Moral Agents Unlike Us (Extended Abstract)

Jen Semler
Cornell Tech
jensemlerphilosophy@gmail.com

Suppose AI developers succeed in creating genuine artificial moral agents. That is, suppose AI systems will be reliable and competent moral reasoners, capable of extracting the morally relevant features of a situation and responding to moral considerations in their decision-making. They would not simply be following moral rules or mimicking moral agents. These systems would act *from* morality, rather than merely *in accordance with morality*.

Assume further that these future artificial moral agents will not be phenomenally conscious—they will have no first-personal experiences or affectively felt emotions. However, despite their lack of phenomenal consciousness, these systems will have all the necessary capacities underlying moral agency. For instance, we can assume such systems will be responsive to moral reasons. They will be merely cognitive moral agents.

Insofar as it is possible to create non-conscious artificial moral agents, these agents will be moral agents that are, in many ways, unlike the paradigm case of moral agency, namely prototypical adult humans. As such, it's not immediately clear what place these artificial moral agents should have in the moral community. Specifically, it's not clear whether artificial moral agents should have the same roles and responsibilities as human moral agents.

It might be tempting to adopt the following view:

> *Indifference*: For any moral decision-making context, there is no reason to prefer a conscious moral agent to a non-conscious moral agent as the decision-maker.

Indifference is motivated by the thought that, put simply, a moral agent is a moral agent, full stop—all moral agents, in virtue of being moral agents, should occupy the same moral roles.

For instance, suppose there are two human doctors, equal in all medically relevant ways. As doctors, part of their role involves making value judgments and moral decisions. In this case, it seems that we should be indifferent between the two doctors in this role. Both are competent doctors and moral agents, and so we have no reason for preferring one over the other. If we should be indifferent between two human moral agents, then denying *Indifference* seems to amount to speciesism—preferring a human moral agent just because she is human.

In this paper, I argue against *Indifference*. I argue that we should, in some cases, discriminate between conscious (i.e., human) and non-conscious (i.e., artificial) moral agents—even if non-conscious moral agents are genuine moral agents. This is not because human moral agents are better at or more justified in making moral decisions than artificial moral agents. And it is not because of speciesism. Rather, it is because many moral decision-making contexts require more than moral agency. Sometimes, moral agency is not all that matters.

In Section 1, I outline the notion of a non-conscious moral agent in moral detail, and I situate my claims within discussions of algorithmic decision-making, artificial moral agency, and the ethics of consciousness. In Section 2, I present two cases to evoke intuitions about when conscious moral agents may and may not be preferred over non-conscious moral agents.

In Section 3, I explain two underlying asymmetries between human moral agents and artificial moral agents that stem from their asymmetry in phenomenal consciousness. The *moral status asymmetry* is that human moral agents have moral patiency, while artificial moral agents do not. The *affective asymmetry* is that human moral agents have affective emotions, while artificial moral agents have (at most) functional or behavioral equivalents of emotions. In Section 4, I identify two ways in which these asymmetries manifest as factors that bear on *Indifference*: in cases involving relationships and some forms of responsibility.

In Section 5, I show how these factors help us understand when it's impermissible to allow artificial moral agents to make moral decisions. In Section 6, I consider near-term implications for the moral role of existing AI systems as well as corporations.