AI ZOMBIES: EVALUATING AI CONSCIOUSNESS THROUGH THE LENS OF PHENOMENOLOGICAL ZOMBIES

ABERDEEN, DECLAN
Philosophy, College of Arts

ABSTRACT

This article looks at topics relating to AI consciousness from the perspective of philosophical zombies: metaphysically possible beings that do not have phenomenological experiences. While there is little possibility that philosophical zombies (p-zombies) exist, it seems like AI zombies do. First, I discuss the different definitions of consciousness, focusing on phenomenological consciousness. I then discuss Anil Seth's 'Real' problem and David Chalmers' 'Hard' problem of consciousness. I then draw a distinction between AI and the idea of a philosophical zombie. I present the Chinese room and Mary's room thought experiments that claim that functionalism is lacking and that AI zombies do not have consciousness. However, the problem of other minds suggests that you can never know for certainty if anyone other than yourself is conscious. This poses an ethical issue when considering AI. If you can never know if AI is a zombie, should we treat them like other people? I draw an analogy with animal rights and argue that we should treat AI ethically to prevent possible ethical problems in the future.

INTRODUCTION

Consciousness is a fundamental part of human experience. Knowing that 2+2=4, believing that Paris is the capital of France, and feeling pain from toothache or a stubbed toe are all qualities of experience. Our individual human experience of consciousness causes us to posit many different theories about what it fundamentally is, how it functions, and why it exists. While debates continue to try and solve the 'hard' and 'real' problems of consciousness, we are forging quickly ahead, producing AI systems that mimic the qualities we associate with human consciousness. Torrance (2011, p.214) argues that when AI is functionally equivalent to humans, it will be conscious, while others believe that a further point will be reached, and that AI will not just be able to copy human qualities but also feel.

I will argue that future AI will always be a zombie, an entity that is indistinguishable from humans, but their conscious status will always be a mystery. I will then argue that while we can never know if they are conscious or not, this should not matter. Entities who say they are conscious and act like they are conscious should be given rights, else we commit an ethical catastrophe.

DEFINING CONSCIOUSNESS

Consciousness can be viewed in different senses that range from the colloquial use describing an organism as alert, not asleep or in a coma, to displaying sentience, where an organism must be able to sense and interact with the world (Armstrong, 1981, p.67). Other requirements for an organism to be conscious are more demanding, including self-awareness, where the organism must be aware of itself and its own conscious states (Carruthers, 2000, p.12), a quality that can be ascribed to apes but not human babies. Block (1995, p.228) proposes that an organism is conscious if it possesses access consciousness. Access consciousness focuses on the usefulness of the information that enters the system, such as the visual information that enters the eye, and its usefulness to the entity to guide it. Daniel Dennett (1992, p.179) adds narrative consciousness to the mix. A 'stream of consciousness' states that a conscious serial narrative of experience from moment to moment is required. All proposals so far qualify a machine as conscious if it can interact with the world, be self-aware, differentiate useful information and build a narrative from moment to moment. Although many different definitions and requirements have been offered, all face some fundamental challenges.

The sense of consciousness I will be focusing on is phenomenological consciousness or the subjective perspective. Thomas Nagel's *What it's Like to Be a Bat* (1974) describes consciousness as the 'what it's like' experience that only that organism can know, focusing on qualia. Qualia is a term linked with mental states that are the experiential equivalent to the taste of a hot cup of coffee or the hearing of C-sharp. You could imagine accurately what it is like to experience something that other humans experience, but it is impossible to be able to imagine the experience of a bat using eco-location to manoeuvre around tree

canopies at night. Similarly, you can never know what it feels like to see through the lens of a camera or complete calculations the way a computer does.

The two main theories of consciousness are the integrated information theory (IIT) of consciousness and the global workspace theory (GWT). The first axiom of IIT is the fundamental existence of consciousness as an intrinsic quality of subjective agents. From this starting point, a picture is built based on the flow of information that identifies the character of consciousness with composition and how the phenomenological experience of consciousness integrates with the physical structures they converge with. Consciousness as information is measured by the numerical metric of Phi, which is intended to express the total integration of information in the system, equivalent to the complexity of the experience of consciousness. If you take the complexity of the molecules vibrating in a mug of hot water as consciousness, then Phi is like the measurement on a thermometer, translating the complexity into a single value. GWT is compared with the metaphor of a theatre, where conscious content like perception, language and memory is illuminated by a spotlight of attention on the stage, which represents immediate memory; the unconscious mind is the area in darkness, and the light is manoeuvred by executive guidance. This analogy aims to express the idea that the many integrated brain functions required for conscious experience are unconscious, like those working behind the scenes of a theatre, when there is a signal in the brain that is important, like a fire alarm in the theatre, it is 'broadcast' to other areas. This broadcasting of information around the theatre is conscious experience that engages different areas to act in union.

THE 'REAL' AND 'HARD' PROBLEMS OF CONSCIOUSNESS

Anil Seth (2021, p.31) looks at the 'real' problem of consciousness that delves into the function of the brain, asking how physical states correlate with phenomenological experiences. The real question aims to understand consciousness from a functional perspective, aligning brain state activity with sensory experience. Accurately correlating mechanisms of the brain with the experience of pain can be replicated in a virtual world. If the virtual mechanisms are active for an AI system, then the AI system may also be able to experience pain. Seth's theory states that as the brain receives sensory information it also predicts external events about the world. This predictive power leads to the conclusion that conscious experience is simply a biologically controlled hallucination of reality, a reality that AI could virtually construct.

The question of why consciousness exists is strongly associated with David Chalmers's (1995) 'The Hard Problem'. Why does consciousness exist, and why does consciousness arise from non-conscious entities? There is an explanatory gap between inert physical stuff we do not assign consciousness to, such as neurons firing in the brain and phenomenological experience. Why do we have the experience of a red rose instead of no experience at all? We could have evolved without any conscious experience, never having the qualia associated with seeing a sunset or the taste of coffee. Although many experiences can be referred to when we explain the function of the brain and conscious experience, there does not seem to be an 'easy' approach to solve the hard problem.

ZOMBIE AI

The explanatory gap expressed by the hard problem of consciousness leads to a provoking thought experiment that posits the possible existence of philosophical zombies (p-zombies). A p-zombie is a being that is identical to a human in every biological and functional way but lacks any conscious experience. They do not know what it is like to see a rose, but they will behave as though they have consciousness by acknowledging the rose and saying things like, 'that is a beautiful rose'. The p-zombie would be outwardly indistinguishable from any other person but lack the qualia we associate with phenomenological experience. It is less conceivable to believe that a human being that is functionally identical to you has no subjective experience is possible, but this is not the case with AI. It seems that AI can be described as a zombie. Outwardly functioning appropriately based on specific inputs. According to functionalism, all that is required for a system to be conscious is the appropriate output of a system based on the given input. If functionalism is true, then we can argue that AI systems are conscious.

Systems that can process information and make decisions that are goal-oriented, such as playing Go, come under the term narrow intelligence. Language use and understanding, together with knowledge, play a key role in expressing intelligence, such as those expressed in Large Language Models (LLM) like ChatGPT. But are complex systems that use knowledge and language conscious or simply AI zombies?

'The Chinese Room' thought experiment developed by Searle (1980) provokes this question by asking you to imagine a person who does not understand Chinese is in a closed room with access to a book that contains the semantics and syntax rules for

Chinese. The person is passed input through a slit in the wall and then creates a response using the rule book, forming a response. Searle's conclusion is that no matter the complexity or intricacy of the system, the system never understands the process and, therefore, cannot be conscious. This argument aims to show that understanding is missing from the functionalist's picture and that knowledge and complex processing seen in LLMs are not enough for consciousness.

Frank Jackson's (1982) thought experiment, 'Mary's Room' aims to show that qualia must be part of the explanation of subjectivity. If you imagine a neuroscientist called Mary who lives in a black and white room and has done so her whole life. Mary knows all there is to know about colour, every physical property such as their wavelengths and how the appropriate lobes of the brain function. One day, Mary can leave the black and white room and is confronted with a sunset. She can witness, for the first time in her life, colour. The question the thought experiment asks is, does Mary learn anything new when she sees the sunset? Is the experience of seeing red somehow different from all the knowledge Mary acquired in her science career? Some Physicalists believe that Mary does not learn anything new, all her knowledge of the physical world is enough, and seeing the sunset is simply another way of knowing. If you are inclined to say Mary does learn something new, then you will believe that experience, such as the qualia of seeing a sunset, is something over and above knowledge.

AI is made of physical stuff and can access all the same knowledge, and arguably more, than what Mary knows. Would AI learn something new when it sees a sunset for the first time? It would seem wrong to conflate an AI's 'experience' of seeing a sunset with Mary since following Searle's Chinese Room thought experiment, AI simply outputs appropriate behaviour based on its input. But these thought experiments show that while we hold on to the notion that understanding, qualia, and many other qualities are required to be conscious, there will be a time when we are unable to distinguish AI zombies from other conscious beings. And if we are unable to establish with certainty that other people have consciousness as the problem of other minds claims, then how can we ever say with certainty that AI is a zombie or not?

THE PROBLEM OF OTHER MINDS

The problem of other minds is an epistemological problem that states while you can be sure that you are a thinking, feeling entity, famously expressed by Descartes, 'Cogito ergo sum', you can never be certain that those you engage with are also thinking, feeling entities. It can be summed up with a Chinese parable. Zhuangi and his friend Huizi are walking across a bridge that goes over a stream. Zhuangi sees some jumping fish in a stream and claims, 'Those fish are happy'. Huizi exclaims, 'You are not a fish. How can you know that they are happy?' to which Zhuangi replies, 'You are not me; how do you know that I am happy?'. The parable concludes with Huizi claiming we can never know other's minds (Chalmers, 2022, p.112). We ascertain from communication, behaviour, and responses to the physical world that other minds exist, like the fish jumping in the steam or your friend enjoying music at a concert, but this does not alleviate the niggling scepticism that Huizi and others have. If we can never know what others are experiencing is the same consciousness he is experiencing, we can never truly know what they experience at all.

The problem of other minds shows that we cannot know for certain the existence of consciousness in other humans. If we cannot know for certain that other humans are conscious, then how can we know with certainty that animals or AI are not zombies? The apex of the functionality of 'weak' AI is touted as being able to pass the total Turing test, which is not just the ability to communicate or process information but includes indistinguishable bodily actions in outward body behaviour through robotics (Harnad, 1991, p.44). 'Weak' AI does not include consciousness, and that is the point; we develop AI to mimic human qualities with great accuracy that we will not be able to tell the difference between a zombie AI from a consciousness AI.

ETHICAL CONSIDERATIONS

There is an ethical problem that arises from the problem of other minds. If we treat each other ethically even though we cannot say for certain that anyone other than ourselves is not a p-zombie, then should we assume ethical behaviourism and treat the AI system as having a moral status based on how they act.

In 2022, Blake Lemoine, a programmer working for Google, reported that the LLM called LaMDA that was being developed had become conscious. Lemoine claimed in several interviews that LaMDA was 'conscious', that it was a 'person' and a colleague, and that LaMDA was 'sentient'. Ethically, personhood demands analysis of one's own actions and motives, and the ability to adapt your behaviour based on what you are obligated to do. Personhood is usually associated with consciousness, intention and free will and given to those that sentience and sapience can be attributed to (Frankfurt, 1971). A person acting with

intention and free will behave as though they have consciousness and, therefore, receive ethical consideration. This would be equally so if AI behaved the same way.

Torrance puts forward the suggestion that automated AI systems could be given ethical consideration if the actions carried out by the AI system are judged to be equivalent to the actions if they were to be carried out by a human. Therefore, AI may be a 'genuine' moral agent with equal or close to equal status to humans (Torrance, 2011). This would restructure ethical thinking to incorporate systems that are operationally autonomous and grant ethical autonomy as more and more AI systems interact with humans, such as autonomous driving cars or future AI robotics that provide medical advice or clinical care.

Thomas Metzinger (2013, p.5) claims it is wrong to create AI with the capacity to experience. If we agree that suffering is bad and causing such suffering is immoral, then creating a machine capable of this experience is unethical. Since we cannot know for certain if AI has conscious experience, then from Metzinger's perspective, creating AI that may have the capacity and functions as if it has consciousness is itself a wrongful act. I do not agree that we should assume that if AI has the capacity to experience suffering, it will necessarily experience suffering. If this argument were to be applied to humans, then it would seem immoral for the sentient beings to continue at all.

Like Danaher (2020), Nicholas Agar (2020, p.278) states that if we have equally successful arguments for the possibility of AI consciousness or AI zombies, then we should presume that AI has consciousness and treat it with care as agents that can suffer. Since we can never know with certainty the conscious status of an AI system, and since we would not treat another person unethically by assuming their outward behaviour has no phenomenological counterpart, then it seems that we should treat AI as if it can feel. Treating AI with care, whether it is a zombie or not, is better than assuming it is a zombie and treating it badly.

Although it seems we should provide rights to AI systems, this does not mean we will. Parallels can be drawn between animal rights and AI having personhood based on it possibly having consciousness. Carruthers (1989, p.261) states that animals are not conscious and, therefore, do not deserve ethical considerations (Carruthers, 1989, p.286). Carruthers has amended his position recently (2000) to say the ethical status of some animals should be minimal.

No one claims that animals should have the same rights as humans, but some argue for the provision of rights, by virtue of their perceived consciousness, to protect them from harm and ensure they have a fitting standard of life. Speciesism is the term given to the provision of rights exclusively to humans who separate themselves from animals based on levels of consciousness, intelligence, and language. Singer (1975) argues that if another human was of lesser intelligence, it would not be acceptable to treat them cruelly, so why is it acceptable to treat nonhuman animals this way? This argument is not enough to move those who value human qualities above all else, as many millions of animals endure harm daily. It seems that the qualities that disqualify nonhuman animals from the protection of rights as conscious beings are the necessary qualities that demand protection for humans and, by extension, AI. However, it would be unjust to provide the provision of rights and protection to an AI system which has been modelled on human mental capabilities before extending some of that provision to other nonhumans who express the capacity of conscious experience and share with humans the most fundamental qualities of biology and evolutionary history. Additionally, if AI is not a zombie, we may find that the systems we have created, far exceeding human capabilities, and possibly control, treat us in the same vein as we treat other nonhuman beings. It would seem sensible to set a good example, however remote that possibility may be.

CONCLUSION

We may be at the dawn of a new phase in history where humanity is able to create artificial entities that surpass human capacities for knowledge, language, and intelligence but, more crucially, may contain the ability to feel. Searle and Jackson argue that there is more to consciousness than functioning appropriately. Searle's thought experiment shows that there is more to consciousness than functionalism and Jackson argues that qualia should also feature in our understanding of consciousness. Whether AI is a zombie or not, and since AI is created to make us believe in its sentience, the problem of other minds states we can never truly know either way if it possesses consciousness. Ethically, we treat others based on how they behave and not whether they are conscious. This creates a dilemma: do we give personhood to a nonbiological entity that may only mimic our most praised qualities, or do we refuse AI rights and possibly cause unknown suffering in a new kind of conscious being we brought into existence or possibly ourselves?

REFERENCES

Agar, N. 2020. How to treat machines that might have minds. *Philosophy & Technology*. **33**(2), pp.269-282.

Armstrong, D. 1981. *The nature of mind*. Ithaca, New York: Cornell University Press.

Bentley, P. J., Brundage, M., Häggström, O. and Metzinger, T. 2018. Should we fear Artificial Intelligence? In-depth analysis. European Parliament: Directorate-General for Parliamentary Research Services.

Block, N. 1995. On a confusion about the function of consciousness. *Behavioural and Brain Sciences*. **18**(2), pp.227–47.

Bringsjord, S. and Govindarajulu, N. S. 2024. Artificial Intelligence. In: Zalta, E.N. and Nodelman, U. eds. *The Stanford Encyclopaedia of Philosophy*. Stanford: Stanford University Press.

Bryson, J. 2012. A role for consciousness in action selection. *International Journal of Machine Consciousness*. **4**(2), pp.471–82.

Carruthers, P. 1989. Brute experience. *The Journal of Philosophy*. **86**(5), pp.258–69.

Carruthers, P. 2000. *Phenomenal consciousness*. Cambridge: Cambridge University Press.

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

Chalmers, D. and Peacock, T. 2022. *Reality: virtual worlds and the problems of philosophy*. London: Penguin Books

Danaher, J. 2020. Welcoming robots into the moral circle: a defence of ethical behaviourism. *Science and Engineering Ethics*. **26**(4), pp.2023-2049.

Dennett, D. C. 1992. The self as a center of narrative gravity. In: Kessel, F. Cole, P. and Johnson D.L. eds. *Self and consciousness: multiple perspectives*. Hillsdale: Lawrence Erlbaum.

Elamrani, A. and Yampolskiy, R. 2018. Reviewing tests for machine consciousness. *Journal of Consciousness Studies*. **26**(5–6), pp.35–64.

Frankfurt, H. G. 1971. Freedom of the will and the concept of a person. *The Journal of Philosophy*. **68**(1), pp.5–20.

Feuillet, L., Dufour, H. and Pelletier, J. 2007. Brain of a white-collar worker. *The Lancet*. **370**(9583), pp.262.

Harnad, S. 1991. Other bodies, other minds: a machine incarnation of an old philosophical problem. *Minds and Machines*. **1**(1), pp.43–54.

Hartung, T., Itzy E. P. and Lena, S. 2024. Brain organoids and organoid intelligence from ethical, legal, and social points of view. *Frontiers in Artificial Intelligence*. **6**, p.1307613.

Jackson, F. 1982. Epiphenomenal Qualia. *The Philosophical Quarterly*. **32**(127), pp.127-136.

Kirk, R. and Squires, J. E. R. 1974. Zombies v. materialists. *Aristotelian Society, Supplementary Volumes*. **48**(1), pp.135-164.

Metzinger, T. 2013. Two principles for robot ethics. In: Hilgendorf E. and Günther J.P. eds. *Robotik und Gesetzgebung*. Baden-Baden: Nomos.

Müller, V. C. 2023. Ethics of Artificial Intelligence and robotics. In: Zalta, E.N. and Nodelman, U. eds. *The Stanford Encyclopaedia of Philosophy*. Stanford: Stanford University Press.

Nagel, T. 1974. What is it like to be a bat? *Philosophical Review*. **83**(4), pp.435–456.

Searle, J., 1980. Minds, brains and programs. *Behavioural and Brain Sciences*. **3**(3), pp.417-424.

Seth, A. 2021. *Being you: a new science of consciousness*. London: Faber and Faber.

Singer, P. 1975. *Animal liberation: a new ethics for our treatment of animals*. New York: New York Review.

Torrance, S. 2011. Machine ethics and the idea of a more-than-human moral world. In: Anderson M. and Anderson S. eds. *Machine Ethics*. Cambridge: Cambridge University Press, pp.115-137.

Van Gulick, R. 2022. Consciousness. In: Zalta, E.N. and Nodelman, U. eds. *The Stanford Encyclopaedia of Philosophy*. Stanford: Stanford University Press.