DISSERTATION ABSTRACT: *ON ARTIFICIAL MORAL AGENCY*

Suppose Tommy destroys Marvin's house. We might have some questions—whether, for instance, Tommy committed a moral wrong or whether Tommy is morally responsible. The answers to these questions depend, in part, on what kind of entity Tommy is. Our evaluation of this scenario differs if Tommy is a strong gust of wind, or a giraffe, or a human. Specifically, our assessment turns on whether Tommy is a *moral agent.* My dissertation considers what follows if Tommy is an AI system. On my account of moral agency, AI systems can be moral agents in principle, but existing AI systems fail to meet the necessary conditions. As such, our use of AI in moral decision-making should be limited. Moreover, genuine artificial moral agents will be different from human moral agents in normatively significant ways.

Part I—Towards a Theory of Moral Agency—develops a theoretical account of moral agency. In Chapter 2, "Moral Agency Without Consciousness" I preempt an objection to the prospect of artificial moral agency, namely that AI systems cannot be moral agents because they lack consciousness. I argue that phenomenal consciousness is not necessary for moral agency. In Chapter 3, "Two Types of Moral Agency," I propose and defend a distinction: *deontic moral agents* are appropriate subjects of deontic evaluations—their actions can be described as morally wrong—and *responsible moral agents* are appropriate subjects of responsibility ascriptions—they are fully-fledged moral agents. This distinction illuminates difficult cases of moral agency as well as instances of genuine responsibility gaps.

Part II—Prospects of Artificial Moral Agency—evaluates the extent to which AI systems are moral agents. I consider whether existing machine learning methods and empirical results support classifying existing AI systems, specifically large language models and reinforcement learning systems, as moral agents. In Chapter 4, "Artificial 'Agents' are Not Agents," I argue that AI systems lack the kind of agency required for moral agency—namely, the capacity for intentional action—because they lack mental states on both interpretivist and representationalist views. In Chapter 5, "Artificial 'Agents' are Not Moral," I argue that AI systems are far from instantiating the additional necessary capacities for deontic and responsible moral agency: AI systems lack responsiveness to moral reasons and moral understanding.

Part III—Using Artificial (non) Moral Agents—considers how the moral agency of AI systems, or lack thereof, bears on how we should use those systems in moral decision-making. In Chapter 6, "Artificial Moral Behavior," I argue that delegating moral decisions to AI systems is wrong—even if the outputs are reliable and accurate—because doing so replaces moral *actions* with, at best, moral *behaviors*. In Chapter 7, "Moral Agents Unlike Us," I argue that even if AI systems qualify for responsible moral agency, they are different from human moral agents in morally significant ways. While their lack of consciousness is no barrier to moral agency, it *is* a barrier to playing certain roles in the moral community. Moral agency is not all that matters.