Artificial Systems and Moral Agency: Not a Question of Consciousness

There are several explanations on offer for why artificial systems are not moral agents (yet). Variously, all of consciousness, free will, autonomy, emotional engagement, commitment to moral issues, and rationality have had supporters. Recently, Véliz (2021) has argued that consciousness is essential for moral agency because one needs emotions to be able to value things. This finds correspondence with many other accounts (e.g., Coeckelbergh, 2010; Himma, 2009; Johansson, 2010; Moor, 2006; Purves et al., 2015) which claim that consciousness is what artificial systems require to be moral agents.

However, I argue that these accounts are problematic. There is, first, Floridi and Sanders' (2004) objection that intentional states like consciousness are not measurable. They say that one requires an unrealistic 'God's eye perspective' to gain access to another's conscious states, and that access is needed to determine their moral agency (Floridi & Sanders, 2004, p. 16). This objection is often understood as arguing that consciousness *may be* a condition of moral agency but is impractical to measure. However, the difficulty of measuring consciousness instead points towards the fact that consciousness is merely associated with moral agency and not essential to it.

This becomes clearer with a comparison to Véliz's (2021) argument, Véliz claims that machines (she refers to algorithms) cannot be moral agents because they cannot be conscious, which is necessary for emotional experience, which itself is necessary for one to be able to be autonomous and to value things. Accordingly, moral agency only *requires* sufficient autonomy and the ability to value, but Véliz claims that consciousness is necessary for those conditions. However, Consciousness is arguably not necessary for emotional experience, and is unlikely to be necessary for autonomy and evaluative abilities. This follows from the Floridi and Sanders' problems of measuring others' conscious states: we do not habitually measure other's degrees of consciousness, but do habitually measure their level of moral agency, so moral agency is likely to be something that we can measure (i.e., a property independent from consciousness). This is demonstrated further through intuitive functionalist cases in which non-conscious systems can be emotional, value things, and even be autonomous.

My diagnosis for the confusion of those who emphasise artificial consciousness as a requirement for moral agency (or other abilities) is that it is easy to mistake consciousness for the true conditions of moral agency due to them being strongly associated in our minds. The source of that association is anthropocentric bias: we think that consciousness is a necessary condition for moral agency because humans always have both properties simultaneously. As long as we are moderately functionalist about machine ethics (as we should be) then we should leave consciousness behind in discussions of machine moral agency.

References

Coeckelbergh, M. (2010). Moral appearances: Emotions, robots, and human morality. *Ethics and Information Technology*, *12*(3), 235–241. https://doi.org/10.1007/s10676-010-9221-y

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/b:mind.0000035461.63578.9d

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, *11*(1), 19–29. https://doi.org/10.1007/s10676-008-9167-5

Johansson, L. (2010). The Functional Morality of Robots. *International Journal of Technoethics (IJT)*, *1*(4), 65–73. https://doi.org/10.4018/jte.2010100105

Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, *21*(4), 18–21. https://doi.org/10.1109/MIS.2006.80

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous Machines, Moral Judgment, and Acting for the Right Reasons. *Ethical Theory and Moral Practice*, *18*(4), 851–872. https://doi.org/10.1007/s10677-015-9563-y

Véliz, C. (2021). Moral zombies: Why algorithms are not moral agents. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01189-x