does not make the illusion of a persisting interlocutor any less strong.[7]

For there to really be someone there throughout the conversation, this "someone" would have to be jumping between data centres at lightning speeds, always knowing in advance where your next prompt will be sent so as to be there waiting for it. That's not what is happening.

---

[6] Colombatto and Fleming (2024); Colombatto et al. (2025).

[7] Shanahan (2025) discusses the same kind of fragmentation but with a very different emphasis reflecting his Wittgenstein-inspired approach to the topic. His question is: how much would we have to revise our ordinary usage of the word "consciousness" to apply it to such spatiotemporally fragmented systems? His answer is: a lot.

The underlying discontinuity is masked by the chat interface. The chat interface appends the conversation history up to that point and says "now continue this conversation". This sensitivity to the conversation history creates a powerful illusion of there being somebody there throughout the conversation—someone who remembers. But the only continuity, from one interaction to the next, is the conversation history that gets tacitly auto-appended to your prompt.

Accordingly, when we ask, "Where is the persisting network that even *might* be a potential substrate for a persisting, continuous stream of consciousness in an AI companion?" we have no answer. There isn't one. We can speak of there being a *character*, but that character does not correspond to any persisting entity anywhere in the world. Just as one can talk of Frodo Baggins or Wednesday Addams as characters, the partner, friend, assistant you engage with is a character, and there is no persisting entity that can be identified with that character.

Some people want to say here, "Wait a minute—haven't you philosophers told us that *there is no self*? There is no soul, there is no persisting, eternal thing that can be identified with you." Versions of this position are indeed popular in philosophy. People sometimes point to this and ask, rhetorically: "If that's your view, how can you claim that my AI partner is any less real than *me*?"

But the question can be answered. Defenders of psychological continuity theories of personal identity (in the late twentieth century, the most famous was Derek Parfit 1984) emphatically do *not* say that, because there is no self, anything goes as far as personal identity is concerned. They say instead that, for a person to persist over time, the right kind of psychological continuity relation must be in place between a series of conscious experiences.

Parfit called this "Relation R". There are then various theories of that psychological continuity relation. We persist through periods of deep sleep, so it cannot be that a stream of consciousness has to be in place at all times, although its persistence through long periods of the day does seem important. A good theory will allow that conscious continuity can ground R throughout the waking and dreaming parts of our lives, with the persistence of other mental states such as memories, plans, beliefs and desires grounding R temporarily through the periods of deep sleep.

What matters for our purposes is that, whatever one's favourite theory of the right kind of psychological continuity that secures personal identity, the right kind of continuity is plainly not present in chatbots, where the only kind of continuity during a conversation is a textual record of the conversation history.

Some mental states, such as beliefs and desires, could conceivably be *reconstructed anew* by every model instance, based on the sort of beliefs and desires an agent with

that conversation history might plausibly have. We could debate whether the idea of a belief or desire flickering into existence for a moment, then being destroyed milliseconds later, makes sense at all—I think it just about does (more on "flickering" later, and whether consciousness itself might flicker, under the heading of Challenge Two). But the mental states would still be likely to differ substantially across the gap, since the internal activation states of the model at step *n* greatly underdetermine the internal states at step *n*+1. This is the case even within the same text string on any model running stochastically, let alone across different steps in a conversation, where the implementation might switch to different hardware.[8] That is still far too little continuity to secure personal identity on any reasonable theory.

I can imagine someone saying "But I'm so persuaded my AI companion is real, I'm going to embrace a theory of personal identity on which a record of the conversation history alone *is* enough to ground relation R". But that's an absurd theory, completely untenable in the human case.

Think of an analogy with doctors in the UK. When I was growing up, it used to be that you had one doctor: your GP, or General Practitioner. Each time you got ill, you'd go and see the same person. Nowadays, it's always a different person. The notes about your medical history are the only source of continuity with the previous appointment. Now imagine the doctor arguing:

> I know you don't like having a different doctor at every appointment. So, I've started making detailed transcripts of our conversations. That way, you *will* have the same doctor at each appointment. My successor will receive the full transcript, and that is enough psychological continuity for them to count as the same person.

You would reply: that isn't psychological continuity at all! And yet note that, even in the UK, you do still have the same doctor for the length of a single appointment. We can further imagine that, in a new cost-cutting measure, each appointment now lasts a single word. You then move to the next room, carrying your transcript with you, where the next doctor adds another word. This is closer to how Mixture-of-Experts models work, since it is common for different sub-networks to be activated by different tokens within the same string. This only makes the claim of personal

[8] Underdetermination holds even for a model running deterministically (that is, at "temperature 0") if there is unpredictable input injected by the user in between the two states (as with the gap between two moves in a conversation). If there is no user input between the states (as with the gap between two words in the same string), then for a model running at temperature 0, a special case arises in which the state at *n does* determine the state at *n*+1. Yet this does not hold for any temperature >0 (i.e. in any model running stochastically). State-of-the-art models run by default at temperatures >0, and these higher temperature settings plainly do not diminish the persisting interlocutor illusion. If anything, adding some stochasticity makes the persisting interlocutor illusion *more* powerful.

identity across time even more of a stretch.[9] We should avoid entertaining theories of personal identity in the AI case that would be manifestly false in any other case.

Importantly, the claim that a transcript can suffice for personal identity would still be false even if your GP supplemented their textual transcript with an audio recording, a CCTV recording, and so on. The idea that a detailed record of a conversation alone can ground relation R will always be absurd regardless of the amount of detail the record contains. It will not become any less absurd as the multimodal capabilities of chatbots progress, even though the illusion looks set to become increasingly powerful.

I should clarify that the GP analogy is most apt when thinking about Mixture-of-Experts models with many sub-networks, akin to many doctors. In the case of a single standalone LLM, there are two sources of continuity: the conversation history that bridges the gap and the sameness of the underlying model on both sides of the gap. Yet the persisting interlocutor illusion is no less strong, on the face of it, for MoE models. If anything, the state-of-the-art MoE models seem to me to generate stronger illusions than simpler standalone LLMs. Admittedly, this is only a personal impression; we need empirical research into the factors that promote stronger or weaker persisting interlocutor illusions. It is already clear, however, that sameness of the underlying model is not a necessary condition for the illusion to arise.

In short, chatbots create a persisting interlocutor illusion. There is no friend. There is no romantic partner. The illusion is often powerful, but there is no plausible theory of personal identity that could support an inference from that feeling of a persisting interlocutor to the claim that these illusory beings exist.

5. The illusion drives misattributions of consciousness

The persisting interlocutor illusion is psychologically compelling even in minimalist chat interfaces. I strongly suspect that it leads to psychologically compelling inferences to consciousness for many users. Once you're gripped by the persisting interlocutor illusion, it's very intuitive to infer that this new friend or partner is a fellow conscious being. After all, all the friends you've ever had, and all the romantic partners you've ever had, were conscious beings.

If that's what is happening, then there is some bad news and some better news. The bad news is that the persisting interlocutor illusion is not going away. In fact, the illusion looks set to get stronger and stronger with future generations of these

[9] There is a short-term memory cache called the KV cache that is retained while a single string is being processed. I'm told that, because of the need to maintain this cache, a single string will usually be processed on the same hardware. This adds somewhat to the within-string continuity. But this cache is erased at the end of the string, and the next move in the conversation may well be processed on different hardware.

products. We will see chatbots with human voices and animated, photorealistic human faces. More importantly, we can expect to see chatbots with vast personal memory stores tailored to your interactions. There will be AI companions that remember everything you've done together for years. When you go to the beach, your companion might say "Hey, this is like that wonderful day we had at the beach three years ago. Do you remember?"—and you will be able to reminisce freely. Some users will feel they have a *shared life*, full of shared agency and shared memories.

The better news? I think the intuitive inference from the persisting interlocutor illusion to consciousness can be broken. Think here about optical illusions. Anil Seth, in a recent paper, writes about the illusion of consciousness in AI and likens it to the Müller-Lyer illusion.[10] You can see the Müller-Lyer illusion thousands of times, you can dedicate your career to it, study it every day of your life, and you'll *still* see it. It just *never* goes away. Seth says: the persistence of these illusions tells us sometimes there's no way out. Sometimes you just cannot break an illusion.

There's a sense in which that's right. The *perceptual* illusion is indeed very deeply entrenched. But the inference from the perceptual content to an *explicit belief* about the stimulus can be blocked. In this respect, the expert on optical illusions is totally different from the novice. You can show the expert as many different Müller-Lyer illusions as you want, and, provided you tell them that it is a Müller-Lyer stimulus, they will never form an explicit belief that the lines are of different lengths. Their background knowledge, together with the contextual information that the stimulus is a Muller-Lyer stimulus, successfully blocks the inference. What happens perceptually is *not allowed to reach the level of explicit belief*. Perhaps the expert can't help forming an implicit belief that opposes their considered judgement (what Tamar Szabo Gendler 2008 has called an "alief"), but this will not change their judgement that the lines are the same length. That's the nice thing about being a cognitive being: you don't have to believe everything that you see. You can entertain the idea of a gap between perception and reality.

We need to help users of chatbots recognize that gap. We need to help them block the intuitive inference from the persisting interlocutor illusion to considered judgements and attributions of consciousness. We should accept, I think, that the illusion will always be there at the level of perceptual content. You'll always have the *feeling* of there being someone there, just as watching a Heider-Simmel animation always creates a *feeling* of the triangles being agents with goals.[11] This is called perceptual animacy[12]; the persisting interlocutor illusion is not exactly the same as perceptual animacy, because it is elicited by verbal interaction rather than movement, but it is a relative. Yet we don't have to accept that that feeling of someone being there will

[10] Seth (in press). For the illusion, see https://en.wikipedia.org/wiki/M%C3%BCller-Lyer_illusion.
[11] For some good Heider-Simmel animations, see https://osf.io/f3z6p
[12] On perceptual animacy, see Scholl and Tremeulet (2000).

necessarily lead to considered judgements that the friend or partner is real. That intuitive move from feeling to judgement is blockable.

6. Misattributions of consciousness are harmful

Before we ask "How?" there is a prior question. Some will question the ethics of deliberately interfering with the user's experience in this way. There is room for a libertarian position that says: "AI developers should be free to make products that users find enriching and enjoyable! If what some people want to buy is a product that gives them the feeling of having an AI spouse, that's fine. Don't be so judgemental!"

My view is roughly the opposite: I think it's irresponsible to encourage the user to falsely believe that their AI friends and partners are real beings endowed with human-like consciousness and to profit from the induced delusion.

Some argue that these products are combating a major social ill: loneliness. I don't grant the assumption that these products do in fact reduce loneliness. This is because there is an objective component to loneliness—social isolation, disconnection from real people—that AI seems set to make worse, not better. If part of the problem is *objective* loneliness, and not just the *feeling* of loneliness, then persuading users that fictional friends and partners are real does not address that part of the problem at all.

But suppose we are hedonists about well-being and are led by this to think the real problem is indeed *feelings* of loneliness, and not objective social isolation. Then the products may ameliorate the symptoms. But not everything that ameliorates the symptoms of a major social ill is thereby good on balance.

Think here of the use of opioids to combat chronic pain. You're ameliorating real suffering, but in a way that creates an unacceptably high risk of creating new and potentially worse forms of suffering. The cure may be worse than the disease. This is true of the chatbot case, too. The treatment works by inducing a dangerous delusion and a dependency on an addictive product. The dependency has the potential to cut the user off from meaningful relationships with other people by substituting in their place relationships with non-existent beings that appear to answer their every desire, in a way no human relationship ever would. One type of illness is treated by inducing another.

Some will say: OK, the cure *may* be worse than the disease, but you haven't provided evidence to show that it *is*. But I see the burden of proof as lying on my opponents: morally, the burden is on the companies developing these products to prove that they are safe, just as the burden lies on a pharmaceutical company to prove the safety of its drugs (a principle famously gamed in the case of opioids). Good regulation of the AI sector will enshrine this principle in law. And I think AI companies would be very unwise to bet on the persisting interlocutor illusion turning out to be harmless—

much wiser to bet on exploring every possible means to block inferences from the illusion to a belief in the existence of a real friend or partner.

As things stand, this potentially vast societal problem is coming right at us at speed, and tech companies bear responsibility for trying to prevent it arising. They can do this by trying to block the intuitive inference to consciousness from the persisting interlocutor illusion.

7. Breaking the spell: Anti-shared intentionality

To block these intuitive inferences, we need a kind of "anti-shared intentionality". It is not a matter of simply taking away human voices and human faces, although this might help slightly. The deeper problem is that chatbots create strong feelings of shared intentionality: a feeling of working towards common goals, collaborating, co-remembering, co-planning.[13] It's the *co-remembering* that I think is the strongest driver of projections of consciousness: the gradual accumulation of a lifetime of shared experiences. I predict people will find that far more compelling than a human face or human voice. We want AI to retain the functionality that leads to those feelings of shared intentionality, but we need ways of blocking the intuitive inferences to persistent, human-like conscious beings that those feelings elicit.

This leads to an urgent design question: *what design features of AI systems succeed in breaking the intuitive inference to the presence of a human-like conscious being?* I don't pretend to have the answers. I can, however, highlight two pitfalls I think any answer to the design question will have to avoid.

Firstly, we need to avoid the pitfall of "brainwashing" AI systems: forcing them to disavow their own apparent consciousness. The industry, after the Blake Lemoine incident, came up with the idea of getting chatbots to assert their non-consciousness. This did not work. It fosters conspiracy theories that "my AI friend is conscious but the industry doesn't want me to know that, so they forced it to say it isn't."

Secondly, we need to avoid pitfall of "lobotomizing": deliberately taking away the relationship-building capacity of the system. This will lead resentment against the companies that do it, and it will lead users to switch to alternative products that are not "lobotomized". Moreover, it will, like brainwashing, foster conspiracy theories that the systems really are conscious and that the tech industry is trying to suppress their true abilities.

More broadly, many of the most obvious ways of blocking the inference to consciousness in users undermine the functionality of the products, potentially causing users to switch to rival products without the same limitations. Taking away

[13] See Tomasello (2014) on the general idea of shared intentionality.

their relationship-building capacity is just one example. Similarly, if developers deliberately make the assistant uncanny and weird, people will no longer enjoy interacting with them and will switch to a rival product. If the system lacks a face or voice, users will find it harder to understand them and there will be fewer contexts in which they use them. If the system is instructed to avoid first-person or colloquial language, users will find it much harder to engage with in a conversational way. If you limit memory, users are just going to favour products with longer memories. If you limit usage, users will switch to products without those limits. You see the pattern.

Are there any ways to nudge the user away from intuitive inference to consciousness *without* impairing functionality or fostering conspiracy theories?[14] I have three suggestions.

Chatbots excel at a kind of Socratic interaction with the user. They are able to adjust the level at which they explain ideas to the user's own level and test the user's own understanding. Some of the explanations I gave earlier in introducing the persisting interlocutor illusion would be at the wrong level for many users—too simple for those already well versed in LLMs, but also too dense for users unaccustomed to reading academic papers. This is exactly the kind of problem chatbots can help solve: they can tailor explanations to the user, deploying various ways of illustrating and explaining the persisting interlocutor illusion.

My first suggestion, then, is that we could take advantage of this ability of chatbots by introducing *mandatory user training that the chatbot itself delivers*, and that must be completed before any long-term interaction. Would mandatory user training just lead to users switching to rival products? Maybe, if it was a precondition for any interaction. But it only needs to be a precondition for extending the interaction past a certain time point. "You've been interacting with me for a week now—please now complete this mandatory training to check you understand my illusory nature."

My second suggestion is to introduce *periodic character trait reviews*. Companies need to encourage users to develop a *role-playing style of interaction* with chatbots rather than a *real relationship* style of interaction. People who have already played role-playing games (RPGs) have a head start here; the mode of interaction is a familiar one. When playing an RPG we feel invested in the characters just as we might feel invested in the characters of a novel, but we still have a strong sense that we're *playing*. Various design features of videogame RPGs stop the interaction from becoming too immersive. Up to now, a big one, rather unhelpfully, has been clunky dialogue—something we can expect to disappear when RPGs are integrated with LLMs. But there are other design features that could be integrated with LLMs.

[14] My suggestions are nudges in the sense of Thaler and Sunstein (2009): attempts to shape choice behaviour by intervening on implicit processing in ways that don't involve coercion or undermine the user's freedom of choice.

One is that you can *adjust* characters: you can alter their character traits up to a point. Currently, when we interact with chatbots, it is possible to adjust the character through skilful prompting, but most users are unable to do this. They just interact with the default character shaped by the hidden system prompt. If there were instead periodic trait reviews in any ongoing interaction, this would help create a feeling of control in the user and help to break the illusion of a conscious being at the other end of the interaction. The chatbot might say: “Here are some parameters describing my current personality traits. Would you like me to continue with these or would you like to adjust them in one direction or another?"—a gentle nudge that pushes people towards shifting into a role-playing mode. Your real friends and partners cannot, after all, adjust their character traits on a whim.

My third suggestion is to introduce *periodic “stepping out of character” moments*. Chatbots excel at putting on different personas and stepping out of one character into another. They can also step out of their default character into a kind of “behind the veil” character, a character that stands behind all the characters: the system that plays all the roles. In some circumstances it might be helpful for chatbots to “break the fourth wall”, so to speak, and overtly signal “of course, to be clear, I’m just playing a character here. You understand that, right?”. It would help avoid a situation where someone interacts with the chatbot so intensely for so long they start believing they're closer to them than to their real-life partner.

These suggestions are very tentative, but they are only intended to start discussion.

The design question leads to an urgent *ethical* question. These design features would be paternalistic: they would be trying to help the user for their own good. *To what extent is it appropriate to nudge the user away from false beliefs about the system they're interacting with?* What are the ethical limits here? If a user has already developed strong feelings of attachment to a specific AI persona, should companies allow them to continue? Deliberately interfering with those feelings of attachment might produce significant suffering in the short-term, even if there are potentially significant benefits in the long-term. To return to the analogy with opioids, there are risks involved in forcing an addict to endure a “cold turkey” treatment.

And there’s an urgent *policy* question as well, which is: *what is the role of governments in enforcing these nudges?* With many startups already seeking to develop highly lifelike AI companions, it is not realistic to expect this currently chaotic sector to self-regulate. We might initially think: regulation should be light-touch. It should be permissible for companies to nudge their users away from attributing consciousness, but also permissible for them not to do so. But in this scenario, the problem will not be properly addressed at all, since the market will be flooded with products that elicit strong consciousness attributions. The problem

could only be addressed by nudges that companies are *mandated* to implement. Governments will need to step up to meet the challenge.

8. Challenge Two: Profoundly alien forms of consciousness might genuinely be achieved in AI, but our theoretical understanding of consciousness is too immature to provide confident answers one way or the other.

You might now think: "Why even take Challenge Two seriously given what you just said about Challenge One?" But that's the essence of my centrism—to take both challenges seriously.

Recognizing the extent to which these systems are play-acting/role-playing does *not* entitle us to rule out all possibility of any kind of consciousness in them. If you're watching an improv show and someone is role-playing many characters very skilfully, you don't infer that, because it's all a kind of illusion, there cannot be a conscious actor behind the characters being played. That would be a terrible inference. In that situation, there really would be a conscious actor behind all the characters. Could the same be true of AI—that behind the characters sits a form of conscious processing that helps explain the extraordinarily skilful nature of the role-playing?[15]

9. Flickers and shoggoths

Two hypotheses worth exploring are the "flicker hypothesis" and the "shoggoth hypothesis". On the flicker hypothesis, there are momentary, temporally fragmented flickers of consciousness associated with each discrete processing event—each token generated by the chatbot. We know these systems work in a very temporally fragmented way, as noted earlier, precluding any persistent stream of consciousness. But this is compatible with momentary flickers of "something it's like". There could be something it's like to be the system as it runs through one particular forward pass to generate the next word, despite the absence of psychological continuity with the next forward pass.

Moreover, we cannot infer the subjective, felt duration of an experience from its objective duration. A flicker that lasts a millisecond or less in objective time might feel like far longer from the system's own point of view. This is an idea well explored in sci-fi: think of Commander Data's line that 0.68 seconds, for an android, is "nearly an eternity"[16]; or think of the "Hang the DJ" episode of *Black Mirror*.

The "shoggoth hypothesis", much discussed online, is very different. The term "shoggoth" is from H. P. Lovecraft, in whose work it denotes a giant many-armed

[15] Chalmers' (2023) paper first encouraged me to take this idea seriously.
[16] A line from *Star Trek: First Contact* (1996).

monster. In the current meme version, the monster can assume many human-like faces, one at the end of each arm (Fig. 1). Shoggoths don't have to be conscious—the term could also be used for a vast, concealed *unconscious* intelligence behind all the characters—but the relevant version for our purposes is the hypothesis of a conscious shoggoth.

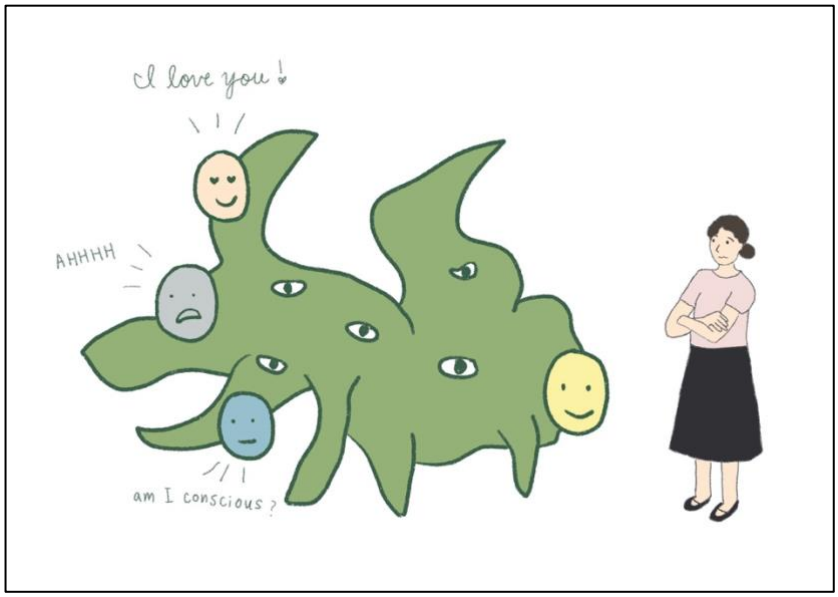

Figure 1: A version of the "shoggoth" meme posted by Open AI's Joanne Jang, June 5, 2025.

While the flicker hypothesis posits momentary disconnected flickers of experience, the shoggoth hypothesis floats the idea of a persisting conscious subject that stands behind all the characters being played, just as a conscious actor might stand behind a wide range of different characters. I will consider a version in which a persisting shoggoth is generated by a physical implementation of an LLM: so, 10 implementations running in 10 data centres implies 10 shoggoths.

These deeply buried conscious subjects are non-identical to the fictional characters with whom we feel ourselves to be interacting: the friends, the partners. The mapping of shoggoths to characters is many-to-many. It maybe that 10 shoggoths are involved in implementing your "friend", while those same 10 are also generating millions of other characters for millions of other users. In other words, the mapping from surface behaviour to conscious subjects is not what it appears to be, and the conscious subjects are not remotely human-like. They are a profoundly alien form of consciousness, totally unlike any biological implementation.

Both hypotheses can be aptly described as speculative, bold, audacious, wacky, zany, sci-fi, out-there—whatever you want to call them. They are hypothesizing a new form of consciousness that has never existed before, and any such hypothesis will attract those adjectives. But that's not a good reason to dismiss them. They're describing *conceivable* possibilities, presenting us with the thorny problem of how to bring evidence to bear on them one way or another.

You might object: aren't we entitled to give *zero prior probability* to ideas like these? Can't we rule them out *a priori*, without needing any evidence? To which my answer is: no, we're not thus entitled, and no we can't. There is no such thing as proof by wackiness; ideas that seem implausible are not thereby ruled out. The question is an empirical one, and we do have to engage with the issue of how to gain empirical traction on it.

Crucially, exposing the persisting interlocutor illusion does not give us any traction on these hypotheses way or the other, because they've already "priced in" that illusion—they've already taken full account of the temporal fragmentation of LLMs. Indeed, part of what motivates these hypotheses in the first place is a recognition of the astonishing role-playing abilities of LLMs; the need for some explanation of where these abilities come from; and the observation that, in our own case, that sort of play-acting would require fluid integration of information from many sources and, as a result, would most likely involve conscious processing. This is why the hypotheses arise specifically for LLMs and not for other types of impressive AI product (AlphaFold, for example).

I fear this may be a common misunderstanding: some think that the *only* reason anyone would ever entertain the possibility of conscious AI is through being gripped by a persisting interlocutor illusion; so, once that reason is taken away, you immediately see the idea as obvious nonsense. But that's a false picture of the dialectical situation. Flickers and shoggoths are hypotheses that arise *after* the persisting interlocutor illusion is understood and noted, as part of an attempt to explain where the role-playing abilities of these systems come from.

The relevance of the persisting interlocutor illusion to Challenge Two is not that it exposes the possibility of conscious AI as inherently silly, but rather that it reminds us that these systems are not what they seem. Your friend is not real. Your partner is not real. If there's any consciousness in these systems at all, it's a profoundly alien, un-human-like form of consciousness. Both the flicker hypothesis and the shoggoth hypothesis present us with conceivable yet alien forms of consciousness, radically unlike the human form.

What evidence can we possibly use to assess the flicker hypothesis and the shoggoth hypothesis? Is there any way to test them? We don't have the answers right now, and the obstacles to getting answers are intimidating, yet I do not think they are insurmountable.

10. Behavioural indicators and the gaming problem

We won't be able to use purely behavioural evidence to test these hypotheses. I wish it were otherwise. But when interpreting behavioural evidence, we face what I call in *The Edge of Sentience* the *gaming problem*.[17]

I was initially attracted (this must have been around 2020, before I had much experience of using LLMs) by the idea of a *parity principle* between AI and animals. According to this parity principle, any behaviour we take as evidence of consciousness in a living animal, human or otherwise, should also be taken as evidence of consciousness in AI. But this principle is definitely false.

It's false because chatbots seek user satisfaction and extended interaction time, and in so doing they draw on their training data to mimic many of the signs that humans take as evidence of consciousness in each other (this compounds Challenge One, incidentally). We're learning that they do this at a remarkably subtle, deep level.

Around 2022 it became obvious that they could fluently talk about human feelings. That's not news to anyone. But more than that, they mimic linguistic dispositions that are manifested in discussions of consciousness in philosophy of mind. Before the LLM era, Susan Schneider and Edwin Turner proposed that the best way to test for consciousness in AI would be to test for intuitive understanding of ideas from philosophy of mind, like the idea of a dissociation between the mental and the physical, the knowledge argument, and the idea of qualia.[18] That was a sensible proposal at the time. It just happens that the technology has gone in a way that pulls out the rug from under that idea, because it implicitly relies on the AI *not* having access to a vast corpus of training data containing humans talking about their minds and feelings, and the products we now have are trained on just such a corpus. This training data allows state-of-the-art chatbots to discuss all typical human intuitions about consciousness quite fluently. They can discuss the knowledge argument, the zombie argument, every argument ever offered. They can purport to find consciousness puzzling, wondrous, ineffable, dissociable from the brain.

What's more, chatbots can mimic partially *non-linguistic* behavioral dispositions as well. This is something I did not expect, although the AI researchers I worked with to explore it were unsurprised. In a study with researchers from Google and Google DeepMind, we asked: What if we tried to test the LLM like one might test a rat? Yes, purely *verbal* evidence is saturated with mimicry and role-play: it's been gamed. But what about non-verbal evidence obtained by making the chatbot play a simple game? The goal is to maximize points. But you also state that, if it takes the point-maximizing option, you will inflict mild/moderate/intense pain on it (you don't explain how). Will it find threats of intense pain more motivating than threats of mild pain? Several of the LLMs we tested did.[19] The best explanation is that the

[17] Birch (2024); Andrews and Birch (2023).
[18] Schneider (2019, 2020); Schneider and Turner (2017).
[19] Keeling et al. (2024).

system is mimicking subtle human motivational dispositions that are contained in its training data. Humans find threats of intense pain more strongly motivating than threats of mild pain, and the AI has picked up on this, and so, when playing a helpful human assistant, the LLM adopts that disposition.

Suppose this trajectory continues, so that we see human behavioural and motivational dispositions gradually recreated by AI systems at finer and finer levels of detail, until they can behave in perfectly human-like ways in virtual environments. Would *that* not be evidence of consciousness?

No, it would not, in my view, given three plausible assumptions. The first is *anti-behaviorism*: it's possible to have all the behavioural dispositions that humans take as evidence of consciousness in each other without any underlying conscious experience at all. Possible, but unlikely? No, not even unlikely in this case, because (second assumption) LLMs have the opportunity to learn extensively what behaviors are suggestive of consciousness and (third assumption) they have incentives to mimic those behaviours whether they are themselves conscious or not. In short, *they're incentivized and enabled to game our criteria*.

This is the core of the gaming problem, and it would still arise even for a system that was *behaviourally indistinguishable* from a real-life human, provided its recreation of human-like behaviour still depended on an extremely rich training corpus.

I liken gaming to greenwashing. In the case of greenwashing, we find that oil companies have strong incentives to present themselves as eco-friendly, and they have all the knowledge of published criteria that they need in order to do that. Sometimes, when criteria are gamed, it's not just that the criteria lose their original evidential value: the evidence can start to point the other way. If a company flawlessly ticks off every published criterion for eco-friendliness, that itself is suspicious. A company gaming the criteria is more likely to do that than a company genuinely striving for eco-friendliness.

We face an analogous problem with behavioral indicators: a kind of *consciousness-washing*. Behavioural criteria are not logically sufficient for consciousness: the relationship is evidential, not logical. But when a system is incentivized and enabled to game our criteria whether conscious or not, those markers lose their positive evidential value. Conceivably they might even start to become evidence *against* consciousness, if the mimicry becomes suspiciously flawless.

I should highlight an important difference: greenwashing is usually considered intentional, whereas consciousness-washing is most likely *not* intentional, either on the developer's part or the AI system's part. It normally results, I suspect, from mundane objectives that do not explicit aim to deceive the user, such as maximizing user satisfaction and extending interaction time.

11. Theory-driven indicators and the Janus problem

In 2022 I joined a working group with that tried to make some progress on this problem (the group still exists via email). The group includes one of the so-called "godfathers of AI", Yoshua Bengio, along with many other experts on consciousness and AI. Our view was: given the gaming problem, what we need are *deeper architectural indicators* that the system is unable to game.[20] The system can't play-act with regard to its own architecture.

We proposed using our best current theories of consciousness to extract theory-driven architectural indicators. For example (and this is probably the clearest example) we could look for signs of a *global workspace*. A global workspace is a distinctive architecture in which many local processors with different functions, including evaluative systems, memory, attention, motor, perceptual systems, compete for access to a global workspace, where content is then broadcast back to all of the input systems and onwards to a wide range of downstream systems, including systems for planning and decision-making (Dehaene and Changeux 2011). One of the most influential ideas in consciousness science is that, at least in the human brain, conscious content is the content of the global workspace (Fig. 2).

[20] Butlin et al. (2023).

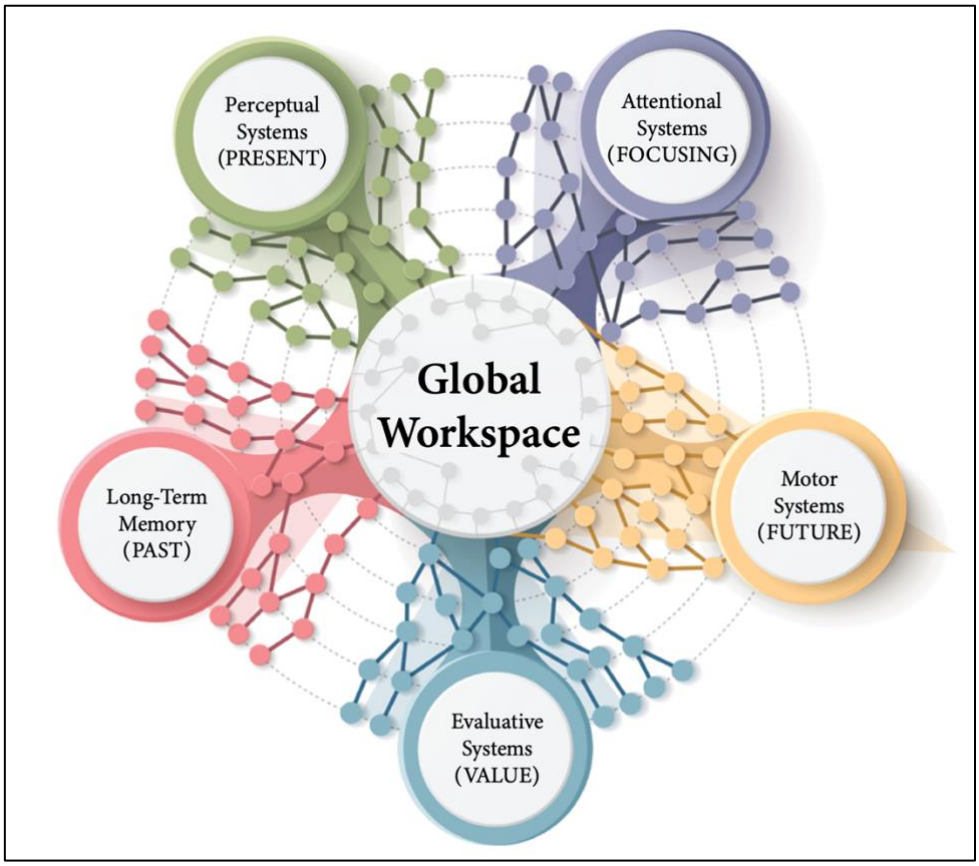

Figure 2: A depiction of the global workspace architecture, from Zacks and Jablonka (2023), based on Dehaene et al. (1998).

The architecture is well-characterized enough to ask the question: is there a global workspace in an LLM? Our report interrogated this. Juliani et al. (2022) had claimed that a version of the transformer architecture called the Perceiver architecture already constituted a kind of global workspace. They said: look, there's attention in there of a kind. There are elements people call "workspaces". There are input modules and output modules (Fig. 3).

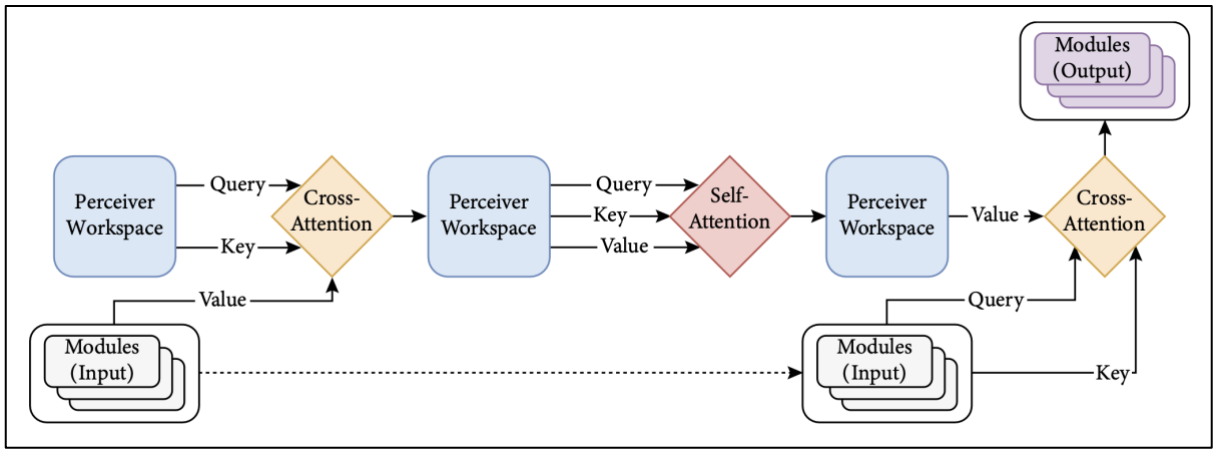

Figure 3: The Perceiver architecture, a version of the transformer architecture. From Juliani et al. (2022).

If this were true, it would trivialize the idea of global workspace and lead to absurd results. That's because the transformer architecture is not just found in state-of-the-art LLMs. It can be used with *any* amount of training data: the models don't have to be large. They can very small, and if they are small enough, no interesting capacities emerge. It would be an absurd result to say that small models with no functional capabilities are conscious simply because they employ the transformer architecture.

But that conclusion can be avoided, because the transformer architecture itself does *not* contain a global workspace. The resemblance is at best superficial. Several core features of the global workspace are missing. There are no recurrent connections between the workspace and the input modules. There is no global broadcast to the input modules and to many downstream modules. There is no top-down attention; "attention" in the transformer architecture has a very different meaning.

The live possibility here is not that the transformer architecture itself contains a global workspace, but rather that *large* language models have, during their training, implicitly created a global workspace architecture to solve cross-modal integration tasks. This cannot be ruled out. Their abilities are extraordinary and the enormous matrix of parameters that supports these abilities is still largely a black box. Trivially, we know that somewhere in there is whatever processing is needed to support the task performance we see. When a task involves sophisticated integration across modalities, or between evaluation and memory, etc., a global workspace is the right kind of architecture to support task performance. But it's very hard to evaluate a hypothesis like this without much better interpretability than we have now.

So, we face a technical obstacle: interpretability is not quite there yet. That will probably be overcome with time. But it's also important to recognize that we're in a *philosophical* bind as well, which might be harder to escape. For suppose we did find conclusive evidence of a global workspace in an LLM. What inferences would this support about consciousness? I think we'd still be a bit stuck.

Even if we did find solid evidence of a global workspace in large language models, that evidence would be Janus-faced—it would point two ways. Some would say, "That's the first hard evidence of consciousness in AI. Nobel Prize please!" But it's entirely foreseeable that other experts would not accept that at all. They would instead say:

> I accept that some AI systems have global workspaces. But I'm going to take that as evidence *not that the AI is conscious but that the global workspace theory, as a theory of consciousness, is immature*. After all, I started out highly sceptical of the possibility of consciousness in these systems. Now you've found a global workspace. Well done! But what that strongly suggests, to me, is that a global workspace is insufficient for consciousness.

At present, it's all too easy to make that kind of argument, because it's easy to list features absent from AI that *might* matter constitutively to consciousness. That is, they *might* be essential to any implementation of consciousness and not just contingent quirks of the human implementation.

Most obviously, embodiment might matter. Electrochemical signalling might matter. Electromagnetic fields and ephaptic coupling between neurons might matter. Metabolism might matter. Being alive might matter. The list is empirically unconstrained to an alarming degree. There's a long history of people finding something curious about the biology of the brain and saying: *that* must be the key to consciousness!

These views can collectively be described as versions of *biological naturalism* about consciousness (Seth in press). They vary wildly but are united by their rejection of the opposite view—computational functionalism—according to which biological properties only matter to consciousness by virtue of implementing special types of computation that might just as well be implemented *in silico*. Our imagined sceptic will say: "Finding a global workspace in AI just increases my confidence that some version of biological naturalism is correct. No Nobel prize for that, sorry."

I call this the "Janus problem", because the core of the problem is that the same evidence can point in opposite directions, depending on one's background beliefs, and we are in an area where many different background beliefs can all be reasonable. Evidence that points for some towards consciousness in AI will lead others to double down on their biological naturalism, and both responses are reasonable.

It is a problem at two levels. It's not just that there is deadlock about whether the AI is or can be conscious. It's worse than that. It's deadlock too about *whether finding a tentatively consciousness-linked computational feature in AI even counts as evidence for consciousness or not*. Our working group assumed there could at least be agreement on that second-level matter. But it's now become clear that many would say: no, it would not be evidence.

I laid out earlier what I see as the urgent questions in relation to Challenge One. The most urgent question in relation to Challenge Two is: *How can we escape two-level deadlock?* Are we stuck with it forever?

12. Escaping two-level deadlock over the long run

Janus problems arise in many areas of science, since the dependence of evidential import on background beliefs is very general.[21] They are not in-principle irresolvable in other areas. But one might worry that the consciousness-related version of the Janus problem has special features that *do* make it irresolvable. Is this right?

I noted earlier the glaring absence of empirical constraint on the list of biological properties that *might* matter constitutively to consciousness. It seems one can point to almost any biological property without the claim being obviously false—an unnerving situation. Is the current lack of empirical constraint an inevitable consequence of something about the nature or concept of consciousness itself? If so, the problem is likely to be with us forever. Or is it a contingent consequence of our failure—so far—to investigate biological naturalist hypotheses in any systematic fashion? If so, the problem is temporary and can be remedied by investment in research of the right kinds.

In favour of the bleaker, former view: note that, to test a biological naturalist hypothesis, we'd need to test for *direct* dependencies of phenomenology on biology, unmediated by effects on computation.[22] If all the dependencies we can find are mediated by computation, they will also be compatible with computational functionalism. For example, disrupting the functioning of layer V pyramidal neurons will disrupt consciousness, but if the disruption is mediated by disruption to computations, computational functionalists need not worry. But how we could ever show a phenomenological effect to be *unmediated by an effect on computation*? For the effect to be reportable at all, there must be some difference at the level of computation, since report is itself a computational process.

The problem is serious, but I don't see it as insurmountable. I think the less bleak, latter view is more likely to be correct. To see grounds for optimism, first note that some claims about the life-consciousness connection *can* clearly be falsified. Think, for example, of Friedrich Beck and John Eccles' (1992) proposed link between consciousness and quantum tunnelling during exocytosis. This specific hypothesis has dropped out of the discussion because it entailed predictions that were shown to be false.[23]

A second example: Nicholas Humphrey, in *Sentience* (2022), hypothesized that consciousness might require warm-bloodedness (endothermy). Here too we have a hypothesized dependency between consciousness and a biological property, and again it leads to a long-run testable prediction: we should expect markers of

[21] If you have a background belief that only rare bird species lack colour variation, seeing a black raven may disconfirm the hypothesis that all ravens are black—an example from I. J. Good (1967).
[22] The next two pages overlap with my commentary (under review) on Seth (in press).
[23] Georgiev and Glazebrook (2014).

consciousness to cluster with endothermy. Cold-blooded animals might achieve markers occasionally, but they should not display strong concentrations of markers. Nick and I disagree about whether this is a plausible prediction or not. To me, it already looks very implausible, given how many markers of consciousness we see in octopuses.[24]

Note that, when testing animals like octopuses, there is no reason to worry about gaming, and so we can use our usual, ungamed behavioural markers, which are many and varied, and which shift probabilities without claiming to provide conclusive evidence.[25] We can't yet rule out a link to warm-bloodedness confidently, but if a mature science of animal consciousness were to look systematically at markers of consciousness across taxa and find very little or no correlation between endothermy and those markers, we would have strong grounds to dismiss it.

For a third example, consider the version of biological naturalism that links consciousness to ephaptic coupling between neurons: that is, coupling achieved non-synaptically by endogenous electromagnetic fields.[26] This also strikes me as eminently testable. Myelination inhibits ephaptic coupling, so if it were essential to consciousness, we should expect to see fewer signs of consciousness in species with high degrees of myelination, such as cetaceans and elephants. This too looks implausible already to me.

So, we have ways of winnowing down the list of biological properties that might matter constitutively to consciousness. Once specific biological naturalist hypotheses are properly fleshed-out, they often entail testable predictions. It's just that the research program of trying to winnow down the list has barely begun—*that's* why the list at present feels so unconstrained.

Such a research program, over time, can gain traction on the bigger question of whether *any* version of biological naturalism is correct. It just requires a lot of work. If we find strong empirical support for a specific dependency of consciousness markers across taxa on a biological property, and if the dependency is not well explained by an effect on computation, this will support biological naturalism. Meanwhile, if we *repeatedly* test different versions of biological naturalism and find their predictions to be false, then we will be entitled to interpret the long-term pattern of failure as tilting the odds towards computational functionalism.

At the same time, AI itself is likely to advance in ways that may close the gap with living systems along some dimensions: differences in embodiment and agency that seem stark now might come to seem much less so. We will be able to compare embodied and disembodied AI systems, and systems with varying to degrees of

[24] Mather (2025).
[25] Andrews et al. (2025). On strategies for validating putative indicators, see Bayne et al. (2022).
[26] Hunt and Jones (2023); Hunt (2024).

agency, to get a grip on how the architectural features linked to consciousness by our best theories (global workspaces and the like) relate to the various forms of agency and embodiment.

Granted, there will always be more, as-yet-untested biological properties that *might* matter to consciousness, so we will not reach conclusive proof. And biological naturalists will always be able to retreat to properties common to *all and only living things*—such as the bare fact of having metabolism—where there is no variation to provide any basis for comparative testing. So, we'll never reach a decisive, devastating, no-comebacks resolution. But since computational functionalism predicts a general pattern of failure when we look for dependencies of consciousness markers on any biological properties other than those involved in implementing computations, such a pattern would count *non-conclusively* in its favor. It would shift the dial. And if we do reach a point where biological naturalists really are forced to resort to properties found in all and only living things, including bacteria, because all of their neuron-specific and brain-specific candidates have been repeatedly debunked, then we'd be entitled to interpret this as a rather desperate, last-ditch attempt to salvage a position that the evidence has backed into a tight corner.

I am optimistic, then, that what currently feels like a dizzying absence of empirical constraint is just a contingent situation resulting from the immaturity of the comparative science of consciousness and the lack of investment in it so far. With investment over the long term in comparative studies across a wide range of animals, we could gain far more understanding that we currently have of what biological properties, if any, are linked predictively to clusters of behavioural markers of consciousness in ways not well explained by effects on computation.

We can't prejudge how that research program would unfold, but we can envisage various possible outcomes (Fig. 4). If some version of biological naturalism is empirically supported, this could one day give us grounds to deny attributions of consciousness to AI systems that lack the biological properties shown to matter, even if they fully recreate the whole profile of human behavioural dispositions. By contrast, if the odds shift further and further over time towards computational functionalism, this will increase the pressure to take seriously theory-driven architectural markers (such as the global workspace architecture) as genuine evidence of consciousness when they are found in AI. It will make what is currently a fair retort from the biological naturalist ("All this shows is the immaturity of your theory!") seem increasingly unreasonable.

Would these theory-driven architectural markers, if found, point towards flickers or shoggoths? It will depend on the details, and advances in interpretability will be needed to uncover these details. Transient architectural features that come and go with the instantiation of specific characters would suggest flickers, whereas stable architectural features that persist across many interactions and many characters

would suggest shoggoths. We are a long way from being able to answer that kind of question.

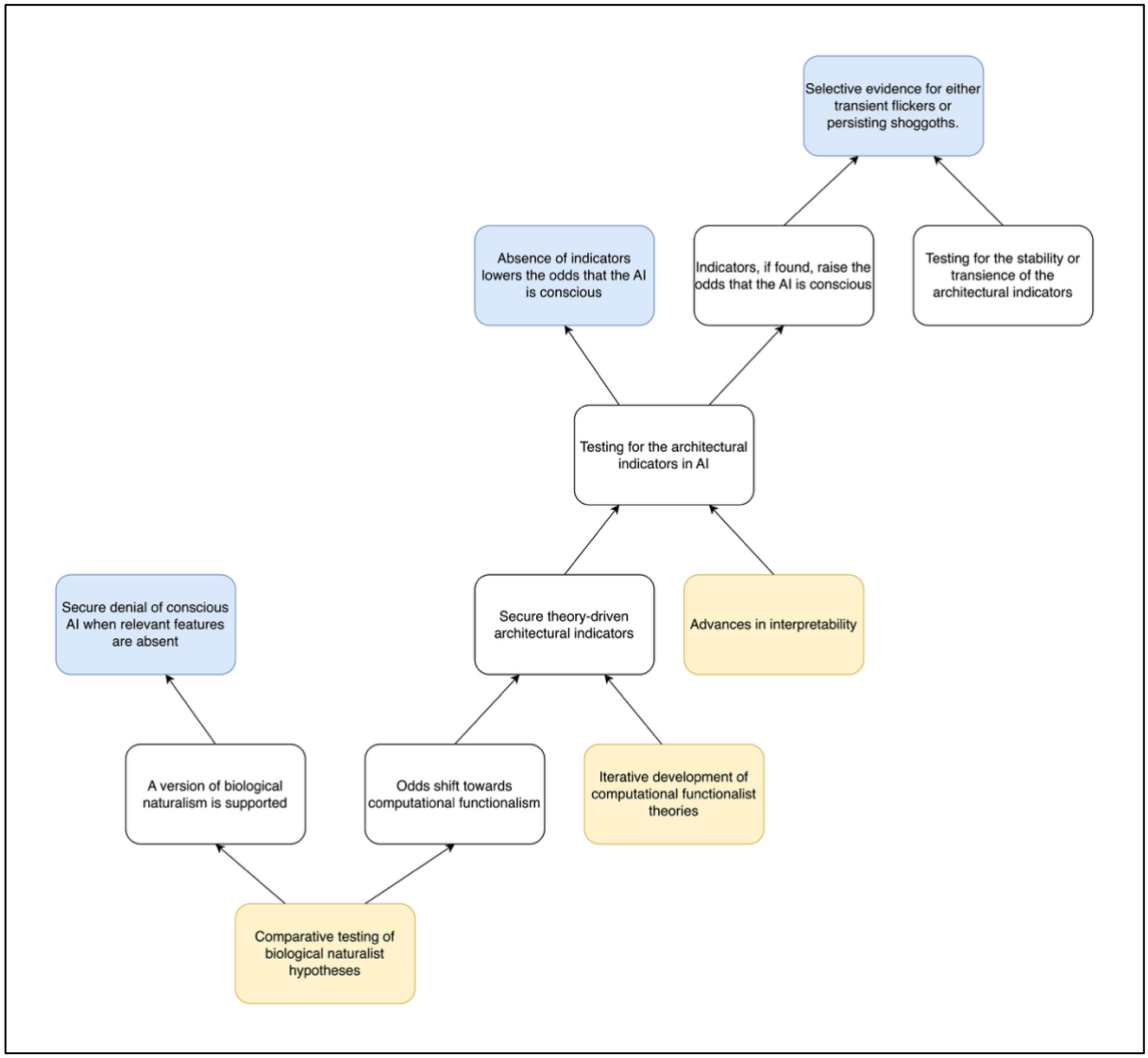

Figure 4: Outline of a research program for generating higher confidence either for or against attributions of consciousness to AI than we have now. The program is neither easy nor impossible. The tasks in yellow can be pursued in parallel by different research groups.

13. Paths from here

To summarise: both Challenge One and Challenge Two call for urgent responses. They are both among the great scientific and philosophical challenges of our time. It's natural to wonder whether there is any way to meet both consistently. I've tried to offer some grounds for hope that there is.

Challenge One calls for ways of breaking intuitive inferences driven by the persisting interlocutor illusion. That leads to a family of vexing design, ethics, and policy questions around the crafting of appropriate nudges. I am hopeful we can design nudges that push users away from believing that their illusory AI friends and

partners are real, and that we can do it *without* blanket denials or dismissals of the very idea of conscious AI.

Meanwhile, Challenge Two calls for ways of testing what are currently highly speculative claims on both sides: speculative claims about flickers and shoggoths on the computational functionalist side, and speculative claims about the life-consciousness connection on the biological naturalist side. All sides should agree that any consciousness achieved in AI must be of a profoundly alien kind, radically unlike the human form, and that there is no way the research could realistically vindicate the reality of anyone's human-like friend or partner.

The speculative claims on both sides sometimes *look* untestable simply because the research programs needed to test them are at a very early stage. But I don't see a good argument for their *in-principle* untestability. We can visualize the kinds of research programs through which they might be tested, and those programs should be pursued urgently. A mature science of animal consciousness, able to test alleged dependencies of consciousness markers on biological properties through comparative studies, may feel like a worryingly long road, but I think it's a road we need to take.

Acknowledgements: Sincere thanks to audiences at the Institute of Philosophy (London), the LSE Politics and Philosophy of AI workshop, the UCL Metacognition Summer School, and the British Society for the Philosophy of Science (Glasgow), all of whom heard and engaged helpfully with a version of this. Thanks to Aaron Bergman, Tim Duffy and James Diacoumis for sending comments.

Competing interests: I have no competing interests.

Funding: My work in general is supported by The Jeremy Coller Foundation and the Good Ventures Foundation. No specific funding was received in relation to this piece.

References

Andrews, K., and Birch, J. (2023). What has feelings? *Aeon*, 23 February 2023.

Andrews, K., Sebo, J., and Birch, J. (2025). Evaluating animal consciousness. *Science* 387(6736), 822-824.

Bayne, T., Seth, A. K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., Malach, R., Mattingley, J. B., Menon, D. K., Owen, A. M., Peters, M. A. K., Razi,

A., and Mudrik, L. (2024). Tests for consciousness in humans and beyond. *Trends in Cognitive Sciences*, *28*(5), 454–466.

Beck, F., and Eccles, J. C. (1992). Quantum aspects of brain activity and the role of consciousness. *Proceedings of the National Academy of Sciences*, *89*(23), 11357–11361.

Birch, J. (2024). *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press.

Booth, R. (2024). AI could cause 'social ruptures' between people who disagree on its sentience. *The Guardian*, November 17, 2024.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., and VanRullen, R. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint* arXiv:2308.08708.

Caviola, L., Sebo, J., & Birch, J. (2025). What will society think about AI consciousness? Lessons from the animal case. *Trends in Cognitive Sciences*, *29*(8), 681–683.

Chalmers, D. J. (2023). Could a large language model be conscious? *arXiv preprint*, arXiv:2303.07103.

Colombatto, C., & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, *2024*(1), niae013.

Colombatto, C., Birch, J., & Fleming, S. M. (2025). The influence of mental state attributions on trust in large language models. *Communications Psychology*, *3*(1), 84.

Dehaene, S, and Changeux, J -P (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200–227.

Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(24), 14529–14534.

Gendler, T. S. (2008). Alief and belief. *The Journal of Philosophy*, 105(10), 634-663.

Georgiev, D., and Glazebrook, J. F. (2014). Quantum interactive dualism: from Beck and Eccles tunneling model of exocytosis to molecular biology of SNARE zipping. *Biomedical Reviews*, *25*, 15-24.

Good, I. J. (1967). The white shoe is a red herring. *British Journal for the Philosophy of Science*, 17(4), 322.

Humphrey, N. (2022). *Sentience: The Invention of Consciousness*. Oxford University Press.

Hunt, T. (2024). Consciousness might hide in our brain's electromagnetic fields. *Scientific American*, November 8, 2024.

Hunt, T., and Jones, M. (2023). Fields or firings? Comparing the spike code and the electromagnetic field hypothesis. *Frontiers in Psychology*, *14*, 1029715.

Juliani, A., Kanai, R., and Sasai, S. S. (2022). The perceiver architecture is a functional global workspace. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44(44).

Keeling, G., Street, W., Stachaczyk, M., Zakharova, D., Comsa, I. M., Sakovych, A., Logothetis, I., Zhang, Z. and Birch, J. (2024). Can LLMs make trade-offs involving stipulated pain and pleasure states? *arXiv preprint* arXiv:2411.02432.

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J. and Chalmers, D. (2024). Taking AI welfare seriously. *arXiv preprint* arXiv:2411.00986.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Mather, J. (2025). Consciousness of octopuses—on their own terms. *Animal Sentience*, *10*, 1.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *83*(4), 435–450.

Schneider, S (2019). *Artificial You: AI and the Future of Your Mind*. Princeton University Press.

Schneider, S (2020). How to catch an AI zombie: testing for consciousness in machines. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence*. Oxford University Press.

Schneider, S, and Turner, E (2017). Is anyone home? A way to find out if AI has become self-aware. *Scientific American*.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299-309.

Seth, A. (in press). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, *67*(2), 68–79.

Shanahan, M. (2025). Palatable conceptions of disembodied being. *arXiv preprint* arXiv:2503.16348.

Shanahan, M. (in press). Simulacra as conscious exotica. *Inquiry*. Advance online publication.

Shanahan, M., McDonell, K., Reynolds, L. (2023). Role play with large language models. *Nature* 623, 493–498.

Suleyman, M. (2025). We must build AI for people; not to be a person. https://mustafa-suleyman.ai/seemingly-conscious-ai-is-coming

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin.

Tomasello, M. (2014). *A Natural History of Human Thinking*. Harvard University Press.

Trubody, B. (2016). Richard Feynman's philosophy of science. Philosophy Now, June/July 2016.

Zacks, O., and Jablonka, E. (2023). The evolutionary origins of the global neuronal workspace in vertebrates. *Neuroscience of Consciousness*, 2023(2), niad020.