What Is It Like to Be a Robot?

Philosophical Issues of Computer Science,
Politecnico di Milano, July 2023

Matteo Monti

Abstract

The recent years' fascinating rise of AI, has rekindled the interest of researchers towards defining the relationship between Artificial Intelligence and consciousness. In this work, I have provided compelling evidence to support the following thesis: *a conscious AI system won't help us understand human consciousness*. To accomplish this, I have referred to the challenges put forth by the mind-and body problem and the hard problem of consciousness. Through the use of the "Conscious Android" thought experiment, I delved into the core of the argument and highlighted the difficulties intrinsic to detecting machine consciousness. Then, by following the natural development of the thought experiment, I have detailed how the subjective nature of the human consciousness, strictly correlates and is ultimately indivisible from our first person human point of view; outlining how this conclusion inevitably renders the use of a conscious AI system to unravel the mysteries of our consciousness a dead end. In my investigation of the subject, I've also pointed out what I consider to be some potentially fruitful research directions that may require further attention in the near future.

1 Introduction

In this work, I will argue that a conscious AI system won't help us understand human consciousness. I'll explore this thesis by analyzing the impact that a hypothetical conscious AI system would produce on human endeavors towards a deeper, more comprehensive, understanding of our own conscious mind. Would the existence of a conscious Artificial Intelligence be of any significance in regards to our species' understanding of itself? Would humankind find itself any closer to unveiling the truths of the mind if such AI systems were designed?

I won't delve into any of the issues regarding the technical details and mechanisms related to the question of *how* consciousness could manifest itself in an AI's artificial mind; I will focus, instead, on the effects of such occurrence, by first addressing the difficulties insisted in the non-trivial task of recognizing consciousness in an artificial system; and then by discussing how potential findings and conclusions obtained from such artificial consciousness are not so easily transferable to our human counterpart. Doing so, my aim is to highlight the fundamental gap between our conscious mind and the artificial one; demonstrating how the creation of the latter is not beneficial for the understanding of the former.

1.1 Definition of consciousness

To pursue the above purpose, I'll embrace the same definition of conscious organism famously adopted by Thomas Nagel, which states:

> ≪An organism has conscious mental states if and only if there is something that it is like to *be* that organism - something it is like *for* the organism≫ [Nagel, 1974, 436]

This particular definition captures the more subjective nature of being a conscious organism. It allows us to frame the problem at hand in terms that are more relatable to our personal experience of the world. *I know to be conscious since I can describe extensively what it would be like to be me.* As detailed by [Van Gulick, 2022, Section 2.1], Nagel's definition enables us to accept the presence of consciousness in external creatures, different from ourselves, by observing that "there is something that is like" to be that creature, even though we'll never be able to perfectly replicate and experience their exact state of consciousness from our physically restricted human point of view.

1.2 Motivations

Ultimately, why should we even care about any of this? Since the question was first posed, human consciousness has been one of the most fascinating, unconquered bastions of our quest for knowledge. We have never seemed able to grasp its fundamental mechanisms or way of manifesting itself:

> ≪There is nothing that we know more intimately than conscious experience, but there is nothing that is harder to explain.≫ [Chalmers, 1995, 1]

Coincidentally, Artificial Intelligence has integrated into our lives, affecting many areas of society and research. As we grow more accustomed to it, it's only natural to wonder whether a particular declination of AI might be able to assist in resolving the challenging problems we struggle with, like understanding our own consciousness. We'll see how this road leads to a dead end.

1.3 Overview on the logical structure

In order to make my reasoning clearer I will structure my argument as follows: we will first set the foundations of the topic by introducing the Mind-Body problem (Section 2.1), and the hard problem of consciousness (Section 2.2). Later, we'll delve into the core of the issue by outlining

the "Conscious android" thought experiment, which will help us state what I believe are the two main challenges that arise as a direct consequence of our approach (Section 3). We will then address two objections on the subject that may stem from my arguments (Sections 4.1 and 4.2). Finally, I will point out what might be some of the directions of interest for further research on the topic (Section 4.3).

2 Laying the groundwork

In order to tackle the issue at hand, it's important to first understand the context surrounding the question; in particular it's important to understand the difficulties raised by the *Mind-Body problem*.

2.1 The Mind-Body problem

In its most straightforward formulation the Mind-Body problem is a philosophical dilemma that attempts to investigate the connection between the mind and the body. As discussed by [Chalmers, 1997], it involves understanding how mental states, such as thoughts, emotions and subjective experiences, relate to the physical processes that take place in the brain and in the rest of our body. At the problem's core, lies the question of whether the mind and the body are two fundamentally distinct entities or if they are two aspects of the same underlying reality.

According to Chalmers, the mind-body problem can be divided into two parts: an easy one and a hard one. The easy part of the problem concerns the psychological aspects of the mind, which include all its functional properties like memory, its learning ability and, in general, every property of the mind that plays a causal role, such as the ability to react to environmental stimuli, to focus the attention or to deliberately control our behavior. Each one of these aspects can be potentially broken down as measurable physical states of the mind and explained in terms of neural mechanisms, thus becoming a problem concerning biology and cognitive science. While presenting enormous technical difficulties, this part of the problem offers a clearly-defined research path to find the answers to them.

The hardest part of the problem concerns the phenomenal aspects of the mind, the subjective "what is like" aspect of mental states [Nagel, 1974]. This part of the problem ultimately, culminates in the unresolved question: how could a physical system give rise to conscious experience?

2.2 The hard problem of consciousness

If we imagine to strip the mind of all aspects related to its psychological properties and everything that is objectively determined, what we're left with are its phenomenological properties. Our subjective experience of life, our sense of self, the emotions that color each one of our conscious experience, the sudden rush of pleasure that we feel when we get a joke. we are basically left with our personal window that overlooks the world. Why is it the case that every physical and material stimulation we receive from the external world, is accompanied by our own personal experience of it? The hard part of this dilemma has been reformulated by [Chalmers, 1995] as the hard problem of consciousness:

How do physical processes in the brain give rise to the subjective experience of the mind?

Various approaches have stood out over the years and have been widely recognized by the research community as the main attempts to shed light on the problem, with the two main ones being dualism and the more modern materialism, particularly its declination found in *reductionism*. The progress done so far in understanding the link between the psychological

mind and the phenomenal mind has almost entirely limited itself to the functional explanation of behavior, leaving the question of conscious experience almost untouched.

3 Imagining the conscious AI system

The hard problem of consciousness helps us highlight what we are lacking in terms of understanding of our own conscious mind, and the existing gap between mind and brain. By looking at the recent advancements in AI and at some of its most striking applications (i.e. in the field of natural language modeling), one might think that we are getting closer to developing an AI so powerful and complex that it will somehow manifest consciousness. Others have begun to advocate that consciousness might be a desirable implementation to AI in order to enhance its effectiveness [Esmaeilzadeh and Vaezi, 2021], [Hildt, 2023, Section 5]

Although reflections of a potential future of conscious AI systems may seem purely speculative or far stretched, one must acknowledge that researchers of the field are indeed trying to mimic aspects of human consciousness in their work [Lipson, 2019], so the assumption that a fully conscious artificial system could exist at some point in the future doesn't seem that detached from reality.

3.1 The conscious android experiment

Let us now consider the hypothetical. Assume we set out to shed light on our own mind. We are especially determined to solve the hard problem of consciousness. To that end, we decided to create a conscious android that we believe will aid us in understanding how human consciousness might emerge from the physical brain, using an inference approach and analogy. Imagine now that after years of constant endeavor, we were ultimately successful in developing such an artificially conscious android in our laboratory – we will call him Roy to honor his dignity as an allegedly aware entity. We are now faced with two non-trivial challenges that separate us from fulfilling our purpose.

1. The first challenge addresses the question of how can we know whether Roy is conscious.
2. The second challenge is more subtle and possibly more pressing: given that Roy is conscious, are the assumptions and conclusions we might draw about the functioning of his consciousness really applicable to humans?

3.2 The first challenge

As humans, we are intimately familiar with our own form of consciousness due to our first-person experience of it; it's clear from our perspective, that consciousness is something that characterizes our being. However, we only have access to Roy from a third person perspective, and this clearly represents an issue. Is it even possible to establish with an adequate level of certainty that he possesses a subjective experience? The challenge that we face taps into a very well known and studied topic in philosophical literature, it is in fact a particular illustration of the *problem of other minds* [Avramides, 2020, Section 1], applied to a synthetic entity. The problem highlights the difficulty of justifying the commonsensical and widespread notion that individuals other than oneself have minds and are capable of thinking or feeling in the same way one does.

Although being a theoretical non-trivial challenge, we do not find it troublesome to project the attribute of consciousness to other human beings, in fact, we do it all the time. We assume other humans, that behave and act like us, to possess the same degree of consciousness as we do, even though there is no absolute proof of that being the case other than their behavior and spoken word. Human societies are built upon the assumption of shared consciousness. We form

relationships, build families, and establish communities based on the understanding that others have subjective experiences and consciousness. We connect with others on an emotional level, form bonds, and develop trust because we perceive others as conscious beings like ourselves.

One might argue that a similar approach could be applicable in the case of Roy. Accordingly, we could simply state that if Roy was to show complex human-like behavior then, the assumption of it being a conscious entity would be considerable plausible and even likely. Known works in this direction tried to propose tests that aim at detecting the presence of consciousness in a machine. In particular, Aida Elamrani and Roman Yampolskiy conducted a detailed analytical overview on the matter [Elamrani and Yampolskly, 2019], by reviewing a multitude of existing tests proposed over the decade 2004-2014. Interestingly, they state that each one of the reviewed tests can be grouped in one of two grand categories: tests that infer the presence of consciousness by analyzing the complexity of the machine's *architecture*, and tests that deduce its presence by observing the machine *behavior*. They conclude their study by stating that each test category comes with its own strengths and weaknesses. In particular, architecture-based tests presuppose a detailed understanding of the artificial systems and their architecture while behavior-based tests seem to depend in large extent on the human ability to interpret the behavior exhibited by the machine.

From these considerations it is clear that the first challenge presents numerous difficulties in of itself, and the scientific communities are far from reaching consensus in its regards [Elamrani and Yampolskly, 2019, 24].

3.3 The second challenge

For the sake of our experiment, let's assume that relying on a behavior-base test is enough to determine that Roy is indeed a conscious android. We are fundamentally relying on Nagel's definition of consciousness, by acknowledging that Roy is experiencing "something that is like to be him". We observe him acting exactly like a human, questioning his existence in an inquisitive manner, and reacting to external stimuli coming from the environment by manifesting emotions consistent with a conscious experience of reality. He is even able to communicate, interact and form a bond with us, exactly like a human would do.

We then start analyzing Roy's mind, determined to locate the source of its conscious experience to deepen the understanding of our own. But as we study Roy's artificial brain fissures, we are bitten by an ever-tightening question: are our discoveries on this artificial mind's consciousness transferable to our human biological mind?

Looking back on the mind-body problem we've detailed in section 2, we realize that one of its two pillars has been compromised. We built an incredibly complex artificial mind in order to achieve a level of consciousness equivalent to that of humans, but in doing so, we inevitably created something completely different from ourselves on the physical level. Something that might very well possess human-like consciousness, while also being vastly dissimilar from us. How could a brain this fundamentally different from ours assist us in filling the gap highlighted by the mind-body problem? It simply could not.

Bridging the gap between our physical and phenomenological mind, intrinsically necessitates both variables of the equation to be human-like. Given our structural differences, the fact that Roy manifests a human-like consciousness does not imply that his subjective experience is in any way reflective of ours. We would be examining Roy's brain processes from an external point of view, in hope that by objectively describing the physical processes that give rise to his consciousness, we might extrapolate useful insights about our subjective point of view; unfortunately this would only provide us with an objective understanding of them, leaving us with no additional information concerning our subjectivity. We hit again what seems like a

phenomenological brick wall.

Our approach is in fact, fundamentally flawed: by investigating Roy's brain we tried to explain our subjective character of experience by detaching ourselves from our own first-person viewpoint, in favor of one of greater objectivity, external from ours. This approach towards the understanding of consciousness is the one advocated by *reductionism*, which states that by gaining a complete and objective characterization of the physical processes of the brain we could ultimately understand its phenomenological properties.

> ≪If the subjective character of experience is fully comprehensible only from one point of view, then any shift to greater objectivity - that is, less attachment to a specific viewpoint - does not take us nearer to the real nature of the phenomenon: it takes us farther away from it.≫ [Nagel, 1974, 445, lines 1-4]

Nagel's words underline the main issue encountered by applying the reductionist approach: by adopting an external point of view, we leave behind the subjective experience of consciousness, which is exactly what we're trying to justify! It would be analogous to attempting to explain the experience of the color red just by means of objectiveness; one might be successful in conveying an effective depiction of the color solely through verbal description, but the subjective experience of it would remain inaccessible from an external perspective.

It's evident that both challenges present some glaring difficulties that are hard to ignore and maybe, impossible to overcome.

4 Objections and future directions

So far, I've argued what I believe to be the reasons why the creation of a conscious AI system would not help us find the answers to our questions regarding human consciousness. Specifically, I've highlighted the two main problems that would arise by following this particular path. Now, I would like to address some of what might be the objections to my point.

4.1 The "physical equivalence" objection

One possible objection could stem from some skepticism with respect to my take on the second challenge covered in section 3.2, highlighting for example that, since we're willing to accept the hypothetical, one could build an AI system which, in addition to manifesting human-like consciousness, is also provided with an *exact* replica of the biological human brain, in order to achieve a physical equivalence with respect to our human counterpart. By accepting this possibility one could go on and argue that both elements of the mind-body problem are now present, and thus it would now be possible to draw significant conclusions about consciousness by conducting a comparative study.

To answer this objection I will point out that, yes, this would indeed get around the problem I've addressed in section 3.2 about the physical dissimilarity existing between the two minds, and yet, creating such an artificial brain, would defeat its aiding purposes. This scenario would be equivalent to studying the mind of a human being; it would be exactly like trying to motivate human consciousness by pointing out the complexity of our own mind. We would have run in a circle straight back to square one.

4.2 The "consciousness' general properties" objection

Another objection that I would like to address is the following: "By studying a conscious artificial system, we might still be able to draw some general properties about consciousness that, given their general connotation, might still be applicable to the human case".

While this argument might seem a compelling one, it misses a central point: how could we establish the *generality* of a consciousness' property that we've derived from only one specific instance of it? This would only be possible if we had some prior knowledge about our own consciousness and its functioning; only then would we be able to recognize specific recurring properties as general, but of course, we have nothing like that. To offer a suiting analogy, it would be like taking an alien to the movie theater and showing him the screening of a horror movie. Having no prior knowledge of human costumes or movie culture, he could deduce, for example, that "being frightening" or "making people in the theater scream", are general properties of every movie. I think that we can ultimately refute this particular objection by saying that: it would not be possible to confer the attribute of *generality* to a property of consciousness since the limited knowledge of our own, doesn't yield a proper term of comparison.

4.3 Future directions of research

What is clear in the context of machine consciousness, as highlighted in section 3, is a lack of consensus with respect to what could constitute a feasible way to determine whether or not a synthetic system is conscious or not. This is fundamentally due to a multitude of dispersed opinions with regards to what could be an appropriate and realistic definition for artificial consciousness [Hildt, 2023, 19, Section 10]. I strongly feel that this particular aspect is in urgent need of clarification in order to allow for progress in this field.

Machine consciousness should be pursued and treated as a separate field. Research on the matter should be driven by motivations other than gaining a deeper understanding of ourselves. Such motivations could focus on establishing if a conscious machine would indeed coincide with an increase in overall performance, or analyzing the ethical implications that a morally relevant form of machine consciousness would signify.

5 Conclusions

In this work, I've illustrated how a conscious AI system won't help us understand human consciousness. Throughout Section 2 we have set the groundings that helped us frame the problem and understand the issue at hand; then, though the use of a thought experiment, I've outlined the two fundamental arguments for my claim: in Section 3.2, we have addressed the difficulties that surround the task of recognizing consciousness in an artificial system, concluding that a definitive test able to gather the consensus of the research communities has yet to be designed, and that the question remains very much an open one. Subsequently, in Section 3.3, we have faced the challenge that represents the core of the dilemma, and seen how trying to understand the subjective character of one's consciousness by leaving his first-person point of view in favor of an objective, yet external, one it's ultimately inconclusive. In Sections 4.1 and 4.2, I've addressed what I believe to be two possible objections concerning my view, by also providing pertinent counter-objections. Finally in Section 4.3, I've pointed out some possible directions for future research on the matter of machine consciousness in light of the obtained results.

References

[Avramides, 2020] Avramides, A. (2020). Other Minds. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition.

[Chalmers, 1995] Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–19.

[Chalmers, 1997] Chalmers, D. J. (1997). *The Conscious Mind*. Oxford Paperbacks.

[Elamrani and Yampolskly, 2019] Elamrani, A. and Yampolskly, R. V. (2019). Reviewing Tests for Machine Consciousness. *Journal of Consciousness Studies*, 26.

[Esmaeilzadeh and Vaezi, 2021] Esmaeilzadeh, H. and Vaezi, R. (2021). Conscious AI. *CoRR*, abs/2105.07879.

[Hildt, 2023] Hildt, E. (2023). The prospects of artificial consciousness: Ethical dimensions and concerns. *AJOB Neuroscience*, 14(2):58–71. PMID: 36409517.

[Lipson, 2019] Lipson, H. (2019). Robots on the run. *Nature*, 568(7751):174–175.

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

[Van Gulick, 2022] Van Gulick, R. (2022). Consciousness. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.