v2: 20 May 2025

Preprinted: 28 February 2025
Peer-approved: 10 April 2025

Qeios, Vol. 7 (2025)
ISSN: 2632-3834

Commentary

The Hard Problems of AI

Olegas Algirdas Tiuninas[1]

1. Charles University, Prague, Czech Republic

There is currently an enlivened debate regarding the possibility of AI consciousness and/or sentience, as well as arguably more partial capabilities we associate with consciousness such as intelligence or creativity. The debate itself can be traced back to the inception of computing, but its current revitalisation is powered by recent advancements in the field of artificial intelligence that have seen a swift increase in its capabilities to act in seemingly human-like ways. I argue that the debate is methodologically flawed, as it approaches the question of AI consciousness, intelligence, etc., as a decidable question dealing with matters of fact. Those engaged in the debate are driven by a desire to find a suitable definition of, e.g., consciousness that would allow them to definitively settle the question of whether a particular AI system is conscious. However, drawing on Ludwig Wittgenstein's later philosophy, I argue that no such definition exists because the predicates in question are inherently vague (meaning that any verdicts they yield are bound to be vague, too). Moreover, the impression that we might be dealing with directly unobservable matters of fact is itself a flawed generalisation of the practice of observation reports to the practice of sensation reports[1]. In reality, third-person consciousness (sentience, agency, etc.) attributions are independent of a stipulated internal process happening inside those persons (or systems, in the case of AI). Therefore, the only sense in which the question of, e.g., AI consciousness can be meaningfully asked is a pragmatic sense: what is it best to *think of such systems as?* But this question is subject to sociological and psychological factors, not conceptual ones. Therefore, it cannot be decided by the aforementioned strategies.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

Introduction

With the rising capabilities of artificial intelligence, there is a sense of increased urgency regarding the questions of AI consciousness, sentience, agency, intelligence, capability to be genuinely creative, and so on. I will call those the 'hard problems' of AI. The name is meant to draw an analogy with the Hard Problem of consciousness in the philosophy of mind, popularized by David Chalmers. The distinguishing feature of what I call 'hard problems', just like in the case of consciousness, is that it is widely believed that such problems, if solvable at all, can only be solved by in-depth conceptual reflection. Just like in the case of consciousness, no amount or kind of empirical evidence is seen as ultimately relevant to the resolution of the problem. In contrast to the 'easy' problems, which are more or less the question of obtaining the requisite empirical results and best summarizing them, the 'hard' problems are seen as the proper domain of philosophy. I argue that such problems are red herrings

doi.org/10.32388/13B814.3

that arise from unwarranted essentialist assumptions regarding entities central to our practices and world-modelling. Despite seeming to pose serious ontological questions, they merely distract from the very real challenges to do with artificial intelligence[1]. I will also speak about the sense in which these questions *are* potentially important, but the answer to them is not the kind of ontological discovery that many philosophers would want, but rather a pragmatic negotiation of a linguistic convention.

The Context of the Hard Problems

It is worth pointing out that the pioneers of computing themselves have been quite dismissive regarding such 'hard' questions. In his seminal[2] paper, Alan Turing stated that the question 'Can machines think?' is 'too meaningless' to merit serious discussion. Edsger Dijkstra quipped that 'the question of whether a computer can think is no more interesting than the question of whether a submarine can swim.' A similar sentiment has been voiced more recently by the likes of Noam Chomsky, John Pollock[2], and Blaise Agüera y Arcas[3], all of whom find philosophical speculation on the possibility of machines possessing consciousness, understanding, etc., misguided. On the other side, you have thinkers like David Chalmers, Patrick Butlin, Jaan Aru, John Searle, Patricia Churchland, or Hubert Dreyfus, who treat these questions as substantial, even as their conclusions differ.

While I am sympathetic to those who find the hard problems to be pseudoproblems, I find that many such critiques lack a coherent articulation of why exactly such questions are misguided. It appears that not only are the questions too senseless to deserve discussion, but their senselessness is seen as too obvious to deserve an explanation. This needlessly fosters a disconnect between the two camps and causes the people engaged with the questions to continue in what appears to me to be a fruitless endeavour. Therefore, what I attempt to elucidate in this article are the fundamental issues inherent in such approaches.

The Framework of the Hard Problems

For those who find these questions meaningful, the fundamental strategy of trying to answer them can be modelled as a two-step process. First, the proposed criteria for predication of a certain quality are outlined. Secondly, it is assessed whether an actual or potential AI system could meet those criteria, thus possessing the quality in question; most usually, 'consciousness'. Therefore, in the remainder of this article, I will speak of the problem of AI consciousness as a synecdoche for the hard questions more generally. As I will further argue, such a substitution is justified, as the conceptual issues present in the case of consciousness can also be seen to be damning in all the other cases.[4]

For example, in Chalmers' recent article 'Could a Large Language Model be Conscious?'[3], he considers various candidate theories for the necessary conditions of consciousness. Such candidates include self-report, seeming conscious, conversational ability, general intelligence, senses and embodiment, word-modelling, etc. Chalmers asks to what extent current or potential LLM systems can instantiate these properties and contends that the question of whether these models are or could be conscious hinges on that. Similarly, Butlin et al. draw on computational functionalism to suggest that implementations of the kind of computations in neural networks that are analogous to the ones producing human consciousness would provide an indication of AI consciousness[4]. They then analyse, at length, whether artificial agents do or could exhibit such structural properties.

There is a two-fold problem with this approach. The first is that the potential fact of AI consciousness is treated as an *additional* fact over and above the proposed indications of consciousness. What the authors attempt to discover is a certain hidden property of these systems that is merely manifested through the observable properties. While highly intuitive, this is a misleading picture of our third-person consciousness attributions. In fact, it is a consequence of what, following Horwich, I will call the extended private arena model of experience[1]. The second is the assumption that there is a particular set of definite features that make something conscious or non-conscious (not *what* such a set is, but *that* there is such a set). This falls prey to Wittgensteinian objections to do with the vagueness of definitions. Mistaken perceptions of these two issues intertwine to make the problem of AI consciousness seem like a genuinely puzzling issue, while in reality, it lacks meaningful content.

The Private Arena Model

First, I will address the 'additional something' picture of third-person consciousness that leads us to believe that there is a meaningful question about hidden AI qualia that can be inferred from certain observable properties of the system. It is a result of construing the practices of first-person consciousness attributions and third-person consciousness attributions as being essentially the *same* practice, albeit with different degrees of certainty – self-evident in the first-person case, merely inferred in the third-person case. This employs what can be called a private arena model of consciousness.

The private arena model is essentially a way of conceptualising qualia by analogy with external phenomena. When experiencing the world, I notice that some of my perceptions are readily corroborated by outside observers, while some – like, say, the pain in my left knee – only seem to be directly accessible to me personally. While the experience of my own pain is as real and immediate to me as an experience of an external object, it is not to others. A tempting picture, therefore, is to see these experiences as being located in a sort of quasi-space that only I am able to observe – a private arena, if you will. By extension, since I conceptualise others as having minds akin to mine, it seems natural to extend this picture to them also – to locate their experiences in a sort of quasi-space that only they are privy to (the 'extended private arena' model). And then, since there is a fact of the matter as to whether there are certain objects in regions of space that are inaccessible to me, there seems also to be a fact of the matter regarding whether the experiences of others do exist in those quasi-spaces. Therefore, what is at stake when we try to decide whether an AI system is conscious is deciding what entities are *out there,* i.e. the question of whether there is an entity like the qualia of a particular AI system X which is 'part of the world's furniture'.

However, this is an overextension of the spatial metaphor. It leads us to believe that the attribution of mental events to others is a result of making reasonably sure, through indirect means, that their respective quasi-spaces contain mental events of a certain kind.[5] But this mixes up two quite separate concepts: the properties of our mental representation of X and the properties of X that we perceive in order to bestow a certain mental representation.[6] Particularly, the determination of whether a particular entity has qualia is in no way present when *deciding* whether it is conscious. It is rather that *once* we decide that it is conscious, we are picturing it as having qualia in this sort of private arena, because it is the most convenient representation for us.

In *Philosophical Investigations*[5], Wittgenstein proposes a thought experiment: suppose that everyone has a box that only they can see the contents of. I look into my own box, and I can see that there is a beetle in it. While I cannot see the content of other people's boxes (though I can hear some buzzing coming from them from time to time), they also tell me that they have a beetle in their box. Therefore, as a community, we effectively agree to call the thing in the box 'beetle'.

Wittgenstein urges that in this scenario, it does not matter if everyone has the same thing in the box or if some or all of the other boxes are empty. 'Beetle' is the name we agreed to give to the thing in the box. The language-game is independent of the actual contents of the boxes. It is sufficient that *modelling* the other boxes as having beetles in them is helpful to me to explain and predict many of their external characteristics, including people's reports about the contents. The question of whether other people *actually* have a beetle in the box, rather than being difficult to decide, is simply meaningless to me; an answer either way affects nothing.

This is analogous to our relation to the internal experience of others. The basis on which we build a model of their behaviour as being produced by consciousness has nothing to do with ascertaining that the behaviour is actually produced by consciousness akin to ours (as Wittgenstein puts it in §303, 'just try – in a real case – to doubt someone else's fear or pain'. The overwhelming certainty that is often appropriate is not warranted by induction – from a single case, no less). The 'hidden ingredient' of another's qualia is simply not required for the system to work: the model of our own consciousness fitting another's behavior is sufficient[7]. Thus, what it effectively *means* for another to be in pain is to act in specific ways that are conducive to attributions of pain-states, nothing more.[8] And the same is true for other internal states.

Appreciating this difference between first-person sensation reports and third-person consciousness attributions – practices that, though interlinked in terms of mental representations, have principally different rules – dispels the impression that more than word-use is at stake in our investigation of the 'hard questions'.

The Vagueness of Definitions

The proponent of the hard question might still insist that while there are no facts about AI qualia that we can hope to discover, there is still a question of the *correctness* of calling them conscious. For example, we consider it correct to call other people, or certain animals, conscious, even though the debunking of the extended private arena model suggests that we are not discovering something about their having 'same as we do'. The question can thus be asked about AI agents in the same sense.

However, the principal issue is that we are treating a predicate like 'conscious' as more than a 'mere' word, more than a move in a language-game. Rather, we treat it as a label for an entity that has measurable properties. This assumption, though intuitive, can be shown to be unwarranted. Specifically, we are making a mistake in assuming that there are some objectively extant necessary and sufficient conditions for attributions of the predicate, which, even if they are not pre-reflectively clear to us, we can discover by looking at the concept 'carefully enough'.

Dijkstra's parallel is particularly instructive here. Let us pretend that someone indeed seriously posed a question of whether submarines can swim - or whether they are merely floating. How are we to answer such a question? No facts, or even potential facts about submarines, are uncontroversially qualifying. For there to be such facts, we would need first to have a clear and unambiguous definition of what 'swimming' means. Yet what we will find is that, just like in Wittgenstein's example with games, no such definition is readily available to us. What we do when we classify things as swimming versus floating is tied to certain family-resemblance type characteristics. In most cases that we are concerned with, they are present to a large enough degree to put the objects or creatures in question squarely in one category or the other. A couple of reasonable candidates for such characteristics in the case of 'swimming' might be intentionality and self-propulsion. Things that swim cause their aquatic motion through their own means, and they do intend to go in the direction they are, in fact, going. The case of submarines thus appears problematic because, unlike most things that swim, they possess one but not the other – they produce their own motion, but have no intention of going where they are going. In other words, 'swimming' is a term we use every day when speaking of a range of phenomena without thinking twice about it because most of those phenomena are striking enough for us not to bother with a precise definition. The behaviour of submarines lies on the outskirts of that range, and so we may wish, if we feel the need to relegate them to a particular category, to negotiate a linguistic convention that enables us to do that. But we will not be discovering anything philosophically interesting by such a process, neither about swimming nor about submarines.[9]

The only difference between submarines swimming and AI consciousness is that while seriously asking such a question about submarines appears to be silly, the notion of 'consciousness' is so central to our way of life that it appears to be 'real': the word seems to be pointing to something tangible. There is a kind of urgency in resolving the ambiguity regarding such predicates, which further contributes to the impression that there must be a matter-of-fact answer that is not a mere negotiation of a linguistic convention. But when we try to flesh out the reason for believing this – that entities like consciousness, intelligence, etc., have some essence, which ordinary words like 'swimming' do not – we turn up empty-handed. In reality, we apply the term 'conscious' to a range of phenomena that happen to be distinctive enough for us not to acknowledge that we are always operating with only an approximate demarcation of what is subsumed under 'conscious'. (To put it differently, the membership in the set of conscious things appears to be binary, even though it is continuous – because in most practical cases, it is significantly closer to 1 or to 0 for us to ignore the fact that it is never either). AI systems possess some, but not all, properties of things we usually deem 'conscious', and thus we are puzzled – since usually the attribution of the predicate is quite effortless. But to attempt to resolve this puzzlement by insisting that we must simply 'zoom in' on what is the essential property of consciousness is misguided.

I believe that the only reason we want there to be an essence to such entities is because the notion is so central to our world-modelling that we want it to have clear and distinct boundaries – making it more 'real' in our minds, more tangible (this is the same reason why the Humean account of causality as something continuous with correlation is unintuitive at first, as it challenges the perception of the central notion of causality as a separate entity). Thus, we try to artificially separate it from a melee of ambiguous terms with only approximate boundaries by insisting that the boundaries of *this* particular notion are sharp. This is a bias that was pointed out as far back as in Plato's *Dialogues*[6], where in the *Parmenides* dialogue, Socrates cheekily remarks that if things like beauty and justice are to have their essences, then so should hair and mud – which do not appear to be natural candidates for a lofty essentialist. Of course, a modern-day essentialist need not believe that consciousness exists in some kind of Platonic realm; many would perhaps shirk the notion – but it is the intuition that *something like* the Platonic essence picture is true that underlies attempts at definitions in terms of necessary and sufficient properties. To ask the question 'What does it mean to be conscious *exactly?*' is to view the word 'consciousness' as more than a move in the language-game – but that is all we have good grounds to believe it to be. Once we free ourselves from this unwarranted picture, the perceived potential for discovery in the case of AI consciousness dissipates.

Clearly, both the objection to the 'extended private arena' model and the vagueness-of-definitions objection can be applied equally well to other predicates that typically feature in the hard problems[10]. It is misguided to try to determine facts about some internal hidden process – there is no meaningful ontological question. And, once that is conceded, what is left is a linguistic question, which cannot be decided non-arbitrarily. Therefore, the hard problems of AI are pseudoproblems.

The Hard Problems Beyond Philosophical Speculation

So, what *should* we say about such questions? Should we think of AI systems as conscious or as genuinely creative, or can those only be the attributes of humans and some animals? 'Say what you choose, as long as it does not prevent you from seeing the facts'[5]. As long as we have a clear picture of what is happening with our relationship to AI systems and that it is the friction between the usual ease of attribution of 'hard question' predicates and the ambiguity in the case of AI systems that perplexes us, the philosophical quandaries are resolved. Once we realise this, we may want to call AI systems, e.g., conscious, or we may not. That may depend on the perceived comparative utility of both options, or on the habits that we will naturally fall into, a consensus that will naturally emerge (perhaps in a sort of way that took place regarding the consciousness of non-human animals)[11]. Right now, there is an intuitive resistance to granting AI systems human-like qualities such as intelligence or creativity, evident in the persistent moving of goalposts when testing for them – the Turing Test, once considered the gold standard for 'thinking', appears to many to be insufficient now that several LLMs are capable of passing it (see, e.g., [7]), and there is a rejuvenation of efforts to move away from the 'originality and effectiveness' model of creativity that held sway for decades, as that makes it potentially all-too-easy for certain AI systems to be seen as creative (this is referenced in footnote 10). I believe that as time passes and the capabilities of these systems become ever more advanced, it will just be too convenient to think of them as possessing such typically human qualities for anyone to seriously ask the question[12] (in that sense, I side with John Pollock [see footnote 1]). We will naturally choose a world model that gives us the greatest predictive ease. But whatever the case may be, this is not the kind of answer to the hard questions that philosophers engaged with the hard problems of AI hope for.

I am not denying that thinking of AI systems as conscious or intelligent may have significant implications for how we will practically interact with them. In that sense, the question may indeed be considered important. For example, Geoffrey Hinton, who shared the Nobel Prize in Physics in 2024 for his pioneering work in neural networks that form the backbone of many advanced AI systems we know today, tends to favour agential language in his interviews, perhaps to draw attention to the fact that the capabilities and potential dangers of such systems are not limited by them being mere 'imitations' of understanding. Given that John Searle, the author of the famous 'Chinese Room' argument against AI consciousness, has brazenly dismissed the possibility of an AI takeover on the basis that since AI is not conscious, it will never 'want' to take over, this concern is not misplaced[13]. (Hinton occasionally mentions, though, that the question of whether the machines really 'think' is quite unimportant, echoing Turing's sentiment.) It is true that our attitudes tend to naturally be quite different towards things we perceive as having consciousness versus things that we perceive not to (this is most obvious in the types of rights we are willing to grant creatures once we consider them conscious)[14]. We may thus wish to engage in a kind of conceptual engineering that will try to define the notions of consciousness, agency, etc., in a way that results in the most beneficial decisions regarding particular systems. However, these decisions would not be science-like statements about the world, but instrumental decisions about how it is most useful to model our experience given the particular tendencies of humans when interacting with differently modelled entities.

Conclusion

The question of whether an AI system is conscious, creative, intelligent, agentive, etc., when asked in the *strict* sense, is meaningless. What I mean by a *strict* sense is that the person asking the question is driven by an underlying assumption that an unambiguously correct answer to the question can be deduced by examining the limits of what falls under the predicates carefully enough. Since the predicates are vague, and the reason we even ask the question in the first place is because the behavior of an AI system lies on the outskirts of the applicability of a given predicate, no definitive answer can be given. On the other hand, if one asks the question in a pragmatic sense, i.e., in the sense of 'what is it most helpful to think of AI systems as?', some answers may turn out to be more useful than others. However, this is subject to psychological and sociological factors, not conceptual ones that philosophers are typically preoccupied with.

Statements and Declarations

Author Contributions

The sole author conceived the argument, wrote the initial draft, and reviewed and edited the final manuscript.

Footnotes

[1] Discussing those is beyond the scope of this paper. However, I mean the familiar problems of alignment, interpretability, security, bias, copyright, etc.

[2] ‘Once [my artificial intelligence system] OSCAR is fully functional, the argument from analogy will lead us inexorably to attribute thoughts and feelings to OSCAR with precisely the same credentials with which we attribute them to human beings. Philosophical arguments to the contrary will be passe’ [8].

[3] ‘it is unclear how we would distinguish “real understanding” from “fake understanding.” Until such time as we can make such a distinction, we should probably just retire the idea of “fake understanding.”’ [9]

[4] There is an outpouring of literature on the hard problems, so while in many cases consciousness is discussed first and foremost, the breadth of the discussion is also illustrated by the fact that some articles can be seen to disregard consciousness and instead focus on more specific facets of LLMs like creativity [10]., agency [11], or sentience [12], among others. As I said, the pitfalls of these questions are ultimately the same as in the case of consciousness, so I do not discuss them separately. My rationale is the same as that of Phil Hutchinson [13], who writes, ‘(...) I shall proceed in a manner that might seem to suggest that one can treat “thinking” and “understanding” interchangeably. I do so advisedly; the grammar of – i.e. the sensical uses to which we put – these terms is different. However, *the sorts of confusions we are led to in our philosophical considerations of these (grammatically distinct) terms stem from our being unconsciously in thrall to a particular picture; they are confusions that often have the same source and character.*’ (Emphasis mine)

[5] As Wittgenstein writes [14]: (...) In general I do not surmise fear in him - I *see* it. I do not feel that I am deducing the probable existence of something inside from something outside; rather it is as if the human face were in a way translucent (...)’.

[6] To highlight the difference between the two: observing certain physical phenomena, such as heat, might lead us to have a mental representation of atoms as little balls – a quite useful explanatory model. But this is different from having perceived them to be little balls or even earnestly believing them to be little balls, thus asking questions like ‘what color is an atom?’ etc.

[7] One might wish to insist that there is *still* a fact of the matter about whether others have qualia, just like there is a fact of the matter about whether other boxes contain beetles or not. Leaving aside the existence of objects in spaces that are in principle inaccessible – let alone in inaccessible ‘mind-spaces’ that are mere conjectures – is a clumsy proposition; what matters is that nothing that we *say* about others’ consciousness depends on whether they are conscious. We would say all of the same things even if everyone else were a philosophical zombie – a mindless creature who only externally acts in ways that feel minded. Thus, as long as we agree that calling other people conscious is correct, it can be seen that assenting to ‘X is conscious’ has zero relation to its having qualia, whether or not the notion of it having qualia can be seen as meaningful.

[8] This is where the more neuroscientifically-driven approaches to machine consciousness, such as the aforementioned Butlin et al.’s ‘Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.’[4] or Aru et al.’s ‘The feasibility of artificial consciousness through the lens of neuroscience.’[15] go wrong. The paradigmatic case of consciousness attribution – that of other humans – depends on *nothing like* ascertaining that there are specific types of processes in their brains, but on the usefulness of modelling them as conscious. To impose a different criterion on AI systems is highly arbitrary.

[9] The same dynamics can be seen in Wittgenstein's Moses example, analysed in *Philosophical Investigations*[5]. We know, more or less, what we mean by 'Moses': a person who lived in such and such a time, who was fished out of the river by the Pharaoh's daughter, and who led the Israelites out of the desert, etc. However, there is no definitive answer to *which* of those propositions should turn out to be false for us to concede that 'Moses did not exist'. We usually do not think about that since a substantial number of propositions about Moses are not normally doubted. But, if it were, reasonable speakers could disagree about whether 'Moses did not exist' was true, and no non-arbitrary resolution of the question would be available.

[10] In fact, one could argue that there is no 'private arena' component to, e.g., creativity. I think it is rather in a gray area: certainly, we do not intuitively see it as crucially dependent on internality in the way consciousness is, but there is still a sense in which it can seem like a *particular type* of internal process. Paradoxically though, rather than making it a clear(er)-cut case of a misguided question, as one of the pillars of what should make it sensical is (at least in part) removed, it leads some commentators to suggest that the question is *more* answerable, since we are in possession of better tests for creativity that rely on externally measurable characteristics (e.g., the 'originality and effectiveness' model). However, I think the fact that we have such tests is a result of creativity being something we found interesting to measure in humans – as it is not possessed by all humans to the same degree – rather than the notion being somehow inherently less subject to vagueness (the same is true of intelligence). To do that, it was necessary to employ some definition that approximately captures how we use the word. There is nothing in principle preventing us from developing a test for consciousness that would rely on external behaviors – it is just that there was no use for such a test. The issue of vagueness is the same, however: once there is a disagreement about whether a particular AI system is, e.g., creative, picking a particular way of testing for it will presume the answer (as Mark Runco points out in his review of an early draft of this paper, 'suggestions that AI can be creative have rejuvenated efforts to define creativity in a meaningful way'. It is difficult to believe that such a 'meaningful way' was not heavily informed by the authors' pre-existing conviction about whether AI should be deemed creative). Therefore, while the asymmetry between having operational definitions of creativity, but not consciousness (with tautological-sounding criteria like 'seeming conscious' (Chalmers) being suggested) is real, this should not be taken to imply that facts about AI creativity are more decidable. Such definitions are abstractions from ordinary language use made for practical purposes, and their role is approximately descriptive, rather than legislative. To then use them as being somehow *more* authoritative in settling the question of AI creativity than ordinary language is to put the cart before the horse. It is to make the same mistake that a computational functionalist makes when he thinks he has a privileged conceptual grasp of 'consciousness' that allows him to settle its application to AI (although such definitions can admittedly be instrumental in deciding if we should *think* of AI as creative, as it is easier to see if it satisfies, more or less, the criteria we are content to use for humans).

[11] To some extent, the direction we are likely to go in is elucidated by research like that of Colombatto & Fleming[16].

[12] Since the theory of mind is so essential to modelling our social environments, the bias to anthropomorphize is very strong (akin to pareidolia). The strength

doi.org/10.32388/13B814.3

and affective power of the projection are evident if we consider how we involuntarily apply it even to entities we do not think of as conscious, like ventriloquist dummies. As Graziano writes, ‘The model is automatic, meaning that you cannot choose to block it from occurring.... With a good ventriloquist ... [the] puppet seems to come alive and seems to be aware of its world.’[17]. It is hard to believe that a system that would persistently give one this impression could be classified as anything but conscious.

[13] ‘If the computer can fly airplanes, drive cars, and win at chess, who cares if it is totally nonconscious? But if we are worried about a maliciously motivated superintelligence destroying us, then it is important that the malicious motivation should be real. Without consciousness, there is no possibility of its being real.’[18].

[14] More generally, this has to do with projecting onto entities we model as conscious the kind of features we are familiar with in the case of the ‘default’ conscious entities – first and foremost, other humans. We pre-reflectively presume that the ‘new’ conscious entity is likely to have the kinds of reactions to various untested scenarios that the familiar conscious entities do. On one hand, such anthropomorphisation is distorting since there are many respects in which many of the systems we have today deviate from human-like reasoning. On the other hand, since many of them are explicitly designed to mimic human behaviour, there is clear utility in adopting such a model. Engaging with the cost-benefit analysis of such projections onto the different kinds of systems we have today is beyond the scope of this paper.

References

1. a, b*Horwich P (2011). Wittgenstein's metaphilosophy. Oxford: Oxford University Press.*
2. ^*Turing AM (1950). "Computing machinery and intelligence." Mind.* ***59****(236):433–460. doi:10.1093/mind/lix.236.433.*
3. ^*Chalmers DJ (2023). "Could a large language model be conscious?" Boston Review.*
4. a, b*Butlin P, Long R, Elmoznino E, Bengio Y, Birch JC, Constant A, Deane G, Fleming SM, Frith CD, Ji X, Kanai R, Klein C, Lindsay GW, Michel M, Mudrik L, Peters MA, Schwitzgebel E, Simon J, VanRullen R (2023). "Consciousness in artificial intelligence: Insights from the science of consciousness." ArXiv. doi:10.48550/arXiv.2308.08708.*
5. a, b, c*Wittgenstein L (1953). Philosophical investigations. G. E. M. Anscombe, ed. Oxford: Wiley-Blackwell.*
6. ^*Plato (1984). The dialogues of Plato. New Haven: Yale University Press.*
7. ^*Jones CR, Bergen BK (2025). "Large language models pass the Turing test." arXiv. doi:10.48550/arXiv.2503.23674.*
8. ^*Pollock J (1995). Cognitive carpentry: A blueprint for how to build a person. MIT Press.*
9. ^*y Arcas BA (2022). "Do large language models understand us?" Daedalus.* ***151****(2):183–197. doi:10.1162/daed_a_01909.*
10. ^*Franceschelli G, Musolesi M (2024). "On the creativity of large language models." AI & Society. doi:10.1007/s00146-024-02127-3.*
11. ^*Swanepoel D (2021). "Does artificial intelligence have agency?" In: Clowes RW, Gärtner K, Hipólito I, Eds. The mind-technology problem. (Studies in Brain and M*

ind; vol 18). Springer; p. 77–94. doi:10.1007/978-3-030-72644-7_4.

12. ^*Dung L (2023). "How to deal with risks of AI suffering." Inquiry.* ***1****(29):1–29. doi:10.1080/0020174X.2023.2238287.*
13. ^*Hutchinson P (2010). "Thinking and understanding." In: Jolley KD, ed. Wittgenstein: Key Concepts. Key Concepts. Acumen Publishing; p. 92–108.*
14. ^*Wittgenstein L (1980). Remarks on the Philosophy of Psychology. Vol II. Oxford: Blackwell.*
15. ^*Aru J, Larkum ME, Shine JM (2023). "The feasibility of artificial consciousness through the lens of neuroscience." Trends Neurosci.* ***46****(12):1008–1017. doi:10.1016/j.tins.2023.09.009.*
16. ^*Colombatto C, Fleming SM (2024). "Folk psychological attributions of consciousness to large language models." Neuroscience of Consciousness.* ***2024****(1):niae013. doi:10.1093/nc/niae013.*
17. ^*Graziano M (2013). Consciousness and the social brain. Oxford University Press.*
18. ^*Searle J (2014). "What your computer can't know." The New York Review of Books.*

Declarations

Funding: No specific funding was received for this work.
Potential competing interests: No potential competing interests to declare.