Does it (morally) matter whether the AI machine is conscious?

Kamil Cekiera, University of Wroclaw (Poland)

International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: consciousness, conscious machines, AI ethics,conceptual engineering,the concept of human

Extended abstract

An unprecedentedly rapid pace of the development of artificial intelligence (AI) technologies, as arguably is what we are experiencing now, is on par with a growing theoretical interest in issues related to AI's functions, status, influence on society and changes it makes in our everyday life. Philosophers are especially keen on analyzing how the advances in AI technologies and design shape our thinking and impact the way we take things to be in many different areas.

One of such areas is ethics, the field traditionally concerned with questions about human (and non-human animals) (inter)actions. In the face of the ever-growing significance and advancement of AI technologies, philosophers also needed to take them into account. Broadly speaking, moral philosophers are interested in one of the two types of ethical challenges to AI. The first addresses all kinds of worries one may have as to how intelligent machines can harm or otherwise negatively influence people, society, the environment etc., and what we should do to avoid it or diminish such unwelcome effects. The second concerns the machines themselves – their moral status, potential rights or the way others are responsible for them (for an overview see e.g., Bostrom, Yudkowsky 2014; Müller 2021).

Since the outset of philosophers' interest in AI, consciousness was one of the hugest topics catching philosophers' attention. Even though attention was paid to it also from the ethical perspective, it has usually concerned one of the topics from the first cluster mentioned above. Thus, for instance, Susan Schneider analyzing "possible ethical implications of synthetic consciousness" in the context of AI, emphasizes the worries that "the inadvertent or intentional development of conscious machines could pose existential or catastrophic risks to humans – risks ranging from volatile superintelligences that supplant humans to a human merger with AI that diminishes or ends human consciousness" (Schneider 2019, p. 67). In his recent book, David Chalmers (2022) proposes, however, a different approach. According to him, consciousness is what bestows moral status for a given subject. If an AI machine is conscious (and Chalmers argues forcefully that it could be), then it has a moral status – just as any other conscious creature.

The possibility of granting moral status for AI machines or robots used to be analyzed in philosophy in terms of e.g. intentionality (see Anderson 2013), being able to feel pain or act rationally (for an overview, see Liao 2020). However, the view defended by Chalmers gives a completely new perspective on that issue. If the AI machines could be conscious and that would grant them moral status comparable to – or in fact identical to – that of human beings, that changes the way we should think not just about the moral status of artificial beings but about the concept of humanity itself. For instance, in his recent paper Kamil Mamak argues that "robots will never have full human rights, even if it is decided they should have a moral standing similar to that of humans" (Mamak 2022: 1). However, as I am going to show, if Chalmers

is right, then the moral standing of robots is not just similar to that of humans, but instead robots should be considered humans when it comes to ethical stance. Thus, in such a case, we would need to engineer the concept of humans.

The issue of engineering concepts of philosophical importance has become popular in recent years thanks to the development of the so-called conceptual engineering movement (for an overview, see e.g., Cappelen 2018; Burgess, Cappelen, Plunkett 2020). Although conceptual engineering deals with a huge number of issues, the basic idea is simple: in cases when we encounter a given notion or concept that seems to be problematic in one way or another – e.g. epistemically defective, functioning differently than we intend it to function, referring too narrowly or too broadly – we can engineer or ameliorate the concept in question as to avoid such undesirable consequences. The very idea of conceptual engineering as understood dates back to Carnap's notion of explication, according to which we can either get rid of the problematic concept and propose a new one instead or we can propose changing its extension (Carnap 1950).

In my talk, I will first show how Chalmers' argumentation inevitably leads to the conclusion that the concept of human needs to be changed and what consequences it has for a number of social institutions, such as human rights. Secondly, I will try to elucidate what functions we want the concept of humans to fulfill and how one can ameliorate that concept in accordance with those functions. Thirdly, I am going to argue that Chalmers' argumentation is flawed as it is not that clear that conscious machines are possible. That all leads to the conclusion that we must pay much more attention to the role of consciousness in the moral domain and what it takes to be a human being.

References

Anderson, D.L. (2013) Machine intentionality, the moral status of machines, and the composition problem, [in:] Philosophy and Theory of Artificial Intelligence, V.C. Müller (ed.), Springer.

Bostrom, N., Yudkowsky, E. (2014) The ethics of artificial intelligence, [in:] The Cambridge Handbook of Artificial Intelligence, K. Frankish, W.M. Ramsey (eds.), Cambridge University Press.

Burgess,, A., Cappelen, H., Plunkett, D. (eds.) (2020) Conceptual engineering and conceptual ethics, Oxford University Press.

Cappelen, H. (2019) Fixing language: An essay on conceptual engineering, Oxford University Press.

Carnap, R. (1950) Logical foundations of probability, University of Chicago Press.

Chalmers, D.J, (2022) Reality+: Virtual worlds and the problems of philosophy, W.W. Norton & Company.

Liao, S.M. (2020) The moral status and rights of artificial intelligence, [in:] Ethics of Artificial Intelligence, S.M. Liao (ed.), Oxford University Press.

Mamak, K. (2022) Humans, Neanderthal and rights, “Ethics and Information Technology” 24, 33, https://doi.org/10.1007/s10676-022-09644-z.

Müller, V.C. (2021) Ethics of Artificial Intelligence and Robotics, [in:] Stanford Encyclopedia of Philosophy, E.N. Zalta (ed.), https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/.

Schneider, S. (2019) Artificial you: AI and the future of your mind, Princeton University Press.