Qeios

Peer Review

Review of: "The Hard Problems of AI"

Sabine Koeszegi[1]

1. Technische Universität Wien, Austria

The hard problem of AI

This review is a joint achievement of Martin Pool and Sabine Koeszegi.

The paper argues that attributions of third-person consciousness are independent of any fixed internal processes within these persons (or systems, in the case of AI). Consequently, the only meaningful question can be a pragmatic one: "What is the best way to think about such systems? Olegas Algirdas Tiunina's central argument, from a Wittgensteinian - linguistic - perspective, is that we do not have a concrete and commonly understood definition of consciousness and qualia; since they have no objectively measurable properties, they can only be a matter of words - of negotiating meaning in a "word-game". He further suggests that the discussion of whether AI systems have qualia and consciousness is a red herring that distracts from the very "real challenges of AI". Consequently, he claims that the "hard problem" of consciousness cannot be solved empirically or even conceptually and attempts to shift the question from the ontological to the linguistic.

This seems an unwarranted conclusion - consciousness clearly exists, even if we don't have a common definition for it. We agree, however, that the further argument that consciousness is a "mixed bag" of things and a continuum rather than a binary "yes-no" property is entirely valid, though not argued here.

We argue that the question of consciousness in general, and in AI systems in particular, is central because, first, if we want to understand the universe, we need to understand qualia. This is an end in itself; secondly, the assumption of qualia in third parties has a strong influence on our actions:

Referring to Wittgenstein's thought experiment with the beetle in the box, the question of whether other systems (humans, AI systems) actually have a "beetle in the box", i.e. have qualia and consciousness, may be meaningless in terms of predicting and explaining the system's behaviour - as long as our model of the system is conclusive - but it makes a fundamental difference to how we interact with the system: for example, we will find it much harder to harm something if we assume that it feels pain than if we do not.

doi.org/10.32388/CGEGSF

Olegas Algirdas Tiuninas "does not deny that thinking of AI systems as conscious or intelligent may have significant implications for how we will interact with them in practice" and acknowledges that the question is essential. Indeed, this is the central point: we are affected differently by whether we assume that other systems are simulations of consciousness or are conscious.

Nor does the example of the submarine and whether it swims or not support the idea that the ontological question is irrelevant. Not only does the submarine not have intentionality, as the author rightly points out, and therefore does not swim, but it also does not have the bodily movements associated with swimming. An octopus, for example, has intentionality but does not swim; it propels itself. If a submarine had intentionality, it would not swim but propel itself. So, a submarine is not on the edge of the "swimmer" set since it clearly does not swim.

It is pointed out, however, that there are indeed things on the edge of a set where the definition is unclear. However, it is not at all clear how this makes the question of what the submarine does "silly". If defining what a submarine does is important, we need to find new terms for it or work on the definition.

But the more fundamental problem with the analogy is that it is by no means irrelevant to think about what the submarine does, just because we don't have the language to describe what it does.

The author's statement at the end of this part, that negotiating a new linguistic convention about this would neither tell us anything philosophically interesting, rather torpedoes the validity of the linguistic approach of which he is a proponent. Instead, it nicely emphasizes the importance of thinking about what things really are. Because things clearly do have an existence which goes beyond name-calling.

One could also take issue with the relevance of the verb in this case. Is not "swim" what the submarine *does* much like "talk" is what an LLM *does*? Is it not mixing up unclarity about the definition of observable traits – traits which the author says have nothing to do with consciousness – with those unobservable third-party attributes (consciousness) about which he says there is so much unclarity? Indeed, is there not a science of submarines, which will look into, which will look inside, submarines to see how they work? Is there not indeed a science that will look into the intentionality of submarines, i.e., into the decision-making processes of the crew together and into the head of the captain in particular? And what would this have to do with "language-games"?

Despite the difficulty with the analogy, the implication that sets have "fuzzy edges" is entirely true. From this, he goes on to posit that consciousness is indeed a more loose definition than this – it is on the one hand a continuum, and on the other hand, there could be several differing and overlapping definitions or

attributes of consciousness. This is a fascinating proposition that merits further exploration. Unfortunately, the author is being somewhat untrue to himself in this, since this seems in fact to be the basis for an ontological exploration – and as such, one which is here believed to be fruitless.

Following on from this, we believe it is valid to pursue all of these investigations – ontological, pragmatic, societal, psychological, and linguistic – and that one line of investigation in no way precludes the others. We would encourage the author to pursue his very valid arguments about the weaknesses and difficulties of other approaches, while acknowledging that many of his arguments can be very helpful in pursuing other avenues.

Olegas Algirdas Tiunina claims that the discussion about whether AI systems have qualia and consciousness distracts from the "real challenges of AI". However, he never addresses the "real challenges of AI". It would be interesting to learn more about the real challenges of AI, following the pragmatic approach suggested by the author.

In conclusion, we think, among others, one real challenge is to understand whether AI systems might have a form of consciousness beyond our current ability to describe what it is.

Declarations

Potential competing interests: No potential competing interests to declare.