Submission for the workshop on "*AI Awareness and Ethics*" at AISB-2026

Extended abstract (Word count: 980 without references)

Can we imagine an ethical status of AI systems without consciousness?

In view of the great diversity of theories of consciousness and the lack of established methods for justifying its attribution, the topic of AI consciousness is marked by enormous indeterminacy. Still, answers to questions of whether AI systems could, should, or cannot have an ethical status in human-machine interactions are urgent.

I suggest approaching those questions in a somewhat unusual way: Instead of solving the two hard problems—which theory of consciousness is the right one and how we can justify attributions—I pose the provocative question of whether consciousness is a necessary condition of all attributions of an ethical status. Are unconscious entities conceivable that have to meet certain obligations, and for which social norms deliver guidance on how humans interact with them?

For humans and other living beings, it seems out of the question that, besides agency and intelligence, consciousness is an essential feature. Moral patiency is only attributed to beings that can suffer, and suffering presupposes phenomenological consciousness. Also, concerning moral agency, phenomenological consciousness is taken as a necessary condition. For example, Carissa Véliz argues that in order to understand that there are moral reasons that prevent one from inflicting pain on others, one must know what it means to feel pain: *"But entities that do not feel cannot value, and beings that do not value cannot act for moral reason"* (Véliz, 2021, 487). Consequently, it seems there is only one way to argue for an ethical status of AI systems: to find arguments that they may possess phenomenological consciousness and thereby argue that we should expand our moral circle to include artificial life.

Pointing to gradualist approaches that aim to extent the set of possible candidates for moral agency to non-human animals, like the one of Paul Shapiro who argues for a wider spectrum of moral agents: *"As such, it will be helpful to view our issue not in terms of a black and white dichotomy between moral agents and non-moral agents, but rather as a broad continuum with acting on principle at one end, the bare ability to act virtuously at the other, and reciprocally altruistic animals who can't act on moral principles somewhere in the middle"* (Shapiro, 2006, 358), I ask whether a more radical gradualist approach could argue for a spectrum that might also overcome the sharp boundary between conscious (living) and non-conscious (nonliving) entities.

Before tackling the hard question of whether artificial systems may deserve something like a quasi-ethical status, to which I do not yet have an elaborated solution, I turn to the question of a potential social status of AI systems. Drawing on Shapiro's idea of a wider spectrum, I suggest imagining a multidimensional spectrum of social interactions to avoid the black-and-white dichotomy between mere tool use and social interactions. At one edge of this spectrum, I locate single-sided sociality; cases in which the sociality of the human is tossed into the void because the other entity (the artificial system) has no capacity for social uptake and is in no respect a social partner. There, the artificial system is used as a tool by the human, as social interactions require at least some reciprocity. On the other edge of the spectrum, I locate full-blown, cooperative social interactions as we know them from human-human interactions. Inbetween, we can start to investigate various instances of social interactions that I label as asymmetric quasi-social interactions. Those interactions are asymmetric, as they do not impose the same conditions on all participants; they involve unequal partners. I characterize those interactions as quasi-social to avoid that one would mistakenly equating them with full-fledged social interactions. Established gradualist approaches that characterize the sociality of children and nonhuman animals have already shown that it is not necessary to impose the same demanding conditions on all participants in a social interaction, as intellectualist approaches do. I will argue that a radical gradualist approach based on the idea of family resemblance can come up with a distinct set of conditions for AI systems to qualify as quasi-social interaction partners in another instance of asymmetric quasi-social interaction. I will suggest that consciousness, sentience, and reflective rationality might be conditions that can be questioned for artificial quasi-social interaction

partners and propose that minimal forms of agency and intelligence (socio-cognitive abilities enabling coordination with the interaction partner and instrumental rationality) could be sufficient. Thereby, I claim that agency and intelligence can be disassociated from consciousness.

Returning to the more far-reaching question concerning the potential ethical status, I will discuss considerations that may motivate us to use the idea of a multidimensional spectrum to characterize the potential quasi-ethical status of AI systems. There, I will point to, in my view, not very attractive consequences of a sharp divide between a full-fledged ethical status of assumed conscious entities and the lack of any ethical status of assumed non-conscious entities. If one assumes that consciousness is a necessary condition for qualifying for moral agency and patiency and applies this to AI systems, this can lead to two extreme positions, which I label as *Hard-core instrumentalist* and *In-expectation of AGI* view. Either one takes a position that denies AI systems any form of phenomenological consciousness in principle, meaning that they cannot be attributed any ethical status, or one argues that they do possess consciousness and thus can meet all the conditions for ethical status, including agency and intelligence. However, following a *Hard-core instrumentalist* view, one will either need to address an increasing number of responsibility gaps or extend human responsibility in cases where they were excused for good reasons. Arguing for an *In-expectation of AGI* view, one has to be prepared under certain circumstances to sacrifice humans for machines. I will conclude with some considerations of how assuming a multidimensional spectrum of moral agency and moral patiency could help to prepare the ground for another solution.

References

Shapiro, P. (2006). Moral Agency in Other Animals. *Theoretical Medicine and Bioethics*, 27(4), 357–373. https://doi.org/10.1007/s11017-006-9010-0

Véliz, C. (2021). Moral Zombies: Why Algorithms are not Moral Agents. *AI & SOCIETY*, 36(2), 487–497. https://doi.org/10.1007/s00146-021-01189-x