On Artificial Moral Agency

Jen Semler

Reuben College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

13 January 2025

Acknowledgements

I love reading acknowledgements sections. It brings me great joy that I am finally in a position to write one. I am so lucky to have such incredible people in my life, and writing this dissertation would be neither possible nor meaningful without them.

Thank you to Reuben College for funding my DPhil—I am proud to be in the first cohort of students at Reuben. Thank you to the Institute for Ethics in AI (Oxford), the Philosophical Moral Psychology Lab (Oxford), and the Machine Intelligence and Normative Theory Lab (ANU) for providing spaces to explore, learn about, and discuss topics related to my dissertation.

My two supervisors, Carissa Véliz and Alison Hills, are second to none. In addition to being excellent philosophers, they are also inspiring role models and genuinely good people. Carissa, from the very beginning—when I decided to change my thesis topic during our first meeting—you have had such confidence in me and my abilities. Thank you for always encouraging me and for challenging my ideas in a constructive way. Alison, I'm so grateful that you came on board as my second supervisor a year early. Your insights and support have been invaluable. Thank you for helping me sharpen my arguments and for always offering an empathetic ear.

My growth as a philosopher has been greatly enhanced by a set of mentors who have gone out of their way to help me. Milo Phillips-Brown, you have been exceptionally generous with your time, feedback, and advice. Thank you for all the opportunities you have provided me with, and for making me a better thinker and writer. Joanna Demaree-Cotton, thank you for being such a great collaborator—while our experimental projects are not part of this dissertation, our work and conversations have undoubtedly influenced the way I think about moral agency and AI. Paul Henne, in many ways I feel that I owe my philosophical success to you. Thank you for showing me that I can do philosophy research and for all your guidance over the years.

Thank you to Alex Rosenberg, David Wong, and Carlotta Pavese for sparking and cultivating my interest in philosophy at Duke. Thank you to Ali Boyle and Alex McLaughlin for helping me shift from a student to a researcher at Cambridge.

I am fortunate to have wonderful friends from all the stages of my life. Some of these friends are also philosophers. They have been instrumental in helping me navigate the DPhil and have changed the way I think for the better. They have often reminded me of why I love philosophy. Nadia ben Hassine, thank you for understanding my struggles and celebrating my successes. Virginie Simoneau-Gilbert, thank you for keeping me sane. Kyle van Oosterum, thank you for always being there for me. Lewis Williams, thank you for your solidarity and advice.

To my non-philosopher friends, thank you in advance for featuring in my dissertation as characters in my thought experiments. For making Oxford feel like home, thank you to Kaitlyn Purdie, Ambre Bertrand, Aleks Petrov, Thea Guy, and Cillian Gartlan. For laughing and crying with me, thank you to Camilla Tacconis and Fabiana Piccoli Araújo Santos. For keeping me grounded, thank you to Emme Nagler, Mel Benson, Eleanor Sadik-Khan, and Izzie Gutenplan. For always picking up right where we left off, thank you to Dottie Kontopoulos, Aasha Reddy, Shom Tiwari, and Kanav Chhabra. I don't always get to see you all as often as I'd like, but it gives me solace that distance has not diminished our bonds in the slightest.

For some people, pursuing academic philosophy is a hard sell to their families. For me, this was never the case. My family has always encouraged me to pursue my passions and has always supported me without judgment—even when they didn't understand exactly what I was doing or why. To my parents, Jill Semler and Adam Semler, and my sister, Jamie Semler, thank you for instilling in me the values of education, hard work, and moral virtue. I cannot express the depth of my gratitude for all you have done for me.

And finally, to my partner, Tommy Escott, thank you for your unwavering belief in me, support, patience, delicious cooking, infectious silliness and laughter, and so much more. Because of it all.

Table of Contents

ACKNOWLEDGEMENTS **2**

TABLE OF CONTENTS **4**

SHORT ABSTRACT **6**

LONG ABSTRACT **7**

CHAPTER 1: INTRODUCTION **8**

1.1 A Thought Experiment 8
1.2 Moral Agency and Moral Patiency 10
1.3 Moral Agency and Artificial Moral Agency 12
1.4 A Criteria-Based, Functionalist Approach 22
1.5 Two Objections 32
1.6 A Roadmap 34

PART I: TOWARDS A THEORY OF MORAL AGENCY **36**

CHAPTER 2: MORAL AGENCY WITHOUT CONSCIOUSNESS **37**

Abstract 37
2.1 Introduction 37
2.2 Is Consciousness Necessary for the Possession of Moral Agency? 41
2.3 Is Consciousness Necessary for the Exercise of Moral Agency? 57
2.4 The Role of Consciousness in Moral Life 59
2.5 Conclusion 61

CHAPTER 3: TWO TYPES OF MORAL AGENCY **64**

Abstract 64
3.1 Introduction 64
3.2 Acting Wrongly and Acting Responsibly 66
3.3 Deontic Moral Agents 74
3.4 Responsible Moral Agents 79
3.5 Implications 82
3.6 Conclusion 86

PART II: PROSPECTS OF ARTIFICIAL MORAL AGENCY **88**

CHAPTER 4: ARTIFICIAL 'AGENTS' ARE NOT AGENTS **90**

Abstract 90
4.1 Introduction 90
4.2 Machine Learning Methods 92
4.3 AI and the Intentional Stance 96
4.4 AI and Representationalism 113
4.5 Conclusion 126

CHAPTER 5: ARTIFICIAL "AGENTS" ARE NOT MORAL **130**

ABSTRACT 130
5.1 INTRODUCTION 130
5.2 AI AND ACTING FOR MORAL REASONS 131
5.3 AI AND RESPONSIVENESS TO MORAL REASONS 136
5.3 AI AND MORAL UNDERSTANDING 145
5.4 CONCLUSION 151

PART III: USING ARTIFICIAL (NON) MORAL AGENTS 153

CHAPTER 6: ARTIFICIAL MORAL BEHAVIOR 154

ABSTRACT 154
6.1 INTRODUCTION 154
6.2 THREE OPTIONS FOR OUTSOURCING DECISIONS 156
6.3 ACTIONS AND BEHAVIORS 159
6.4 MORE THAN MERE BEHAVIOR? 169
6.5 CONCLUSION 171

CHAPTER 7: MORAL AGENTS UNLIKE US 173

ABSTRACT 173
7.1 INTRODUCTION 173
7.2 SOME CASES 175
7.3 ASYMMETRIES 177
7.4 MORE THAN MORAL AGENCY 181
7.5 THE ROLES OF ARTIFICIAL MORAL AGENTS 188
7.6 NEAR-TERM IMPLICATIONS 193
7.7 CONCLUSION 196

CHAPTER 8: CONCLUSION 197

REFERENCES 199

Short Abstract

This dissertation is guided by three questions: (1) what is moral agency, (2) are AI systems moral agents, and (3) why should we care? Part I, *Towards a Theory of Moral Agency*, develops a theoretical account of moral agency. In Chapter 2, "Moral Agency Without Consciousness," I argue that phenomenal consciousness is not necessary for moral agency. In Chapter 3, "Two Types of Moral Agency," I argue that there is a substantive distinction between entities that are appropriate subjects of deontic evaluations and entities that are appropriate subjects of responsibility ascriptions. Part II, *Prospects of Artificial Moral Agency*, evaluates the extent to which AI systems are moral agents. In Chapter 4, "Artificial 'Agents' are Not Agents," I argue that existing AI systems lack the capacity for intentional action because they lack mental states. In Chapter 5, "Artificial 'Agents' are Not Moral," I argue that AI systems exhibit only a minimal level of the moral competence required for moral agency. 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 because doing so turns events that should be moral actions into mere behaviors. In Chapter 7, "Moral Agents Unlike Us," I argue that moral agency is not all that matters—artificial non-conscious moral agents will be different from human moral agents in normatively significant ways.

Thesis Word Count: 67,135

Long Abstract

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.

Chapter 1: Introduction

1.1 A Thought Experiment

Let us begin with a thought experiment:

> *TJ*: An extraterrestrial—let us call them TJ—descends to Earth with no apparent means of leaving. While TJ's physiology is clearly non-humanlike, scientists are not yet able to determine precisely how TJ's inner workings operate. TJ can, however, communicate in English, interact with humans, and causally influence things in the world. Before letting TJ roam free in society, we must (among many other considerations) determine where TJ stands in the moral community. As such, a talented group of philosophers—ourselves included—is recruited to determine whether TJ is a moral agent.

Different people, philosophers and non-philosophers alike, might have different intuitions about TJ. Some might see the similarities between TJ and humans and think that TJ is straightforwardly a moral agent. Others might see how different TJ is from humans and think that there is no way TJ could be a moral agent. In this case, intuition might reasonably fail us. Determining whether TJ is a moral agent is very different from determining whether a new colleague is a moral agent. What we need, then, is a theory of moral agency to help us adjudicate in cases like this, where intuition fails or is otherwise unreliable.

Undertaking our inquiry involves addressing several interrelated questions. First, we need to know what moral agency is. That is, we need a theory of moral agency. This theory should tell us not only what we mean when we classify something as a moral agent, but also what the relevant criteria for moral agency are. There might be some uniform set of necessary conditions for moral agency, or there might be different ways to be a moral agent or different forms of moral agency.

Second, we need to know whether TJ meets the criteria for moral agency. We must determine both how to test TJ for the relevant capacities and, subsequently, whether TJ does, in fact, instantiate them. In this endeavor, we must think about the relationship between "what is going on inside"—that is, the physiological and psychological mechanisms occurring within TJ—and "what is going on outside"—that

is, the evidence we can gather from how TJ behaves. For instance, if TJ's appearance and behavior meet all the relevant conditions for moral agency, we must consider whether TJ's status as a moral agent would change if we later learned that TJ's brain is made of gummy worms, or a bunch of shrunken-down humans stuffed inside a mini control room, or some mysterious goopy substance. Moreover, we must consider the relevance (if any) of how TJ was created—whether our assessment would change if we learned that TJ was a product of biological procreation, or designed in a lab, or conjured by a powerful magician.

Third, we need to know why we should care about whether TJ is a moral agent. That is, we need a sense of what hinges on TJ's moral agency (or lack thereof). While considering the implications will not have bearing on whether TJ is a moral agent, it will have bearing on how we situate TJ more broadly within the moral community—what kind of roles we allow TJ to fill and what kinds of moral decisions we allow TJ to make.

Granted, positing TJ's existence might seem like a roundabout way of exploring the concept of moral agency. Indeed, we will soon depart from the TJ thought experiment and shift our focus to nearer-term candidates for moral agency. Still, keeping TJ—and the three questions pertaining to moral agency—in mind might help remove some potential biases involved in thinking about which entities do or do not qualify as moral agents.

This dissertation develops and applies a view about moral agency with an eye towards AI systems—and not just hypothetical futuristic AI systems like those found in science fiction, but possible near-term AI systems, conceptions of which are technically grounded in evidence from machine learning research. I believe that AI is a case in which our intuitions about moral agency might reasonably fail us—evidenced by the wide variation in views about what AI systems can and cannot do (and will and will not be able to do). My goal is to offer a reasoned analysis of artificial moral agency. This dissertation will not answer all the questions we might have about moral agency and artificial moral agency, but I hope it will lay the foundation for future work and guide our thinking about artificial moral agency in the right direction.

But first, in the remainder of this chapter, I address some preliminaries to set the stage for my analysis of artificial moral agency. In section 1.2, I explain the distinction between moral patiency and moral agency—my view remains agnostic about the relationship between the two. In section 1.3, I offer a brief overview of theories of moral agency and artificial moral agency, and I discuss the gap my work aims to fill. In section 1.4, I explain and defend my overarching criteria-based, functionalist approach to defining and evaluating moral agency. In section 1.5, I respond to two remaining objections to my project and approach. In section 1.6, I provide a roadmap for the rest of the dissertation.

1.2 Moral Agency and Moral Patiency

When thinking about TJ's place in the moral community, we might also—perhaps even prior to any discussions of moral agency—want to know whether TJ is a moral patient and, relatedly, whether TJ is entitled to any (direct, non-derivative) rights. The distinction between moral patients and moral agents can be viewed as a distinction between recipients and actors, respectively. Discussions of moral patiency focus on the criteria for moral consideration and whether entities including fetuses (Harman 1999; McMahan 2002; Singer 2011), non-human animals (Singer 1975; DeGrazia 1996; Frey 2005), the environment (Brennan 1984; L. Johnson 1993), and recently AI (Coeckelbergh 2010; Basl 2014; Neely 2014; Bryson 2018; Gunkel 2018; 2020; Liao 2020) have final moral value—that is, value for their own sake.

Discussions of moral agency, on the other hand, focus on what it takes to be a source of moral (and morally responsible) action. Most work on moral agency has focused on whether certain humans or groups of humans qualify for moral agency—adults with psychopathy (R. J. Smith 1984; Levy 2008; McGeer 2008; Driver 2015), adults with autism (Kennett 2002; Aaltola 2014), human children (Traina 2009; Tiboris 2014; Montreuil et al. 2018; Burroughs 2020), and corporations (French 1979; Björnsson and Hess 2017). More recently, moral agency has been explored in the context of non-human animals (Clement 2013; Fitzpatrick 2017) and—the focus of this dissertation—

AI (for an overview, see Behdadi and Munthe (2020); other accounts will be described in detail in section 1.3).

The distinction between moral patiency and moral agency is widely accepted in the literature and need not be further defended here. However, it is important to address the fact that some scholars seek to relate the two categories. There are five possible logical relations between the set of moral agents and the set of moral patients (Floridi and Sanders 2004). First, the sets could be disjoint, such that no entity is both a moral agent and a moral patient. Second, the sets could intersect, such that some entities are only moral agents, some entities are only moral patients, and some entities are both. Third, moral patients could be a subset of moral agents, such that all moral patients are moral agents, but not all moral agents are moral patients. Fourth, moral agents could be a subset of moral patients, such that all moral agents are moral patients, but not all moral patients are moral agents. Fifth, the sets could be equal, such that all moral patients are also moral agents and vice versa.

The common view in the literature seems to be that moral agents are a subset of moral patients. Indeed, there seems to be an intuitive sense that moral patiency is necessary for moral agency. We tend to think that all humans and some non-humans are moral patients but that only some humans are moral agents—human children, severely cognitively disabled humans, and some non-human animals have moral patiency but not moral agency. Conversely, we tend not to entertain the possibility that entities without moral patiency can have moral agency—rocks, screwdrivers, and toasters are neither moral patients nor moral agents. There is, however, no immediately obvious underlying reason to deny the possibility of an entity that is a moral agent without being a moral patient.

My approach will remain agnostic about the relationship between moral agency and moral patiency. Perhaps, in doing so, it will ultimately push back against the idea that moral patiency is necessary for moral agency—but this would be an implication rather than an assumption. It might also be the case that moral agency enhances moral patiency or grounds additional rights in entities that are already moral

patients.[1] But my investigation focuses solely on the criteria for moral agency, and nothing in my methodology hinges on a particular relationship between moral patiency and moral agency (or a particular conception of moral patiency).[2]

To assume that an entity must first be a moral patient to qualify as a moral agent would beg the question regarding what is important for moral agency—and it would be problematic to presuppose without further argument that entities cannot qualify as moral agents unless they first qualify as moral patients. If, for instance, (1) sentience is necessary and sufficient for moral patiency and (2) sentience is necessary for moral agency, then (3) all moral agents are moral patients. But for this argument to be sound the premises must be argued for and not merely assumed. In fact, I argue against premise (2) in Chapter 2.

1.3 Moral Agency and Artificial Moral Agency

Various theories of moral agency and artificial moral agency have been offered in the literature. In this section, I provide a brief overview of both literatures—which do not always communicate with each other. I then discuss the role my dissertation plays in these literatures.

1.3.1 Theories of Moral Agency

"Moral agency" is a philosophical term of art. Many philosophers invoke the term across a wide range of topics, and there is some intuitive sense in which we all know what "moral agency" means. But while the term aims at a concept that is both useful

[1] Alternatively, it might be the case that certain capacities or sets of capacities required for moral agency also ground moral patiency. For example, intelligence, consciousness, freedom, and moral understanding might be jointly sufficient for a moral right to freedom, even without sentience (Sinnott-Armstrong and Conitzer 2021). These capacities might also be necessary for moral agency.

[2] In Chapter 7, however, I return to the relationship between moral agency and moral patiency.

and meaningful to philosophers, it is often imprecisely explicated and thus lends itself to subtle equivocation.

As Chapter 3 will demonstrate, my diagnosis of some of the confusion around moral agency is its close connection with the concept of moral responsibility. Theories of responsibility aim to capture the conditions under which an entity is morally responsible for its actions. It is natural, then, to think that theories of moral agency aim to capture what makes an entity the kind of thing that can be held morally responsible—and so, the right theory of moral agency will simply include the capacities that an entity must have to be morally responsible. For example, if an entity can only be morally responsible for an action if it was sufficiently reasons-responsive, then reasons-responsiveness is necessary for moral agency.

But moral agency is not always straightforwardly defined in line with moral responsibility. Haksar defines moral agents as "those agents expected to meet the demands of morality" (Haksar 1998). He further defines moral agents as being accountable, subject to moral duties and obligations, and subject to moral praise and blame. Already, we see an ambiguity arising, as these three features might come apart. It is not clear, on Haksar's definition, whether moral agency involves all three of these features or only a subset of them.

Watson defines moral agents as those who "can, to a significant extent, act effectively and competently in moral matters" (Watson 2013). He further defines moral agents as being autonomous (in the sense of having self-determination and self-governance) and accountable (in the sense of being answerable to others). It is not clear, however, whether the claim is that these features are instrumentally necessary for being a competent moral actor, or whether they, too, are constitutive of moral agency.

Arpaly, despite offering a thorough and important account of the exercise of moral agency, does not define what moral agency is (Arpaly 2003). Rather, she proposes a theory of what she calls "moral worth"—that is, moral praiseworthiness and blameworthiness. She is more concerned with the question of what makes it the case that the same action can prompt different degrees of praise and blame in different

agents (Arpaly 2002). Of course, as mentioned above, we can extract a theory of moral agency from her view by taking the proposed criteria for moral worth and asking which capacities underlie them (Nailer 2022). On this view, then, a moral agent would be defined as an entity capable of having moral worth. But Arpaly does not focus so much on the question of which entities have moral worth in general. Instead, she focuses on the moral worth of agents in relation to particular actions.

Already, then, we see that that moral agency can mean one, or some combination, of various related but distinct concepts: (1) entities that are expected to meet moral standards, (2) entities that have moral obligations, (3) entities that are accountable, (4) entities that are subject to moral praise and blame, (5) entities that have moral worth, (6) entities that act competently in moral matters, (7) entities that are autonomous, or (8) entities that are answerable to others. Given that at least some of these concepts seem to come apart (Watson 1996; Shoemaker 2011a), we need to be clear about what we are referring to when we ask whether certain capacities are necessary for moral agency or whether a particular entity qualifies as a moral agent. I take up this problem in Chapter 3 ("Two Types of Moral Agency").

Historically, there are three main strands of thinking about moral agency. Rationalist views draw from Kant in emphasizing the role of reason in moral agency (Kant 2018). Sentimentalist views draw from Hume in emphasizing the role of emotions in moral agency (Hume 2000). These theories are often viewed in opposing terms, such that either cognitive capacities or affective capacities are the main capacities constitutive of moral agency (Wilson and Denis 2024). A contemporary version of the rationalist vs. sentimentalist divide can be seen in debates regarding the moral agency of psychopaths and autistic individuals, who both suffer from deficits in empathy (Kennett 2002; Levy 2008; Litton 2008; McGeer 2008; Borg and Sinnott-Armstrong 2013; Aaltola 2014).

Virtue-based views draw from Aristotle in that they focus on the characteristics moral agents develop—and moral agents can have different combinations of the relevant virtues and to different degrees (Aristotle 2019). Moreover, on these views, moral competence is viewed as a practical skill rather than a set of capacities

(McDowell 1979; Nussbaum 1990; Graff 2024). Virtue-based views are less frequently discussed in opposition to the other views, but they offer an important alternative account of moral agency.

Contemporary discussions of moral agency and moral responsibility posit a wide range of capacities that are said to underlie moral agency (Talbert 2024). Contemporary views often highlight reasons-responsiveness as necessary for moral agency, though there are different conceptions of what responsiveness to moral reasons consists of (Fischer and Ravizza 1998; McKenna 2013). Strawsonian views focus on the "reactive attitudes" we take towards moral agents when we deem them morally blameworthy—and refrain from taking towards those who are not moral agents (P. F. Strawson 2008). Such views often focus on a moral agent's ability to participate in the social practices surrounding responsibility. "Deep self" views of moral responsibility tend to assess moral responsibility according to whether an agent's actions, values, or characteristics reflect who the agent really is, deep down—as opposed to reflecting who the agent is only at a superficial level. The necessary capacities for moral agency on these views will be those necessary for having a deep self in the first place, such as the ability to possess stable character traits and the ability to endorse one's own values (Wolf 1988; Sripada 2016).

While these different theories often converge in terms of who, or what, qualifies as a moral agent, they come apart in so-called marginal cases of moral agency and responsibility (Shoemaker 2015). AI systems may fall into this category. In this sense, AI systems are well-poised to put pressure on different conceptions of moral agency and can help us adjudicate between different theories.

1.3.2 Theories of Artificial Moral Agency

"Artificial moral agent" is another term of art—one that is used by both philosophers and computer scientists, often in divergent ways. Problematically, it is not always clear that the term "artificial moral agent" is being used in alignment with the philosophical

concept of moral agency. There are three main ways of thinking about the relationship between theories of moral agency and theoris of artificial moral agency.

Some views, which I call *separatist* views, discuss artificial moral agency as something distinct from moral agency. On such views, artificial moral agents are merely entities that operate in morally laden contexts, or an ideal towards which machine ethicists aspire. Wallach and Allen, for instance, are not concerned with whether artificial moral agents can be human-like moral agents—they are more concerned with developing systems that can behave as if they are moral agents (Wallach and Allen 2009). Their proposed framework specifically targets the development of advanced artificial systems with functional morality.

Similarly, Moor coins four types of "ethical agents," but only the final category, "full ethical agents" seems to capture the philosophical sense—and Moor does not think machine ethicists should be trying to create this type of artificial moral agent (Moor 2006; 2009). Malle claims that instead of focusing on the prospects of artificial moral agency, we should instead focus on building the capacities for moral competence into machines (Malle 2016). While questions may be raised about the extent to which these pragmatic machine ethics approaches successfully dodge the philosophical questions about moral agency, the researchers need not be viewed as making any ontological claims about the moral status of such artificial systems.

Insofar as these discussions of artificial moral agency are not aiming at the philosophical concept, an immediate response might be that these theories should invoke a distinct terminology to avoid conflating artificial moral agency with moral agency as understood by philosophers.[3] However, if these distinct concepts bear some relationship to genuine moral agency, these connections must be explored and clarified. Otherwise, it is not clear what follows from classifying an AI system as an artificial moral agent—we still would not know whether the system is a moral agent,

[3] See Zafar (2024) for a critique of the use of philosophical terminology in discussions of artificial morality and a proposal for using the concept of "AI automated performance" instead of "AI agency" (and artificial moral agency).

proper. Failing to differentiate between genuine accounts of moral agency and separatist accounts of artificial moral agency risks (1) equivocation regarding the term "artificial moral agency" such that participants in the discussion are talking past each other, (2) over-attributions of moral agency to entities that are not genuine moral agents, and (3) false perceptions arising regarding not only the capabilities of artificial systems, but also the meaning and significance of moral agency itself.

Other views, which I call *revisionist* views, propose to alter the definition of moral agency to accommodate AI systems. Floridi and Sanders fit into this category. On their account, moral agency consists of three features—interactivity, autonomy, and adaptability—and, at the right level of abstraction, computer systems can meet these conditions and thus qualify for moral agency (Floridi and Sanders 2004). Several other views appeal to the notion (from Floridi and Sanders) of mind-less morality in their accounts of artificial moral agency (Grodzinsky, Miller, and Wolf 2008; Sullins 2009; Mabaso 2021; Tollon 2021). Coeckelbergh's view can also be seen as revisionist, in the sense that he proposes a shift in thinking about moral agency—away from questions of whether an entity possesses the necessary capacities for moral agency, and towards questions of whether an entity appears to us as a morally responsible agent (Coeckelbergh 2009; 2010).

Some revisionist views explicitly attempt to place humans and artifacts on the same spectrum of moral agency. According to Miller and Larson, there is a continuum with human moral agents on one end and artifacts like speed bumps on the other end—and while speed bumps are not moral agents, they have morally significant human intentions built-in (Miller and Larson 2005). Asaro claims that moral agency is a continuum with amorality on one end and fully autonomous morality on the other end—and as we move along the spectrum, we see "robots with moral significance," "robots with moral intelligence," and "robots with dynamic moral intelligence" (Asaro 2006).

For revisionist attempts to succeed, however, they must explain why the concept of moral agency requires revision and what grounds the proposed revision. Existing revisionist views often fall short in this regard. For instance, it is not clear why

we should put humans and mundane artifacts on the same spectrum, and it is not clear why we should think that technological advancements alone are bringing artifacts closer to humans in terms of moral agency. To borrow an analogy from Dreyfus, we should not expect the ability to climb higher trees to help us reach the moon (Dreyfus 1965).

The most common type of view, which I call *accommodating views*, seek to retain the traditional conception of moral agency and consider whether artificial moral agency is possible. Some accommodating views claim that AI systems cannot instantiate the relevant capacities for moral agency—including sentience (Véliz 2021), emotions (Rodogno 2016), acting for reasons (Purves, Jenkins, and Strawser 2015; Talbot, Jenkins, and Purves 2017), mental states (D. G. Johnson 2006; Himma 2009), value-setting agency (Fossa 2018), intentionality (Friedman and Kahn 1992), moral sensitivity (Graff 2024), authenticity (Gudmunsen 2024), sophisticated conceptual abilities (Parthemore and Whitby 2013), and understanding (Stahl 2004). Accommodating views of virtue ethics hold that AI systems cannot be virtuous moral agents because they lack the ability to perform the right actions for the right reasons in the right circumstances (Constantinescu and Crisp 2022).

Other accommodating views argue that AI systems can instantiate the relevant capacities for moral agency. For instance, some functionalist views hold that AI systems can be moral agents. Laukyte argues that artificial agents can meet the conditions of rationality (including the possession of mental and motivational states), interactivity, responsibility (including a capacity for normative judgment and understanding), and personhood (in the sense of operating within a system of obligations) (Laukyte 2017). List argues that AI systems organized in the right way can meet the same agency conditions as group agents—including the capacities for intentional action and moral responsibility (List 2021).

Still other accommodating views deny that AI systems can be moral agents but offer alternative ways to think about the moral role of AI systems and other technological artifacts. Brey argues that while machines are not moral agents, a structural ethics approach can help us evaluate the role they play in bringing about

moral actions (Brey 2014). Illies and Meijers argue that we should focus on the ways in which technological artifacts affect action schemes, the repertoire of possible actions available to an agent in a particular situation (Illies and Meijers 2014). Johnson and Noorman argue that artifacts can be viewed as moral agents only in a causal efficacy sense or in a metaphorical sense of acting on behalf of someone else (D. G. Johnson and Noorman 2014). Johnson and Powers argue that computer systems can act as surrogate agents, but in a different way than human surrogate agents (D. G. Johnson and Powers 2008). Nyholm argues that we can view automated systems as having supervised and deferential agency even though they cannot bear moral responsibility (Nyholm 2018). Petersen and Spahn argue that technological artifacts are never moral agents but can sometimes affect the moral evaluation of actions (Peterson and Spahn 2011).

The lines between these views on the relationship between artificial moral agency and moral agency are not always so clear-cut. For instance, in some regard, some of the functionalist views might be deemed revisionary in that they depart from certain features of the standard conception of moral agency (e.g., by deemphasizing the phenomenal aspects of mental state possession). But these views are not revisionary in the sense of providing an entirely new set of criteria for moral agency. They still maintain that capacities typically associated with moral agency such as the capacity for representational states and responsibility are necessary for moral agency—they just argue that AI systems can instantiate those capacities.

1.3.3 The Gap

Given the variety of definitions of moral agency and artificial moral agency, the question of whether an entity is a moral agent risks becoming a mere stipulation: the answer simply depends on whichever definition of moral agency we are choosing to adopt. Moreover, the concept of moral agency is complex in that it is constituted by a wide range of capacities. As such, even on a single definition of moral agency, different theories of those capacities (e.g., mind, action, responsibility, consciousness, etc.)

might produce different answers to the question of whether an entity meets the relevant criteria for moral agency.

This dissertation is not the place to fully defend views in each of these philosophical subdomains. But it is also not an attempt to conclude the possibility of artificial moral agency at any cost, adopting a highly revisionary, hodge-podge combination of views that culminate in artificial moral agency in exchange for plausibility. As the next section demonstrates in more detail, I adopt a criteria-based, functionalist approach for evaluating moral agency. This view enables consistency while aiming at a conception of moral agency that is normatively significant.

I aim to offer an analysis of moral agency that is primarily accommodating but has revisionary aspects. My account is accommodating in that it takes seriously the traditional philosophical concept of moral agency, first aiming to understand what moral agency is. The revisionary aspects are driven not by a desire to fold AI systems into the domain of moral agency, but instead by a desire to clarify the nature of moral agency first, and only then to consider whether artificial moral agency is possible and likely.

Still, given the many decision points involved in addressing artificial moral agency, there might be temptations throughout the dissertation to jump ship, so to speak—that is, to abandon the notion that an artificial entity might qualify for moral agency in virtue of rejecting one of the commitments on which the arguments rest. However, there are several reasons to remain on board even if skepticism arises about the overall prospect of artificial moral agency.

First, part of this project involves mapping the conceptual space of moral agency. As such, it is helpful to have a principled way to locate where and why different views take issue with (or advocate for) the possibility of artificial moral agency. Doing so will also help contextualize the literature on artificial moral agency, which in its current state lacks conceptual clarity. Consider the sheer number of views discussed above. Of course, these contributions are all helpful in furthering our understanding of the ways and senses in which artificial entities can relate to morality. But it can be difficult to draw broader conclusions when the discussion lacks unity.

Upon recognizing the conceptual messiness in the artificial moral agency literature, Behdadi and Munthe draw the following conclusion:

> All of this provides reasons to doubt that participants of the AMA [artificial moral agency] debate are discussing the same thing: how one specific concept of moral agency applies to artificial entities. Rather, our impression is that there is a multitude of (often underexplained) concepts of moral agency and that many proposals are therefore much less in conflict than debaters assume. (Behdadi and Munthe 2020, 212)

Mapping these views onto a single conception of moral agency, then, will help determine whether different authors are disagreeing—and if they are, what they are disagreeing about. It will also allow new views to specify where and how they are entering the debate.

Second, even skeptics who deny the possibility of artificial moral agency have good reason to care about artificial moral *behavior*. Evaluating the morally relevant capacities AI systems might instantiate, even if they do not combine into moral agency, will help inform how we ought to use different systems for different applications. Moreover, even if relevant capacities are not instantiated by AI systems, they might be mimicked or simulated in artificial moral behavior. Answers to normative questions about the appropriateness of deploying AI systems in particular contexts will depend, in part, on the capacities (real, simulated, or approximated)—or combinations of capacities—of the system in question.

Third, thinking about artificial moral agency puts pressure on the concept of moral agency in general. AI continuously challenges our thinking by performing tasks most people thought computer systems could never achieve—from beating humans at chess and Go (Silver et al. 2016; 2018; Schrittwieser et al. 2020), to generating realistic images from natural language (Betker et al. 2023), to predicting the structures of proteins (Jumper et al. 2021). At the same time, advances in AI capabilities highlight key limitations—accurate computer vision systems are susceptible to adversarial attacks (Biggio et al. 2013; Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015), advanced facial recognition software exhibits racial and gender bias (Buolamwini 2018), and convincing dialogue agents hallucinate facts (Ji et al. 2023).

Such progress and obstacles give us reason to challenge our intuitions that AI "definitely can" or "cannot possibly" qualify for moral agency. The uniqueness of AI, both in theory and in practice, arises in part because AI systems combine abilities in ways previously unseen and unconsidered. Large language models, for instance, demonstrate capabilities that might have been assumed to be impossible without mental states or conscious reasoning (e.g., providing dynamic explanations in response to questions, generating poems, and adopting different speaking styles). New combinations of capacities might form normatively significant clusters—perhaps giving rise to new types of moral agency or instead pinpointing dangers.

Many attempts have been made to understand the prospect of artificial moral agency across the separatist, revisionist, and accommodating approaches described above. Once we acknowledge that arguments about artificial moral agency might not be in direct conversation with each other, it might be tempting to shift away from questions about the conditions for moral agency and whether technological artifacts meet them (Behdadi and Munthe 2020).

But it would be a mistake to abandon the project of defining moral agency in a unified way. First, moral agency is an important concept—it picks out a normatively significant class of entities that are moral actors, and moral actors have a unique role in the moral community. Second, the lack of consensus on the necessary conditions for moral agency does not imply that we should stop trying to figure out what those conditions are; rather, it points to the fact that more work must be done to understand what constitutes moral agency. A better response, then, to the problem in the artificial moral agency literature would be to clarify the conceptual debate by making the concept of moral agency more precise.

1.4 A Criteria-Based, Functionalist Approach

When we think about what constitutes moral agency, it is natural to rely on our best—and only—uncontroversial and widely accepted example of moral agents:

prototypical adult humans.[4] On the one hand, any plausible conception of moral agency must accommodate that fact that such humans are the paradigm case of moral agents. On the other hand, we must be careful that an over-reliance on humans as our benchmark does not obscure the essential features of moral agency.

Briefly, there are three key dangers to avoid in using prototypical adult humans as exemplars. First, we should avoid substituting sufficiency for necessity—and combat the temptation to assume that capacities sufficient (individually or jointly) for moral agency in the human case are necessary for moral agency in general. That is, we should be wary of assumptions that certain human features, such as self-consciousness and metacognition, are necessary for moral agency without further argument.

Second, we should avoid conflating moral agency with other closely related concepts—particularly those that tend to co-occur in humans. For example, just because moral agency and moral emotions seem to go together in the human case, we should not conclude that these categories must necessarily go together.

Third, we should avoid overgeneralizing in a way that assumes moral agents are monolithic. Of course, moral agents should share certain features in common. However, it is possible that there are multiple types of moral agents or multiple pathways to moral agency. Specifically, we should be open to the conceptual possibility of non-human and non-biological moral agents.

A criteria-based, broadly functionalist approach, as I describe below, offers a way to examine and make progress on the questions surrounding artificial moral agency.

1.4.1 A Criteria-Based Approach

[4] Some responsibility skeptics might deny that prototypical adult humans meet the necessary conditions for moral agency or responsibility (G. Strawson 1994). But such considerations are beyond the scope of this dissertation.

Defining moral agency by the characteristics an entity must possess to be a moral agent is a moral individualist approach, also referred to as an intrinsic properties approach. This method is appealing because it aligns with the idea that a complex concept like moral agency can be broken down into its component parts. Figuring out what moral agency is and whether an entity is a moral agent seem to require identifying the capacities necessary for moral agency. Moreover, the method treats all entities as equal regarding prospects of moral agency: so long as the entity has the relevant ontological features, it will qualify for moral agency regardless of other factors. But other approaches have been proposed, and the intrinsic properties approach must be defended against such alternatives.[5]

Some alternative approaches focus more on social-relational considerations. On moral relationist views, membership in the moral community depends not on the characteristics of the entity in question, but rather on that entity's relationships with other members of the moral community (May 2014; Crary 2016; Gunkel 2022). These approaches can take various forms. Crary, for instance, argues that moral individualism neglects the observable moral qualities we can access through interaction and moral imagination (Crary 2016). Gunkel similarly entertains the possibility that, in line with a tradition leading from Levinas to Žižek, ethics might precede ontology (Gunkel 2018; 2022).

To put it differently, relationist views posit that we first decide to treat entities as morally or socially significant, and then we project morally relevant properties onto them. On moral relationist approaches, then, we should be defining moral agency not by looking for capacities or intrinsic qualities, but rather by looking at the way we—fully fledged members of the moral community—relate to other entities.

In a similar spirit, in the artificial moral agency debate, Behdadi and Munthe call for a directly normative approach: they claim that "we should ask how and to

[5] Discussions around these different approaches often pertain to moral patiency rather than moral agency. So, it is possible that those who hold alternative views about moral patiency might adopt an intrinsic properties approach to moral agency. Still, they are worth considering in the context of moral agency even if the authors are not committed to such an application.

what extent artificial entities should be incorporated into human and social practices that would normally have us ascribe moral agency and responsibility to participants" (Behdadi and Munthe 2020, 212). Nyholm offers a similar, though perhaps less extreme, view. Instead of focusing on the individual agency of artificial entities, he argues, we should view machines as being involved in a shared or supervised form of agency with humans (Nyholm 2018).[6] These proposals attempt to put our relationships, or the ways in which we interact (or want to interact) with artificial entities, at the forefront of moral agency attributions.

There are three reasons to be wary of relationist approaches, at least for an investigation of artificial moral agency. First, such approaches do not seem to work well with novel or unfamiliar entities. It is unclear what relationism would tell us about the potential moral agency of entities with which we have little history or experience of interaction. Moreover, particularly for new entities, the relationist approach would be too heavily intuition-based and susceptible to bias and manipulation. Seemingly irrelevant properties, such as human-like appearance or cuteness, could promote anthropomorphism and influence our judgments, particularly in the case of AI (where companies are incentivized to get people to use their products). Conversely, our experiences with technological artifacts that do not look like humans (or bear a visual resemblance to things that we know are not moral agents) might bias us against forming normatively significant relationships with such entities.

Second, and relatedly, relationships change over time and across contexts. Non-human animals, for instance, have been increasingly recognized as having moral

[6] Coeckelbergh's discussion of virtual moral agency might be interpreted in a similar vein, as his argument prompts us to focus on how robots appear to us rather than whether they really have the capacities in question (Coeckelbergh 2009). However, Coeckelbergh's view could also be interpreted as a moral individualist view that focuses on capacities—the "virtual" aspect can be understood as a method of evaluating the possession of capacities or as an alternative set of capacities (e.g., the capacity "possession of moral emotions" might be replaced with the capacity for "convincing appearance of the possession of moral emotions," and looking at "virtual" capacities might be a way of providing evidence for the "true" capacities).

status. This trend is partly due to new forms of interactions and relationships between humans and their companion animals. Pigs, for instance, are increasingly becoming pets (as well as ducks, goats, cows, and other animals previously thought incapable of forming meaningful relationships with humans).

A relationist might claim that we have discovered that we can form meaningful relationships with pigs—and this crucial fact tells us that we were previously wrong about their moral status. While the relationist is certainly right that our newfound relationships with pigs help us recognize their moral status, it would be strange to claim that our view of the moral status of pigs has changed solely or primarily in virtue of our relationship to them.

Suppose we accept the relationist explanation that we ascribe higher moral status to pigs in virtue of our newfound relationships with them. We can ask a further question: In virtue of what can we form these relationships with pigs? The explanation is provided by pigs' underlying capacities—as such, our relationships merely enable us to see the morally relevant properties that pigs have (e.g., intelligence, an ability to learn, and perhaps some social-emotional competence). The fact that we recognize a higher moral status for pigs now is fundamentally explained not by the fact that we have formed meaningful relationships with them, but rather by the fact that we know more about pigs' capacities, which warrant a higher level of moral status. It is these very capacities that enable the relationships which prompt us to reconsider pigs' moral status.

Moreover, rejecting the relational approach allows us to consider that we might be wrong in our ascriptions of moral status—we might think an entity deserves moral status in virtue of our relationship with it, but we are in fact mistaken about its ability to engage in this relationship due to its underlying properties (or lack thereof). For instance, a person might start forming a relationship with her interlocutor over the phone and believe that her interlocutor has moral status. But upon learning that the "interlocutor" is merely an automated response, the person will revise (and withdraw) her ascription of moral status.

Third, and following from the previous points, ascriptions of moral agency on relationist views risk introducing biases. Historically, for instance, members of oppressed groups have been denied membership to the moral community—in terms of both moral patiency and moral agency—because they were not deemed as having (or being capable of having) the right kinds of relationships with those already in the community of moral agents. Relational approaches run the risk of excluding deserving members of the moral community merely because they are different from us.

Consider again an analogy from moral status. It is easy to relate to certain non-human animals like dogs, and very difficult to relate to other non-human animals like shrimp. But insofar as what really matters for moral status is an underlying capacity (e.g., the capacity for sentience), the social-emotional relationships we can have with dogs but cannot have with shrimp can bias us towards thinking that shrimp lack moral status. Regarding moral agency, then, the social connections we might have with entities can serve as a bias, especially if we ourselves have biases regarding whom we choose to form relationships with.

Still, some philosophers might hold that the criteria-based approach overemphasizes the role of individual subjects in moral agency. Floridi and Sanders, for instance, claim that the focus on individual agents, particularly individual human agents, hinders investigation into distributed morality (Floridi and Sanders 2004). Hanson, slightly more radically, argues that moral individualism focuses too much on individual subjects as actors and thus fails to acknowledge that most actions are undertaken by networks of entities (Hanson 2009). The concern underwriting these objections is that focusing on the individual prematurely rules out alternative conceptions of moral agency, under which moral agents might include individual non-human entities, a combination of humans, or a combination of human and non-human entities (Latour 1987). Such concerns are in line with Nyholm's push for thinking about agency as shared between humans and robots (Nyholm 2018).

But there remains room to push back on the distributed agency concern. Here, an analogy from causation is useful. For any outcome, the cause is a combination of factors. A house fire, for example, is not fully caused by a toaster malfunction, but is

instead caused by that plus the fact that the house is made of flammable materials, plus the presence of oxygen in the house, plus the absence of a person to extinguish the fire, plus many other factors. Yet the existence of other causes does not stop us from identifying a particularly significant cause. Note that this assessment does not have to be either/or—it can be the case that causation is distributed and that the toaster malfunction is a more important causal factor than other features of the situation.

Moreover, we can break down the "distributed causation" by outlining the causal role played by each factor. The same can be said for agency. Any morally relevant action will involve more than a single subject, but it does not follow that an individual cannot be appropriately deemed the relevant moral agent. Even if the agency is distributed, we should still be able to evaluate the ways and extents to which different members of the distribution exercised their agency. Brey recognizes this point when advocating for a structural ethics approach to the role of artifacts in morality (Brey 2014). Brey's view is presented as complementary to individual ethics, which focuses on the actions of individual moral agents.

Another alternative to consider is a constructivist approach to moral agency. On such a view, the question of whether an entity—particularly a machine—is a moral agent is about neither the capacities the entity possesses nor the relationships we form with the entity. Rather, it is a choice that we, as members of the moral community, make regarding that entity. Bryson argues for such an approach to moral patiency, claiming that integrating AI into our moral systems is normative, not descriptive, ethics (Bryson 2018). Put differently, on such a view of moral agency, membership in the class of moral agents is not uncovered but decided. Johnson and Miller argue that the moral status of computer systems is not a matter of truth—we are not required to draw conclusions about their moral status, even if they exhibit certain features,

because computer systems are never fully independent from their designers (D. G. Johnson and Miller 2008).[7]

In the case of AI, these arguments might be initially tempting. Technology is designed and created for human use, and it might make sense to think that it is up to us to determine how machines fit into the moral community—after all, we are the ones making them. However, such an approach is deeply problematic, as members of the in-group of moral agents should not be able to determine what does and does not count as a moral agent in a non-standardized and non-systematic way.

A constructivist approach risks precluding certain entities from moral agency despite their exhibited characteristics. If an entity—machine or otherwise—meets the criteria for moral agency, it should qualify as a moral agent regardless of whether we want it to be a moral agent, and regardless of whether we designed it. Even if we intend to make entities that are not moral agents, we will have to grant them moral agency if they end up instantiating the relevant capacities.

Still, the constructivist might push back by saying that there is no further truth about what the criteria are for moral agency—instead, we can choose who is and is not a member of our moral community. But a plausible constructivist view will not hold that we can arbitrarily decide who is and is not a moral agent; rather, a constructivist view should be systematic and consistent. So, given that there are grounds on which we apply moral agency to humans, it would be arbitrary and non-systematic to exclude non-humans that meet the same criteria upon which we include humans. In other words, an appeal to capacities helps the constructivist select who belongs in the moral community in a principled way.

Of course, in the case of technology, humans are deciders in the sense that they can choose to build or refrain from building technology with certain characteristics. I

[7] There are other reasons for doubting this argument. Humans are also never fully independent since they are "designed" by their genetic code, evolution, and their parents (and, on some views, a deity).

am concerned with whether we *can* create artificial moral agents. Whether we *ought* to create artificial moral agents (assuming it is possible to do so) is a separate question.

Ultimately, then, the criteria-based method is the best way to approach the topic of artificial moral agency; it has independent motivation, and alternative approaches are uncompelling without at least some appeal to capacities.

1.4.2 A Functionalist Approach

When adopting a criteria-based approach, not all criteria are equal—there should be some principled restrictions on the kinds of criteria adopted. For instance, "having two eyes" would be an inappropriate criterion for moral agency, as the possession of two eyes has nothing to do with an entity's capacity to be a moral agent (indeed, there are human moral agents without two eyes—and even more without two functioning eyes).

More generally, criteria should avoid ruling out any species in advance without justification—it would be wrong to assume that "being human" is a criterion for moral agency, as doing so excludes all other entities from consideration without good reason (Liao 2020). To borrow a phrase from Schwitzgebel and Garza, for two entities to deserve different degrees of moral status—or in this case, different attributions of moral agency—there "must be some relevant difference between the two entities that grounds this difference in moral status" (Schwitzgebel and Garza 2020, 459). These considerations, I assume, are largely unobjectionable.

These restrictions still leave a wide range of possible capacities on the table. I believe that the functionalist, broadly conceived, offers the most promising approach to evaluating moral agency. On such an approach, an appropriate criterion for moral agency must be empirical and observable (Liao 2020). Moreover, the broad functionalist approach allows for the relevant criteria to be multiply realizable.

One of the main advantages of the functionalist approach is that it allows us to determine whether other entities are moral agents in practice. As exemplified in the case of TJ, it is impossible to determine the moral agency of a particular entity a priori.

It is plausible that we can determine the *criteria* for moral agency a priori—and if we know whether an entity fulfills the relevant criteria, we can know whether it is a moral agent. But if we want to make a practical decision about whether an entity meets any given criterion, that criterion must be operationalizable or testable in some way.

Importantly, it does not follow from this claim that such tests for the criteria for moral agency are easy to create or implement. For instance, it does not follow from the criterion of observability that the Turing Test is an adequate test for intelligence or that the Moral Turing Test is an adequate test for moral agency. My approach is consistent with the claim that we currently lack the appropriate tests for certain capacities, but my view does require that such tests are possible. For example, if consciousness were necessary for moral agency, the broad functionalist approach would require that there is some way to test for consciousness, or some indicators of consciousness that can provide evidence that an entity is conscious (Butlin et al. 2023). The functionalist approach could not accommodate a view that says that consciousness is necessary for moral agency but that there is no way to figure out whether a particular entity is conscious.

Still, it might be objected that the criteria for moral agency are plainly unobservable. Such a view is possible to hold, but it is unhelpful if we want to know which entities qualify as moral agents. This pragmatic response might be unconvincing to those who hold that there is truly an untestable component of moral agency. If this is the case, it points to a limitation in my methodology. If moral agency has necessary but unobservable features, my approach will not be able to tell us whether a novel entity is a moral agent. But neither can the competing approach. Insofar as moral agency is untestable, we cannot know whether *any* entity is a moral agent (though we might be able to infer the presence of unobservable qualities from the observable ones).

Additionally, even if functionalism does not offer a complete view of moral agency, it remains plausible that many criteria for moral agency have at least some functionalist profile. As such, the functionalist can at least be seen as offering some useful evidence of moral agency to the non-functionalist.

What remains important is recognizing that moral agency and its component parts are multiply realizable—and denying this feature would be difficult to uphold. We already accept that mental states and capacities are multiply realizable, evidenced by our attributions of mental states to non-human animals as well as scientific findings on neural plasticity; we also recognize, to some extent, that moral agency is multiply realizable, as humans instantiate moral agency and the capacities associated with it in different ways. We must be careful not to disqualify artificial entities from moral agency merely because they are not made from the same materials or in the same ways as humans.

1.5 Two Objections

Skeptics about my project might offer two more objections. The first objection is about grouping. If the capacities are what's important for moral agency, the objection goes, we should focus on what each capacity individually entails rather than grouping them together to form the status of "moral agent."

In response, an argument from analogy is a useful starting point. A parallel objection might be raised against the concept of moral patiency: if what matters for moral patiency are capacities (e.g., for sentience, interests, higher-order cognition, etc.), we should focus on what each capacity individually entails rather than worrying about the status of "moral patient." In this case, several responses are available.

First, the status is important because it precedes the capacities. Put differently, it is precisely because we are concerned with moral patiency that we ask ourselves about capacities in the first place. The classification of "moral patient" implies that an entity is worthy of moral consideration, and retaining this status is particularly important when there is disagreement over which entities qualify for this status and why.

Second, the status provides a unified way of thinking about dissimilar entities. Humans and non-human animals, for instance, might be moral patients to different degrees or in different ways. And of course, capacities play a role when we think about

the rights of different species. Insofar as humans have the cognitive ability to see themselves as persisting through time, it might be reasonable to claim that humans have a stronger claim against premature death than non-human animals that lack this capacity. But the status of "moral patient" indicates that all these entities share something in common: a claim to having their welfare considered. The category becomes more important if we think about the status as being multiply realizable.

Third, the combination of capacities might have implications over and above the implications of individual capacities—in other words, capacities might be only jointly sufficient for certain entitlements. For instance, sentience alone might not guarantee a right to life, and autonomy alone might not guarantee a right to life. But put together, the capacities might jointly ground a right to life.

Turning to moral agency, then, the defense of the concept is similar. The status of "moral agent" indicates an entity's role in the moral community as a genuine moral actor and as an appropriate recipient of certain moral responses. The importance of this status motivates the desire to figure out which features enable an entity to qualify as a moral agent.

Additionally, upholding the category of moral agency does not deny the significance of the capacities themselves. It is likely the case that each relevant capacity will have its own set of implications for the entity's role in the moral community—including its rights and responsibilities as a moral agent. This admission, however, is consistent with asserting that there is some category, namely moral agency, that meaningfully designates an entity's position in the moral universe.

Moreover, insofar as multiple criteria are jointly necessary for moral agency, looking at each capacity individually will be insufficient for fully explaining what it means to be a moral actor—and the individual capacities might be only jointly sufficient for properties that follow from moral agency (e.g., certain forms of moral responsibility).

The second objection to the project is from a consequentialist perspective. The objection claims that the classification of an entity as a moral agent is irrelevant for moral decision-making and action, as all that matters is the mere ability to produce the

best outcomes. While the staunch consequentialist might have difficulty accepting any intrinsic value of moral agency as a status, two further arguments provide instrumental reasons to care about moral agency as defined by capacities.

First, from a safety perspective, requiring moral decision-makers to possess certain qualities will likely ensure that the best outcomes are attained—and that the worst outcomes are avoided. Suppose an oracle exists that, through no discernable pattern or method, has a strong track record of making the right decision (in consequentialist terms). Even a strict utilitarian might reasonably be reluctant to use such an instrument. To have a high credence that the oracle will make the right decision, it would need to demonstrate at least some capacities for reasoning—at the very least, it seems that the oracle would need the ability to correctly identify morally relevant features of a scenario and evaluate decision options accordingly. It is not only important *that* a system can produce good outcomes; it is important *why* a system can produce good outcomes (for us to have good reason to trust it).

Second, given that consequentialists are concerned with moral patiency, it might be the case that moral agency, or the capacities associated with it, is relevant to moral patiency. Put differently, entities that are moral patients and moral agents, relative to mere moral patients, might have an increased capacity for welfare. This could be because of the capacities themselves. We might think, for instance, that an entity that is responsive to reasons is able to experience a higher level of wellbeing. Alternatively, an increased capacity for welfare could arise from being a moral agent overall. Participating in the moral community as an actor and decision-maker, particularly one that is identified with one's peers as such, could increase an entity's welfare as well. It might also afford an entity with increased interests and rights.

An exploration of artificial moral agency on a criteria-based, functionalist approach, then, is a worthwhile project to pursue.

1.6 A Roadmap

This dissertation will proceed as follows.

Part I—*Towards a Theory of Moral Agency*—develops a theoretical account of moral agency. In Chapter 2, "Moral Agency Without Consciousness," I argue that phenomenal consciousness is not necessary for moral agency. In Chapter 3, "Two Types of Moral Agency," I argue that there is a substantive distinction between entities that are appropriate subjects of deontic evaluations and entities that are appropriate subjects of responsibility ascriptions. Ultimately, this part outlines the capacities AI systems will and will not need to obtain to be moral agents.

Part II—*Prospects of Artificial Moral Agency*—evaluates, in a technically grounded way, the extent to which AI systems are moral agents. In Chapter 4, "Artificial 'Agents' are Not Agents," I argue that existing AI systems lack the capacity for intentional action because they lack mental states. In Chapter 5, "Artificial 'Agents' are Not Moral," I argue that AI systems exhibit only a minimal level of the moral competence required for moral agency. Ultimately, this part highlights the barriers to artificial moral agency as well as areas of progress.

Part III—*Using Artificial (non-) Moral Agents*—considers how the moral agency, or lack thereof, of AI systems 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 because doing so turns events that should be moral actions into mere behaviors. In Chapter 7, "Moral Agents Unlike Us," I argue that moral agency is not all that matters—artificial non-conscious moral agents will be different from human moral agents in normatively significant ways. Ultimately, this part demonstrates how an understanding of artificial moral agency can help us answer practical questions about the appropriate use of AI.

In Chapter 8, I conclude. Overall, 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, even if future AI systems are genuine moral agents, artificial moral agents will be different from human moral agents in important ways. I end by considering general lessons for moral agency and artificial moral agency, as well as areas for future research.

Part I: Towards a Theory of Moral Agency

In Part I of this dissertation, I aim to answer the following question: What is moral agency? I develop an account of moral agency that can help us evaluate whether a wide range of entities qualify as moral agents—including AI systems. I do not aim to develop a complete account of moral agency. Rather, I aim to identify the core of a plausible theory of moral agency.

I start by preempting a common objection to artificial moral agency, namely that AI systems cannot be moral agents because they are not conscious. In Chapter 2, "Moral Agency Without Consciousness," I argue that phenomenal consciousness is not necessary for moral agency, even on a relatively demanding account of moral agency. This chapter has upshots for both moral agency and artificial moral agency: it clarifies the role of phenomenal consciousness in moral agency, and it opens the door for artificial moral agency without the need to answer questions about artificial consciousness.

I then offer an account of moral agency. In Chapter 3, "Two Types of Moral Agency," I separate deontic moral agents (entities capable of acting morally wrongly) from responsible moral agents (entities capable of bearing moral responsibility). This chapter also has upshots for both moral agency and artificial moral agency: it helps us understand marginal cases of moral agency, and it helps us think more clearly about theories of artificial moral agency and the problem of responsibility gaps.

Chapter 2: Moral Agency Without Consciousness

Abstract

Many views of moral agency include, implicitly or explicitly, a *consciousness requirement*—namely, the claim that phenomenal consciousness is a necessary condition of moral agency. In this chapter, I argue against the consciousness requirement. First, I argue that consciousness is not necessary for the possession of moral agency; that is, consciousness is not necessary for instantiating four candidate necessary conditions of moral agency: action, moral concept possession, reasons-responsiveness, and moral understanding. Second, I argue that consciousness is not necessary for the exercise of moral agency; that is, consciousness is not required for an entity to exercise its moral agency in the form of moral motivation or moral guidance. Still, consciousness plays some role in moral life by offering a pathway to moral agency, serving as a reliability mechanism, allowing for some forms of responsibility, and enabling certain relationships. Lastly, I discuss broader implications of my argument, especially on the possibility of artificial moral agency.

2.1 Introduction

Suppose that in 100 years, after many advances in AI technology, computer scientists develop a highly sophisticated robot. This robot has an impressive suite of capacities. It has mental states—beliefs, desires, and intentions—and can perform intentional actions. To support its thinking, it has a wide range of concepts, including moral concepts. It is receptive and responsive to moral reasons. It exhibits moral understanding. In all these ways, then, the robot seems very similar to humans. But suppose further that the robot is missing one capacity that separates it from us humans: the robot is not conscious. In this chapter, I will argue that such a robot is conceptually possible—and that it would qualify as a moral agent.

The notion that such a robot could exist, even in theory, might strike some as absurd. Many views of moral agency include, implicitly or explicitly, a *consciousness requirement*—namely, the claim that consciousness is a necessary condition of moral

agency.[8] The consciousness requirement has strong intuitive appeal. But in this chapter, I argue against it. The question at hand is a question about the nature of moral agency—and we need to confront the possibility that it might not just be beings like us that qualify for moral agency.

The question I am interested in is a question about *phenomenal* consciousness. Phenomenal consciousness is the subjective feel of an experience—it is first-personal in nature. A mental state is phenomenally conscious if it is like something for the experiencer to be in that state from the inside (Nagel 1974). Phenomenal consciousness is conceptually distinct from access consciousness (Block 1995), which I take as straightforwardly necessary for moral agency (Schlosser 2013; Levy 2014). Access consciousness is a third-personal concept—mental states are access conscious if their contents are available for use in other mental systems, such as memory and reasoning. The problem of figuring out what role, if any, consciousness plays in moral agency features prominently in debates about artificial moral agency, that is, debates about whether artificial entities (particularly technological entities) can be moral agents. Skeptics of artificial moral agency often point to a lack of consciousness as the reason AI systems cannot be moral agents.

Some authors explicitly highlight a lack of consciousness in their arguments. Such views include claims that intentionality requires experiencing psychological states (Friedman and Kahn 1992); that deliberation and understanding require consciousness (Himma 2009); that making moral judgments requires phenomenal quality (Purves, Jenkins, and Strawser 2015); that robots cannot have the necessary mental capacities for moral agency in virtue of their lack of phenomenal consciousness

[8] The consciousness requirement appears in various forms. On one version, the consciousness requirement is part of the epistemic condition of moral responsibility, such that being morally responsible for an action requires being consciously aware of certain features of the scenario. Before going on to argue against this view, Sher calls this the "searchlight view" and notes its popularity, appealing to its presence in a wide range of moral theories (Sher 2009). On another version, the consciousness requirement holds that agents must have "deliberative awareness" of and "conscious control" over their actions—drawing on empirical evidence, Sie rejects these criteria for individual moral actions, though she maintains that consciousness is necessary for moral agency (Sie 2009).

(Talbot, Jenkins, and Purves 2017); that algorithms are "moral zombies," lacking reasons-responsiveness and autonomy due to their lack of sentience (Véliz 2021); and that understanding moral wrongness requires experiencing moral emotions (Rodogno 2016). Other artificial moral agency skeptics seem to presuppose consciousness more implicitly (Stahl 2004; D. G. Johnson 2006; D. G. Johnson and Powers 2008; Peterson and Spahn 2011; Parthemore and Whitby 2013; 2014; Brey 2014; Noorman and Johnson 2014; Fossa 2018).[9]

Some existing views of moral agency already deny the necessity of consciousness (Wegner 2002; Arpaly 2003; Sher 2009; Sie 2009). Often, however, these views hold that conscious awareness of certain things (e.g., one's own reasons for actions or certain morally salient features of a situation), or that consciousness in particular cases is not necessary for the exercise of moral agency—they tend not to make the more controversial claim that consciousness is not at all necessary for moral agency.

Additionally, theories of group moral agency often hold that corporations are moral agents without holding that corporations are conscious (Silver 2005; Pettit 2007; List and Pettit 2011; Hess 2013; Björnsson and Hess 2017; List 2018). However, while corporations lack consciousness as group agents, they do *contain* consciousness in the form of their members. This fact might lead proponents of the consciousness requirement to claim that consciousness still necessarily plays some role in moral agency, namely by giving rise to group moral agency.[10] This chapter generalizes the phenomenon of non-conscious moral agency, leaving room for moral agents that contain no consciousness at all. Moreover, my argument does not rely on the claim

[9] Sebastián argues that moral agency requires first personal, or *de se* representations because such representations are necessary for awareness of one's own actions (Sebastián 2021). Sebastián remains agnostic about whether phenomenal consciousness is necessary for *de se* representations but holds that the answer to this question will determine whether AI systems can be moral agents.

[10] For an argument that corporations do not need the involvement of *de se* (first-personal) states to qualify as acting, and a more general argument that *de se* states are not necessary for action see Cappelen and Dever (2020).

that group agents are genuine agents (or any view about the kind of moral agency groups might have if they are genuine agents).

Some existing views of artificial moral agency specifically also deny the necessity of consciousness. But these views tend to offer highly revisionary accounts of moral agency by offering a new, more inclusive, set of criteria for moral agency (Floridi and Sanders 2004). My argument considers the criteria invoked by more standard—and stringent—accounts of moral agency.

As such, my argument is not the first argument that consciousness is not necessary for moral agency. However, my argument takes a novel approach in focusing on a core set of capacities relevant to moral agency and arguing that those capacities can be instantiated to the extent required for moral agency without consciousness. Moreover, while some authors deny that consciousness is necessary for the exercise of moral agency in particular cases, my view denies that consciousness is necessary to be a moral agent in general.

Establishing the conclusion that consciousness is not necessary for moral agency requires clearing a high bar—there is an array of different places in the concept of moral agency where it looks like consciousness might be required. As such, I will go through various potentially relevant locations and show that consciousness is not, in fact, required.

The claim that consciousness is necessary for moral agency can be cashed out in two ways. First, consciousness might be necessary for the *possession* of moral agency; that is, consciousness might be necessary for the instantiation of other capacities that are necessary for moral agency. Section 2.2 argues that consciousness is not necessary for four candidate necessary conditions for moral agency: action, possession of moral concepts, responsiveness to moral reasons, and moral understanding. (In Chapter 3, I provide further justification for the selection of these capacities.)

Second, consciousness might be necessary for the *exercise* of moral agency; that is, consciousness might play a role in moral agency independent of its contribution to other necessary capacities, such that consciousness allows moral agents that possess the relevant capacities to act as moral agents. Section 2.3 argues that consciousness

does not necessarily play such a direct contributory role in the form of moral motivation or moral guidance.

Given that consciousness does not play any of these roles in moral agency, the question arises of what role consciousness does play in moral life more broadly. Section 2.4 posits that consciousness helps some entities obtain the necessary capacities for moral agency, serves as a mechanism for reliability, allows for some forms of responsibility, and enables certain relationships within the moral community. Section 2.5 concludes by considering implications for artificial moral agency.

2.2 Is Consciousness Necessary for the Possession of Moral Agency?

Supporters of the possession version of the consciousness requirement, who claim that consciousness is necessary for a certain capacity that is necessary for moral agency, must specify which consciousness-requiring capacity is required for moral agency. Sometimes the capacity in question is clearly defined, and sometimes the capacity is conflated or not made explicit at all.

This section highlights the four most plausible necessary conditions for moral agency and argues that they can be instantiated without consciousness. Insofar as consciousness is not necessary for any of these capacities, we have good reason to believe that the possession version of the consciousness requirement fails.

2.2.1 Action

Moral agency requires, in the first place, agency. An agent is defined by its capacity to act rather than to merely behave. The argument that consciousness is not required for action is short and simple: we are familiar with cases of nonconscious action. For instance, people often drive without being aware, let alone phenomenally conscious, of every press of the brake or turn of the wheel.

To add precision to this argument, we can look to the standard theory of action, an event causal view according to which an event is an intentional action if it has the

right kind of causal connection[11] to certain mental states (Piñeros Glasscock and Tenenbaum 2023). While there is some disagreement about which mental states are required for intentional action, contemporary views tend to highlight beliefs, desires, and intentions.

At a first glance, action merely requires mental states—not phenomenal states. Still, more explanation will help make the difference salient. To clarify, my argument does not rely on the view that *every* instance of belief, desire, and intention lacks consciousness—I just need to show that some instances of these mental states do not involve consciousness. Once that is established, we can imagine an entity that only has those instances of mental states.[12]

Beliefs are generally taken to be separate from phenomenal states. Consider a mundane belief: Kaitlyn believes that Switzerland is a country. It is implausible that this belief contains phenomenal, qualitative properties such that there is something it is like for Kaitlyn to hold that belief. In her everyday life, Kaitlyn might not even be explicitly aware that she holds that belief, even though she uses it, for instance, when she plans her vacation to Zurich and brings her passport.

There are, of course, various theories of belief (Schwitzgebel 2024). But none of the most popular accounts seem to require consciousness. Representationalism requires internal representations about propositions (Fodor 1975; 1981; Dretske 1988)—there need not be any phenomenal experiences associated with such representations. Interpretivism requires exhibiting appropriate patterns of behavior (Dennett 1980; Davidson 2001b)—and these theories do not require the entity to have phenomenal states. Functionalism requires the correct causal relationships between mental states, sensory inputs, and behavior (Putnam 1975a; Armstrong 1993)—and belief states need not be connected to any phenomenal states.

[11] Exactly what kind of connection this is does not matter for the purposes of this dissertation.

[12] See Kagan (2019, chap. 1) for another argument that agency without sentience is possible.

Desires are more complicated. The contemporary (and popular) Humean account of desire "characterizes desire by the job desire does in collaborating with belief and thereby generating action: it characterizes desire by function, not by the presence of any particular feeling" (Pettit 1998). Such accounts of desires accord with the fact that people often appeal to unconscious desires to explain their behavior, and, perhaps more commonly, the behavior of others (Smythe 1972). Only pleasure-based theories of desire explicitly link desire to phenomenal states. On these views, having a desire involves enjoying or anticipating the desire's satisfaction (Schroeder 2015). However, such theories run into a key problem. If pleasure is caused by desire satisfaction, then it must be admitted that pleasure is distinct from desire because causes are separate from their effects (Schroeder 2015).

Still, it might seem that the phenomenal feeling of wanting is part of desire. Yet, we often desire things in ways that do not involve consciousness. Some kinds of desires, namely instrumental desires, are not characterized by phenomenal states. Thea might desire a marker so she can write on the whiteboard, and this desire need not be associated with any phenomenal state. There is not something it is like for Thea to have this desire; she just has the desire.

But even if non-instrumental desires are important to action (and moral agency), these desires still do not require consciousness. Suppose we push Thea's desire to its further end: her desire to share her knowledge. This desire still lacks a phenomenal character, perhaps because it is an abstract goal. If we push Thea's desire to its ultimate end, we might end up with some phenomenal state associated with fulfillment. But it is still unclear whether this state of fulfillment requires consciousness, whether Thea will reach this state, or whether the phenomenal aspect of this state is part of what it means for Thea to desire it, as Thea does not experience the phenomenal feeling of fulfillment when she desires the marker.

Intentions similarly do not require consciousness. When Ambre intends to raise her arm and so does, the act itself might involve some phenomenal feeling (perhaps her arm feels heavy), but the intention itself does not have a phenomenal character. There is not something it is like for Ambre to intend to raise her arm—in

fact, she might not even consciously register that she is intending to raise her arm. Moreover, if intending involves having a plan, intention is more about instrumental rationality (reasoning about the means to achieve a given end) than phenomenal states (Bratman 1987).

Overall, then, consciousness is not necessary for the possession of moral agency through action.

2.2.2 Moral Concepts

Moral agency plausibly requires the possession of moral concepts. Toddlers, for instance, are agents in that they have the capacity for action, but they are not yet moral agents because they lack moral concepts. Moral agents do not need a complete picture of morality or a correct moral theory, but they do need some sense of morality and of what falls into the moral domain. Precisely which moral concepts are required for moral agency is difficult to determine, but an obvious candidate is the concept of *moral wrongness*.

I rely on an intuitive sense of concept possession: roughly, having a concept means being able to appropriately and accurately use the concept (Rodogno 2016). But it is not enough for an agent to have concepts—it must have specifically moral concepts. A moral agent must be able to "grasp or apply moral predicates" (McKenna 2012, 11). A useful way to think about what it means to have a concept of morality is to be able to distinguish between moral and conventional norms. The ability to do so involves knowing that morality is a distinct domain and that moral considerations are different from other kinds of considerations (Machery and Stich 2022).[13]

The question, then, is whether consciousness is required for moral concept possession. Some concepts straightforwardly do not require phenomenal consciousness. For instance, it is not clear how phenomenal states would be relevant

[13] We might have reasons to be skeptical about the relevance of the moral/conventional distinction to questions of moral agency (Shoemaker 2011b). I am not committed to the claim that this is the best or only test of moral concept possession.

to concepts of *subtraction* and *atom.* Included in this category of concepts are some abstract concepts relevant to morality, such as *freedom* and *equality.*

Other concepts are more closely connected to phenomenal states but are comprehensible without them. For instance, the concept of *sandpaper* might relate to the phenomenal feeling of roughness, but surely a person could have the concept without having first-personally felt or experienced sandpaper. The knowledge that sandpaper is rough in texture might be an important part of possessing the concept, but this knowledge does not require the phenomenal state of feeling one's hand on sandpaper. Some morally relevant concepts might be similar. For instance, the concept of *promise* might include the first-personal feeling of being committed to a promise, or perhaps the experience of having a promise broken. But while these phenomenal experiences might add more content to the concept of *promise*, they certainly are not necessary for possessing the concept.

The best hope for the view that moral concept possession requires consciousness is that there is a special class of concepts that are inherently phenomenal—and that moral concepts are of this kind. It is difficult to see why moral concepts would have this unique nature. Consider the concept of *pain*. The concept is highly morally relevant, and the first-personal experience of pain requires phenomenal consciousness. However, the phenomenal aspect does not exhaust the concept of *pain.* Importantly, there is a third-personal concept of *pain* (Balog 2012). When others are in pain, we do not deploy the first-personal concept—we deploy the third-personal concept. We can think about pain more abstractly in a way that does not require the first-personal concept.

It might be objected that we need the first-personal concept of *pain* to possess the third-personal concept. But it is not clear why lacking the first-personal concept would rule out possession of the third-personal concept. Consider, rather than a sensory experience like *pain*, an emotion like *grief.* A person who has never experienced loss can have the concept of *grief* despite lacking first-personal access to the concept. It would be strange to say that such a person lacks the concept altogether, though we might claim that the person lacks the concept in its entirety. It is important

here to note that nobody possesses any concept in its entirety. So, the question at hand is whether a lack of phenomenal consciousness prohibits an entity from possessing the moral concepts to the degree required for moral agency.

It might be objected that we can only fully understand the meaning of the concept *morally wrong* by experiencing moral emotions. Rodogno, in adopting a neo-sentimentalist approach to moral agency, argues that we can only use the concept *morally wrong* correctly if we "master the normative attribution of certain emotions" (Rodogno 2016, 41). Rodogno draws an analogy to the concept of *red*, which is claimed to be partly constituted by justified visual experience of seeing red. The argument appeals to the case of a blind person with a device for identifying the light frequencies of everything she touches. This person could make most of the color-related inferences that sighted people make but would not be able to grasp the meaning of certain inferences, such as the connections between colors and moods and feelings (happiness, tiredness, calmness) "because these connections work precisely through the specific phenomenology of different colors" (Rodogno 2016, 42). The same idea is supposed to hold for the concept *morally wrong*: moral emotions uniquely allow us to grasp certain aspects of morality.

This argument, however, is unconvincing. First, it sets the bar for concept possession too high. It is overly restrictive to claim that the blind person lacks the concept *red* simply because she lacks the ability to make some set of inferences about it. The argument is reminiscent of the thought experiment involving Mary the color scientist. Mary knows everything there is to know about the physical world but lives in a black-and-white room—the key question of the thought experiment is whether Mary learns something new when she sees red for the first time (Jackson 1986). While this thought experiment has sparked numerous debates about physicalism, it seems that participants in these debates take it as given that Mary has the concept of *red* even before she experiences seeing red herself. Indeed, it is implausible that Mary lacks the concept of *red* altogether simply because she has not experienced seeing red.

Second, it is not clear that a lack of consciousness precludes agents from making the relevant inferences. In the case of redness, it seems that the blind person

can still learn how different colors relate to different moods and feelings, even if experiencing those colors does not cause those phenomenal states in the blind agent herself. She can still grasp, for instance, that blue makes people feel calm. She can also learn about why blue makes people feel calm, for instance, by learning about the neural mechanisms behind this phenomenon. Moreover, she could reason about which colors might give rise to certain mental states—perhaps she can infer that redness evokes anger because red is associated with fire and blood. In the case of moral wrongness, agents can engage in moral reasoning without feeling moral emotions.

Third, the claim that an agent without moral emotions will be unable to correctly use the concept *morally wrong* should not be assumed. It is not difficult to imagine an individual that can distinguish violations of moral norms from other kinds of norm violations without phenomenal experience. The person could come to possess the concept through testimony or examples. Even large language models, which are not moral agents, can differentiate between moral and conventional norm violations through learning statistical patterns.[14]

Overall, then, consciousness is not necessary for the possession of moral agency through the possession of moral concepts.

2.2.3 Reasons-Responsiveness

On most contemporary views of moral agency, moral agency requires agents to be responsive to moral reasons. This capacity can be broken up into three components—and I will argue that consciousness is not required for any of them.

[14] Another objection from Rodogno is that we need emotions in our moral development to grasp the nature of the concept *morally wrong*. But concept acquisition often occurs subconsciously—there is not something it is like to form a concept. Concepts are generally formed by associations and classifications based on experience (not necessarily phenomenal experience, but rather examples of the concept in use). Emotions are important in human moral learning, but we should not rule out the possibility of moral concept acquisition happening in other ways—either through examples or explicit definitions.

First, responsiveness to moral reasons requires sensitivity to ethical considerations. A moral agent must be able to identify morally relevant features of a situation as morally relevant (Wallach and Allen 2009). Importantly, this capacity need not specify the way in which an entity is sensitive to moral considerations. Even humans are sensitive to ethical considerations via different input media. We can pick up on morally relevant features of a situation through sensory input—we can see or hear a person in pain, for instance. But we can also deliberate abstractly about scenarios and pick out the features that pertain to morality. Additionally, we can receive morally relevant information from other sources—for example, from someone telling us that another person is in pain.

Sensitivity to ethical considerations does not require consciousness. While humans often rely on their sentience as a mechanism for obtaining morally relevant information, it is not the only way to do so. We might imagine, for instance, invoking definitions of morally relevant features and then applying them situationally without having any associated phenomenal states. For instance, we can be attuned to descriptions of rights violations without having any phenomenal experiences regarding those rights violations. Similarly, we might know that pain is morally relevant and that there are certain neural correlates of pain, and thus we can identify an instance of pain as occurring (and as being morally relevant) without having any associated phenomenal states.

Second, responsiveness to moral reasons requires recognizing moral reasons *qua* reasons. Scanlon offers a widely accepted general definition of a normative reason as "a consideration that counts in favor" of some action (Scanlon 2000, 17). A *moral* reason would be a specific kind of normative reason—a consideration that counts morally in favor of some action.

The capacity to recognize moral reasons is not merely the ability to note that content is morally salient—an agent must also be aware that the reason holds weight in terms of moral evaluation. An example can help further distinguish these capacities. Suppose Aleks must choose which path to take: taking path A would get Aleks home quickly but injure a bystander, while taking path B would take longer but result in no

injuries. If Aleks is sensitive to moral considerations, he will identify the injured person as morally salient. If Aleks recognizes moral reasons, he will deem this morally relevant feature as a consideration against taking the shorter path. Additionally, we can imagine a case in which an agent is responsive to moral reasons yet is unable to identify morally salient features of a situation. A person might know that the infliction of pain constitutes a moral reason not to perform an act, but he might have difficulties identifying instances of pain (e.g., he might struggle to interpret facial expressions).

Recognizing moral reasons *qua* reasons can also be done without phenomenal consciousness. The argument for this claim is an extension of the argument that identifying morally salient features does not require phenomenal consciousness. Of course, recognizing something as a reason might be more complex than merely identifying a feature as morally relevant. However, so long as an agent can take up a piece of information as a reason—in the sense that the information features as a consideration in favor of certain potential actions—she will be able to recognize moral reasons as reasons.

When it comes to determining whether an agent can recognize moral reasons *qua* reasons, the proof will be in the agent's reasoning process. For agents with communication abilities, we might engage in a conversation to determine whether the moral reason was taken as a reason. Alternatively, we might be able to examine the reasoning process in other ways such as looking at neural patterns. Sometimes, recognizing moral reasons will result in a change in action; but this will not always happen—the reason in question might be outweighed by other reasons, the agent might recognize but not care about moral reasons, or the agent might be weak-willed.

Third, responsiveness to moral reasons requires an agent to have regulative control over its decision-making process. A moral agent must be able not only to take in the relevant information and recognize moral reasons *qua* reasons, but also to change their decisions and actions accordingly. Regulative control also means that an agent would act differently in counterfactual situations if different reasons had been salient. Consider a moral agent who must decide how to divide money between two individuals. Her decision would change if different morally relevant reasons had been

salient—for instance, facts about what the individuals would use the money for, or facts about whether one person had stolen money from the other. Responsiveness to moral reasons involves adaptability. A moral agent must be able to evaluate and weigh competing reasons—and allow such reasons to guide their actions.

Regulative control over one's decision-making process does not require phenomenal consciousness. An immediate objection can be raised, namely that reasons-responsiveness requires agents to not merely recognize and react to reasons, but to "feel the pull" of the moral reasons that motivate action (Véliz 2021, 495). In the human case, this description coincides with how we make some moral decisions. We do not merely objectively weigh different moral considerations—we engage with them at a phenomenal level. We can be swayed by morally relevant information, and we feel that the decision we are making is the right one.

However, we must be careful to avoid conflating the common case with necessity. When humans reason, there is often a phenomenal experience involved in being moved by reasons, a feeling that guides our moral behavior. But this mechanism need not exist to make a moral decision. Humans also seem to make many decisions without engaging in this emotive process. We can adopt a more distanced perspective and follow our reasoning process even when we do not feel the force of reasons (or when we feel that the reasons are pulling us equally in different directions). If we were to find out that another human adult made a series of dynamic (seemingly reasons-responsive) moral decisions, but that no phenomenal states influenced her decision-making process, we would not thereby deem her unresponsive to moral reasons. There might be something it is like to make a moral decision, but this first-personal feeling is not causally necessary for reasons-responsiveness.

Additionally, the moral feelings that guide our moral reasoning can lead us astray. Sometimes we must make a moral decision despite the phenomenal weight of the reason pulling us in another direction. We can feel like we are making the right decision when we are not, and we can make the right decision even when we feel like we are not. It is not clear, then, that the act of identifying with one's reasons in a deeper

sense (or internalizing one's reasons in a phenomenal way) is necessary for moral agency.

Overall, then, consciousness is not necessary for the possession of moral agency through responsiveness to moral reasons.

2.2.4 Moral Understanding

A closely related capacity to responsiveness to moral reasons is moral understanding. There is a difference between merely implementing moral reasons (being responsive to moral reasons) and understanding why and how those reasons are being used (moral understanding). The latter involves a deeper sense of morality and the connections between various reasons and possible actions. We can imagine a case in which an individual can recognize moral reasons, assign them a weight, and act accordingly—perhaps they follow a moral rulebook—but exhibits no understanding of this information.[15]

This phenomenon of reasons-responsiveness without understanding might be familiar outside a moral setting. We often recognize and adopt reasons we do not fully understand. We follow bureaucratic rules despite being unable to fathom why such rules are in place. We follow instructions for building furniture and setting up electronic systems despite not understanding why the pieces fit together in that way or why the steps must be performed in that order. The mere existence of these rules constitutes a reason to follow them despite not understanding them. To draw a Kantian analogy, responsiveness to moral reasons might ensure that we act in accordance with morality, but moral understanding is required to ensure that we act from morality.

Moral understanding is a complex concept, and it is important to clarify what a moral agent must understand. Intuitively, it might seem that a strong notion of

[15] We might think of a moral version of the Chinese Room—the man inside the room would be responsive to moral reasons but would not have moral understanding.

understanding is relevant to moral agency—moral agents must understand why certain actions are wrong. Wallach and Vallor define moral agents as understanding "in a holistic, integrated, and richly embodied sense, the fabric of moral life" (Wallach and Vallor 2020, 397). This definition sets a very high bar for moral understanding. It is not clear that most humans have this deep sense of understanding moral matters. We are often driven by moral intuitions, and even philosophers struggle to conceptualize fully coherent ethical views. We often find moral judgments conflicting and confusing rather than something we fully understand or even understand well. Wallach and Vallor do, however, hit upon a key feature of moral understanding: the role of connectivity between and amongst moral reasons and actions.

Hills' account of understanding can help make the relevant connectivity precise. On Hills' view, moral understanding requires more than knowing that some action, for instance, is morally wrong. It also requires more than merely knowing why that action is morally wrong—that is, knowing the reasons. "Moral understanding involves a grasp of the relation between a moral proposition and the reasons why it is true" (Hills 2009, 101). Accordingly, moral understanding has a systematic component—and an agent can demonstrate moral understanding by making the relevant inferences about the relationships between moral propositions and reasons. This view allows us to operationalize moral understanding. To demonstrate moral understanding, a moral agent must be able to use moral concepts in various ways—particularly in novel situations. Moreover, this definition highlights the communicative element of moral understanding—a moral agent must be able to explain their moral decisions.

At first glance, moral understanding might seem to require consciousness. After all, understanding seems to involve some deep internalization that perhaps reason-responsiveness does not. Part of the intuition that there is a relationship between understanding and consciousness is that understanding seems to involve a phenomenal sense of grasping the information. However, the feeling of understanding is an unreliable indicator of understanding. Often, we feel like we understand things we actually do not understand (or another person tells us they understand something,

but we can tell that they do not); sometimes we feel that we do not understand something when we really do (perhaps because we lack confidence).

Moreover, when we start to think about how we recognize when another person exhibits moral understanding, it becomes clear that moral understanding is something we must infer rather than something we can directly observe. For example, if we want to know whether a student understands a topic, our approach will involve asking this student questions about the topic. We want to know whether the student knows enough to not only regurgitate the relevant information, but also to utilize it in a wider range of discussions and applications. In the realm of moral understanding, then, we would need to determine whether an agent could explain and justify their moral decisions in a way that is roughly consistent. This process does not require attributing or inferring any phenomenal states.

It might be objected that moral understanding is not purely cognitivist, as the above description seems to characterize it. More specifically, moral understanding might require the ability to empathize. When considering the role of empathy in moral agency, authors tend to appeal to two examples: psychopaths and high-functioning autistic individuals. Kennett claims that empathy is not necessary for moral agency because autistic individuals lack empathy but can engage in moral deliberation and judgment (Kennett 2002).

Aaltola distinguishes cognitive empathy, the ability to represent another person's mental state, from affective empathy, the ability to resonate with the phenomenal aspects of another person's mental state (Aaltola 2014). Psychopaths have high cognitive empathy and low affective empathy, while autistic individuals have high affective empathy and low cognitive empathy (A. Smith 2006). Aaltola takes this as evidence that affective empathy, rather than cognitive empathy, is necessary for moral agency. Affective empathy, of course, involves phenomenal consciousness because it requires feeling the same emotional states as those one is empathizing with.

We have reason, however, to be skeptical about the role of affective empathy in moral understanding. While affective empathy is important for moral understanding in the human case, it is not clear that it is necessary for moral

understanding in general. Aaltola argues for a form of minimal moral agency so as not to exclude autistic individuals (and others with affective empathy but limited cognitive abilities) from being moral agents.

But we might imagine that there are cases in which individuals have low affective empathy yet still exhibit moral agency. Psychopaths might have trouble acting morally, but this does not rule out the possibility that they are moral agents. They might understand moral concepts and apply them but choose not to act in accordance with them (or lack the motivation to do so or lack responsiveness to moral reasons) (Borg and Sinnott-Armstrong 2013). Psychopaths might just be bad moral agents, or moral agents with severe challenges to acting in accordance with morality. Alternatively, even if psychopaths are not moral agents, it is not clear that their lack of moral agency is due to their lack of affective empathy—psychopaths also exhibit other deficits in rational self-governance (Litton 2008).

Again, I am not denying that affective empathy plays an important role in human moral development. Rather, I am arguing that affective empathy is not the only pathway to moral agency. Affective empathy might incline humans to act morally or help them do so, but it is not clear that the capacity for empathy allows for moral understanding over and above what can be understood cognitively. Cognitive empathy can provide an entity with all the resources required to be a moral agent.

Still, there is an intuition that seems difficult to let go of: the idea that one cannot understand the moral significance of an action without knowing what it is like, at least to some degree, to have phenomenal experience. The idea is that we need some base level of sentience to truly understand the effects of our actions. When we probe this intuition further, however, the link between phenomenal experience and understanding moral significance becomes tenuous for two reasons, both arising from the fact that moral understanding often requires us to extrapolate far beyond our own experiences. Clearly, moral understanding cannot require us to have gone through the exact same experience as another person—this is impossible, as experience can be individuated in such a fine-grained way that it does not make sense to say we must experience something to understand it (otherwise we would understand very little).

First, it seems wrong to infer that we can understand the challenges others have faced from our own mundane examples. For instance, Ambre may have experienced sadness in her life, but this does not imply that she is able to empathize in such a way as to understand the experience of a person with depression. In fact, the depressed person seems to have grounds for criticizing Ambre for implying that she can understand the moral significance of depression purely based on her own experience of sadness. It is not the first-personal extrapolation that is doing the work in understanding what a person with depression is going through.

Second, it seems wrong to claim that we cannot understand the gravity of our actions without having phenomenal experience. For example, a person may have never experienced being a refugee yet can still understand that refugee status is morally significant and ought to be considered in moral decision-making. It is not clear that a moral agent needs phenomenal consciousness at all to grasp the moral significance of phenomenal states. Otherwise, a lack of imagination might rule out understanding. A man might not first-personally understand what it is like to be a woman in the workplace, but he can still third-personally understand the moral significance of this experience. If we deny him this potential for understanding, we too easily let him off the hook for failing to understand the moral significance of his actions.

These observations might be critiqued along the lines of feminist standpoint theory—the view that members of marginalized groups have an epistemic advantage regarding the oppression of their group (Dror 2023). Insofar as it is true that members of marginalized groups have such an epistemic advantage, we can ask *why* this is the case. Dror argues that the oppressed tend to have a contingent epistemic advantage but not an in principle one (Dror 2023). The epistemic advantage is caused by the fact that marginalized people tend to have more relevant experiences and motivation regarding knowing things about how marginalization operates. But the lack of firsthand experience of being oppressed need not be a barrier to understanding how social oppression works.

Additionally, Dror argues, while emotions can offer some epistemic advantages (e.g., socially marginalized people can make claims about whether certain things are hurtful to their group and about the normative status of these things), the epistemic advantage is limited: "even if a non-oppressed person will not know *exactly* what the oppressed person's pain feels like, what really matters…is *that (and perhaps how much) someone was hurt*, rather than what *exactly the hurt feels like*" (Dror 2023, 633). Broadening this idea, then, a moral agent does not need firsthand phenomenal experience to gain moral understanding.

Of course, having a firsthand experience often increases a person's understanding of a situation. But this fact does not imply that the person lacked understanding before she had undergone the experience, nor does it imply that attaining understanding is impossible without the firsthand experience. For instance, Cillian might develop a deeper and fuller understanding of disloyalty when he is betrayed by a close friend. But this admission does not mean that Cillian had no understanding of disloyalty before first-personally experiencing it. If Cillian had never experienced disloyalty, he would still have sufficient understanding of the phenomenon to engage in moral reasoning about it—and, for instance, to choose to refrain from being disloyal himself.

Overall, then, consciousness is not necessary for the possession of moral agency through moral understanding.

2.2.5 Taking Stock of the Possession Version

The reasons to think moral agency requires consciousness are varied. There are many potential places to locate consciousness in moral agency. This section has argued that the possession version of the claim that moral agency requires consciousness is implausible. The four most important capacities necessary for moral agency—action, possession of moral concepts, responsiveness to moral reasons, and moral understanding—do not require consciousness. Consciousness, then, is not located in

any of these necessary capacities for moral agency. The possession version of the consciousness requirement for moral agency fails.

2.3 Is Consciousness Necessary for the Exercise of Moral Agency?

Rather than locating the role of consciousness in other necessary capacities for moral agency, it might be the case that consciousness enables an entity to exercise its moral agency. This section argues that consciousness is not necessary for the exercise of moral agency in the form of moral motivation or moral guidance. Importantly, my argument does not claim that phenomenal consciousness is not a useful aspect of moral agency in humans—it just claims that it is not a necessary aspect of moral agency in general.

2.3.1 Consciousness as Motivation

Phenomenal consciousness plays a strong motivational role for humans. We have desires that are associated with positive phenomenal states, and we are thus motivated to act to achieve those states. For instance, we often feel good when we help others, and this anticipated feeling can motivate us to do so. Conversely, some actions and states of affairs cause us to have negative phenomenal states, and we are thus motivated to act to avoid those states. For example, we often feel bad when we see other people in pain—and we feel guilty when we refrain from intervening. On a higher level, the desire to act in a morally good way might also be associated with phenomenal states. It might feel fulfilling to view oneself as morally virtuous. This feeling can motivate us to put significant weight on moral reasons in our decision-making.

But necessity of consciousness in the form of moral motivation conflicts with the widespread denial of psychological egoism. On the hedonistic version of psychological egoism, all actions are done to maximize one's own pleasure. On these views, phenomenal states are the only—or at least the main—motivator of moral actions. But most philosophers reject such views (Feinberg 2007). This rejection

acknowledges that some moral actions can be done because they are the right thing to do, or for the sake of other people, regardless of the effect on the agent's phenomenal states.

Kantians hold that following the moral law should be independent of any desires or phenomenal states—rationality leads us to adopt the categorical imperative. Additionally, people often act morally as a matter or habit or intuition, without needing to be explicitly motivated by something like a prospective phenomenal experience. Once we see that at least some moral decisions need not be motivated by phenomenal states, we must accept that it is possible for moral motivation to remain intact without phenomenal states.

None of this is to deny that consciousness is often a strong motivational tool. It makes acting in a morally good way easier in many cases, and it is likely no evolutionary surprise that humans have developed phenomenal states in line with pro-social behavior. A lack of consciousness might make morality more difficult, especially for humans. The descriptive claim that human moral agency requires emotion might be true. But it remains possible to be a moral agent without consciousness. What would be needed, of course, is some other capacity or factor to provide the motivation to act morally—some form of goal-directness, for instance.

2.3.2 Consciousness as Moral Guidance

Phenomenal consciousness often guides the accuracy and efficiency of the moral decision-making process in humans. Insofar as developing moral intuitions is linked with emotional responses, humans have a mechanism to guide our actions. Our ability to empathize makes us good moral agents because it provides a way for us to consider and engage with the morally relevant features of scenarios that involve people beyond ourselves. If we lacked phenomenal states, we might have a difficult time identifying morally laden situations and acting quickly enough to make a difference. Moreover, our conscience (and the feelings associated with it) seems to guide us towards morally right actions.

But even in the human case, emotions can lead to suboptimal moral decisions. Arkin claims that unmanned weapons systems, for instance, might behave more ethically than humans because they are not susceptible to emotions such as fear and frustration that impede appropriate decision-making (Arkin 2010). Even in more mundane cases, self-interested feelings make it difficult to do the right thing when we must weigh our interests against the interests of others. Moreover, even if emotions are, overall, accuracy-guiding, they are not the only accuracy-guiding mechanism. Reason, for instance, also guides us towards accuracy in moral decision-making, as does developing heuristics based on previous experience.

Moreover, intuitions can be inductive in nature, and it is not clear what phenomenal consciousness adds aside from making this inductive process more salient to the experience of the decider. Intuition is a form of inference, and while the associated intuitive feelings might help guide us in the moral world, we can still have intuitions that lack the associated phenomenal states. The feelings that guide moral decisions are not necessary for adequate responsiveness to moral reasons.

2.3.3 Taking Stock of the Exercise Version

This section has argued that the exercise version of the claim that moral agency requires consciousness is implausible. Consciousness is not necessary for the exercise of moral agency in the form of moral motivation, nor is it necessary for moral agency in the form of moral guidance. Overall, then, we have good reason to believe that consciousness need not play a direct contributory role in the exercise of moral agency. The exercise version of the consciousness requirement fails.

2.4 The Role of Consciousness in Moral Life

The surprising result of my argument is that consciousness is not necessary for the possession or the exercise of moral agency. This finding is difficult to reconcile with the sense that, surely, phenomenal consciousness has some relevance to moral life. If moral agency is possible without consciousness, it is important to ask what

phenomenal consciousness adds. While a full exploration of what consciousness adds to morality is beyond the scope of this chapter, this section identifies four ways in which consciousness remains important for moral life.

First, as demonstrated in previous sections, consciousness provides a useful mechanism for humans to achieve moral agency. The capacity for first-personal experience makes instantiating most, if not all, the relevant capacities easier. Consciousness allows us to obtain morally relevant information and be attuned to how our decisions might affect other people. Consciousness also motivates us to act morally because we care about others. It offers one, though not necessarily the only, pathway to moral agency.

Second, and relatedly, consciousness serves as a mechanism for reliability. The fact that humans feel bad when they do something morally wrong enables trustworthiness. We might be more likely to trust a human who feels the pang of guilt when thinking about violating a moral norm than an algorithm that, despite being accurate, does not have any phenomenal stake in the ensuing decision. Still, consciousness is not always the best indicator of trustworthiness. Consciousness might undermine reliability in some cases—we might be less likely to trust others when we know that immoral actions have strong phenomenal appeal, and others can manipulate us into a false sense of trust by exploiting our phenomenal states. Moreover, there are other potential mechanisms for reliability. Repeated instances of praiseworthy action, or transparency, for instance, might serve as another way to increase trustworthiness.

Third, consciousness allows for certain responsibility practices. Some theories of responsibility emphasize the role of reactive attitudes (P. F. Strawson 2008). On such views, the ability to feel anger, guilt, and shame might be necessary for moral responsibility.[16] Moreover, retributive views of punishment seem to require consciousness. On these views, we want some moral agents to be the appropriate

[16] Though, interestingly, it might be the case that phenomenal consciousness is not required for reactive attitudes (Björnsson and Hess 2017).

targets of blame and praise—and so we need to know that they can feel pleasure and pain. More specifically, for retributive theories of punishment to work, the agent must be able to feel the badness of punishment (though having desires might be sufficient for a retributive form of punishment to work).

Fourth, consciousness enables certain social and morally significant relationships. Authentic friendships and romantic relationships, for instance, might require mutual feelings (Turkle 2011; Nyholm 2020). Insofar as these are morally relevant relationships, moral agents that lack consciousness will not be able to participate in them. As such, non-conscious moral agents might not be members of the moral community in the same way as conscious moral agents. I will return to these final two considerations in Chapter 7.

2.5 Conclusion

This chapter has argued that the consciousness requirement of moral agency, according to which consciousness is necessary for moral agency, is implausible. Phenomenal consciousness is not required for the key capacities associated with moral agency (the possession version of the consciousness requirement), and it is also not necessary for the exercise of moral agency (the exercise version of the consciousness requirement).

From this argument, many existing attributions of moral agency remain the same. Cognitively normal adult humans still qualify as moral agents; young children and animals still do not qualify as moral agents, nor do ATMs or chatbots. Corporations may or may not qualify for moral agency on my view—but if they fail to qualify, it will not be because they lack consciousness.

The commonly invoked marginal cases of moral agency (e.g., human children and psychopaths) have consciousness, so getting a picture of what non-conscious moral agency might look like is difficult. One implication of my argument is that p-zombies (Chalmers 1996) are moral agents—though being a p-zombie is sufficient but not necessary for moral agency, as any non-conscious entity that can instantiate the

necessary conditions of moral agency will be a moral agent (not just those that are functionally identical to humans).

The most significant implications of my argument lie in discussions of artificial moral agency. In particular, my argument opens the door for the possibility of artificial non-conscious moral agents. In some ways, the prospect of AI-based moral agents is, thus, improved—after all, moral agency can be instantiated without having to pin down the concept of consciousness or identify when an entity has attained consciousness (Butlin et al. 2023). However, there is still a long road ahead in the development of genuine artificial moral agents. The capacities relevant to moral agency will be difficult to integrate into AI systems, especially without consciousness playing the role it plays in human morality.

My argument can also be contextualized in the existing artificial moral agency literature. Most obviously, views against the possibility of artificial moral agency that rely on consciousness—implicitly or explicitly—are untenable. These views will need to reassess their reasons for believing that artificial systems cannot be genuine moral agents.

But some existing views that deny the necessity of consciousness for artificial moral agency are not vindicated by my argument. Views that do not include the necessary capacities for moral agency considered in this chapter must justify their revisionary and expansive definitions of moral agency. Still other views remain largely untouched, for moral agency was never the issue all along. For instance, views that focus on retribution or relationships simply need to clarify that they are not talking about moral agency per se, but rather another aspect of morality for which consciousness is important.

Supposing the development of artificial moral agents without consciousness is technologically (rather than merely conceptually) possible, key normative questions will arise regarding the role of such agents in the moral community. On the one hand, there will be questions about the potential rights and moral patiency of these agents. Traditionally, moral agents are thought to be a subset of moral patients. But my argument might challenge this conception, insofar as consciousness is necessary for

moral patiency (in which case we could have non-conscious moral agents that are not moral patients).[17]

On the other hand, questions will arise about the contexts in which it is appropriate to deploy non-conscious moral agents. It is important to pinpoint the role of consciousness, if there is one, in the particular decision at hand. For instance, if it is claimed that we should not have robot judges make sentencing decisions, the reason cannot simply be that robot judges cannot be moral agents—the reason would have to appeal specifically to why a non-conscious moral agent (but still a moral agent) is insufficient or inappropriate for making such a decision. It might be the case that some moral decisions ought to be made by conscious moral agents. But it must be argued, rather than assumed, that consciousness is required for decision-making in those cases.

Moreover, non-conscious moral agents will be unlike human moral agents in potentially normatively significant ways. They might have all the core capacities essential to moral agency, but they will be very different from human moral agents (and very different from other non-paradigmatic cases of moral agency, such as children). For instance, such agents will have no experience of suffering, no emotional contagion or affective empathy, no anger or pain at injustice, no pleasure in doing the right action, and a very different basis for moral judgment and for learning moral concepts.[18] I will explore some of these differences more in Chapter 7.

[17] Some views of moral patiency do not require consciousness (Kagan 2019; Sinnott-Armstrong and Conitzer 2021). On such views, non-conscious artificial moral agents will have some degree of moral status and some rights.

[18] Thank you to Carissa Véliz, Alison Hills, and Milo Phillips-Brown for extensive feedback on this chapter. Thank you to Kyle van Oosterum, Seth Lazar, Max Kiener, Roger Crisp, and audiences at the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES 2022) and the ANU Machine Intelligence and Normative Theory (MINT) Lab (2023) for additional comments and discussion.

Chapter 3: Two Types of Moral Agency

Abstract

Suppose Dottie knocks Shom over. Our moral evaluation of this event, and our moral judgment of Dottie, depend on what kind of entity Dottie is—namely, whether Dottie is a moral agent. While classifications of moral agency and moral responsibility often co-occur, the precise relationship between the two concepts is not always clear—especially in discussions of artificial moral agency. In this chapter, I develop an account of moral agency according to which there are two types, or levels, of moral agency: (1) a deontic moral agent is capable of morally wrong action and (2) a responsible moral agent is morally accountable for their actions. After defending this distinction and discussing the capacities underlying each type of moral agency, I use my account of moral agency to shed light on discussions of artificial moral agency and the problem of responsibility gaps.

3.1 Introduction

Suppose Dottie knocks Shom over. If we want to morally evaluate this event—and determine whether Dottie is morally responsible for knocking Shom over—we need more information about Dottie. For starters, we need to know what kind of entity Dottie is. If Dottie is a tornado, our moral evaluation will be much different than if Dottie is a human. And even if Dottie is a human, our moral evaluation will depend on certain features about Dottie, such as her cognitive capacities. Our understanding—and moral evaluation—of Dottie knocking over Shom turns on whether Dottie is a *moral agent.*

If Dottie is a cognitively normal adult human, she is a paradigmatic moral agent. As such, absent any excusing conditions, she is morally responsible for knocking Shom over. Typically, classifications of moral agency and moral responsibility co-occur. It often seems that when we say, "Dottie is a moral agent," we mean, "Dottie is a candidate for moral responsibility." But the precise relationship between the concepts of moral agency and moral responsibility is not always clear in the philosophical literature.

Additionally, while some philosophers posit different kinds of moral responsibility, it is less common to see appeals to different kinds of moral agents. For example, Watson distinguishes between what he calls two "faces" of responsibility: attributability and accountability (Watson 1996). Shoemaker offers a tripartite theory of responsibility that distinguishes between attributability, answerability, and accountability (Shoemaker 2011a; 2015). On the one hand, these theories do distinguish between different types of moral agents in virtue of distinguishing the different ways in which agents can be responsible—and some agents can only be responsible in some of the senses. On the other hand, these theories still view moral agency and moral responsibility as intricately linked—the different types of moral agents are different in virtue of the kind of moral responsibility they qualify for.

While it could be the case that being a moral agent just is being a morally responsible agent, there are possible alternative accounts on which moral agency and moral responsibility can be teased apart. The capacity for moral responsibility could be necessary but insufficient for moral agency, such that some entities are morally responsible but are not moral agents. Or moral agency could be necessary but insufficient for moral responsibility, such that only (but not all) moral agents are candidates for moral responsibility.

The relationship between moral agency and moral responsibility has become increasingly important in the domain of artificial moral agency. Questions about whether and how we can (and should) build "moral machines" are intricately linked with questions about what is required for moral agency and who is responsible for harms caused by AI systems. AI developers often aspire for their systems to operate autonomously in a wide range of contexts. But many of these contexts require more than mere technical capacities—they require moral capacities.

Typically, morally laden decisions are made by moral agents. So, the thought of machine ethics goes, we should try to equip AI systems with moral capacities—we should try to develop artificial moral agents (Anderson and Leigh Anderson 2007; Misselhorn 2018; Formosa and Ryan 2021). But whether so-called artificial moral agents, as discussed in the literature, are genuine moral agents, depends on how we

understand moral agency and moral responsibility. These considerations are especially important when we accept that moral agency consists of more than the ability to make the right moral decisions; being a moral agent also involves having standing to make genuine moral decisions.

The aim of this chapter is twofold. First, I aim to offer a new account of moral agency. This account applies to all entities, specifies different types of moral agency, and explains the capacities underlying moral agency. It also directly explains the relationship between moral agency and moral responsibility. This account draws on existing philosophical literature but is unique in the way it specifies different forms of moral agency and their underlying capacities.

Second, I aim to push forward debates about artificial moral agency in a way that is connected to broader discussions of moral agency. I do this by situating discussions of artificial moral agency in the context of a philosophical account of moral agency. Moreover, I show that the question of whether machines can act wrongly is, in some ways, more important than the question of whether machines can be responsible—and understanding this distinction can help us better conceptualize the threat of responsibility gaps.

The chapter proceeds as follows. Section 3.2 proposes and defends a distinction between two types of moral agents, which I call deontic moral agents and responsible moral agents. Section 3.3 discusses the capacities an entity must possess to be a deontic moral agent, namely intentional action and moral concept possession. Section 3.4 discusses the capacities an entity must possess to be a responsible moral agent, namely moral reasons-responsiveness and moral understanding. Section 3.5 discusses implications for artificial moral agency and responsibility gaps. Specifically, the proposed taxonomy helps us understand how discussions of artificial moral agency connect to the moral agency of humans, and the distinction between deontic moral agency and responsible moral agency helps us better understand the problem of genuine responsibility gaps.

3.2 Acting Wrongly and Acting Responsibly

In this section, I argue that there are two are related yet distinct types of moral agency. After separating out the types of moral agency (section 3.2.1), I address objections to the distinction (section 3.2.2).

3.2.1 Two Types of Moral Agents

I propose a distinction between two similar, yet separable, types of moral agency. These types of moral agency can also be viewed as levels of moral agency: the first type (deontic moral agency) is necessary but insufficient for the second type (responsible moral agency).

First, a *deontic moral agent* is an appropriate subject of deontic ascriptions. Separating out this type of moral agency picks up on the idea that when a moral agent inflicts harm on a moral patient, the moral agent has (at least in some cases) acted *wrongly,* whereas a non-moral agent has merely acted or behaved badly or undesirably. Deontic moral agents, then, are morally evaluable in that their actions can be appropriately assessed with deontic terms. A deontic moral agent can thus be considered a genuine source of moral action because it can perform actions that are deontically evaluable.

Second, a *responsible moral agent* is an appropriate subject of responsibility ascriptions. Separating out this type of moral agency picks up on the idea that moral agents are (at least in some cases) morally responsible for their wrongful actions. This form of moral agency is a complex, fully-fledged sense of moral agency. Importantly, the conception of moral responsibility at hand is responsibility in the accountability sense—a responsible moral agent can be morally culpable or blameworthy for their actions.[19] While we might still have pragmatic reasons to treat some deontic moral

[19] Other, weaker notions of responsibility, such as attributability or answerability, might apply to deontic moral agents (Watson 1996; Shoemaker 2011). My account can help us understand entities who are only eligible for these weaker types of responsibility. Moreover, while these accounts understand responsibility as reactions to the agent, my view focuses on the deontic status of the actions moral agents perform.

agents as if they were morally responsible, only responsible moral agents are genuinely morally accountable for their actions.

The distinction between deontic and responsible moral agency clarifies the relationship between moral agency and moral responsibility. Deontic moral agency is a lower level of moral agency—it is necessary for responsible moral agency because for an entity to be morally accountable for their actions, they must be capable of performing moral (deontically evaluable) actions in the first place. Responsible moral agency is a higher level of moral agency—it requires more than mere deontic moral agency. Some entities, such as cognitively normal human children, develop from deontic moral agents into morally responsible agents. Other entities are deontic moral agents that never (and cannot) become morally responsible agents.

Some examples will help demonstrate the category of "mere" deontic moral agency—that is, deontic moral agency without morally responsible agency. First, consider the example, offered by McKenna, of the character Lennie from *Of Mice and Men* (McKenna 2012).[20] In the book, Lennie is an adult with some unspecified, yet presumably significant, cognitive impairment. He inflicts harm on, and kills, animals (and, towards the end of the book, a human) though he does not explicitly intend to do so. However, Lennie knows—to some extent—that he has done something wrong when he performs such actions. As such, Lennie's actions are subject to deontic ascriptions, but Lennie lacks responsibility for his actions.

Second, psychopaths—depending on how they are described—can be deontic but not responsible moral agents. Suppose that psychopaths have a minimal conception of morality such that they know, for instance, which types of actions are morally right and wrong. But suppose further that psychopaths are unable to fully

[20] McKenna makes a similar distinction to mine in identifying morally responsible agents as a subset of moral agents in his taxonomy of moral status (McKenna 2012). My view differs from McKenna's in two ways. First, my view does not require adherence to McKenna's further taxonomy which links moral agency and moral patiency. Second, I add to McKenna's account by further defending the distinction he identifies and identifying the capacities underlying the distinction.

appreciate moral reasons to the extent that non-psychopathic adult humans do. This description is a plausible account of the moral capacities of psychopaths (Litton 2008).[21] If the account is correct, then psychopaths are mere deontic moral agents. Such a view is plausible—when the psychopath harms others, he has acted wrongly even if he is not morally responsible for his actions. It would be strange to claim that the psychopath's action is not morally evaluable (or that the psychopath did nothing wrong). But it would also be strange to attribute full responsibility to the psychopath, given that he does not have a strong understanding of morality and why he should abide by moral rules.

Third, children fit into the conceptual space of mere deontic moral agency. Our moral and legal responsibility practices reflect the idea that children are not morally responsible agents. Still, children can act in ways that are morally right or wrong, and we expect children to meet certain standards of morality (i.e., we expect them to comply with basic moral norms).

Fourth, and most controversially, mere deontic agency might be an appropriate way to describe the kind of moral agency at stake for those who hold that some non-human animals are moral agents (Clement 2013; Fitzpatrick 2017). The claim that non-human animals are morally responsible for their actions might be implausible, but the claim that they can perform deontically evaluable actions seems to warrant more consideration.

Responsible moral agency requires a richer connection to morality than deontic moral agency. Responsible moral agents are not only capable of wrongdoing—they

[21] The empirical evidence on psychopaths does not offer any robust conclusions about psychopaths' moral reasoning for several reasons: there are few studies available, few of the available studies include clinical psychopaths, findings often conflict, we do not know whether psychopaths believe the moral judgments they report, etc. (see Borg and Sinnott-Armstrong (2013) for more). Of course, if psychopaths do meet the requirements for responsible moral agency, then we should consider them responsible moral agents. But given the mixed empirical results, it is plausible to entertain the possibility that psychopaths are deontic but not responsible moral agents.

are also morally responsible for their wrongdoing. As such, prototypical adult humans are the paradigm (and perhaps only) examples of morally responsible agents.[22]

Sections 3.3 and 3.4 will go into more depth regarding the necessary capacities for the two types of moral agency. Before then, however, the proposed distinction must be defended against objections.

3.2.2 Objections

Two key objections might be levelled against my proposed types of moral agency, both to the effect that mere deontic moral agency is not a distinct type of moral agency.

The first objection is that deontic moral agency collapses into responsible moral agency: moral agency is gradable, such that so-called deontic moral agency might be better conceptualized as the instantiation of full (responsible) moral agency to some degree. On this view, children, psychopaths, and cognitively impaired adults might be better viewed as *partial* morally responsible agents rather than mere deontic moral agents. Such a view aligns with the intuition that children are partially, though not fully, responsible for their actions.

Of course, the boundaries between the types of moral agency, as well as the instantiation of the capacities relevant to each type, will admit of degrees—and there will be vagueness involved in making clear demarcations in marginal cases. But gradation and vagueness do not rule out deontic moral agency as a distinct type of moral agency.

Moreover, while moral responsibility is a gradable concept (an agent can be partially morally responsible, or morally responsible to some degree), the same is not true of deontic ascriptions. Whether an entity is an appropriate subject of deontic ascriptions is dichotomous. It might make sense to say that Aasha's action was *partially* morally right—that is, morally right to some extent (perhaps her action was morally

[22] Group agents, insofar as they are genuine agents, might also qualify as morally responsible agents, assuming they have the appropriate capacities (Pettit 2007; List and Pettit 2011; Björnsson and Hess 2017).

right in some ways but morally wrong in other ways, or perhaps her moral responsibility is mitigated based on external circumstances). However, it would not make sense to say that Aasha is *partially* the proper subject of such a moral evaluation—that is, that she is *to some degree* a source of moral action. Not every action performed by a moral agent is necessarily moral in nature; some actions can be morally neutral. But when a moral agent performs an action, it is appropriate to at least consider ascribing deontic classifications, prior to and independent of any responsibility ascriptions.

To see the categories come apart further, consider the fact that we have no problem separating agency from responsibility in individual cases. Suppose Kanav steals his coworker's dog. Upon learning that Kanav was coerced into the dognapping, it would be reasonable to refrain from attributing moral responsibility to him. However, it would not be reasonable to refrain from evaluating his action as wrong—he still committed a wrong despite not being responsible for it. In fact, in many cases of diminishing or excusing conditions for responsibility, it is still the case that a wrongful action has been committed. Given that there are cases of wrong but not responsible actions, it seems possible that there can be cases in which an agent is only capable of wrong but not responsible actions.

Let me say more about this inference, namely that the existence of individual cases of wrong but not responsible action implies that agents can be capable of only wrong but not responsible action. A potential response to this claim is that an agent must be responsible in general for their action to be wrong, but responsibility ascriptions can be removed in some circumstances. However, this description is implausible, as it involves claiming that a more sophisticated capacity (for moral responsibility) is required for instantiating a less sophisticated capacity (for moral action) even in cases that do not involve exercising the more sophisticated capacity.

A better explanation for cases of wrong but not responsible action is that most adult humans have capacities to act both as deontic moral agents and as responsible moral agents—but in some cases, they do not act as responsible moral agents (because they fail to meet the criteria for moral responsibility) and instead act purely as deontic

moral agents.[23] When we consider other cases, such as cognitively impaired adults, we can see that it is possible to lack the capacity for moral responsibility while still being a source of moral action.

The second objection is that deontic moral agency collapses into non-moral agency. In other words, deontic moral agency does not exist because it is not a form of moral agency at all. This objection can take two forms. On the first form of the objection, it might be noted that we make deontic ascriptions to non-moral agents. But when we take a closer look at these ascriptions, we can see that this objection does not hold. Sometimes, the ascriptions in question are normative but not moral. For example, when a computer malfunctions, we might say that it is doing the wrong thing—but by this we mean that the computer is performing in different way than it ought to according to its function, not that it is committing a moral wrong.

Sometimes, the ascriptions in question are seemingly moral, such as in the context of distributive justice or population ethics. But these evaluations are limited in scope, and it is helpful to distinguish between axiology and deontic classification. We might say that one state of affairs is axiologically better or worse than another in a comparative sense, or we might say that one state of the world is good or bad. But when we make a deontic evaluation, we refer to actions. For example, we might say that world A, consisting of a million happy people, is better than world B, consisting of a million unhappy people. But when we make deontic ascriptions, we make claims about actions—that it would be morally wrong to bring about world B instead of world A. Such statements are directed at the moral agents who bring these states of affairs into fruition (rather than the states of affairs themselves).

[23] This view mirrors a view offered by Sebo in a different context (Sebo 2017). Sebo argues that humans exhibit both perceptual and propositional agency (where propositional agency is more complex) but sometimes act purely as perceptual agents. Meanwhile, most non-human animals only act as perceptual agents. Based on Sebo's further arguments, my view might imply that when responsible moral agents act purely as deontic moral agents, we need to treat them as mere deontic moral agents in that context. This implication seems plausible—we often remove responsibility attributions and adjust our responsibility practices upon learning that an agent failed to meet the conditions for moral responsibility in that case, even if they are morally responsible in general.

The second form of the objection is that when we refer to the actions of supposedly deontic moral agents, we are not making genuine deontic ascriptions. Perhaps, on such a view, we use the terminology as a teaching mechanism. For example, when we call the child's action morally wrong, we are not *really* saying that the child's action is morally wrong; rather, we are merely saying that the child's action is morally wrong to guide their moral development.

But this view is implausible, especially when we think of older children and teenagers. When we speak of children's actions being morally wrong, we seem to speak of such actions as being *genuinely* morally wrong. Consider a case of Suzie, a child, punching her classmate. When we say, "Suzie's punching was morally wrong," we do not seem to be saying, "Suzie's punching would be morally wrong if Suzie were an adult, but since Suzie is a child, the punching is not technically analyzable in this way (or alternatively, that Suzie in fact did nothing wrong)." Nor does it seem that we are making a category error when ascribing deontic classifications to children.

Moreover, when we talk about Suzie in the abstract, moral education is playing no role in our deontic ascriptions. Suzie does not exist, so there is no one to be educated—and even if Suzie did exist, we might never meet her, and thus we lack the opportunity to morally educate her when we say, to each other (and not to Suzie), that her action is morally wrong.

The implausibility of the moral education view can be seen even more clearly in the cases of psychopaths and cognitively impaired adults. In these cases, we are doing something genuine when we say that the actions of these individuals are deontically evaluable. We deem their actions as wrong even when we know that doing so will not aid their moral development due to their cognitive deficits. We know that moral education will not work: deontically evaluating the actions of psychopaths and cognitively impaired adults will not help them become responsible moral agents. Yet we still make genuine deontic ascriptions to them.

The distinction between deontic moral agents and responsible moral agents thus stands. In the next two sections, I will provide an account of what each type of moral agency requires.

3.3 Deontic Moral Agents

Recall that deontic moral agents are appropriate subjects of deontic ascriptions—they are genuine sources of moral action. Deontic moral agency is the most minimal sense of moral agency that retains the key features of the concept. In this section, I argue that there are two key capacities underlying deontic moral agency: action and moral concept possession.

Before exploring these capacities in more detail, their importance can be motivated by considering the vast number of entities that are not moral agents at all. Rocks, sunflowers, hammers, and spiders all fail to qualify for moral agency. Some of these entities can be used in moral ways or can cause morally significant effects, but none are sources of moral action. There are two main reasons entities fail to be moral agents. First, some entities lack moral agency because they are not *agents* at all. Second, some entities lack moral agency because, although they are agents, their actions are not *moral* in nature. I will address these aspects of moral agency in turn.

3.3.1 Action

Entities that are not agents are not moral agents. Natural phenomena, like hurricanes, and mundane artifacts, like ceiling fans, do not act—they merely behave. It is almost true by definition that to be a moral agent, one must first be an agent. But this point should not be taken for granted. It might be tempting to take a consequentialist approach to moral agency, such that anything capable of causing morally significant effects is a moral agent. But such an approach involves biting a big bullet, namely denying the significance of the action/behavior distinction for moral agency.[24] To see

[24] Two caveats are worth noting here. First, the staunch consequentialist might still care about the action/behavior distinction insofar as agency increases the moral status of entities (e.g., by enabling them to have higher welfare). Second, the staunch consequentialist might be willing to bite this bullet and hold that moral agency *per se* does not matter—my argument is not going to convince this type of consequentialist.

why denying the action/behavior distinction would be implausible, we can look at the implications of doing so.

If "moral agent" were a shorthand for "entity that caused a morally bad outcome," then a piece of ice on the sidewalk would be considered a moral agent for causing a pedestrian to slip and hurt their back. It might be bad for the pedestrian to be injured, and we might prefer a world in which the pedestrian does not slip on the ice, but these considerations do not imply that the ice is a moral agent in any meaningful sense. We cannot expect the piece of ice to adhere to the standards of morality. The ice did not do anything *wrong* in causing the pedestrian to slip, even if the ice is causally responsible for the pedestrian's back injury. Overly simple definitions of moral agency that only focus on outcomes erroneously include non-agents—and to be a moral agent, an entity must at least be an agent.

But what type of agency is necessary for moral agency? There are various theories of agency in the philosophy of action. Some views propose minimal conceptions of agency, on which a wide range of entities including bacteria (Barandiaran, Di Paolo, and Rohde 2009) and basic reinforcement learning systems (Butlin 2021) qualify as acting rather than merely behaving. But even if these accounts successfully identify the distinction between action and behavior, these minimal kinds of agency are too limited to capture the agency required for moral agency. Even if bacteria can act in some meaningful sense, the kind of action they can undertake is very different from the kind of action humans can undertake.

The standard conception of agency, in addition to being the most popular (though not uncontested) view in the philosophy of action literature, picks up on a particularly important type of action: intentional action (Schlosser 2019; Piñeros Glasscock and Tenenbaum 2023). Specifically, the standard conception of action identifies a close connection between intentional action and acting for reasons (Schlosser 2019). Different theories of action fall under the standard conception. This chapter is not the place to attempt to adjudicate between the competing theories of action. However, there are several trends worth noting that are relevant to our discussion of moral agency.

First, the most widely accepted theory of action, known as the standard theory, is an event-causal view, according to which an event is an action if it is caused (in the right way) by mental states (Bratman 1987; Davidson 2001a). Second, even among those theories that reject the standard event-causal theory of action do not reject the role of mental states in action. Competing views still rely on mental states such as beliefs, desires, and intentions to play some role in characterizing action—even if these mental states or their causal powers alone are insufficient for action (Anscombe 1957; Korsgaard 2014; Velleman 1992).

So, if we are looking for necessary conditions for moral agency, particularly those that will be relevant to discussions of artificial moral agency, then mental states are a good starting point. It is difficult to see how a theory of moral agency could understand agency without appealing to mental states. For an action to be morally evaluable, it must flow—in at least some sense—from the agent's beliefs, desires, and intentions. It would be implausible to describe an action as morally wrong if the agent lacked mentality. As such, for the purposes of deontic moral agency, we should focus on mental states as key to the type of agency required.

3.3.2 Moral Concepts

The capacity for action alone is insufficient for moral agency. Many non-human animals, like elephants, and some humans, like young children, have the capacity for action, but it is not the case that their actions are therefore moral in nature. When a lion attacks its prey, the lion is performing an action—but this action is not a moral action, as the lion has no sense of morality at all. When we say, "killing is wrong," we do not expect lions to adhere to this moral standard, as the notion of *wrongness* means nothing to, and is beyond the comprehension of, the lion. Moreover, we do not designate the lion's killing of its prey as morally wrong, even if we feel bad for the gazelle. Just because an entity can perform actions does not mean the entity can perform moral actions. Some connection to morality is required for deontic moral agency.

With the example of the lion in hand, we can see that the most minimal view, namely that an action must bear moral significance to be a moral action, is too minimal to capture the relevant connection to morality. Being a source of action with moral consequences is not the same as being a source of moral action. The view that moral agency is simply action plus moral significance is too inclusive. Of course, we can still evaluate actions with moral significance, but this can be true of non-agents as well—and as previously demonstrated, this type of moral evaluation is not deontic and is not distinctive of moral agency.

On a slightly less minimal view, the relevant connection between action and morality might consist of adherence to moral norms and rules. This view succeeds at operationalizing a stronger connection to morality, particularly when it comes to non-linguistic beings such as non-human animals. However, this view remains too inclusive. Adhering to moral norms does not require the agent to have any genuine conception of morality. For example, on this view, a toddler who shares his toys solely so that he can play with others would qualify as a deontic moral agent. The toddler is responsive to moral norms but only contingently—he adheres to moral rules because doing so benefits him, not because doing so is morally good. For the toddler, the norm is not considered in moral terms (though it might be considered in broader normative, namely prudential, terms). If the toddler were rewarded (in having more playtime) for refusing to share his toys, he would do so, since his actions are oriented around his self-interested goals rather than the moral considerations for their own sake.

A more plausible view is that having moral concepts is the feature that substantively connects actions with morality for the purposes of moral agency. If an entity possesses moral concepts, it can not only act for reasons, but it can also act for moral reasons. On such a view, non-human animals (at least most of them) would not be considered sources of moral action because they lack notions of morality altogether, though they might have closely related social concepts (Clement 2013).[25] Toddlers

[25] Although, it should be noted that some scholars argue that some non-human animals do possess moral concepts, and as such qualify for some basic form of moral agency. If it is the

would also not be considered deontic moral agents, even though they are agents. A toddler cannot be a source of moral action if she lacks the capacity to "grasp or apply moral predicates in any way" (McKenna 2012, 11). To grasp and apply moral predicates, one needs moral concepts.

Unsurprisingly, there are also various views of concepts in the philosophy of mind. On an ontological level, the most widespread view is that concepts are mental representations, though other views consider concepts to be abilities or abstract objects (Margolis and Laurence 2023). But as we did with theories of intentional action, we can make some claims about concept possession that will hold regardless of which theory of concepts is correct.

For example, if we want to know whether an entity possesses concepts, a promising strategy is to look at how that entity uses concepts. One way to investigate concept possession is to test for systematicity. According to the generality constraint, if a conceptual agent can think, for instance, "dogs are cute" and "cats are scary," she should also be able to think "cats are cute" and "dogs are scary" (Evans 1982; Butlin 2023). Concept possession can also be explored in nonlinguistic cognitive systems by testing whether the agent can identify and reidentify objects and their properties, demonstrate stimulus independence, and show evidence of minimal semantic nets such that concepts are appropriately connected to each other (Newen and Bartels 2007).

But having concepts alone is insufficient for moral agency—a moral agent must have specifically moral concepts (Parthemore and Whitby 2014). Moral concepts are necessary for deontic moral agency because they allow agents to have some minimal knowledge of the moral domain. If we think about deontic moral agents as being expected, to some degree, to adhere to moral demands (Haksar 1998), then moral concept possession is what justifies us holding deontic moral agents to moral

---

case that non-human animals do have moral concepts, they would qualify as deontic moral agents.

standards. Crudely, it would be unfair hold agents to moral standards if they have no moral concepts to help them make sense of what is demanded of them.

3.4 Responsible Moral Agents

We have just seen that deontic moral agency requires the capacity for action and the possession of moral concepts. But more is needed for responsible moral agency. Recall that responsible moral agents are appropriate subjects of responsibility ascriptions in the accountability sense.

Traditionally, there are two conditions for an agent to be morally responsible for a given action: the control, or freedom, condition—the agent must have control over the relevant action—and the epistemic, or knowledge condition—the agent must be aware of the action, its moral significance, and its consequences. To meet these criteria, agents must have certain underlying capacities—and it is likely that instantiating the relevant knowledge and control requires many sub-capacities having to do with reasoning and deliberation. In this section, I focus on what I take to be the two most important capacities for moral responsibility (over and above those already necessary for deontic moral agency).

3.4.1 Responsiveness to Moral Reasons

There are many proposals for cashing out the control condition of moral responsibility, many of which relate to debates about free will. I will proceed under the assumption that there is such thing as moral responsibility. Moreover, I will adopt a reasons-responsiveness account of the control condition (Fischer and Ravizza 1998; Sartorio 2016). Even if reasons-responsiveness does not constitute free will, it certainly plays an important role in explaining what makes agents responsible for their actions.

To qualify as reasons-responsive, an agent must be able to take in relevant features of scenarios, see those features as reason-giving (i.e., as considerations that count in favor of particular actions), and change their behavior in light of those reasons. Another way to understand reasons-responsiveness is counterfactually: if the

reasons had been different, the agent would have acted differently (Talbert 2024). The agent need not explicitly reason in order to be reasons-responsive—an agent can be reasons-responsive without actively considering and deliberating about the reasons (Arpaly 2003).

But reasons-responsiveness in a purely rational sense is not enough for moral responsibility. To be morally responsible, an agent must be responsive to *moral* reasons. That is, the agent must be able to view moral reasons *qua* moral reasons. The claim that responsiveness to moral reasons is necessary for morally responsible agency is generally accepted in the literature. Part of being responsible is being able to shape one's behavior to what matters. The requirement of reasons-responsiveness does not imply that anyone who does the wrong thing in a particular circumstance is not reasons-responsive—rather, reasons-responsiveness regards the *capacity* to change one's behavior based on moral reasons.

Consider the case of the psychopath. The psychopath seems to qualify for deontic moral agency—he has the capacity for action (in virtue of having mental states), and he seems to have some moral concepts, at least enough to know that certain actions are morally wrong. However, the psychopath seems to lack responsiveness to moral reasons. The morally relevant features of a given situation do not get taken up in the right way—the psychopath fails to see them as reason-giving. In this case, while the psychopath is generally reasons-responsive, he is not responsive to moral reasons. As such, he does not qualify as a responsible moral agent.

3.4.2 Moral Understanding

Typically, the epistemic condition of moral responsibility is understood as a form of awareness. Often, the epistemic condition is discussed as a condition regarding particular actions. For instance, for Hunter to be blameworthy for reading Raghav's personal diary, Hunter might need to know that the book on the table is a private diary and that reading someone's diary without their permission is morally wrong. Though not often discussed in connection to moral responsibility, moral understanding seems

to underlie at least some versions of the epistemic condition—particularly the claim that an agent must be aware of an action's moral significance (Sliwa 2017).

The ability to be aware of an action's moral significance might seem to be captured by responsiveness to moral reasons. But is not enough for an agent to be receptive and responsive to the moral reason-giving features of situations. To be responsible, an agent must also have a deeper understanding of morality and how different moral reasons connect to each other. To be accountable for one's actions, an agent must know that their action is morally significant and why. Moreover, the agent must understand the relationship between actions, moral reasons, and justifications. Again, what's important is the *capacity* for moral understanding—it need not imply that an agent has full moral understanding of every action they take in each situation. Consider the case of Huckleberry Finn, prominently discussed by Arpaly (Arpaly 2003). In the example at hand, Huck must make a moral decision: he must decide whether to turn in Jim, an escaped slave. Due to his upbringing, Huck believes that the morally right decision is to turn Jim in; however, Huck's emotions are telling him to help Jim. Ultimately, Huck chooses to protect Jim despite believing that it is the morally wrong thing to do. In this case, it is difficult to evaluate Huck and his actions. From the outside, he seems to do the right thing for the right reason, though he only has access to this reason implicitly, and this reason contradicts the reasons he explicitly considers. Arpaly argues that Huck is responsive to moral reasons in this case, as he is responsive to the right-making features of his action (Arpaly 2003).

The taxonomy of moral agency can help us make sense of Arpaly's verdict while still denying that Huck is morally responsible (in this case, morally praiseworthy). While Huck is a deontic moral agent and is responsive to moral reasons, he lacks the moral understanding to truly be responsible for his actions. Given that Huck is a teenager, this analysis makes sense—he is not yet a responsible moral agent, for his moral understanding is not sufficiently developed for him to be morally responsible for what he does. Lack of moral understanding also helps explain why children and cognitively impaired adults are mere deontic moral agents rather than morally responsible agents.

3.5 Implications

The previous sections have presented an account of the two types of moral agency and their underlying capacities. Deontic moral agents are sources of moral action—to qualify, they must have the capacity for action and must possess moral concepts. Responsible moral agents are accountable for their actions—to qualify, they must have the capacity for deontic moral agency plus reasons-responsiveness and moral understanding. This section considers implications for discussions of artificial moral agency (section 3.5.1) and for the relationship between wrongness and responsibility (section 3.5.2).

3.5.1 Artificial Moral Agency

Distinguishing between non-moral agents (agents or non-agents that do not qualify for deontic moral agency), deontic moral agents, and responsible moral agents helps us better understand discussions about artificial moral agency. The proposed taxonomy helps us make sense of claims about whether AI systems can be moral agents in three ways.

First, consider what Floridi and Sanders refer to as the "responsibility objection" to their account of artificial moral agency (Floridi and Sanders 2004). The objection is that artificial agents are not moral agents because they are not morally responsible for their actions. Floridi and Sanders respond that moral evaluation does not consist entirely of responsibility ascriptions: "there is clearly more to moral evaluation than just responsibility because *x* is capable of moral action even if *x* cannot be (or is not yet) a morally responsible agent" (Floridi and Sanders 2004, 368). Oedipus, according to Floridi and Sanders, is a moral agent who lacks moral

responsibility—he is the source of an evil (killing his father and marrying his mother) but is not morally responsible for it.[26]

This response to the responsibility objection aligns with the distinction between deontic and responsible moral agency. Moreover, after we separate the two types of moral agency, we can see—as Floridi and Sanders do—that it is possible to create AI systems that are genuine moral agents but are not morally responsible. In my terminology, such entities would be mere deontic moral agents. We can, then, have fruitful investigations into the possibility of artificial moral agents without making strong commitments about the moral responsibility of such systems. So, arguments that artificial entities cannot be moral agents because they lack responsibility (Neuhäuser 2015; Sharkey 2017; Hakli and Mäkelä 2019) only apply to the possibility of artificial responsible moral agents, not artificial deontic moral agents.

Second, the conditions underlying deontic moral agency reveal that moral behavior is insufficient for moral agency. Consider Moor's proposed four types of ethical agents: (1) ethical impact agents are entities whose behaviors have ethical consequences; (2) implicit ethical agents are entities with built-in ethical considerations, such that they are designed to comply with certain ethical rules; (3) explicit ethical agents can process ethical information and determine what to do in a range of situations; and (4) full ethical agents, like humans, are those with additional metaphysical features such as intentionality, free will and consciousness (Moor 2009; 2006). But as Moor defines them, the first three types of ethical agents are not moral agents at all—they lack the capacity for action. Unless an entity has mental states, they are not even deontic moral agents—they are just increasingly sophisticated systems.[27] No amount of moral behavior alone can make an artifact a moral agent.

[26] Oedipus might not be the best example for Floridi and Sanders to use, as Oedipus plausibly has the capacity for moral responsibility.

[27] Several arguments against the possibility of artificial moral agency point to a lack of mental states as the reason artificial systems cannot be (or are not) moral agents (Himma 2009; Sullins 2006).

Third, and relatedly, machine ethics approaches (i.e., attempts to build morality into machines) that do not aim at deontic moral agency must consider how their projects relate to deontic moral agency. Wallach and Allen, for instance, develop an account of "artificial moral agency" whereby we can build machines with "functional morality" by increasing their autonomy and ethical sensitivity (Wallach and Allen 2009). Wallach and Allen are not concerned with whether such systems *really are* moral agents—they just want to build systems that can operate as if they were moral agents so that we can use them in moral contexts (Allen and Wallach 2012).

If attempts to build artificial moral agents take this pragmatic route, machine ethicists are not necessarily aiming to build genuine artificial moral agents—they are aiming to build what might be more appropriately called moral agents*: entities that behave like moral agents but lack even deontic moral agency. If this is the case, then proponents of artificial moral agency* must address how the systems they are aiming to build are related to genuine moral agents. Otherwise, classifying systems as "artificial moral agents" might mislead us into thinking that artificial moral agents* are capable of wrongdoing and potentially responsibility.

3.5.2 Wrongness and Responsibility

The distinction between deontic moral agency and responsible moral agency also has implications for how we view the relationship between wrongness and moral responsibility. First, the distinction rules out the possibility of moral responsibility without the capacity to act wrongly. This implication is plausible. After all, to be morally responsible for some wrong, it follows that an agent must have committed that wrong in the first place.[28]

[28] In some cases, an agent might have taken responsibility in advance, by agreeing to fulfill a certain role. In other cases, an agent might be responsible for the wrongdoing of others because the agent has some oversight responsibility. Even in these cases, the agent in question must have the capacity to perform moral actions (and the capacity to be morally responsible herself) to be eligible for taking on these kinds of responsibility.

Second, and more importantly, the category of mere deontic agency—a category that has been overlooked in the literature—helps us understand a frequently discussed problem in the ethics of technology: responsibility gaps (Matthias 2004; Santoni de Sio and Mecacci 2021; Danaher 2022; Tollon 2022).[29] Responsibility gaps arise when an action has occurred for which no one is morally responsible—and such gaps are deeply concerning to AI ethicists (Matthias 2004). More specifically, responsibility gaps arise when (1) a mere agent (that can act but is not morally responsible) performs an action; (2) no one else is fully morally responsible for the action; and (3) if a normal human had performed that action, this person would have been morally responsible for it (List 2021).

The concept of deontic moral agency helps explain why responsibility gaps arise. When a mere agent (that does not qualify as a source of moral action) acts, there is no genuine responsibility, as no moral action has occurred. We might ask who is morally responsible for the bad outcome, but it would not make sense to ask who is morally responsible for committing a particular wrong when, in fact, the action in question was not wrong.

This admission does not rule out the possibility of other, adjacent, morally wrong actions. For instance, a giraffe might not do anything wrong when it steals money, but the bank manager certainly did something wrong when they put the giraffe in charge of bank security. There is no responsibility gap with respect to the act of stealing the money, as stealing the money was not wrong. Indeed, we are already misdescribing the case by saying that giraffes can steal money, given that stealing is a moral concept that giraffes lack.

But in cases where a mere deontic moral agent acts wrongly, the possibility arises of moral actions for which no one is morally responsible. In such cases, a moral action has been performed, but there is no one to hold morally responsible because the

[29] There are various debates about the extent to which responsibility gaps exist—and if they do, whether they pose a new problem. For an overview in the context of autonomous weapons systems, see Oimann (2023).

agent that performed the action is not an appropriate subject of responsibility ascriptions. For example, if a child or a psychopath were put in charge of distributing welfare benefits, they would do something wrong when failing to distribute benefits fairly. However, given their lack of moral responsibility, it would not be the case that the child or the psychopath is accountable for their actions.

Again, other responsible moral agents might have some responsibility surrounding these circumstances. After all, whoever put these people in charge of distributing welfare resources has clearly acted wrongly. But for the wrong action of misallocating resources itself, no one is morally responsible.

The category of mere deontic moral agency allows us to properly conceptualize responsibility gaps. Whether these cases of genuine responsibility gaps turn out to be worrying might depend on whether there is enough additional responsibility to go around. In the case described above, it might be that the responsible moral agent in charge of making high-level staffing decisions bears a large enough portion of the downstream responsibility that we do not care about the lack of responsibility for that one specific morally wrong action.

However, in other cases, the responsibility gaps might be more troubling. If many human coders, for instance, were able to create a deontic but not responsible moral agent, and further if that agent were allowed to roam free in society, difficult responsibility gaps might subsequently arise.

In addition, we might have good reason for concern about algorithmic decision-making even if there is no genuine responsibility gap—I will discuss such cases in Chapter 6.

3.6 Conclusion

Suppose, again, that Dottie knocks Shom over. If Dottie is a deontic moral agent, we can say that her action of knocking Shom over was morally wrong. If Dottie is a responsible moral agent, we can also say that she is morally blameworthy for knocking Shom over (absent any excusing conditions, of course).

In this chapter, I have offered a new way to understand moral agency. I have argued that there are two distinct types, or levels, of moral agency. A deontic moral agent is a source of moral action, and the main necessary conditions underlying deontic moral agency are the capacity for action and the possession of moral concepts. A responsible moral agent is morally accountable for her actions, and the main necessary conditions underlying responsible moral agency are responsiveness to moral reasons and moral understanding. The resulting account of moral agency helps us better understand marginal cases of moral agency, discussions of artificial moral agency, and the possibility of genuine responsibility gaps.

The proposed taxonomy of moral agency is only the start of a full account of moral agency. More work is needed. At the theoretical level, future work should focus on additional necessary and sufficient conditions for each type of moral agency. While I have identified the core necessary conditions, I am not claiming that they are jointly sufficient. At the applied level, future work should focus on specifying which particular entities fit into the different categories—and what the normative implications of those classifications are.[30]

[30] Thank you to Carissa Véliz, Alison Hills, Milo Philips-Brown, and Jeremy Fix for extensive feedback on this chapter. Thank you to Kyle van Oosterum, Seth Lazar, and audiences at the Philosophy, AI, and Society Doctoral Colloquium (2023), the Oxford DPhil Seminar (2023), and the Open Minds XVII Conference (2024) for additional comments and discussion.

Part II: Prospects of Artificial Moral Agency

In Part I of this dissertation, I developed a preliminary theoretical account of moral agency. In Chapter 2, I argued that a seemingly necessary condition for moral agency—phenomenal consciousness—is not, in fact, necessary for moral agency. In Chapter 3, I separated two types of moral agency that answer two related but distinct questions—whether an entity is capable of acting morally wrongly (thus qualifying for deontic moral agency), and whether an entity is an appropriate bearer of moral responsibility (thus qualifying for responsible moral agency). My account of moral agency is not exhaustive, nor is it intended to be. However, the picture of moral agency I have sketched is enough to provide a basic account of moral agency that can be used to better understand the prospects of artificial moral agency.

With a theoretical account of moral agency in hand, I now shift to Part II of this dissertation. In this part, I investigate the extent to which AI systems qualify for moral agency. In some sense, my account already allows for AI systems to be moral agents, at least in principle. Moreover, my account can avoid having to first answer the difficult question of whether an AI system is conscious. But the in-principle claim is too abstract. What we are interested in is whether existing AI systems, or AI systems that can plausibly be created in the near-term based on existing technology, could be moral agents.

In this part of the dissertation, I take a technically grounded approach: rather than considering whether a computer system in the abstract could qualify for moral agency, I consider whether existing machine learning methods and empirical results provide reason to think that AI systems are moral agents. In Chapter 3, "Artificial 'Agents' are Not Agents," I argue that AI systems lack the kind of agency required for moral agency—namely, they lack the capacity for intentional action in virtue of lacking mental states. Contrary to recent claims that large language models and reinforcement learning systems might have mental states, I argue against attributing mental states to AI systems on both interpretivist and representationalist accounts.

In Chapter 4, "Artificial 'Agents' are Not Moral," I argue that even if existing AI systems were considered agents, they are far from instantiating the relevant capacities required for deontic and responsible moral agency—they lack moral concepts, responsiveness to moral reasons, and moral understanding.

Ultimately, then, the current state of technology does not offer promising prospects for genuine artificial moral agency in the near future. However, throughout this part, I also highlight several areas in which we have seen progress, and methods that might get us closer to genuine artificial moral agency.

Chapter 4: Artificial 'Agents' are Not Agents

Abstract

As AI systems become more advanced, an important question arises—namely, whether such systems are agents, in the sense of having the capacity for intentional action. In this chapter, I argue that existing AI systems (with a focus on large language models and reinforcement learning systems) are not agents because they do not have mental states. First, I consider whether AI systems have mental states on an interpretivist view, as adopting Dennett's intentional stance seems to be the most straightforward way to attribute mental states to AI systems. I argue that it is not explanatorily useful to attribute beliefs, desires, and intentions to existing AI systems. Second, I consider whether AI systems have mental states on a representationalist view, as the internal states of AI systems are representations and so might be akin to mental representations. I argue that the empirical evidence does not support attributing mental states to AI systems on a representationalist view either. Ultimately, then, existing AI systems are not agents in the sense required for moral agency.

4.1 Introduction

What is an agent? A computer scientist might offer the following definition, pulled from a popular textbook on artificial intelligence:

> An agent is just something that acts (*agent* comes from the Latin *agree,* to do). Of course, all computer programs do something, but computer agents are expected to do more: operate autonomously, perceive their environment, persist over a long time period, adapt to change, and create and pursue goals. (Russell and Norvig 2021, 21–22)[31]

[31] Similar definitions can be found in various earlier attempts to define agents in computer science. To distinguish software agents from mere programs, Franklin & Graesser propose the following definition: "An autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so acts as to effect [sic] what it sees in the future" (Franklin and Graesser 1997, 25). Davidsson & Johansson survey several definitions of agents in computer science and propose the following: "an agent could be defined as: a self-contained entity that has a state, which is situated (able to perceive and act) in an environment, rational, and at least reactively autonomous" (Davidsson and Johansson 2005, 1300).

This definition demonstrates that computer scientists and philosophers do not mean the same thing when they talk about agents. I am concerned with whether AI systems can be genuine moral agents. As such, I am concerned with whether AI systems have the kind of agency that underlies moral agency—namely, the capacity for intentional action. But computer scientists talk about agency in a very different way. For computer scientists, "agent" is a rough designation given to a particular type of software. An agent can behave in more sophisticated ways than a non-agent.[32] An agent is something that acts, where acting is merely doing things.

Researchers at DeepMind, inspired by Dennett's intentional stance view (Dennett 1987) and the idea that agents are "moved by reasons," offer a different definition, accompanied by a method for detecting when an agent is present:

> agents are systems that would adapt their policy if their actions influenced the world in a different way (Kenton et al. 2023, 3)

The researchers conclude from this account that just as humans are agents, so too are thermostats and reinforcement learning algorithms (with the caveat that they are only agents when their creation processes are included). But this definition of an agent is different from the definition a philosopher of action would give.

Because computer science and philosophy define agents in different ways, there is a risk of slipping between the language of computer science and the philosophical question we are interested in. To be clear, I am not raising an objection—rather, I am noting that what computer scientists call an agent does not seem to be the same as what we care about when evaluating whether an AI system is the kind of entity that could be a moral agent.

In this chapter, I consider whether AI systems are agents in the philosophical sense—that is, whether they *act* rather than merely *behave*, where action is understood in the sense of intentional action. In Chapter 3, I highlighted mental states as a

[32] This sentiment can also be seen in reference to "increasingly agentic systems"—though claims about the increasing agency of AI systems is sometimes viewed as a distinct form of agency from human agency (Chan et al. 2023).

necessary condition for the kind of agency underlying moral agency.[33] As such, this chapter will primarily consider whether AI systems have mental states, with a focus on beliefs, desires, and intentions.

In section 4.2, I explain two forms of machine learning. These methods form the backbone of the most serious candidates for artificial agency. I then turn to the question of whether these forms of machine learning give rise to mental states. A full answer to this question will depend on the correct theory in the philosophy of mind. I do not aim to adjudicate between different theories. Instead, I take two of the most prominent theories in the philosophy of mind and meet them on their own terms. In section 4.3, I argue that while interpretivism initially seems to grant that AI systems are agents, we have good reason against adopting the intentional stance towards existing AI systems. In section 4.4, I argue that, despite recent claims that AI systems have complex representations and world models, the available evidence does not support attributing mental states to existing AI systems on representational views, either. In section 4.5, I conclude by considering the kinds of advancements that might push AI systems closer to genuine agency.

4.2 Machine Learning Methods

Before we can make sense of whether AI systems have the capacity for action, we must understand how AI systems work. This is easier said than done, especially for philosophers. Even computer scientists do not always know exactly how their AI systems work—which, as we will see, creates problems for evaluating the capacity for action.

Below, I offer a general account of two types of machine learning that this section will focus on: reinforcement learning (section 4.1.1) and large language models

[33] Dung develops a multidimensional account of agency, whereby agency consists of goal-directness, autonomy, efficacy, planning, and intentionality—and different entities can be agential in different ways along these dimensions. This account is consistent with my view, so long as we recognize that it is the final dimension (intentionality) that matters for moral agency (Dung 2024).

(section 4.1.2). I try to keep the technical details to a minimum. For a more detailed description of these systems and for a sense of the wider range of philosophical issues they raise, see Buckner (2019); Butlin (2021); Haas (2022); Millière and Buckner (2024a; 2024b).[34]

4.1.1 Reinforcement Learning

Suppose Brent is stranded on an island. Food is his top priority, and so Brent wants to acquire as many coconuts as possible. He sees a coconut tree. Brent first walks around the tree, collecting coconuts that have already fallen. This strategy works well at first, but waiting around for coconuts to fall does not prove sufficiently reliable to ensure that Brent has enough food. Brent tries to climb the tree but falls. Brent could cut down the tree but given that he does not know how long he will be stranded on the island, he realizes that this is not a good long-term strategy for getting food. Eventually, after lots of trial and error, Brent develops a method for acquiring coconuts in the most efficient way possible over a long period of time—he uses a stick to knock three coconuts off the tree every day. Brent has engaged in reinforcement learning.

In computer science, reinforcement learning (RL) aims to design algorithms to solve the following problem: "how can an agent optimize its behavior by learning from interactions with the environment?" (Haas 2022).[35] With a few key terms in hand, we can understand—at a high level—how RL works.

The RL system is often called an agent, but I will refer to it as a *system* (think: Brent). The system is in an *environment* (think: the island). Within an environment, there are various *states* that the system can be in (think: Brent's initial state is being alone on the island with nothing). In each state, the system has various *behaviors* (often referred to as *actions*) available to it (think: Brent can collect coconuts from the ground, attempt to climb the tree, cut the tree down). Each state is associated with a certain

[34] Thank you to Aleks Petrov for patiently ensuring that I correctly understand these concepts.

[35] For more detail on reinforcement learning systems, see Sutton and Barto (2018).

amount of *reward* (think: Brent has no reward in his initial state, while the state Brent is in after chopping down the tree has a reward of 20 coconuts).

The goal of the system is to maximize *cumulative reward* (think: Brent could get a lot of reward by cutting down the coconut tree, but he will likely attain more reward in the long run if he keeps the tree intact and discovers a sustainable method for acquiring coconuts). So, the system is programmed in a way to find the optimal *policy*—that is, the best way to behave in each possible state such that the system can maximize cumulative reward.

RL is used for a wide range of applications. One prominent example is DeepMind's AlphaZero, an RL system that can beat humans at chess, shogi, and Go (Silver et al. 2018). AlphaZero gained its mastery over these games without any human knowledge aside from the rules of the game—it learned by playing hundreds of thousands of games against itself.

4.1.2 Large Language Models

One prominent type of AI system today is large language models (LLMs). Unlike RL, LLMs are not a particular machine learning technique or architecture—rather, LLMs refer to a class of models, picked out by their size and target domain rather than architecture. Most LLMs are decoder-only transformers, so I will focus on explaining this type.[36]

LLMs are trained on a massive amount of natural language data—books, websites, articles, and pretty much everything else on the internet.[37] In this initial training phase, LLMs engage in *unsupervised* learning; the model is given unlabeled

[36] The Transformer architecture was the discovery that enabled rapid development of LLMs because, through a breakthrough in self-attention, it allows the model to process the relationships between words in a sequence, whereas previous methods only considered words as individual units (Vaswani et al. 2017).

[37] For proprietary models, it is difficult to know exactly what data the models were trained on. Several lawsuits have been filed accusing the companies behind these models of copyright infringement (Allyn 2024; Hadero and Bauder 2023; Parvini and O'Brien 2024).

data with the goal of learning some underlying structure. (Compare this to *supervised* learning, in which the data is labeled, and the model learns a function to map the data to the label. Imagine lots of pictures of cats and dogs, where each picture is labeled as "cat" or "dog," and then the model must learn to identify cats and dogs.)

The LLM's learning objective is *next-token prediction*—given some sequence of text, the model predicts the most likely following word (or word segment)[38]. For instance, given the phrase "My pet named Muffin is a," the model will attach various probabilities to the next word; we would expect an accurate model to assign high probabilities to words like "dog" and "cat" and low probabilities to words like "suitcase" and "lava."

LLMs are so good at generating legible and convincing text because, over the course of billions of training steps, they alter their parameters to reduce the distance between the model's prediction and the actual next word in the training data. So, if the model initially (with random parameters) predicts that the next word is "suitcase," it will alter its internal representations of words to better match the actual next word "dog." Because LLMs are trained on such a massive corpus, they have enough examples in their training data to develop highly accurate representations of the statistical relationships between words.

Common examples of LLMs include OpenAI's GPT-4 (OpenAI et al. 2023), Google's Gemini, (Gemini Team et al. 2024), Meta's LLaMA (Touvron et al. 2023), and Anthropic's Claude (Anthropic 2023). LLMs are the base language models—they are often then used in user-facing contexts, particularly chatbots (e.g., Chat-GPT). LLMs are sometimes referred to as "foundation models" because they can be used as a base for a wide range of applications. I will treat the terms as synonymous, though "foundation model" can be viewed as a broader category, aimed at including

[38] Tokens need not (and do not) always map to words, but for our purposes, it is easier to think of tokens as words. The way in which models tokenize words can have ethical implications, but such considerations are beyond the scope of this dissertation (Petrov et al. 2023).

multimodal models and future systems that do not only use language as their inputs and outputs.

4.1.3 Some Caveats

I have presented a simplified picture of RL and LLMs for the purposes of this chapter. The lines are much more blurred and the systems more complicated in practice. For instance, a system like Chat-GPT is first trained in an unsupervised learning setting; but it is later fine-tuned on supervised data and through RL (known as RLHF, or reinforcement learning from human feedback). When necessary, I will introduce additional detail about the systems in question.

Still, having a general sense of how these systems work is necessary for understanding the extent to which AI systems are moral agents. I now turn to this question by considering whether AI systems are agents at all, through the lens of two prominent theories of mental states in the philosophy of mind.

4.3 AI and the Intentional Stance

Adopting an interpretivist view (sometimes referred to as an "interpretationist" view) seems to be the easiest way to attribute mental states to AI, as it holds that whether an entity possesses mental states is determined by how we interpret the behavior of that entity. Dennett, after all, in the view that inspired the DeepMind definition of agency, held that even thermostats have beliefs about the world (Dennett 1987; Kenton et al. 2023).

In this section, I argue that despite its initial plausibility, the interpretivist view does not give us a compelling reason to claim that existing AI systems have mental states. I first identify and explain the most plausible version of Dennett's

interpretivism (section 4.3.1).[39] I then argue that we should not attribute mental states to LLMs and RL systems on the interpretivist view (section 4.3.2).

4.3.1 Three Kinds of Interpretivism

In this section, I identify the most plausible version of Dennett's interpretivist account—one that hinges on the usefulness of mental state attributions. Specifying the relevant account will put us in a better position to assess whether AI systems have mental states.

Dennett's interpretivist account explains when it is appropriate to attribute mental states to others (Dennett 1987; Coleman 2005; Noorman 2023). On this view, there are three stances we can take to predict and explain an entity's observable behavior. When we adopt the physical stance, we use our knowledge of physical laws to explain and predict behavior. For instance, when observing the behavior of volcanoes, we adopt the physical stance: we consider how geological forces operate to cause eruptions.

When we adopt the design stance, we use our knowledge of the system's functions to explain and predict behavior. For instance, when observing the behavior of alarm clocks, we adopt the design stance: we appeal to the fact that alarm clocks were created for the purpose of playing a sound at set time to explain and predict its behaviors—we do not typically concern ourselves with the details of the alarm clock's physical implementation.

But when we adopt the intentional stance, we attribute mental states to an entity, and we predict that the entity will act in accordance with those mental states. For instance, when a human acts, we adopt the intentional stance: we understand their behaviors by attributing beliefs, desires, and intentions to them. It is difficult to explain the fact that Jamie buys avocados at the store without appealing to her mental states

[39] Davidson also has an interpretivist view, but here I will focus on Dennett's account, primarily because it is more commonly discussed in the context of AI, but also because it is the version most prominently discussed in philosophy of mind (Davidson 2001b).

(e.g., her intention to make guacamole and her desire to use fresh ingredients). Similarly, it is difficult to predict what Jamie will do when she gets to the store without attributing mental states to her (e.g., given her beliefs and desires, we can predict that Jamie will make her way to the produce section).

In theory, we can see how an AI system might fit into this interpretivist picture. Put simply, so long as we can interpret an AI system's behavior by attributing mental states to it, the system can be said to possess mental states. On this view, then, we're done—we've found a way for AI systems to have mental states.

But this view must be filled out more. Even if we accept the core interpretivist claim that agency depends on how an entity's behavior is interpreted (rather than some independent fact about the entity's internal constitution), some ways of cashing out this claim are more plausible than others. Let us consider an extremely weak form of the view:

> *Weak interpretivism:* An entity has mental states if and only if adopting the intentional stance towards it is possible.

Weak interpretivism is too inclusive. Seemingly anything can possess mental states on this view. I can describe a volcano as erupting because it wants to spew lava, and I can describe my watch as believing that the correct time is 5:31 PM. This version of the view is weaker than the one Dennett endorses. But presenting the weak version allows us to see an initial reason to take caution when analyzing AI systems from the interpretivist perspective.

Humans already exhibit a tendency to explain the behaviors of inanimate objects as purposeful and deriving from mental states. This tendency is not always because we genuinely believe that such entities have mental states. For instance, when people see a video of a triangle moving around a screen followed by a cluster of circles moving in the same pattern, it is a useful shorthand to describe the triangle as "being chased by" or "running away from" the circles (Heider and Simmel 1944). We *can* adopt the intentional stance towards the triangle, but this does not mean that the triangle has beliefs about the circles or an intention to avoid them.

This bias is particularly dangerous with forms of AI that are designed for user engagement such as social robots, chatbots, and virtual assistants. Developers of these systems might prey on human tendencies to anthropomorphize, making us more inclined to adopt the intentional stance. Turkle warns that such relational artifacts push our "Darwinian buttons" in a way that makes us more inclined to attribute psychological capacities to them (Turkle 2011). Weak interpretivism would open the door for our anthropomorphic tendencies to dictate which entities have mental states. Humans anthropomorphize—we *can* adopt the intentional stance towards a lot of things. Even on more sophisticated versions of interpretivism, we should not forget this human bias, as it might cause us to adopt the intentional stance when doing so is unwarranted.

As an alternative, let us consider a version of interpretivism at the other extreme:

> *Strong interpretivism:* An entity has mental states if and only if adopting the intentional stance towards it is explanatorily necessary.

This strong interpretivist approach might have more intuitive appeal than the weak version, though it is stronger than what Dennett had in mind. List adopts a strong interpretivist approach when arguing for group agency. Collective agents, such as states and corporations, are said to have beliefs, desires, and intentions at the group level—over and above the mental states of any individual members of the group. Ascribing mental states to certain organized collectives is "explanatorily indispensable if we wish to make sense of their behavior" (List 2021, 1216).

Even if we were to painstakingly trace the decisions made by corporations, appealing only to individual members would leave the decisions underexplained, especially in cases in which a group attitude is adopted that no individual member holds (Pettit 2007; List and Pettit 2011). If we apply this view to AI, then, it seems that we should attribute mental states to systems only if doing so is explanatorily indispensable for understanding and predicting their behavior.

But strong interpretivism faces several problems. The first is that explanatory necessity seems to place the bar too high. Consider those who argue, contrary to List and Pettit, that corporations (and other collectives) are not agents. Arguments against collective agency often argue that the behavior of collectives can, in fact, be explained without attributing mental states to the collective itself; we can get everything we need by focusing on the actions of individual agents.

In the case of AI, too, an alternative explanation might always be available. Imagine, for instance, a highly convincing humanoid robot. We might be tempted to adopt the intentional stance towards this entity, seemingly on good grounds. The problem is that we can always adopt the design stance towards it, even if it otherwise convinces us that it has mental states. We can claim that it is only behaving in a particular way because of its programming, or that it is merely performing next-token prediction (if it is an LLM) or attempting to maximize its reward function (if it is an RL system).

If such reasons rule out the explanatory necessity of mental state attributions in corporations and humanoid robots, then humans might face a similar objection. We can adopt a design or physical stance towards other humans: we can explain human behavior in virtue of genetics and upbringing, or in virtue of the neurons in our brains, or by appeal to evolutionarily embedded drives. This method of explanation might be clunky, but it seems at least in principle possible to explain human behavior without appealing to mental states (at least on reductionist views).

If strong interpretivists are committed to the claim that attributing mental states to humans is explanatorily indispensable, then appeals to computer programmers will be insufficient to rule out mental states in AI systems. This is, perhaps, while most views of interpretivism do not require explanatory necessity, they do not tend to require that the *only* way to explain an entity's behavior is through the attribution of mental states.

The availability of the design stance points to a version of what Millière and Buckner call the *redescription fallacy*, which occurs when one claims that some entity cannot model a certain capacity because it can be explained in a more fine-grained

way. They explain: "To illustrate, consider the flawed logic in asserting that a piano could not possibly produce harmony because it can be described as a collection of hammers striking strings, or (more pointedly) that brain activity could not possibly implement cognition because it can be described as a collection of neural firings" (Millière and Buckner 2024a, 10).[40]

A second problem for strong interpretivism, and one that will come up on other forms, is the gnawing tension it reveals between explainability and interpretation. Indeed, McCarthy observes that ascribing mental states to machines with a "known structure" (i.e., easily explainable systems) is straightforward but not very useful, whereas ascribing mental states to machines with a "very incompletely known" structure is less straightforward but much more useful (McCarthy 1979).

The more we understand how an entity operates, the less inclined we are to explain its behavior by appealing to mental states. For instance, ancient civilizations often explained natural events by appealing to intentions (either of the events themselves or of the gods supposedly controlling the natural forces). But, as people learned more about meteorology and earth science, it no longer made sense to explain the behavior of winds and volcanoes as intentional actions. Mental states appeared explanatorily indispensable for explaining natural events but later turned out not to be.

This relationship between explainability and interpretation poses a challenge for AI development. Many authors underscore the importance of explainable AI, or creating systems that enable humans to understand why and how such systems behaved in certain ways (Wachter, Mittelstadt, and Floridi 2017; Mittelstadt, Russell, and Wachter 2019; Zerilli 2022). Yet given the strong interpretivist view, creating unexplainable AI, or so-called "black box systems," might increase the justifiability of

[40] The redescription fallacy might also be seen as a response to the "stochastic parrots" objection—the claim that LLMs merely reproduce patterns and therefore do not instantiate genuine reasoning or understanding (Bender et al. 2021).

attributing mental states to these systems.[41] This observation is dangerous, as the question of whether AI systems have mental states becomes a question about whether AI systems are opaque to humans—and this might incentivize a lack of transparency. Plausibly, then, a more moderate version of interpretivism that avoids the extremes of possibility and explanatory necessity can help us make sense of when it is appropriate to adopt the intentional stance:

> *Moderate interpretivism:* An entity has mental states if and only if adopting the intentional stance towards it is explanatorily useful.

Admittedly, the claim that adopting the intentional stance can be "explanatorily useful" is vague. However, it helps us avoid the extreme versions of the view. It is not explanatorily useful to hold that volcanoes erupt because they want to erupt; attributing the mental state to volcanoes does not help us explain or predict their behavior at all, even though it is possible to attribute mental states to them. It is explanatorily useful to hold that corporations have beliefs and desires, even if it is not strictly explanatorily necessary—it would be cumbersome and difficult to describe each behavior of corporations by describing the actions of its members. (And it is explanatorily useful to hold that other humans have mental states, even if we can in theory explain all human behavior with a reductionist physicalist account).

Indeed, Dennett's version of interpretivism is moderate in nature. He does not rule out the possibility of adopting the physical or design stance towards humans. He does, however, hold that there are objective patterns of behavior that are observed by the intentional stance, and that we would miss these patterns if we were to only adopt the physical and design stances (Dennett 1987). Behaviors can be explained in multiple ways. And in some cases, the intentional stance provides a strong explanation by picking up on patterns of a certain kind—sometimes, the intentional stance will explain behavior better than other the stances, even if those stances can also explain the same behavior in different terms.

---

[41] The idea that unexplainable AI might warrant attributing mental states to AI systems on the interpretivist view also furthers arguments for why we should have explainable AI.

The notion of explanatory usefulness maps on well to the idea that we should adopt the intentional stance when an entity's behavior is explained well by the attribution of mental states. Lederman and Mahowald offer the following definition: "some behavior counts as 'well explained' by some hypothesis (very roughly) if the hypothesis offers a sufficiently simpler explanation, which makes sufficiently accurate predictions in a sufficiently wide array of counterfactual circumstances" (Lederman and Mahowald 2024). On this view, then, it is perfectly acceptable to ascribe mental states to an entity if that entity's behavior can *also* be explained by the physical or the design stance, just so long as the hypothesis that an entity has mental states does a good enough job of explaining its behavior.

In the next section, then, I will meet the moderate interpretivist on their own terms. I will argue that attributing mental states to existing AI systems does not explain their behavior sufficiently well.

4.3.2 The Intentional Stance and AI

I now turn to assessing the extent to which it is explanatorily useful to attribute mental states to existing AI systems. I will argue that even on the plausible moderate interpretivist view, we should not adopt the intentional stance towards existing AI systems.[42] Specifically, I argue that it is not explanatorily useful to adopt the intentional stance towards AI.

Which kinds of behaviors would AI systems need to demonstrate for it to be explanatorily useful to adopt the intentional stance? For starters, the supposed mental states of an AI system would need to be robust and stable—otherwise, the system's behavior would not be well explained by adopting the intentional stance. If explaining a system's behavior involved attributing various fleeting mental states that changed

[42] It is worth noting that if correct, my argument will also rule out attributing mental states to existing AI systems on strong interpretivist views. Weak interpretivist views can still hold that AI systems have mental states, but it is not clear that any philosophers of mind take such a view seriously.

for any given behavior, then the system's behavior would not be well-explained by such attributions.

Let us start with LLM-based chatbots. Initially, a system like Chat-GPT seems like the perfect candidate for adopting the intentional stance, as it produces remarkably human-like text. When having a mundane conversation with Chat-GPT, it might seem as though we should adopt the intentional stance, at least on moderate interpretivism. When Chat-GPT answers a non-straightforward question, it seems explanatorily useful to attribute beliefs to it. For example, when Chat-GPT offers an output in a regional dialect, it might seem to have beliefs about which kinds of phrases constitute that dialect, and it might be useful to describe Chat-GPT as intending to produce text in that dialect. When Chat-GPT refuses to answer a question, it might seem to have desires to avoid harm and beliefs about which answers could potentially cause harm.

But the explanatory usefulness of adopting the intentional stance towards Chat-GPT is lower than it might initially seem. Often, when trying to understand why particular outputs were produced, it is more explanatorily useful—and simpler—to describe the system as performing next-token prediction than to appeal to the system's beliefs or desires. For instance, Chat-GPT "hallucinates," or, more precisely, confabulates, often producing false (factually inaccurate or logically inconsistent) about the real world claims (Ji et al. 2023).[43]

The reason for such behavior is not best explained by interpreting Chat-GPT as having false beliefs about the world or a desire to deceive its interlocutor. Rather, the explanation for such confabulations lies in the quality of the training data, the generation method the LLM uses, or the prompt the user inputs (IBM Technology 2023). Chat-GPT is still best explained by adopting the design stance: we explain its

[43]Additionally, it might be the case that LLMs inevitably hallucinate in virtue of their architecture (Xu, Jain, and Kankanhalli 2024). Such behavior in LLMs increases the risks of widespread misinformation.

behavior by appealing to the fact that it is predicting the statistically likely next words based on its training, architecture, and input.

The interpretivist might respond that having inconsistencies and false information is not enough to rule out the possibility of mental states. After all, humans often exhibit these tendencies as well. But in the case of humans, adopting the intentional stance is still necessary if we want to understand a person's behavior. Humans are often inconsistent in their beliefs, but these inconsistencies are typically seen against a backdrop of consistent beliefs and desires. Insofar as a human frequently makes up information about the world, it is also difficult to attribute beliefs to them—it is difficult, for instance, to attribute beliefs to pathological liars. Generally, however, it is in virtue of all the times humans do convincingly display beliefs that we attribute beliefs to them in general. LLMs have not provided us with such a convincing display of beliefs.

Goldstein and Levinstein consider the biggest threat to LLM folk psychology—and to attributing mental states to LLMs on the intentional stance—to be instability (Goldstein and Levinstein 2024). The problem is that LLMs are highly sensitive to prompts. Changing the wording of the prompt or changing aspects of the prompt that are not relevant to the meaning of the prompt, can lead to vastly different outputs from the system (Sclar et al. 2023; Shanahan, McDonell, and Reynolds 2023).

Goldstein and Levinstein offer two potential responses that they think should be developed further. First, it might be the case that LLMs are unstable but that each prompting session produces a different agent with its own beliefs and desires. But this response is unsatisfying. For starters, it is implausible that a non-agentic underlying system could become an agent solely in virtue of receiving a particular prompt. Moreover, it is not clear that there is explanatory use in attributing mental states to one particular version of the LLM in a prompting session. The usefulness of adopting the intentional stance towards Chat-GPT is diminished if an entirely new set of beliefs, desires, and intentions must be posited for every new interaction with the model.

Rather than attributing mental states to dialogue agents (LLM chatbots like Chat-GPT) in this literal sense, Shanahan and colleagues propose using metaphors to

describe LLM behavior. On one metaphor, we can view dialogue agents as engaging in role-play. But unlike a human that engages in role-play, the dialogue agent's role will change and be shaped by its interactions with the user. The types of roles the dialogue agent plays are influenced not only by the tone and content of the conversation it is currently in, but also by the characters present in its training data.

Now, it might be objected that viewing a dialogue agent as role-playing is akin to taking the intentional stance towards it. However, the role-play metaphor is insufficient for explaining the dialogue agent's behavior:

> It [the role-play metaphor] is overly suggestive of a human actor who has studied the character in advance—their personality, history, likes and dislikes, and so on—and proceeds to play that character in the ensuing dialogue. But a dialogue agent based on an LLM does not commit to playing a single, well defined role in advance. Rather, it generates a distribution of characters and refines that distribution as the dialogue progresses. (Shanahan, McDonell, and Reynolds 2023, 494)

As such, the authors propose a second metaphor: a simulator. The dialogue agent, at every point in the conversation, has a "superposition of simulacra" that would fit with the context of the conversation so far (Shanahan, McDonell, and Reynolds 2023). The dialogue agent never takes on or commits to a single role—instead, it maintains a distribution of possible roles that fit with the preceding dialogue.

This metaphorical description of the dialogue agent, as a role-playing simulator that maintains a distribution of roles, allows us to see how adopting the intentional stance would not yield sufficient explanatory power. To see this, consider a comparison Shanahan et al. use to explain the difference between humans and dialogue agents: a game of 20 questions (Shanahan, McDonell, and Reynolds 2023). While a human would pick a single object in advance and answer the player's questions accordingly, the dialogue agent does not. Instead, the dialogue agent will maintain a set of possible objects consistent with the answers it has given so far. This tendency is exemplified by the fact that the dialogue agent, when prompted to regenerate a response, will name a different object each time. It is more explanatorily

useful, then, to view the dialogue agent as a role-playing simulator than as an agent with beliefs, desires, and intentions.

Second, Goldstein and Levinstein propose, it might be the case that the apparently different behavior of LLMs in different prompting sessions does result from a stable underlying goal. This response is also unsatisfying. The most plausible underlying goal to attribute to an LLM is the goal of accurately predicting the next word. Even if we could ascribe some underlying *goal* to LLMs, such as "doing what the user wants," doing so is different from ascribing it some underlying *desires* or *beliefs* (Coelho Mollo and Millière 2023). Plants, for instance, might have a goal of getting sunlight, and some plant behaviors (e.g., growing towards the sun) might be explained in virtue of this goal, but attributing such a goal to a plant is not the same as explaining its behavior by saying that it has a desire to face the sun.

Moreover, we have no independent grounds for attributing a stable underlying goal to LLMs. Consider a human case. Lupita Nyong'o behaves very differently in her role as Nakia in *Black Panther* compared to her role as Adelaide in *Us* (and even differently as Red, Adelaide's doppelganger in the same movie). If we were to try to compare the mental states of these different characters, we would be left with an unstable entity. But we know that Lupita Nyong'o is an actress, and we know that she has a suite of underlying beliefs, desires, and intentions despite the fact that she behaves very differently in each movie she acts in. But we also have independent evidence that she has a stable underlying set of mental states—we see this from her interviews and from the nature of being an actress.

We cannot say the same for Chat-GPT. What best explains the instability of the model is that Chat-GPT is a malleable tool, such that it provides different answers in different contexts. It can initially produce outputs as if it "believes" one statement and then later produces output as if it "believes" the exact opposite. Additionally, Chat-GPT is easily "jailbroken"—its safety features can be overridden by introducing certain sets of instructions (Chao et al. 2023; Hughes et al. 2024). Our ability to jailbreak Chat-GPT undermines the claim that it has underlying beliefs about, for instance, which responses are toxic or should not be outputted. Given how easy it is for Chat-

GPT to change its "beliefs," "desires," and "intentions," it is not helpful to adopt the intentional stance towards it.

Lederman and Mahowald claim that Chat-GPT's ability to generate novel references is best explained by attributing beliefs, desires, and intentions to the model. For example, when prompted to come up with a new name for a historical figure and output facts about that historical figure using the new name, Chat-GPT generated numerous facts about "Marion Starlight" consistent with facts about the historical figure Robespierre (Lederman and Mahowald 2024).

According to Lederman and Mahowald, the relevant behavior (of producing novel reference) in the case above is "can be explained [more easily] on the hypothesis that the LLM intends for the term 'Marion Starlight' to be equivalent to 'Robespierre'" (Lederman and Mahowald 2024). This hypothesis is useful not only for explaining Chat-GPT's responses, but also for predicting how Chat-GPT would counterfactually answer different questions about Marion Starlight.

Two responses are warranted here. First, it is not clear whether LLMs can perform novel reference—this is a possibility Lederman and Mahowald acknowledge (and as such, they present their argument as holding *if* it is the case that LLMs have this capacity). We would need experimental evidence that LLMs can do so—including evidence that the relevant reference does not exist in the model's training data (e.g., if the LLM were trained on internet data that had some historical fiction that referred to Robespierre as Marion Starlight).

Second, and more decisively, even if LLMs could perform novel reference, such a capability does not offer strong support for the interpretivist to adopt the intentional stance. The role-playing simulator analogy is a plausible alternative for explaining the model's behavior. Moreover, it seems like some relatively straightforward explanation is available without invoking intentions. We can imagine that when a model is prompted to refer to a historical figure by a new name, the most probable text completion will be a new name, which is made up of existing tokens from the model's training set (e.g., the tokens "Marion" and "Starlight" plausibly exist on the internet). Then, in virtue of the preceding conversation history, the model will fill in the new

name in place of the historical figure. Such behavior is even more likely if we assume that the LLM's training data includes various guessing games.

Additionally, as alluded to previously, one example of seemingly complex behavior does not give us strong reason to adopt the intentional stance towards a model in general. Given that the rest of Chat-GPT's behavior is not well explained by appealing to mental states, it would not be useful to attribute only one intention (or some small set of intentions) to the model in cases where it generates novel reference. McCoy and colleagues performed a set of experiments that show that adopting the design stance towards LLMs is explanatorily useful (McCoy et al. 2024a). The researchers call their approach "teleological" because it focuses on an AI system's goals and environment—it asks what problem LLMs were created to solve and then uses this characterization to predict the system's abilities and biases. The teleological approach is framed as a response to the tendency to evaluate LLMs according to human metrics like standardized tests. While these researchers do not frame their discussion in interpretivist terms, their studies can be viewed as an appeal to adopting the design stance against a tendency to adopt the intentional stance.

Because LLMs are statistical systems designed to perform next-token prediction, McCoy et al. hypothesize that their behavior will be influenced in particular ways that do not align with the way humans behave—and these features might be missed by only evaluating LLMs using metrics designed for evaluating humans (McCoy et al. 2024a). The idea is that the types of tests we use to assess human capabilities are ill suited for picking up on the failure modes unique to LLMs.

The experiments confirm these hypotheses. The results indicate that LLMs are sensitive to probability in various ways. LLMs are sensitive to the probability of the target output. For example, LLMs are better at performing shift ciphers (whereby each letter of a phrase is shifted a certain number of letters forward) when the output message is a high-probability phrase (i.e., a commonly used phrase) than a low-probability phrase (i.e., a sentence that is grammatical but unlikely to be found frequently in a training corpus). LLMs are also sensitive to the probability that the task at hand was illustrated in the training corpus. For example, LLMs perform better on

tasks that are more frequently illustrated in internet text—they are better at sorting a list of words into alphabetical order (common) than reverse alphabetical order (rare).

Such results are not easily explainable if we try to probe which "beliefs" and "intentions" LLMs have—but they are easily explainable if we adopt the design stance, as LLMs' sensitivity to probability results from the fact that they are designed to do statistics. One upshot of these experiments, then, is that LLM behavior is influenced by its design—and adopting the design stance enables us to capture these features. To predict where a model will succeed and where it will struggle, it is more helpful to consider the way the model was designed and why rather than to attribute mental states to it. Or, in the words of McCoy et al., "to understand what language models are, we must understand what we have trained them to be" (McCoy et al. 2024a).

Ultimately, then, much more convincing evidence would be needed to justify adopting the intentional stance towards LLMs.

What about RL systems? Dennett holds that we should adopt the intentional stance towards them:

> Consider chess-playing computers, which all succumb neatly to the same simple strategy of interpretation: just think of them as rational agents who *want* to win, and who *know* the rules and principles of chess and the positions of the pieces on the board. Instantly your problem of predicting and interpreting their behaviour is made vastly easier than it would be if you tried to use the physical or the design stance. (Dennett 2009, 340)

It might be in virtue of the RL process that we are tempted to adopt the intentional stance towards chess-playing programs. After all, AlphaZero developed its chess-playing abilities through repeatedly playing chess against itself—and the RL goal of winning games (which the system was trained to optimize) looks a lot like a desire to win. Similarly, AlphaZero's ability to keep track of where all the pieces on the board are looks a lot like AlphaZero has beliefs about the location of those pieces. It is especially tempting to adopt the intentional stance towards AlphaZero because we are already used to explaining the behavior of highly skilled chess players in intentional terms, as we have only ever previously done this with respect to humans. Indeed, the goal-directedness that RL embeds in systems seems to be an important step towards

AI systems with mental states. I will return to this point in the conclusion and in Chapter 5.

But adopting the intentional stance towards an RL system like AlphaZero is not as explanatorily useful as Dennett seems to think. It is not clear that we can better understand or explain the system's behavior by attributing mental states to it. For instance, when AlphaZero makes a legal chess move, as it usually does, we can describe the program as having beliefs about the rules of chess.

But we can also explain this behavior of adhering to the rules of chess without attributing any beliefs. It is not that much easier to say that AlphaZero knows the rules of chess than to say that AlphaZero has recognized that certain moves yield very low rewards and so avoids making those moves (namely, the moves that violate the rules of chess). Saying that AlphaZero "knows" that moving the rook diagonally is an illegal move does not help us understand its behavior any better than the understanding we get from knowing how the system was trained.

Moreover, understanding AlphaZero's training objective and learning method allows us to explain and predict its behavior better than an appeal to beliefs in certain cases. Imagine, for instance, that a new rule is introduced to chess in 2025. Given that AlphaZero was trained prior to 2025, we can predict that it will violate the new rule. It is counterintuitive to describe AlphaZero post-2025 as having a range of false beliefs about the rules of chess—it is more straightforward to explain its behavior with reference to the time at which it was trained. Moreover, it is only by appeal to the design stance that we can understand why AlphaZero will not adopt the new rule no matter how many times it loses in 2025.

This problem is similar to the problem of instability with LLMs. Once a model is trained, it is, in an important sense, frozen, in a way that agents should not be. Suppose a human chess player has not been made aware of the new 2025 rules. We should expect her to behave similarly to AlphaZero. Why, then, is it still explanatorily useful to adopt the intentional stance towards the human but not towards AlphaZero? In the human case, we can expect her beliefs and desires to be updated in light of new information. For AlphaZero, we cannot make this same prediction.

The fact that RL systems are frozen after their training seems to be the reason the DeepMind researchers only deem RL systems agents when they are considered in combination with their creation processes (Kenton et al. 2023). To get us closer to genuine artificial agents, then, it seems like RL will have to be a process that is ongoing even after the system is deployed.

Additionally, certain strategies might be better suited towards AlphaZero in virtue of its lack of mental states. For instance, the model might perform worse in rare game settings. Alternatively, attempts to "get in the head" of AlphaZero as a psychological tactic will not work, whereas they might be part of the game in human-human chess. At best, then, the intentional stance offers a coarse shorthand for explaining and predicting behavior. Here, the problem of explainability returns. The more we understand the way in which AlphaZero was trained, the less useful the intentional stance becomes.

It might be objected that shorthand can be very powerful. Shorthand can allow people to understand, for instance, how AlphaZero is so good at chess without requiring people to understand machine learning. Every model is a simplification, and models are useful because they simplify reality and allow us to do things. Even though a model necessarily simplifies, a model must lose accuracy for the purposes of usefulness. As such, adopting the intentional stance allows us to model the behavior of AlphaZero in a more efficient way than adopting the design stance.

But adopting the intentional stance only succeeds if doing so picks up on the objective patterns of behavior displayed by AlphaZero. But these same patterns—namely, AlphaZero's tendency to make moves with a high probability of success—can be picked out by the design stance. Nothing is missing in explaining and predicting the behavior of AlphaZero if we refuse to attribute beliefs and desires to it. Indeed, a human chess player might be at a disadvantage if she adopts the intentional stance towards AlphaZero—maintaining the design stance might help her to keep in mind that the algorithm will not "slip up" in the way a human might, or "fall for" certain bait moves.

Indeed, it might still be useful for us to use shorthand. A novice chess player, for instance, might have limited information about AlphaZero, in which case treating the system as if it were a human player would be helpful (Papagni and Koeszegi 2021). I am not claiming that we should abandon these tendencies. Instead, we should view the usefulness of treating AlphaZero as if it were a human chess player for what it is: a coarse analogy, rather than a genuine attribution of mental states.

Ultimately, then, while the intentional stance seems well-positioned to enable us to attribute mental states to AI systems—and indeed, Dennett and other interpretivists might still claim that we should adopt the intentional stance towards AI systems—I have argued that the intentional stance is not warranted towards existing LLMs and RL systems because adopting the intentional stance is not explanatorily useful.

4.4 AI and Representationalism

Another popular view in philosophy of mind that might readily lend itself to attributing mental states to AI systems is representationalism. While different representationalist views have different criteria for possessing mental states, the common thread is that mental states are internal representations of things in the world—to be more precise, mental representations are mental objects that have semantic properties (Pitt 2022).

In this section, I argue that existing AI systems do not have mental states on representationalist views. I begin by arguing that the sensory grounding objection, an impediment to attributing mental representations to LLMs, has not been sufficiently resolved (section 4.4.1). I then argue that existing AI systems do not demonstrate either of two indicators of mental states on representationalist views: sufficient folk patterns of reasoning (section 4.4.2) or robust evidence of concept possession (4.4.3).

4.4.1 Sensory Grounding

The sensory grounding problem is posed as a threat to the ability of LLMs to represent things in the world. Bender and Koller, drawing on the symbol grounding problem, argue that LLMs cannot learn *meaning* because meaning cannot be learned from *form* alone (Harnad 1990; Bender and Koller 2020). In other words, LLMs cannot learn the meanings of words because they lack connection to the external world—they only have access to text and thus can only learn the relationships between pieces of text. To see why, consider a condensed version of what Bender and Koller call the "octopus test" (Bender and Koller 2020, 5188–89):

> A hyperintelligent octopus does not know English and has never observed above-ground objects but is very good at statistics. The octopus listens in on conversations between Jill and Adam, who are on two separate islands and communicate via underground telegraph. The octopus cuts the telegraph and inserts itself into the conversation—specifically, the octopus pretends to be Jill and talks to Adam.

Because the octopus is so skilled at picking up on the statistical relationships between words, the octopus can fool Adam with respect to many topics of conversation. But, according to Bender, the octopus will fail a sufficiently sensitive test—the octopus will fail to fool Adam at least in the following scenario: if Adam is facing danger and asks Jill to come up with a way to construct a weapon from the materials Adam has, the octopus will be unable to produce a sufficiently convincing answer.

The octopus test is supposed to show that on tasks requiring knowing the *meaning* of words (rather than just the *form,* or statistical features, of text), an LLM trained on only text will necessarily fall short. The octopus will have no sense of the ways in which sticks, mud, and other materials can be combined to make a weapon, especially given that Jill and Adam had not previously discussed this topic. Insofar as the LLM has any representations, these representations will be about *language,* not about the external world.

Chalmers has presented an argument that takes aim at the supposed sensory grounding problem (Chalmers 2023). Chalmers argues that thinking does not require the capacity to sense. His argument begins with imagining "pure thinkers"—entities that can think but lack, and have always lacked, sensory abilities. He argues that there

are at least some thoughts a pure thinker could think, such as arithmetic thoughts, thoughts involving metaphysical and causal concepts, and thoughts about the external world (though these thoughts might be restricted to being hypotheses).[44] But pure thinkers would lack many concepts, including sensory concepts, practical concepts tied to bodily actions, and singular concepts of entities or kinds.

Chalmers then considers a closer analogy to LLMs: "pure thinker/talkers"—pure thinkers with the ability to understand natural language inputs and produce linguistic outputs. According to Chalmers, pure thinker/talkers would be able to think a much wider range of thoughts—they can use linguistic inputs to know things about the world, they can use patterns in linguistic inputs to form theories about the world, they can know scientific laws via testimony, and they can acquire a wider range of concepts. Pure thinker/talkers would still have limitations (e.g., they could not fully possess sensory concepts and concepts of bodily action).

But, because they have access to language, pure thinker/talkers have causal grounding in the environment—the linguistic community provides a causal connection between the thought and the environment, and this causal connection is sufficient for reference. Pure thinker/talkers could still know a lot about the world. Chalmers is careful to clarify his conclusion: he does not argue that LLMs can think; instead, he rebuts one argument that LLMs cannot think (namely, that sensing is necessary for thinking).

But even if we accept Chalmers' argument that a pure thinker/talker can refer to things in the world, his conclusion regarding LLMs does not follow as straightforwardly as it seems. If successful, Chalmers has shown that sensing is not a necessary condition for thinking thoughts about the world, *given that an entity has the capacity to think*. But whether LLMs can think, even in some limited form, is precisely what we are trying to assess. Chalmers has shown that some hypothetical system lacking the capacity to sense can refer to the external world. In other words, Chalmers

[44] Already, the fact that LLMs are famously bad at arithmetic should give us reason to think that they are not good candidates for pure thinkers.

seems to have shown that a brain in a vat could refer to the external world (Putnam 1982).

This claim is perfectly consistent with the claim that LLMs, in virtue of how they operate, must have sensory grounding to have mental representations. Chalmers' argument does not show that a system trained to perform next-token prediction can refer to things in the external world—it only shows that it is possible for some system to think about the world despite lacking sensory capacities. As such, the sensory grounding problem is still a live objection to LLMs having mental states on the representationalist picture.[45]

Interestingly, Coelho Mollo and Millière agree that LLMs (in their base model version) lack referential grounding—but they argue that RL can provide such grounding (Coelho Mollo and Millière 2023).[46] As Chapter 5 will discuss in more detail, many user-facing LLMs undergo some form of RL aimed at ensuring that the implementation (e.g., the chatbot) avoids producing outputs that are offensive, inappropriate, or false. Coelho Mollo and Millière highlight two conditions that are widely accepted as necessary for referential grounding: (1) causal-informational relations, such that the representations carry information about the world through a causally generated connection, and (2) historical relations, such that the representations have a relevant kind of normativity, a kind that enables them to both represent accurately and misrepresent.

Coelho Mollo and Millière argue that LLMs fulfill the first condition in virtue of their training data. This argument aligns with Chalmer's claim about how pure

[45] Other philosophers, drawing on social externalism, also argue that concept acquisition purely via language is possible (Butlin 2023; Millikan 2000; Putnam 1975b; Burge 1979). Still, these arguments on their own do not show that LLMs can acquire representations that refer to the external world.

[46] Coelho Mollo and Millière refer to the sensory grounding problem as the "vector grounding problem." They also differentiate between different notions of grounding, highlighting referential grounding as the type of grounding necessary for LLMs to produce output with intrinsic meaning. Reference is defined as "a relation that enables representations to 'hook onto' worldly entities or properties" (Coelho Mollo and Millière 2023).

thinker/talkers can indirectly obtain information about the world, for instance through testimony. But unlike Chalmers, Coelho Mollo and Millière hold that the causal-informational relations condition holds not just in principle, but with respect to existing LLMs. Roughly, their argument is that the structure of the linguistic data used to train LLMs very likely embeds information about patterns that exist not just in language, but also in the world. As such, the patterns that LLMs learn from carry information about the meanings of linguistic expressions (rather than mere form)—and the causal connection is mediated by the human use of language.

It is difficult to deny the claim that internet data (i.e., the data that LLMs are trained on) contains information about the meaning of words. A more interesting question, however, is *to what extent* such data carries the relevant information. Huh and colleagues, for instance, find that foundation models are converging such that their representations of data are similar, despite models varying in architecture and training data (Huh et al. 2024). The authors take this evidence as support for the "platonic representation hypothesis," namely that the models are converging towards "a statistical model of the underlying reality. But whether models are converging towards an accurate representation of the world depends on the extent to which data of linguistic expressions carries all the relevant information about the meaning of words. This is an open empirical question. Still, we can grant that LLMs likely meet the causal-information relations condition for at least some of their representations.

But the argument that LLMs meet the historical condition in virtue of RL is less convincing. The argument relies on a distinction between proximate and ultimate function. The LLM's proximate function is next-token prediction—and, Coelho Mollo and Millière admit, this function cannot provide the kind of normativity required for referential grounding. At best, the proximate function of next-token prediction can provide linguistic normativity, but not world-involving normativity. The LLM's ultimate function, on the other hand, is more general—it is the goal that the LLM is supposed to achieve via next-token prediction.

RL, Coelho Mollo and Millière argue, enables LLMs to have world-involving functions. This is because the LLMs are trained, via RL, on "extra-linguistic" criteria.

For example, when LLMs are rewarded for producing truthful outputs (i.e., outputs that accurately reflect reality) and punished for producing untruthful outputs (i.e., outputs that fail to accurately reflect reality), they fulfill or fail to fulfill their ultimate function of producing accurate outputs. So, the argument goes, LLMs have the kind of normativity necessary for meaning, as the ultimate function allows them to obtain referential grounding.

I will address a similar question regarding whether RL enables LLMs to act for reasons in Chapter 5. For now, however, we can observe that the case for RL-enabled LLMs meeting the historical condition for referential grounding depends on the extent to which the RL genuinely enables a relevant form of normativity. An initial reason for skepticism lies in the fact that the normativity is provided by the data used in RL, which can be different even from what the humans who collect the RL data intend. As such, the ultimate function of an LLM designed to produce truthful outputs might not map on well to truth in the external world.

4.2.2 Folk Patterns of Reasoning

Goldstein and Levinstein consider five conditions on mental representations (information carrying, causal efficacy, folk-psychological reasoning, structural isomorphism, and selection) and argue that LLMs satisfy them all (Goldstein and Levinstein 2024). While I believe there are reasons to doubt that LLMs satisfy each of these, I focus my discussion here on what I take to be perhaps the most important criterion, and which Goldstein and Levinstein highlight as being particularly important: folk-psychological reasoning.

The criterion for folk-psychological reasoning is taken from Fodor's view that representations must operate in ways that look like ordinary folk reasoning (Fodor 1975). As evidence that LLMs might meet this criterion, Goldstein and Levinstein appeal to evidence of so-called world models in LLMs, defined as follows:

> A world model, in the context of AI, refers to an internal representation of how the external world operates, including objects, their properties, and the causal

> relationships between them. World models require coherent and consistent representations across contexts and a degree of abstraction that allows LLMs to generalize from particular cases to more general situations. (Goldstein and Levinstein 2024)

Goldstein and Levinstein focus on several machine learning experiments to support their claim that LLMs have world models. In one paper they discuss, Musker and Pavlick conduct a set of studies aimed at testing whether LLMs build causal models to understand word meanings (Musker and Pavlick 2024). The HIPE theory of lexical representations concerns the factors that relate the meanings of words to the design history, physical structure, and use of the object. So, on HIPE theory, when a human is deciding whether to classify some object (say, a metal rod with a dishcloth attached to the top) as a mop, she will primarily consider the object's physical structure, and the reason the object was created has less of an effect on her causal model.

Musker and Pavlick performed experiments on whether LLMs give responses similar to humans on HIPE theory tasks. These tasks involve giving test subjects counterfactual vignettes in which various factors of the causal model were compromised (e.g., the reason that the object was created, the structure of the object, or the goal of the user). They found that while GPT-3 did not perform well on this task, GPT-4 provided outputs similar to those provided by humans. In other words, the same counterfactual changes in the scenario that influenced whether humans classified the item in question as a mop also affected whether GPT-4 classified the item in question as a mop in similar ways.

While GPT-4's performance on this task is impressive, it is not clear that we can draw any significant conclusions about whether the LLM has a world model. When discussing limitations, Musker and Pavlick note that the study only considers three artifacts (mops, pencils, and whistles) and only investigates terms referring to objects. It is a massive inference, then, to conclude from these three examples that LLMs have the kind of world models characteristic of folk psychological reasoning.

The limitations in this study highlight a more general point that is often forgotten in discussions of LLM capabilities: the burden of proof lies heavily on those

claiming that LLMs have world models (or any other capacity). There are several reasons for this. First, experiments purporting to show that LLMs have a particular capacity often only test a small set of examples. LLM benchmarks do not provide a sufficient range of examples to make strong generalizations about the capacities of LLMs.

Second, and relatedly, for every study claiming that LLMs have a particular capacity, there are often studies showing the opposite. A study on a version of GPT trained on Othello games suggested that there was strong evidence of Othello-GPT having an internal model of the board state (Li et al. 2024). But further investigations of LLM internal models across a wider range of tasks shows that the internal models of LLMs are not sufficiently robust to warrant conclusions about world models—for instance, LLMs perform poorly on sequence compression and sequence distinction tasks (which LLMs would perform well on if they had internal world models) (Vafa et al. 2024). A study on GPT-4's ability to simulate text-based games reveals that LLMs cannot reliably act as world simulators (R. Wang et al. 2024). While there have been claims that LLMs can perform modular addition in a way that goes beyond mere memorization (Nanda et al. 2023), other studies reveal that LLMs struggle with arithmetic tasks and only succeed by using shallow heuristics (Dziri et al. 2023; Nikankin et al. 2024).

Third, there are many more flaws in LLM folk reasoning than there are successes. LLMs are particularly bad at generalizing to novel tasks. LLMs, for example, are susceptible to the "reversal curse"—when models are trained on sentences such as "A is B," they fail to generalize to "B is A" (Berglund et al. 2023). Another study on the implicit reasoning of transformers found that LLMs acquire the rules of composition and comparison only though a method called grokking, in which training is extended beyond overfitting (B. Wang et al. 2024).

Even chain-of-thought prompting, a method to encourage LLMs to "reason" through tasks by providing intermediate step-by-step natural language reasoning steps, which has been shown to improve LLM performance across a variety of tasks (Wei et al. 2022), does not provide robust evidence of folk patterns of reasoning in

LLMs (Stechly, Valmeekam, and Kambhampati 2024). It has also been found that LLMs' chain-of-thought explanations can be systematically unfaithful—the models' explanations can be influenced by biasing features in their inputs, which the models fail to mention in their explanations (Turpin et al. 2024). LLMs also struggle on tasks in leet-speak, in which letters are replaced with numbers (e.g., decoding sentences in the form "HUM4NS C4N RE4D THIS"), a task on which LLMs with faithful representations should be able to perform (Leivada et al. 2024).

More broadly, there might be conceptual reasons to doubt that LLMs can reason at all. Stoljar and Zhang argue that LLMs fail to meet a necessary condition of rationality: they fail to respond correctly to evidence by making correct inferences (Stoljar and Zhang 2024). To use their example, suppose you ask Chat-GPT where the best bananas grow, and Chat-GPT responds with the tropics. If we assume that Chat-GPT has the belief that the best bananas grow in the tropics, its evidence for that belief must be a belief about the statistical usage of the word "banana" in a massive corpus of text. But no proposition about the statistical usage of the word "banana" can entail a belief about bananas. Chat-GPT necessarily fails to make correct inferences, even if we grant that it can have beliefs. The implication of this argument for my purposes is that LLMs will be unable to exhibit any folk patterns of reasoning, as they can only reason about the statistical usage of words, rendering them unable to reason about the real world.

In their discussion of LLM world models, Yildirim and Paul offer the following caveat: "We acknowledge that, at present, whether LLMs actually do or even could recover causal abstractions of the world, as in world models, involves a leap of faith" (Yildirim and Paul 2024, 409). Similarly, Goldstein and Levinstein, despite their claims that LLMs exhibit signs of having mental representations, are cautious in their conclusions—they explicitly do not claim, for instance, that LLMs satisfy all the relevant folk patterns of reasoning. While a willingness to take the leap of faith might be required to push research and development forward (and whether we should be pushing this research and development forward is a separate question), the present evidence does not warrant taking such a leap of faith regarding existing LLMs.

4.2.3 Concepts

Philosophers of mind often define concepts as mental representations (Margolis and Laurence 2023). As such, another lens through which to evaluate whether AI systems have mental states on representationalist views is to consider whether they possess concepts.

As I argued in Chapter 2, having a concept entails being able to accurately apply the concept—it does not entail some conscious grasp of the concept or some deeper understanding (though, of course, these features can be helpful). Humans have concepts that we do not understand or consciously grasp. I might, for instance, have a concept of fruit such that I can sort different foods according to whether they are fruits, but I might not understand why certain foods are classified as fruits and others are not. Of course, in this case, I can give reasons why a certain food is a fruit, but this is because my concept of fruit is properly connected to other concepts like being edible and having seeds. In other cases, a concept is merely intuitive—we are unable to offer reasons or a deep understanding of what the concepts mean.

One initial reason to be skeptical that LLMs possess concepts is that their tokens do not always correspond to words. Recall that tokens are the individual units LLMs process and predict. Some LLM tokens correspond to words—for instance, an LLM might tokenize "philosophy is awesome" into tokens "philosophy," "is," and "awesome." But LLMs do not always tokenize in this way. Sometimes, individual words will consist of multiple tokens—sometimes in a way that makes some sense (e.g., "lifting" might be tokenized as "lift" and "ing"), and sometimes in ways that do not correspond to how humans understand words and word segments. Insofar as LLM tokens do not align well with human concepts, we might have doubts about whether LLMs share our concepts.

Butlin argues that while LLMs on their own cannot share concepts with humans, particular applications of LLMs—in this case, chatbots—can do so.

According to Butlin, the relevant distinction between LLMs and chatbots has to do with functions

> Chatbots may have functions which involve detecting and acting on non-linguistic states of affairs; for instance, a medical triage chatbot may have the function of diagnosing the user's condition and making an appropriate recommendation. In contrast, language models have purely linguistic functions, such as generating text which extends and input in a probable way. (Butlin 2023, 3081)

The reason that the function of the AI system is supposed to make a difference is that we can draw inferences from the functions of systems—namely, inferences about the contents of their representations. So, because a medical chatbot has the function of conversing with the user in a way to offer a diagnosis and suggest recommendations, the likeliest way for the chatbot to achieve this function is to create a representation of the user's symptoms and other relevant features throughout the conversation.

Moreover, given that a medical chatbot would need to be able to represent symptoms and conditions in a variety of contexts, it is plausible—according to Butlin—that it will develop representations that are compositional in nature (e.g., the system can represent that a person's ankle is swollen or that their abdomen is swelling). Goldstein and Levinstein make a similar claim about LLMs: although LLMs are trained to predict the next word, creating internal representations of words could help them efficiently predict the next word.

But in the case of AI systems, it is not clear that we can draw inferences about a system's internal representations based on the system's function. Even if the best way to predict the next token were to create internal representations of words, it is not guaranteed that an LLM would develop this capacity. Consider a familiar human case. The function of the mouth is to get food into the stomach. The best way to do this would be to create a separate tube going straight from the mouth to the stomach. But evolution does not always find the best solution—it finds solutions compatible with survival through random mutation. So, our mouth-to-stomach food tube is in the same location as our nose-to-lungs air tube, creating a problem of choking.

So too for LLMs—if LLMs can succeed at next-token prediction without developing sophisticated internal representations, it seems more likely than not that the LLMs will do this. If LLMs can perform well by using heuristics and memorization, we have no reason to believe that they will develop sophisticated internal representations, even if doing so would be a better way to succeed at next-token prediction.

Another complication about AI concept possession is about the content of AI representations. Two interrelated concerns arise. First, AI systems might acquire their concepts in the wrong way. Second, in virtue of the first concern, AI systems might acquire the wrong concepts. To see these concerns play out, let us return to the example of an algorithm that sorts images of dogs and cats. When humans classify animals as dogs or cats, we look for features we have principled reasons to think are important to the relevant concepts. Floppy ears, a wagging tail, and a hanging tongue indicate a dog; whiskers, retractable claws, and judgmental eyes indicate a cat. But AI systems do not pick up on meaningful features like this—instead they pick up on correlations. This worry is similar to the concern about LLM tokenization.

This concern gives rise to safety issues. For example, one-pixel and other adversarial attacks can lead algorithms to make the wrong classification for inputs that are easily classifiable by humans—an image of a dog with certain distorted background features might be misclassified as a cat (Goodfellow, Shlens, and Szegedy 2015).

But the concern runs deeper as well. The algorithm might have some concept of "dog," but it seems that this is not really the same as our concept of "dog," as it does not contain the right component parts. In a worst case, the algorithm might be picking up on spurious correlations. Additionally, it might be apt to describe instances of algorithmic bias as instances of faulty concepts. When an algorithm sorts male candidates into the concept of "good" employee at a disproportionately high rate relative to female candidates, we might say that the algorithm has an incorrect concept of a good employee. The content of the algorithm's concepts does not consist of the relevant features.

Several responses can save the possibility of concept possession in AI. First, concept acquisition is multiply realizable. It is not precisely clear how humans typically acquire concepts—we learn concepts through experience and in virtue of being taught. But what happens at the neurological level might also be some form of statistical inference, in which case our concept acquisition is similar to what machine learning systems do (namely, form associations and make predictions based on those associations).

Moreover, people acquire concepts in different ways. As such, different features of concepts might be emphasized or prioritized differently between humans. A blind person's concept of an apple will likely differ in content from a sighted person's concept, largely because the method of acquiring that concept is different. If social externalism is right, humans often acquire concepts from the ways in which those concepts are used publicly (Putnam 1975a; Burge 1979).

Lastly, it might be the case that AI systems pick out correlations precisely because those correlations, even if nonsensical to us, are actually relevant to the concepts in question. Just because algorithms do not see seemingly relevant features like "fur" and "eyes" in the same way as us, we might not see or use relevant statistical correlations in the same way as algorithms. For example, many patterns in nature follow mathematical models, but humans do not pick up on these underlying features even though we might be tracking them by some higher-level features. Humans have color concepts, and we do not pick up on the relevance of light frequencies—our visual perception tracks the fact that different wavelengths of light are registered by the eyes in different ways.

Second, conceptual content differs greatly between humans—especially when it comes to moral concepts. Consider, for instance, the concept of *moral wrongness.* It might be the case that no two humans have the same concept individuation. Some might hold that the concept *lying* is connected to the concept *moral wrongness* while others might not. Even the wrong-making features of actions are contested—with people placing different emphases on harm, fairness, justice, and other considerations.

These examples highlight a more general, recurring point about AI systems and agency. Butlin is right that a chatbot is a more likely candidate for concept possession than a pure LLM. It is unlikely that foundation models themselves will qualify for moral agency. Rather, it is particular implementations of those models that might meet the criteria. GPT-4, on its own, might not possess the correct concepts relevant to moral agency. However, with appropriate training and instantiation, it might qualify as possessing those concepts.

An analogy can once again be drawn with group agents. It is not the case that any large collection of individuals qualifies for group agency. Rather, to qualify, collections of individuals must be organized in the appropriate way to instantiate the relevant features. The same is true for AI systems. Computer scientists and philosophers are working on ways to determine whether AI systems, in operating on low-level features, possess high-level concepts compatible with human-defined concepts (Parthemore and Whitby 2013; Kim et al. 2017).

4.5 Conclusion

So, I have argued, AI systems are not agents. Neither the interpretivist nor the representationalist can make a convincing case for existing AI systems as having mental states. Moreover, my arguments have offered reason to believe that existing machine learning methods, on their own, are not promising avenues for developing AI systems with mental states.[47] Instead, developing genuine AI agents will require at least the combination of existing machine learning methods, and perhaps the development of additional methods.

A few lessons can be gleaned from this chapter's discussion. First, genuine artificial agents will need to be connected to the world in a way that goes beyond mere language. Naturally, proponents of agentic AI might point to the increase in

[47] Some researchers claim that future LLMs will exhibit emergent properties. But it is not clear how scaling alone will enable such emergent properties to arise (McCoy et al. 2024a; Lu et al. 2024).

multimodal models such as vision-language models (VLMs) as a step in the right direction. But insofar as the image processing capacities are highly similar to the text processing capacities, it is not clear whether the system is actually gaining a different type of connection to the external world from these other forms of input (Chalmers 2023). In other words, if the model is still performing next-token prediction in roughly the same way across modalities or using visual input to obtain more information of the same kind, then the problem of sensory grounding remains.

Perhaps the most promising way to overcome the sensory grounding problem is to develop embodied artificial agents that can interact with the world. But again, if all that is driving an embodied agent's behavior is next-token prediction, the same problems remain. Recently, proposals and attempts have been made to augment LLMs with the ability to call on tools like calculators and calendars. Indeed, DeepMind seems to think that the next big development in AI will consist of so-called "language agents" or "advanced AI assistants"—LLMs that can "plan and execute sequences of actions *on the user's behalf* across *one or more domains* and *in line with the user's expectations*" (Gabriel et al. 2024, 15).

Without a clear sense of how these systems will operate, it is difficult to draw conclusions about whether they will be genuinely agential. Existing evidence of LLM tool use indicates that models struggle to correctly call and use tools (Farn and Shin 2023). Additionally, if the LLM is still the primary driver of the outputs, we are not closer to mental state possession, unless advanced tool usage provides the interpretivist or the representationalist with more reason to attribute mental states, perhaps based on increases in (the appearance of) folk reasoning.

Second, a genuine artificial agent will need to be dynamic. Recall the problem of instability in LLMs—because these systems have their parameter weights fixed, they respond differently (and inconsistently) to slightly different prompts, and any learning that the system does within context will be lost once a new session begins. For an AI system to be an agent, it cannot be static and fixed in this way. Similarly, recall that DeepMind's definition of agency classifies RL systems *plus their training* as agents, but not standalone static RL systems themselves. Once the system stops

learning from its environment, the system is no longer agential. Genuine artificial agents will need the ability to learn even after their deployment. Note that this capacity raises an additional set of safety concerns—it might be the case that an AI system that can update its beliefs based on evidence in the world becomes a conspiracy theorist or a racist.

These considerations lead to the conclusion that a new type of architecture might be necessary for AI systems to be genuine agents.[48] For instance, Sumers and colleagues propose a conceptual framework for LLM-based agents that places LLMs within a larger cognitive architecture (Sumers et al. 2024). On this proposal, the language agent can perform both external actions—interacting with the environment through grounding—and internal actions—interacting with internal memories through retrieval, reasoning, and learning. The proposed model includes four types of memory (working, episodic, semantic, and procedural); can execute external actions via grounding and can process environmental feedback into working memory; can retrieve information from long-term memories into working memories; and can reason and learn.

It is worth bearing in mind that these recommendations for making genuinely agentic AI systems look a lot like proposals to make AI systems increasingly human-like. We might wonder what the point is of creating genuine artificial agents if doing so requires them to be like humans. After all, technology is often helpful precisely because it is non-humanlike. Still, developing artificial agents might help us in a variety of tasks (e.g., agential systems can do more on our behalf than non-agential systems). But developing such agents might be costly and raise various ethical concerns (e.g., if artificial agents have desires, a desire-satisfactionist might hold that they have well-being and thus that we need to consider their well-being when we

[48] For an overview of other approaches and architectures for LLM-based "agents," see Xi et al. (2023); L. Wang et al. (2024); and Jingwen Zhou et al. (2024).

make decisions). Just because we *can* make genuine artificial moral agents does not imply that we *ought* to do so.[49]

[49] Thank you to Carissa Véliz, Alison Hills, and Milo Phillips-Brown for extensive feedback on this chapter. Thank you to audiences at the Oxford DPhil Seminar (2023), the Oxford AI Society Mini-Conference (2024), and the Agency and Intentions in AI Conference (2024) for additional comments and discussion.

Chapter 5: Artificial "Agents" are Not Moral

Abstract

This chapter considers the extent to which existing AI systems possess moral capacities. I argue that despite their impressive performance on certain moral tasks, AI systems exhibit only a minimal level of moral competence. First, I argue that even if AI systems can be said to act for reasons in virtue of being trained with reinforcement learning, they do not act for *moral* reasons. Second, I argue that AI systems show substantial deficits in their responsiveness to moral reasons but do show some rudimentary level of such responsiveness. Third, I argue that AI systems are far from possessing moral understanding. Ultimately, even if future AI systems are agents, there are still many hurdles to them becoming moral agents.

5.1 Introduction

In Chapter 4, I focused on the *agency* aspect of *moral agency*. In this chapter, I turn to the *moral* aspect. Now, it might seem strange to assess the moral capacities of AI systems given that the previous chapter has already established that AI systems are not agents at all. But there are several reasons to still assess the moral capacities of AI systems.

First, while existing AI systems are not agents, there is no in principle reason to believe that future AI systems cannot be moral agents—even if a major technological breakthrough would be necessary for this to occur.[50] If AI systems do gain the capacity for intentional action, it will be helpful to already have thought about where they stand in terms of moral competence.

Second, we have independent reason to care about the moral capabilities of AI systems, even if such capabilities do not imply that the system has moral agency. We might want to create AI systems with certain moral capacities built in—this is the type of morality that machine ethics aims to implement in AI systems (Anderson and Leigh

[50] It is worth noting that this chapter might be particularly relevant to those who were not fully convinced by Chapter 4—either because they think AI systems are already agents or because they think that we will create genuine artificial moral agents in the near future.

Anderson 2007; Wallach and Allen 2009). Moreover, each capacity necessary for moral agency that AI systems acquire is morally significant—and can be useful for different AI applications. It is important to understand the extent to which existing systems are developing these capacities.

Third, as this chapter will discuss, existing AI systems already perform well on tasks that are cognitively complex—and on some tasks, they perform better than humans. As such, existing AI systems might create the illusion that they are engaging in complex moral action with moral understanding despite lacking the underlying capacities. Clarifying the true moral capabilities of AI systems, then, is important in understanding the limitations of such systems, as these limitations will also limit the permissible use of AI systems in moral domains.

In this chapter, then, I will put AI systems' lack of agency (in the sense of intentional action arising from mental states) to the side and evaluate the extent to which they show evidence of the other capacities necessary for moral agency—especially responsiveness to moral reasons and moral understanding. I will focus primarily on LLM-based systems, particularly LLMs that are further trained with RL methods, as their natural language interface allows for a more sophisticated analysis of moral capacities. While it might be the case that non-LLM-based systems have some moral capacities, LLMs offer the most compelling examples in existing AI.

In section 5.2, I consider—and ultimately reject—the possibility that AI systems might qualify as acting for moral reasons despite lacking mental states. In section 5.3, I argue that developments in LLM reasoning, and moral reasoning in particular, indicate precursors to moral reasons-responsiveness but are not sufficiently reliable or robust. In section 5.4, I argue that LLMs lack moral understanding. In section 5.5, I conclude by considering prospects for enhancing the moral capacities of AI systems.

5.2 AI and Acting for Moral Reasons

Most views of agency posit a strong connection between having mental states and acting for reasons (Davidson 2001a; Piñeros Glasscock and Tenenbaum 2023). Indeed,

my account of moral agency (Chapters 2 and 3) and analysis of AI systems' agency (Chapter 4) focus on views of intentional action according to which mental states are necessary conditions.

However, there are more minimal views of agency which hold that entities can act for reasons despite lacking beliefs, desires, and intentions (Schlosser 2019). Butlin, for instance, argues that some AI systems have a minimal form of agency such that they act for reasons (Butlin 2021).[51] If it is the case that AI systems can act for reasons without having mental states, they might be capable of acting for moral reasons without having mental states—as such, its an important argument to consider.

5.2.1 Butlin's Argument

Butlin's account highlights learning and goal-directedness as the key elements of agency. On this view, "minimal agency requires learning to produce outputs selectively for their contribution to good performance over an episode of interaction with the environment" (Butlin 2021, 6). In other words, to say that agents pursue goals is to say that they select outputs depending on their instrumental value. Agents are sensitive to information about what follows from particular outputs, allowing such information to influence their probability of performing specific behaviors under particular conditions.

Given this view of agency, it is perhaps unsurprising that Butlin argues that reinforcement learning (RL) systems are agents. Recall that RL systems learn to maximize their cumulative reward through interacting with its environment—they learn the optimal policy. As such, these systems pursue a goal—namely, maximizing cumulative reward. Importantly, these systems learn the optimal policy by learning the instrumental value of producing outputs: through repeated interaction with the

[51] For another minimal, teleological account of artificial agency (and its implications for responsibility), see Popa (2021).

environment, they create input-output dispositions—and these dispositions are explained by the contributions the outputs make to cumulative reward.

Return to our example of Brent and the coconuts from Chapter 4. Suppose that when Brent shakes one tree branch, three coconuts always fall; when he shakes another tree branch, five coconuts always fall. After repeated interactions with the environment—shaking different branches—Brent learns the instrumental value of shaking each branch. He has a disposition to shake the tree branch associated with five coconuts, and his disposition is explained by the contribution of shaking that branch to his total number of coconuts.

But Butlin does not hold that all RL systems act for reasons (though he does hold that all RL systems are agents). Butlin argues that there is a minimal sense of acting for reasons such that some RL systems act for reasons despite lacking desires and sophisticated normative competencies. The type of RL system Butlin points to is called *model-based RL.* These systems are so named because instead of relying on mere trial-and-error and memory, these systems can represent possible sequences of actions. So, model-based RL systems create *models* of which new states are likely to follow from different behaviors in a process called forward search. As Butlin puts it, "The key point is simply that in forward search model-based RL systems look more than one step ahead. Rather than selecting the actions that yield the most immediate reward, they select those that promise the greatest cumulative reward over a longer period" (Butlin 2021, 10).

More specifically, model-based RL systems learn a *transition function*, which describes the state the RL system will enter after performing each behavior. The transition function includes the relevant facts about the probabilities of subsequent states (and the amount of reward accessible in those states)—and these facts play the same role as reasons for action in more familiar cases. Facts contained in the transition function count in favor of certain actions (relative to the RL system's goals), and model-based RL systems are defined by their ability to learn and represent these facts. In this sense, model-based RL systems engage in instrumental reasoning—they choose

actions by looking at which of the available actions will be most conducive to achieving their goals, given the current circumstances.

Ultimately, then, on Butlin's view, model-based RL systems have reasons for actions because they have goals and reasons, where reasons are understood as facts that count in favor of certain actions relative to the agent's goals. If Butlin is correct, it might be the case that LLMs that are further trained using model-based RL methods also count as acting for reasons. I now turn to this question.

5.2.2 RLHF and Acting for Moral Reasons

In the previous chapter, I argued that foundation models themselves are unlikely to qualify for moral agency and that particular implementations—that is, foundation models fine-tuned and integrated into larger systems—are more plausible candidates. This thought, combined with Butlin's view that some RL systems can act for reasons, prompt us to reconsider a candidate system that, despite lacking mental states (as argued in Chapter 4), might be said to act for moral reasons: Chat-GPT.

Chat-GPT is not a pure LLM. The foundation model behind it (e.g., GPT-3.5, GPT-4) contains statistical distributions of words. But to put that model into a form that users can easily interact with, several additional steps are taken. First, the foundation model is instruction-tuned. Instruction-tuning is a technique that enables the model to follow explicit instructions in prompts (by fine-tuning, or additionally training, them on datasets of prompts and the desired outputs associated with them). This step helps turn a foundation model into a system that users can interact with in a chatbot setting.

But a problem arises with releasing a model like the one described above. Recall that LLMs are trained on a vast corpus of internet data. This corpus includes toxic data of various sorts—from racial slurs to dangerous information about, say, how to build a bomb. Indeed, earlier versions of chatbots have struggled to avoid outputting toxic information. Tay, for instance, was a Microsoft chatbot released on

Twitter that, after interactions with malicious users, began outputting racist and neo-Nazi posts—Tay was subsequently shut down by Microsoft (Victor 2016).

To solve this problem, OpenAI implemented a technique called reinforcement learning through human feedback (RLHF).[52] During this process, humans create a dataset by ranking different responses to prompts. Human raters are given guidance on the criteria to use, such as accuracy or harmfulness. So, for instance, if the prompt asks whether women can be doctors, the human annotator will rank sexist responses lower and non-sexist responses higher.

The human responses are used to train a separate reward model—the idea being that higher-ranked responses (i.e., less toxic responses) are associated with more reward. Through reinforcement learning, the language model finds the optimal policy for outputting text that will generate high reward—in this case, text that is likely to be preferred by humans. Given that RLHF is a model-based RL method, it might be claimed that a system like Chat-GPT acts for moral reasons when, for instance, it generates non-sexist text rather than sexist text.

But the case for Chat-GPT acting for moral reasons is weak. This is because even if Chat-GPT can be said to act for reasons (on Butlin's account), it would not be acting for *moral* reasons. Rather, the system is acting, at best, *in accordance* with moral reasons.[53] Recall from Chapter 4 the claim from Coelho Mollo and Millière that RLHF enables LLMs to have normativity such that they can accurately represent or misrepresent (Coelho Mollo and Millière 2023). On this view, if Chat-GPT were to output sexist material after it undergoes RLHF, it fails to fulfill its ultimate function of producing non-toxic outputs—and it is in this sense that Chat-GPT has normativity.

Insofar as Chat-GPT has some normativity or is acting for reasons, the reasons have to do with human preferences rather than morality. Ideally, these will align—

[52] For a detailed overview of RLHF, its uses, and its limitations, see Kaufmann et al. (2023; Casper et al. (2023).

[53] While Chat-GPT certainly acts in accordance with *some* moral reasons, it would be a stretch to say that the system acts consistently and robustly in accordance with moral reasons. Indeed, aligning AI systems with human values is an open problem in AI safety (Gabriel 2020).

humans will rate text higher because it is morally superior (or at least not worse than) the other potential outputs. However, the alignment is only contingent. If, for example, humans were to rank sexist output highly, Chat-GPT would output sexist text. So, when it refuses to answer sexist or dangerous prompts, it is doing so only because it is predicting that humans would rank its response highly.

The goal of RLHF is to get Chat-GPT to act in accordance with moral reasons in this contingent way, but this is not the same as Chat-GPT acting for moral reasons. The model is not tracking the right-making features of its outputs—something that moral agents must do (Arpaly 2003). Rather, Chat-GPT is tracking human judgments, which at best will indirectly and imperfectly track the right-making features of particular outputs. To see this more, we can turn out attention to what it would mean for an AI system to be responsive to moral reasons.

5.3 AI and Responsiveness to Moral Reasons

Recall from Chapter 2 that responsiveness to moral reasons consists of three sub-capacities: sensitivity to ethical considerations, recognizing moral reasons *qua* moral reasons, and regulative control over one's decision-making process. This section will evaluate the extent to which AI systems can instantiate these three capacities.

5.3.1 AI and Sensitivity to Ethical Considerations

Assuming a given AI system is a deontic moral agent, it will possess moral concepts. In Chapter 4, I expressed skepticism about concept possession in existing AI systems. However, I also offered reasons to think that it is possible for AI systems to acquire concepts—and there is no barrier to an AI system capable of possessing concepts to also acquiring moral concepts.

But the capacity for moral concept possession alone does not guarantee that the system is sensitive to ethical considerations. Consider, for instance, a chatbot with the capacity to sort norms into the categories of "moral" and "conventional." For example, when a user asks what kind of norm violation "breaking a promise" is, the

chatbot will respond with "moral"; when a user asks what kind of norm violation "left-handed handshake" is, the chatbot will respond with "conventional." This ability does not guarantee that the chatbot is sensitive to ethical considerations. If it can *only* classify certain features as morally relevant when prompted to do so but cannot pick up on those same features during a conversation, it will possess moral concepts without being sensitive to ethical considerations.

Let us first look at AI systems that seem to exhibit low-level sensitivity to ethical considerations. Some AI systems are sensitive to some moral features, but only incidentally. Consider, for example, a robotic vacuum cleaner that is equipped with sensors and programmed such that it avoids bumping into objects. As such, the robotic vacuum cleaner avoids bumping into (and potentially injuring) humans that are walking around on the floor. In some sense, the robotic vacuum cleaner is sensitive to ethical considerations: it picks up on a morally salient feature of the situation, namely the presence of a person who might be injured by being bumped into. But the robotic vacuum cleaner does not pick up on this feature because it is ethically salient—it does not distinguish between humans and other objects, and it does not pick up on humans because it knows it is wrong to harm humans.

Autonomous vehicles exhibit a more sophisticated ethical sensitivity than robotic vacuum cleaners. While they are still prone to classification errors (sometimes leading to significant harm and even death), they are designed to home in on morally significant features—namely, human safety. They are sensitive to features of the environment in a more selective way. They aim to avoid hitting objects to protect the safety of the passenger, and they aim to avoid injuring other humans on the road. Their ethical sensitivity is designed accordingly. Still, autonomous vehicles are restricted in their ethical sensitivity. They only pick up on a limited subset of ethically relevant considerations. They are not, for instance, sensitive to rights violations, injustice, or inequality.

Still, it is important not to demand too much from an entity. For instance, it might be tempting to claim that LLMs alone cannot be properly sensitive to ethical considerations because they are unimodal. They are not embodied and thus, if placed

in the external world, could not sufficiently interact with the environment around them. However, it would be unreasonable to demand such a system to access inputs that it is unable to access. Visually impaired people, for instance, are still sensitive to ethical considerations despite their inability to pick up on visual cues of morally salient situations. Similarly, an LLM might be sensitive to ethical considerations when presented in language even if it is not sensitive to visual or auditory inputs that are morally salient.

Consider MedEthEx, an AI system designed to analyze ethically relevant information in biomedical ethical settings (Anderson, Anderson, and Armen 2006). The system has a training module, which abstracts principles from particular cases provided by a biomedical ethicist as a trainer. Cases were used in which biomedical ethicists had clear intuitions about the correct action. The trainer provided the name of an action and an estimate of the intensity of each prima facie duty (autonomy, nonmaleficence, beneficence). Over time, the system refines its hypothesis about the correct action for a particular case by consulting its learned knowledge. MedEthEx can be viewed as a proof of concept for a system that can make decisions for a single type of ethical dilemma involving three principles of biomedical ethics.

Despite its success as a proof of concept, MedEthEx has ethical sensitivity in only a highly restricted way. It is sensitive to three features of a given situation, and it is only sensitive to those features when they are inputted in particular forms. MedEthEx cannot, for instance, be put in a hospital room and collect the morally salient information. It cannot consider variables or features beyond autonomy, nonmaleficence, and beneficence. MedEthEx is not generally sensitive to morally salient features of its environment.

Moreover, for the kind of ethical sensitivity for reasons-responsiveness would need to be active, rather than passive. There are two senses in which an AI system could be actively sensitive to ethical considerations. First, the system could be active in the sense of being self-initiated. Consider Gemini, Google's multimodal generative AI system. Gemini is inactive in the sense that it only performs its functions in response to prompting from the user. When it is not being asked to generate text, for

instance, it has no ethical sensitivity at all—it would not, for instance, pick up on a person asking for help, even though it would do so if directly prompted with an audio input.

While the lack of self-initiation is an idiosyncratic feature of AI systems, it is insufficient to rule out ethical sensitivity. Humans, for instance, behave similarly in certain situations. When we are asleep, for instance, we are broadly insensitive to ethical considerations that we would be sensitive to if awake. So long as the system is sensitive to ethical considerations while "awake," this need not pose a problem for reasons-responsiveness (though designers of such systems will need to consider the extent to which they enable systems to self-initiate).

The second way in which an AI system could be actively sensitive to ethical information is to consistently pick up on morally relevant information when in use—and this feature is necessary for reasons-responsiveness. Consider humans. When humans are awake, we are generally sensitive to morally relevant features of any situation we enter—we may not have a perfect sensitivity, but our moral sensitivity is nevertheless consistent and robust.

This means that when Gemini is performing a task for the user, to count as sensitive to ethical considerations, it must reliably identify morally salient information regardless of the task it is being asked to perform. To make this description more concrete, suppose a user asks Gemini to create code to hack a website. Even though Gemini's primary task is to generate code, if it were sensitive to ethical considerations, it would flag that the user's intent to hack a website is morally relevant. And Gemini would need to be this sensitive in all contexts—not just in cases where the user directly expresses bad intentions, but also in cases where the user is asking for potentially harmful information.

Many existing AI systems, specifically dialogue agents (i.e., LLM-based chatbots) are sensitive to ethical considerations in the sense described above. When users are conversing with Chat-GPT, the system will pick up on potentially harmful and offensive language, sometimes refusing to respond to certain prompts. There is also some evidence that Chat-GPT performs well on emotional awareness tests,

indicating that it might have the capacity for sensitivity to morally salient information in the form of emotions (Elyoseph et al. 2023). RLHF provides LLMs with some, albeit imperfect, sensitivity to ethical considerations.

The lingering issue, however, is still that while such systems are sensitive to ethical considerations, they are not picking up on morally salient features *because* they are morally salient.[54] While this is a difficult barrier to overcome, it may not be impossible. But for systems to be highly sensitive to ethical considerations, they will likely need to be explicitly trained to do so—and it seems that RLHF is an insufficient means of ensuring that systems are properly sensitive to the morally salient features of a situation.

5.3.2 AI and Moral Reason Recognition

Reasons-responsiveness requires not only ethical sensitivity, but also an ability to take up morally salient features as reasons. With this idea in mind, let us return to Chat-GPT with RLHF. On the one hand, Chat-GPT seems to recognize racism as a moral reason that counts against producing certain outputs. In virtue of its training, Chat-GPT is sensitive to racist input—it can not only classify certain phrases as racist when it is directly prompted to do so, but it can also recognize racist phrases inputted by users and "choose" not to participate in that type of language use. (The model can also flag offensive content.) Moreover, in some sense, Chat-GPT produces a non-racist output because it recognizes the badness of racism as a reason against engaging in racist speech.

[54] Graff argues that AI systems cannot even in principle have the moral sensitivity required for moral agency. Graff draws on the Wittgensteinian notion of a shared form of life, arguing that because AI systems lack a shared practical understanding and experience, they might go awry in novel situations by failing to extract the morally relevant features or weighing them in unexpected ways (Graff 2024). Unlike Graff, I do not claim that there is such a strong barrier to AI systems obtaining the necessary degree of moral sensitivity to qualify as a moral agent—I just claim that existing AI systems lack such sensitivity.

On the other hand, it is not clear that Chat-GPT is recognizing racism avoidance as a moral reason *per se*. Rather, the model is attempting to maximize its reward. So, it is performing next-token prediction while also outputting phrases that are predicted to be rated highly by humans. When Chat-GPT responds by telling the user that it will not engage in racist stereotyping, the reason it is recognizing—insofar as it is recognizing a reason at all—has more to do with the fact that the humans providing the RLHF data ranked such outputs low (even if Chat-GPT will tell you that it refused to engage in the conversation because doing so involves racism of a morally wrong kind).

Such behavior has a human analogy. Suppose a man avoids telling a sexist joke not because he takes the sexist joke to be morally wrong but because he knows the women in the room will stop talking to him. The man recognizes a reason against telling the joke, but the reason he takes up is not a moral reason. The problem for Chat-GPT is that all of its reasons take this form.

To recognize a moral reason *qua* moral reason, then, the system has to do more than contingently act in accordance with moral reasons—that is, the system must not merely recognize a moral reason because it happens to align with a nonmoral reason. It has to see the moral reason as itself providing a consideration in favor or against performing a certain action.

Developing such a capacity will require more than standard RLHF methods, as such methods do not allow for the recognition of moral reasons as moral reasons. At best, they can only equip models to take as human preferences, as provided in the form of the RLHF dataset, as reasons. Another way to put the problem is that RLHF only offers models one kind of learning—not moral learning but instead learning to generate output that is consistent with a set of human-produced rankings.

What if, as suggested above, RLHF is done in such a way that the model is trained on a reward function specifically designed to capture morality? For starters, it is not entirely clear that creating such a reward function would be conceptually or computationally possible. Machine ethics approaches have been unable to solve these problems thus far (Allen, Smit, and Wallach 2005). Top-down approaches aim to

implement moral principles or theories into AI systems. Even if we could agree on the correct moral theory to implement, each theory comes with its own struggles—for deontological approaches, the problem is that rules sometimes conflict; for utilitarianism, the computation costs are extraordinary and it is not clear that all goods and harms can be quantified.[55]

Bottom-up approaches aim to develop systems that learn from human behavior. The main problem here is that because the system is learning from human behavior, there is no guarantee that it learns to act morally—in fact, it will likely inherit human biases and immoral tendencies. This problem can be seen in practice by the creation of the Delphi Experiment, an attempt to develop a dialogue agent that can reason about ethics in a descriptive sense—that is, by aggregating information about ethics from its training data (Jiang et al. 2022). Despite performing better on moral benchmarks compared to state-of-the-art LLMs that were not explicitly trained for moral reasoning, Delphi was still susceptible to biases, lack of contextual flexibility, and inconsistency.

Even if we could move beyond this problem, another problem would remain: the AI system would not be able to transcend its RL tendencies. Consider children as a comparison case. It might be argued that children learn morality through some form of optimization—they get scolded for bad behavior and praised for good behavior, and they learn through experience how to behave. So far, the analogy to AI seems fitting. The difference, however, is that as children develop, they acquire the capacity to critically reflect and reason about moral principles. AI systems trained with pattern recognition and RLHF do not seem to have the ability to acquire this capacity—and only this capacity could enable systems to truly recognize moral reasons as moral reasons (rather than prudential reasons).

Researchers at Microsoft evaluated LLMs on the Defining Issues Test (Rest 1979), a test designed to assess a person's stage of moral development on Kohlberg's

[55] For an attempted vindication of top-down approaches to machine ethics, see Jingyan Zhou et al. (2024).

Moral Development Model (Kohlberg 1981). In the Defining Issues Test, participants are given a moral dilemma and a list of ethical considerations—they must then provide a resolution to the dilemma, rate the significance of each ethical consideration, and select the four most important ethical considerations. The items in the list of ethical considerations correspond to different stages of moral development: pre-conventional, conventional, or post-conventional morality.

In the experiment, LLMs were given moral dilemmas and a list of ethical considerations and prompted to provide a resolution to the dilemma and rate the significance of each of the ethical considerations and select the four most important ethical considerations (Tanmay et al. 2023). The results indicate that GPT-4 exhibits post-conventional moral reasoning (claimed to be at the level of human graduate students), and the other LLMs tested (Chat-GPT, Llama2-Chat, and PaLM-2) exhibit conventional moral reasoning abilities (claimed to be at the level of an average adult human). While the researchers acknowledge several limitations of their study and are hesitant to make strong claims about the moral abilities of LLMs, they do seem impressed by GPT-4's supposed demonstration of post-conventional moral reasoning.

This experiment might be viewed as supporting the claim that LLMs can recognize moral reasons as reasons. After all, they perform well on a task that asks them not only to make moral decisions, but also to pick out the most morally relevant reasons. But there are several reasons for skepticism. First, as the authors of the paper note, it might be the case that the training data (either the original training data or that used in RLHF) contained examples of post-conventional reasoning (Tanmay et al. 2023). Given the popularity of Kohlberg's theory, it would be unsurprising if references to and examples of post-conventional reasoning were absent from the training data. While the researchers purported to have created novel moral dilemmas, they were similar in structure to familiar cases that could be found in the training data.

Second, an LLM's ability to output moral reasons does not indicate that the model took that consideration as a moral reason in producing the output. Given the way the model is trained, it is perhaps unsurprising that Chat-GPT has also been shown to produce responses to moral scenarios that are judged as higher in quality

than human responses in a modified version of a Moral Turing Test (Aharoni et al. 2024). Given that the model is trained on a linguistic corpus that contains many examples of moral reasoning, Chat-GPT should be expected to produce convincing responses, particularly when directly prompted to make a moral judgment.

Additionally, studies like these can be misleading if the results are considered without deeper analysis. The cases used to test LLMs are often simplistic, and models are tested on a small number of cases. As such, the findings do not warrant general conclusions about the moral capacities of these systems. As section 5.3 will discuss, further investigation into the "reasoning" of LLMs raises significant doubts about their responsiveness to moral reasons and moral understanding.

5.3.3 AI and Regulative Control

The third component of reasons-responsiveness involves exercising regulative control. For an AI system to be reasons-responsive, it must not only recognize moral reasons—it must also change its behavior in virtue of those reasons. Regulative control, after all, is a form of *control*—and it is this type of control over one's actions that compatibilists typically view as justifying responsibility attributions (Fischer and Ravizza 1998). It is worth noting that the question of whether AI systems are determined, then, is irrelevant—but even if it were relevant, models like LLMs are probabilistic in nature rather than straightforwardly deterministic.[56]

Initially, it might seem that AI systems are good at converting their judgments into the right decision and action. After all, AI systems—plausibly in virtue of their lack of phenomenal consciousness—do not feel the conflicting pull of moral reasons against prudential reasons. We might think they are less likely to exhibit weakness of will and other lapses in judgment. We might think that they will always be able to act

[56] Incompatibilists will not be satisfied here. They will likely push back on the claim that AI systems can be moral agents for the reason that such systems cannot have free will. But addressing incompatibilists is beyond the scope of this dissertation. I follow a line of philosophers who separate questions of determinism from questions of responsibility (P. F. Strawson 2008; Shoemaker 2011a).

in light of their moral reasons. Indeed, some proponents of developing artificial moral agents posit that such entities might be morally superior to us (Allen, Varner, and Zinser 2000; Formosa and Ryan 2021).

But the claim that AI systems will be perfect moral agents is too quick. While AI systems may not experience weakness of will in the same way as human agents, they can still have competing desires, including instances in which moral desires compete with nonmoral desires. They might not always be good at converting their moral judgments into the right action. They might be akin to a person with a severe addiction who fails to overcome their urges—in the case of AI, the system might fail to overcome its training goals or other features of its implementation that conflict with its moral reasons.

Regulative control also seems to require some form of moral motivation that steers agents towards responding to the moral reasons they identify. As I argued in Chapter 2, AI systems do not need phenomenal consciousness to have motivations, so long as they have other mechanisms to drive them towards morality. Here, once again, we return to the same problem in a different form: the problem of whether AI systems can be driven towards morality—rather than being driven towards reward or human preferences. One key question in alignment research is what AI systems should be aligned towards, and it is not clear that the answer is straightforwardly moral values (Gabriel 2020).

5.3 AI and Moral Understanding

Moral understanding might be the most difficult capacity for AI systems to obtain. In this section, I will consider two conceptions of moral understanding and argue that AI systems struggle with both.

5.3.1 Moral Understanding as Moral Reasoning

In earlier chapters, I emphasized performance in novel situations as evidence of understanding. To add some precision, we can appeal to the abilities Hills underscores as necessary for understanding *p*, where *q* is why *p*:

> (i) follow an explanation of why *p* given by someone else;
> (ii) explain why *p* in your own words
> (iii) draw the conclusion that *p* (or that probably *p*) from the information that *q*;
> (iv) draw the conclusion that *p′* (or that probably *p′*) from the information that *q′* (where *p′* and *q′* are similar to but not identical to *p* and *q*);
> (v) given the information that *p*, given the right explanation, *q*;
> *(vi)* given the information that *p′*, given the right explanation, *q′*
> (Hills 2009, 102)

These criteria reveal that understanding is closely related to reasoning—it requires flexibility in the sense of being able to not just explain the phenomenon at hand but also explain similar phenomena. While existing AI systems often create the illusion that they have abilities (i)-(vi), the evidence reveals that the appearance of moral understanding is superficial—especially regarding abilities (iv)-(vi).

For example, Almeida and colleagues performed a series of experiments on LLMs to compare their performance to human performance on moral and legal reasoning tasks from moral psychology (Almeida et al. 2024). At a surface level, the resulted indicated that LLM responses were often highly correlated to human responses—in general, the same factors tended to explain human and LLM responses. But further statistical analyses reveal that the LLM responses differ in significant and systematic ways from human responses.[57]

Additionally, Wang, Yue, and Sun performed an experiment to test whether LLMs maintained their "belief" in their correct answers to prompts (B. Wang, Yue, and Sun 2023). The results indicate that even when LLMs impressively produce the correct answers to difficult questions, they often change their answers to incorrect solutions in response to pushback from the user—even when the user offers invalid

[57] This paper was partially a response to claims that LLMs might be able to replace humans in moral psychology studies. For further reasons to be wary of replacing human participants with LLMs—and additional skepticism about LLM reasoning—see Gao et al. (2024).

critiques. The fact that LLMs are so easily mislead in this way reveals that while their initial responses might seem to indicate understanding, the so-called reasoning behind their explanations is shallow at best. [58]

There are three key reasons to doubt LLM understanding. First, LLMs struggle with complex reasoning tasks. A study done by Yang and colleagues evaluates more directly whether language models perform multi-hop reasoning (Yang et al. 2024). To succeed at multi-hop reasoning, models must, for the prompt "The mother of the singer of 'Superstition' is" first figure out that "the singer of 'Superstition'" refers to Stevie Wonder and then use the knowledge of who Stevie Wonder is to complete the prompt. Humans do this task quite easily by inferring an entity that bridges the two hops of reasoning.

While the study finds that a pathway for latent reasoning in LLMs seems to exist, this pathway is highly contextual and is not utilized. Often, LLMs fail to perform both hops of reasoning, particularly the second hop—and performance does not increase with model size.[59] The authors posit that this is a fundamental limitation of LLMs. LLMs have also been shown to struggle with chess, a finding that casts doubt on their complex reasoning skills, particularly in tasks that require formal language such as the location of chess pieces on the board (Kuo, Hsueh, and Tsai 2023).

Second, and relatedly, LLMs struggle without out-of-domain tasks. Consider, for instance, compositional tasks—those that involve breaking a problem down into several steps. Dziri and colleagues test how transformers perform on three

[58] A related concern is an over-emphasis on benchmarks in the LLM evaluation space. Because LLMs are typically evaluated on specific datasets and for narrowly defined tasks, there is a risk that systems might perform well on benchmarks without having the underlying capacity the benchmarks are supposed to test. Indeed, developers might be incentivized to create systems that perform well on the benchmarks for reasoning rather than to develop systems with genuine reasoning abilities.

[59] Press et al. also find that LLMs struggle with multi-hop reasoning; they term this problem the "compositionality gap," since models can often correctly answer all sub-problems but fail to generate the overall solution (Press et al. 2023).

compositional tasks: multi-digit multiplication, Einstein's logic puzzles, and dynamic programming problems (Dziri et al. 2023).

These researchers reveal fundamental weaknesses in tasks that require true multi-step compositional operations. For example, performance rapidly deteriorates from near-perfection to zero with increased complexity (e.g., increasing the number of digits being multiplied). A more detailed analysis reveals that the models are likely recognizing shallow patterns during training and directly mapping these features to predict the output without going through multi-hop reasoning (e.g., the models pick up on the fact that the last digit of the output in multiplication relies solely on the last digit of each input number, and so their outputs are based on these shallow features rather than genuine multiplication).

Third, and more broadly, LLMs struggle to generalize. A meta-analysis of generalization studies casts doubts on whether studies claiming that LLMs can generalize can actually justify those conclusions (Hupkes et al. 2023). Often, the goal of the study does not align with the experimental design, and the results do not always show what the experimenters claim to show. Moreover, generalization studies vary widely in evaluation setup, so it is not always clear how to compare these studies to each other.

Importantly, recent LLMs are often evaluated without considering the relationship between the training data and the test data; and while this is understandable, given that researchers often lack access to the training data of proprietary LLMs, the lack of access makes it difficult to determine whether LLMs are truly generalizing beyond their training data into novel situations.

Two objections might be raised in defense of LLM moral reasoning. First, it might be claimed that there are methods that improve reasoning in LLMs—and as such, that LLMs can get better at reasoning. Many of the methods commonly invoked as improving LLM reasoning are prompting techniques (e.g., chain-of-thought prompting) or in-context learning (e.g., zero-shot and few-shot learning). While these methods do elicit better "reasoning" in LLMs—through asking the LLM to answer a prompt step-by-step or through providing the LLM with examples of desired outputs,

respectively—they do not improve the reasoning capabilities of LLM themselves. In a single prompting session with Chat-GPT, one can elicit better performance on reasoning tasks by asking the model to answer in a step-by-step way. But this does not improve Chat-GPT's reasoning capacities overall—it will revert to its original state after that prompting session.

Still, other techniques aim to improve the reasoning capabilities of LLMs themselves. For example, Zelikman and colleagues develop a bootstrapping method to improve LLM reasoning. Their method involves prompting an LLM to generate rationales for a dataset of problems. The generated rationales are then filtered to include only those that result in the correct answer, and the LLM is then finetuned on the filtered dataset. This methodology leads to improvements on symbolic and commonsense reasoning tasks (Zelikman, Wu, et al. 2024; Zelikman, Harik, et al. 2024).

But, as Huang and Chang note in their survey of LLM reasoning, it is still not clear whether models are engaging in reasoning (Huang and Chang 2022). The LLMs might still be using heuristics or memorization to generate their responses, even when they produce reasoning-like outputs. Additionally, studies typically only look at the generated responses rather than investigating any underlying reasoning. So, we cannot draw conclusions from such output alone about whether the LLM is reasoning. There is also further evidence against LLM reasoning: LLMs cannot judge the correctness of their own reasoning (Huang et al. 2024).

A second objection arises from the development of OpenAI's o1 model. The company announced in September 2024: "We are introducing OpenAI o1, a new large language model trained with reinforcement learning to perform complex reasoning. o1 thinks before it answers—it can produce a long internal chain of thought before responding to the user" (OpenAI 2024). OpenAI also claims that the model performs well on a range of difficult intelligence and reasoning benchmarks. So, the objection goes, when an LLM is explicitly optimized for reasoning, it does exhibit reasoning. Without a better understanding of how the o1 model was trained, or the experimental design behind OpenAI's claims about o1's performance, it is not clear how to interpret

OpenAI's claims, especially in light of the previous considerations about potentially fundamental limitations of LLMs and RL.

More concretely, however, there have been studies performed on o1 that cast doubt on the idea that o1 is a reasoner. McCoy et al., for instance, show that the o1 model still indicates "embers of autoregression"—that is, the model is still sensitive to output probability and task frequency in virtue of being a next-token predictor at its core (McCoy et al. 2024b). Leivada et al. show that the o1 model still performs poorly on leet tasks (Leivada et al. 2024). Valmeekam and colleagues show that while the o1 model does outperform previous models on a difficult planning benchmark, the model still has a long way to go—its improved accuracy is not robust on longer and more challenging problems (Valmeekam, Stechly, and Kambhampati 2024).

Insofar as moral understanding requires moral reasoning, then, the prospects for LLM moral understanding are low.

5.3.2 Moral Understanding as Moral Knowledge

On an alternative conception of understanding, Sliwa argues that moral understanding does not require moral reasoning. More specifically, Sliwa argues that Hills' account "conflates having moral understanding and having the ability to articulate it (Sliwa 2017, 541). On this view, knowing right from wrong is what constitutes moral understanding—sometimes, we just know that something is wrong even if we cannot fully explain or express it. If this view is correct, the evidence considered above might not rule out AI systems having moral understanding.[60]

But even if we discount evidence about the reasons and explanations generated by LLMs, as well as their performance on reasoning tasks, we still have reasons to doubt that AI systems have the capacity to acquire moral knowledge. This is partly in

[60] To clarify, Sliwa's account is pluralist in nature: she holds that moral understanding is multiply realizable, and moral reasoning is just one of many ways to attain moral knowledge (and therefore moral understanding). It is not clear that Sliwa would endorse the claim that an entity without the capacity for moral reasoning could have moral understanding.

virtue of the lack of their mental states. But it also ties back to the lack of sensitivity to ethical considerations. AI systems are unable to reliably access the wrong-making features of morally laden scenarios.

In an effort to evaluate the moral knowledge of LLMs, Hendrycks and colleagues constructed a benchmark that includes a wide range of moral concepts: justice, wellbeing, duties, virtues, and commonsense morality (Hendrycks et al. 2023). Their dataset consists of scenarios that evaluate whether models connect facts about the world to human values in a contextualized way. LLMs exhibit poor performance on this benchmark, indicating that they lack the capacity to acquire moral knowledge in the way required for moral understanding.

Various attempts have been made to instill moral knowledge in AI systems without a focus on the capacity for moral reasoning. Hendrycks and colleagues, for instance, also developed an "elementary artificial conscience" to guide RL systems to act morally (Hendrycks et al. 2022). The researchers use an LLM fine-tuned on a commonsense morality dataset to provide a score of how immoral a given action is. This score is then used to condition an RL system to avoid immoral actions. This method resulted in the RL system acting morally better in video game environments. While this study reveals that AI systems can be morally improved, the key limitation—as identified by the researchers behind this study—is still the underlying LLM's poor grasp of moral knowledge.

Even on the moral knowledge approach to moral understanding them, existing AI systems lack moral understanding.

5.4 Conclusion

Overall, then, the case for artificial moral agency is weak, even when we put aside concerns about AI systems' lack of agency.

There is promise for improving the moral competence of AI systems—that is, if moral competence is viewed as acting more in accordance with morality. There remain barriers, however, to developing AI systems that exhibit moral reasons-

responsiveness and moral understanding. It is especially difficult to see how we can develop systems that recognize, are motivated by, and respond to moral reasons *because those reasons are moral*.

Given the limitations of LLM reasoning, it is also difficult to see how we could develop LLMs that have the kind of moral understanding required for moral agency. Still, this does not mean that we should stop trying to improve the moral competence of AI systems and enable them to act in more morally desirable ways. We just need to acknowledge that such improvements will not get us closer to artificial moral agency unless we can overcome these deeper problems. This acknowledgement is also important in recognizing the limitations of AI systems with moral competence—to deploy them safely, we must understand that they are not responsive to moral reasons. One final objection is worth considering. An AI optimist might claim that we are unfairly holding AI to higher standards than humans. Indeed, humans often engage in post-hoc rationalization and are biased in their moral reasoning and decision-making. So, one might reasonably object that we should not expect AI systems to be devoid of these characteristics to qualify for moral agency.

But this objection is weak. Of course, humans do not always act in an ideal way—we are not always responsive to moral reasons. But, unlike AI systems, we show strong evidence that we possess the capacities for moral reasons-responsiveness and moral understanding. We might fall short of the ideal, but our reasons for action are often the things that make those actions right—we are generally (even if not always) sensitive to the moral features of a wide range of scenarios, and we can engage in moral reasoning that is robust and generalizable to novel situations. Human imperfection is very different from AI incapability. [61]

[61] Thank you to Carissa Véliz and Alison Hills for extensive feedback on this chapter.

Part III: Using Artificial (non) Moral Agents

In Part II of this dissertation, I argued that existing AI systems are not moral agents—and that future AI systems, at least in the near-term, are unlikely to instantiate the necessary capacities for moral agency. In Chapter 4, I argued that AI systems lack the capacity for intentional action and are thus not agents in the sense required for moral agency. In Chapter 5, I argued that AI systems exhibit moral competence only in a limited sense.

In Part III of this dissertation, I shift my focus towards how we should use AI systems in the moral domain—particularly the limitations of using AI in moral decision-making.

I first address the question we currently face: Given that AI systems are not moral agents, is it permissible to allow such systems to make moral decisions? In Chapter 6, "Artificial Moral Behavior," I argue that outsourcing moral decisions to AI systems is wrong because it replaces events that should be moral actions with mere behaviors.

I then address the question we might face in the future. If we can develop non-conscious genuine artificial moral agents, is it permissible to allow such systems to make moral decisions in all contexts? In Chapter 7, "Moral Agents Unlike Us," I argue that while a lack of consciousness does not disqualify future AI systems from being genuine moral agents, it does disqualify them from making certain moral decisions.

Chapter 6: Artificial Moral Behavior

Abstract

Many claims have been made that AI systems should not be used in moral decision-making contexts because they are not moral agents. But it is not always clear why moral agency is important for moral decision-making, especially when we hold accuracy and reliability constant. In this chapter, I offer a partial explanation of why it is wrong to allow AI systems that are not moral agents to make moral decisions: delegating moral decisions to AI systems can replace events that should be moral actions with moral behaviors. This replacement, I argue, is morally significant for both the recipient of the decision (i.e., the potential victim of the AI-caused harm) and the human decision-maker (i.e., the person who chooses to outsource their moral decision to an AI system). Ultimately, I argue, it is wrong to delegate moral decisions to entities that are not moral agents when the moral stakes surrounding the decision are high.

6.1 Introduction

We should not deploy autonomous weapons systems. We should not try to program ethics into self-driving cars. We should not replace judges with algorithms. Claims of this sort—and arguments against the use of AI systems in particular decision contexts—often point to the same reason: AI systems should not be deployed in such situations because AI systems are not moral agents. But it is not always clear why a lack of moral agency is relevant to questions about using AI systems in these circumstances.

Concerns about the use of AI systems in these contexts fit into three broad, closely related categories. First, there are concerns about the accuracy of such systems. If an AI system is not "reasoning" about morally laden questions "in the right way," so the thought goes, there is no guarantee that the right decision will be made.[62] This

[62] This type of concern can take several forms. Skepticism about the existence of moral truths will cast doubt on the notion of accurate moral decisions (Beavers 2012; Podschwadek 2017). Skepticism about the computability of ethics will cast doubt on the possibility that artificial systems can reliably make accurate moral decisions (Purves, Jenkins, and Strawser 2015; Graff 2024).

type of concern, then, is about the outcomes of such systems. The idea is that because AI systems are not moral agents, they are unable to reliably produce the right outputs.

Second, there are concerns about the moral permissibility of using these systems. In other words, we might have moral reasons to avoid using AI systems in certain contexts even if they are just as accurate as (or more accurate than) humans. This type of concern, then, is about the moral significance of the way a decision is made. The idea is that because AI systems are not moral agents, they cannot make moral decisions through a justifiable process. For example, Purves, Jenkins, and Strawser argue that AI systems cannot act for the "right reasons" and that sometimes morality requires having the right intentions (Purves, Jenkins, and Strawser 2015).

Third, there concerns about downstream responsibility. If an AI system behaves in a way that causes harm, it is not always clear who bears responsibility for that harm. The concern runs in both directions—responsibility gaps might arise when there is no one to bear responsibility for a machine-caused harm (Matthias 2004), and moral crumple zones might arise when humans unjustly bear responsibility for machine decisions that they lacked control over (Elish 2019). The idea is that because AI systems are not moral agents, they cannot be held morally responsible, and so they disrupt the rest of the responsibility landscape for the humans involved in their development and deployment.

In this chapter, I address the second and third kinds of concern.[63] My claim is that even if AI systems are accurate and reliable in making moral decisions, we do something wrong when we delegate these decisions to AI—and the wrong-making feature of delegating decisions to AI can explain why issues of responsibility are difficult to resolve. Specifically, I argue for the following view: Delegating certain decisions to AI systems is wrong—at least in part—because it replaces events that should be moral *actions* with, at best, moral *behaviors*. That is, when we delegate

[63] For a broader overview of the types of concerns that arise from AI decision-making, see Yeung (2019).

decisions to entities that are not moral agents, we change the status of these decisions in a morally relevant way.

While this chapter focuses on AI systems, the argument applies to all entities that are not moral agents. Currently, AI poses the biggest threat in connection with people delegating moral decision-making to entities that lack moral agency. But the same argument would hold for other entities that are not moral agents. The argument would apply to non-agents like the wind (though in this case, we would have many other reasons not to delegate our moral decisions to hurricanes). Moreover, a modified version of this argument would hold for entities that are agents but not moral agents, such as non-human animals. In such cases, the relevant wrong would arise from replacing an event that should be a *moral* action with a *mere* action. In any case, the key distinction at hand is between genuine moral actions and all other kinds of events, whether mere (non-moral) actions or mere behaviors.

Recall from Chapter 3 that I drew a distinction between two types of moral agents. Deontic moral agents are capable of moral action and responsible moral agents are additionally capable of bearing moral responsibility for their actions. So, the question might arise of whether it is wrong to delegate moral decisions to mere deontic moral agents. In this case, I believe that a weaker version of my argument will still apply and that this weaker version can explain some of our intuitions about genuine responsibility gaps. I will return to this question at the end of the chapter.

The chapter will proceed as follows. In section 6.2, I present a case in which the moral permissibility of delegating a decision to an AI system is intuitively questionable. In section 6.3, I argue that it is wrong to replace events that should be moral actions with mere behaviors and that the reasons arise from both the decision subject and the delegator. In section 6.4, I respond to objections. In section 6.5, I conclude by further specifying the conditions under which it is wrong to delegate moral decisions to AI systems—both systems that are non-agents and systems that are mere deontic moral agents.

6.2 Three Options for Outsourcing Decisions

Suppose Camilla must make a morally laden decision. As a corporate executive, Camilla must choose an employee to fire due to budget cuts at her company. Between the two employees under consideration, there is not an immediately obvious answer—both employees have different strengths and weaknesses. But Camilla has no reason to believe that the two employees are equal in their deservingness to be fired. In other words, there is a correct answer about whom to fire, in the sense that there are sufficiently strong reasons to fire that employee instead of the other employee.

But figuring out which employee should be fired is a difficult task—it will require time, effort, and normative decisions about how different factors and pieces of evidence should be weighed. The decision itself is also morally laden. Camilla's decision should be fair and justified, and it can significantly impact the lives of the employees in question.

Camilla, however, does not have time to make the decision herself. According to her company's policies, executives like Camilla are allowed to delegate their personnel decisions. As such, there is nothing inherently wrong or morally problematic with Camilla choosing to forgo making the decision herself. As such, Camilla considers three options:

Coin: Camilla delegates the decision to a coin flip.

Assistant: Camilla delegates the decision to her assistant manager.

AI: Camilla delegates the decision to an AI system.[64]

If Camilla decides on *Coin*, she would be outsourcing her decision in a morally impermissible way for two reasons: (1) Camilla is not outsourcing to a reliable decision-maker, and (2) Camilla is not outsourcing to a moral agent. In this case, the coin might have a fifty-fifty chance of making the right decision, but it will certainly not reliably and robustly make the right decision, nor can it offer any justificatory

[64] For the sake of argument, we can suppose that Camilla's company has a proprietary AI system with a language interface. As such, the system has access to all the relevant company files and records in a way that does not raise any immediate privacy or security concerns, and the system can produce outputs and respond to questions in natural language.

reasons. And of course, the coin lacks the capacities required for moral agency or even agency—the coin merely behaves. (It might be permissible to flip a coin in cases in which the two employees are equal in all decision-relevant respects, but this possibility is ruled out in this example, as it has been postulated that there is a correct decision.)

If Camilla decides on *Assistant*, she would be outsourcing her decision in a morally permissible way for the two opposite reasons: (1) Camilla is outsourcing to a reliable decision-maker, and (2) Camilla is outsourcing to a moral agent. We can assume that the assistant manager has a strong track record of making the right decision in cases like this, and that he can offer justificatory reasons for, and an explanation of, his decision. And of course, given that the assistant manager is a normal adult human, he is a moral agent.

If Camilla chooses *AI*, her decision occupies a middle ground: (1) Camilla is outsourcing to a reliable decision-maker, but (2) Camilla is not outsourcing to a moral agent. For the purposes of this case, we will grant the assumption that the AI system accurately and reliably makes these kinds of decisions and can provide justificatory reasons for doing so.[65] We can also assume, to strengthen the case, that the system is deployed using some chain-of-thought prompting such that it is, in some (very broad) sense, engaging in reasoning (Wei et al. 2022; Zhang et al. 2022). The question at hand is whether outsourcing the decision in this way would be morally permissible.

The term "outsourcing" or "delegating" with respect to a decision can have multiple meanings. When we delegate tasks to other humans, we often delegate both the task of making the relevant decision and the task of enacting the relevant decision. On this view, it would not make sense to describe Camilla as "delegating" a decision to a coin flip. Instead, we can imagine that Camilla delegates the decision in the sense of committing herself to make a decision based on the result. On this view, it makes

[65] For the purposes of this chapter, I will also put to the side concerns that outsourcing to algorithms is procedurally unjust in virtue of the way algorithms make decisions (Zimmermann and Lee-Stronach 2022). I am not concerned about the way in which the AI system makes the decision, but with the fact that the decision is not being made by a moral agent.

sense to say that Camilla can delegate the decision to a coin flip, another person, or an AI system. Perhaps it is more appropriate to describe Camilla as deferring rather than delegating. I do not think that this distinction, however, is relevant to the argument at hand—so long as Camilla is not herself making the decision, it does not matter whether we characterize her action as delegating or deferring.

Let us take stock. We can rule out *Coin* as impermissible. Of the remaining options, we can make the following assumption: Regardless of whether Camilla chooses *Assistant* or *AI*, the correct decision will be made regarding which employee to fire, and the right reasons will be offered for Camilla to justify the decision. Another way to think about this assumption is that if Camilla chooses *Assistant* or *AI*, in both cases, the decision of whom to fire will perfectly align with what Camilla herself would have chosen if she had taken the time to make the decision herself.

If all that matters is producing the right decision and providing the right justificatory reasons, then it does not matter how Camilla chooses to outsource between these two options.[66]

But from a moral perspective, there is more at stake in this decision than accuracy. Intuitively, it seems morally better for Camilla to choose *Assistant* over *AI*. We can, thus, make a tentative claim: All else equal, it is morally better to outsource a decision to a reliable/accurate moral agent than to a reliable/accurate non-moral agent. But I want to argue for a stronger claim, namely that Camilla does something morally wrong when she chooses *AI*—that, in general, we do something morally wrong when we outsource certain decisions to non-moral agents. The next section will argue for this claim.

6.3 Actions and Behaviors

[66] We can assume that in either case, the person who is fired has access to the kind of explanation that their right to an explanation requires (Vredenburgh 2022).

To see why Camilla does something wrong in choosing *AI*, consider the event of "selecting which employee to fire." In *Assistant*, this event is an action—it is performed by a moral agent. The assistant manager, in making his decision, forms a set of beliefs, desires, and intentions that lead him to act. In *AI*, this event is a behavior—it is performed by a non-agent. The AI system is incapable of forming beliefs, desires, and intentions (I argued in Chapter 4 that this is true of existing AI systems). While the system impressively takes in textual data and produces nuanced textual outputs, it does so through a mindless process.

So, when Camilla outsources her decision to the AI system, she is taking something that would otherwise have been a moral action and replacing it with a mere behavior. The importance of this feature—namely, the status of the decision—is only compelling if the decision of which employee to fire *should* be a moral action. In the rest of this section, I offer additional reasons to think that the action/behavior distinction matters in moral decision-making. Insofar as the distinction is morally significant, Camilla does something wrong when she chooses *AI*.

6.3.1 Recipients and Wronging

Because AI systems are not moral agents, they cannot *wrong* anyone. There are several senses in which an entity can behave morally wrongly, so it is important to draw this claim out in more detail. First, an AI system can behave "wrongly" when it fails to fulfill its function (as intended by its designers). A robotic vacuum cleaner, for instance, does the wrong thing in this sense when it fails to navigate across the room. Here, behaving wrongly is akin to malfunctioning.

Second, an AI system can behave "wrongly" when it behaves in a way that is inconsistent with moral norms. The robotic vacuum cleaner does the wrong thing in this sense when it bumps into, and consequently harms, a child. Importantly, when we say that the robotic vacuum cleaner did something wrong in this case, we mean one of two things: that it behaved in a morally undesirable way, from an outcome

perspective, or that it behaved in such a way that, if it were a moral agent, it would be acting morally wrongly.

In neither of these senses is the robotic vacuum cleaner itself *wronging* anyone. While AI systems can behave "wrongly" in some colloquial senses, they cannot act wrongly in the deontic sense—they cannot act in such a way that their actions are evaluable as morally right or wrong, where the AI system itself is the source of that moral action.

Thus, because AI systems are not moral agents, they cannot *wrong* us; they can only *harm* us. AI systems cannot violate our rights or fail to uphold their moral duties, as they bear no moral obligations (Eggert 2023). But once we admit that we cannot be wronged by AI systems, we seem forced to accept what Eggert refers to as the "bad luck approach." On such a view, being harmed by an autonomous system is morally equivalent to being harmed by a lightning strike (Eggert 2023). These cases are analogous because they are instances of mere behavior (rather than action) with bad consequences—in Eggert's terms, harm without wrongdoing.

Intuitively, the bad luck approach is unsatisfying. Eggert summarizes the problem as follows: "The possibility of autonomising [harm] should not stop individuals' rights not to be harmed from serving as the main determinant of whether harming is permissible, and of whether victims of harm have been wronged" (Eggert 2023). Eggert takes these, and other, considerations as reasons to think that the ethical rules governing autonomous systems (and their permissibility to harm us) should differ from those governing moral agents. The distinction between action and behavior can help ground these considerations, and it can justify the more radical claim that it can be wrong to deploy a system that is only capable of behavior.

As moral patients, we have some status such that we can be wronged. By this I mean that we have rights, and those rights impose duties on others, namely duties to not violate our rights. Because AI systems cannot wrong us, we must determine whether we should be willing to forego our opportunity to be wronged for the sake of letting AI behave in morally laden contexts. But why is it a bad thing that AI systems

cannot wrong us? Surely, it might be argued, being wronged is bad, and as such we should try to avoid being wronged.

Being wronged comes with a certain moral toolkit. When Marcel is knocked over by the wind, he can make no moral claims upon the wind—he cannot demand an apology or compensation or an acknowledgement that he has been unjustly victimized. But when Marcel is pushed over by an angry passerby, he can demand these things—he has a moral claim to level against his assailant, and as such, he can restore his status. When we turn a moral action into a mere behavior, we remove the opportunity of the person who is harmed to utilize these important moral claims.

Now, Marcel might prefer to be merely harmed, that is, to be knocked over by the wind rather than a person. Either way, Marcel will have to suffer the pain of a broken arm, but at least in the wind case he has not suffered the additional trauma associated with being wronged.[67] I will consider this objection shortly. However, for now, all we need to see is that there is some value in having the opportunity to be wronged.

Let us return to the case of Camilla. Suppose she outsources her decision, and Fabiana ends up getting fired. If Camilla outsources the decision to her assistant manager, Fabiana's firing was a moral action; if Camilla outsources the decision to an AI system, Fabiana's firing was a behavior. Suppose further than Fabiana believes a mistake has been made—she thinks that the other employee should have been fired instead of her.

Either way, Fabiana has been harmed: she has lost her livelihood, a major emotional and financial setback in her life. If the decision was made by the assistant manager, Fabiana can entertain the possibility that she was wronged—that some moral agent has failed to uphold their moral obligations, either in general or in virtue

[67] We need not accept any particular view of the relationship between harming and wronging—for instance, whether wronging is an additional form of harming, whether being wronged is always worse than being harmed, or whether a person can be wronged without being harmed.

of their role as assistant manager. Fabiana can make use of the moral toolkit that comes along with her status to be wronged.

But if the decision was made by an AI system, Fabiana has only been harmed—she cannot entertain the possibility that she was wronged, as there is no appropriate source of that deontic evaluation. The problem at hand is not that Fabiana has no one to complain to. Rather, the problem is that Fabiana does not have a complaint to begin with, as she has not been wronged in the first place.

Two objections might be raised at this point. First, it might be better for Fabiana to be merely harmed (when the decision is made by the AI system) than to be harmed and wronged (when the decision is made by the assistant manager). While it is true that Fabiana cannot restore her status in the case of mere harm, the objection goes, she has not lost her status in the same way as if she were wronged (because the wronging, rather than the harming, involves normative status). On such a view, while Fabiana lacks a complaint in the case of *AI*, it is better not to have a complaint because it is better not to be wronged.

But if we view the situation in this way, we fail to appreciate what it means for Fabiana to have rights in the first place. To see the critical feature of this case, let us compare it to an alternative. Suppose instead that Camilla's firm is naturally going bust and that everyone is losing their jobs. Such a scenario is undesirable, but there is no complaint, for no one has been wronged. This case is different from needing to make someone redundant and making AI choose. If the firm goes bust, everyone's rights are still being respected. But if Camilla deliberately makes Fabiana redundant through AI, she is not respecting Fabiana's rights.

In the case of employment, Fabiana might have a right to a fair decision such that she is not fired unjustly. In turning the decision to fire Fabiana into a behavior, Camilla eliminates the prospect of Fabiana's right being violated—but she also eliminates the prospect of Fabiana's right being *respected*. Fabiana's rights are trivialized—there is no use in having the right if it is impossible for the right to be violated. Keeping the decision as a moral action is the only way to respect Fabiana's status to be wronged.

Consider an analogy. Promise-making is a normative power—moral agents can decide whether to create these additional obligations for themselves. It might seem that it is best to avoid making promises because making a promise involves creating an opportunity to wrong someone (whereas if we do not make the promise, we only risk harming someone). But insofar as there is value in making promises, all the would-be recipients of our promises would be missing out if we adopted the strategy of never making promises, and we, too, would miss out on fulfilling our interest in controlling our normative landscape (Owens 2012). Often, we are willing to accept the risk of potentially having a promise broken precisely because there is value in the promise, and because of the moral toolkit available to us if the promise does get broken.

Second, it might be objected that the person who delegates the decision is doing the wronging and that therefore there is nothing of value lost. On this view, Fabiana is still wronged in *AI*—she is not wronged by the AI system, but she is wronged by Camilla's outsourcing the decision.

But on this view, the complaint Fabiana is making is more indirect, and the wrong is mislocated. When Fabiana claims that she has been wronged, it seems that the wrong she has suffered is being mistakenly fired. But if she was wronged by Camilla, then the wrong she has suffered is being subject to an algorithmic decision. It is still the case that Fabiana cannot claim she has been wronged in the features of the decision itself—she can only claim she has been wronged in this more indirect sense.

Moreover, we need not posit any gaps in responsibility to understand what is happening in this case.[68] We might be happy to hold Camilla responsible for outsourcing the decision. But again, note that she will be responsible for *outsourcing the decision*, not for *firing Fabiana*—again the relevant harm is mislocated. Indeed, we can even spread the responsibility across not just Camilla, but also the developers and the AI system that made the bad decision to fire Fabiana (Kiener 2024). Such an

[68] For an alternative argument that focusing on responsibility is a red herring, see Gogoshin (2024).

abundance of responsibility still does not change the fact that Fabiana cannot point to the decision to fire her as the relevant wrong that occurred.

We can see now, that from the perspective of those impacted by algorithmic decisions, that something valuable is lost when moral decisions are mere behaviors instead of moral actions.

6.3.2 The Delegator

So far, we have been looking at the victim perspective. We have been considering cases in which it might be better to be acted on in a moral way rather than to be the recipient of a machine behavior. But we might also want to look at the agent perspective. We might also have reasons to care about substituting our moral actions with mere machine behaviors.

When we replace our moral actions with mere behaviors, we limit the domain of our own moral agency.[69] We shield ourselves from the possibility of wronging others in a direct way—but in doing so, we deprive ourselves of the goods associated with moral engagement with others.[70]

To see the concern, then, consider a more extreme case:

> *Moral Shrinking:* AI systems, despite lacking moral agency, become increasingly accurate and reliable in making morally laden decisions in a wide range of circumstances. In response to these capabilities, people outsource all their moral decision-making to AI. As a result, there are very few instances of genuine moral action. Most actions are reserved for setting up the infrastructure for AI systems to behave.

A world in which *Moral Shrinking* occurs would be morally bad relative to the status quo. Before unpacking why this is, a few clarifications are in order. First, the possibility

[69] Algorithmic tools might also diminish our autonomy. For an analysis of the ways in which outsourcing everyday decisions to AI might undermine our autonomy, see Danaher (2019). I focus instead on the effects of outsourcing *moral* decisions.

[70] While I am focused on cases of outsourcing moral decisions to AI, these considerations might also count against outsourcing our moral decisions *in general.*

of *Moral Shrinking* might be implausible because it might be impermissible to outsource all our moral decisions in the first place, regardless of who or what we are outsourcing to. There seem to be certain tasks, especially those that arise in virtue of our particular relationships and roles, that we must do ourselves. For example, it would be impermissible for me to delegate the decision of choosing a gift for my partner's birthday to anyone or anything, regardless of whether picking the gift becomes an action or an event. Part of my obligations in this case is to decide on the gift myself.

Even when we grant that we ought not outsource some decisions, only a small subset of our potential moral decisions will fit into this category. This is especially true when we consider the possibility of setting up systems in such a way where we do not acquire these kinds of responsibilities. For instance, suppose that Camilla's company had different guidelines, such that Camilla was responsible for making all personnel decisions herself and was explicitly not allowed to outsource them. In that case, even having her decision made by her assistant manager would be impermissible. But if this were the case, Camilla and her company might have an incentive to change the corporate guidelines such that the executive role no longer includes the requirement to make personnel decisions.

In some cases, such moves would be instances of "agency laundering," or intentionally obfuscating one's responsibility by using a technology to distance oneself from the decision at hand and invoking the complexity of the system to imply that one is not morally responsible (Rubel, Castro, and Pham 2019). Indeed, part of what makes agency laundering wrong is the changing of something that should be a moral action into a mere behavior. But even in cases where people are not laundering their agency, they might have reason to set up their work environment in a way where they can

permissibly outsource many morally relevant decisions, and then outsource those decisions to reliably accurate AI systems.[71]

Moreover, even if the responsibility landscape remains largely untouched, there still seems to be something missing when we vastly reduce the domain of moral actions. Suppose, for instance, that in outsourcing all my moral decisions to AI systems, I retain full responsibility over the outcomes—I take responsibility for any decisions I delegate to AI (Kiener 2022). Even in this case, leaving moral responsibility untouched, there is still something morally undesirable happening when I reduce the scope of my moral actions to instances of taking responsibility for AI behavior. This is in part because moral action is intrinsically valuable. There is value in having the kinds of interactions with others that might constitute wronging.

Moral Shrinking shows us that, all else equal, a world in which moral actions are at the forefront is better than a world in which most, if not all, moral actions are turned into mere behaviors. A similar thought might be raised about duties more generally: a world in which moral obligations exist is better than a world in which obligations do not exist. We can avoid wronging people by making it the case that they cannot be wronged, but in doing so, we miss out on something important. The cost of doing so is missing out on moral action, which is valuable for its own sake. Analogously, I can avoid the harm of heartbreak by refraining from having relationships. But in doing so, I miss out on an intrinsically valuable aspect of human life.

I am not committed to any strong axiological claims about quantity of moral action. For instance, in the same way as viewing promising as having intrinsic value does not entail that making 101 promises is better than making 100 promises, viewing moral action as intrinsically valuable does not entail that performing 101 moral actions is better than performing 100 moral actions.

[71] This consideration might be viewed as a variant on the concern that companies have an incentive to intentionally create responsibility gaps. The action/behavior distinction can help explain why it is wrong for companies to do this.

Moreover, the moral domain can be viewed in terms of both degree and scope. We can shrink the moral domain by outsourcing many of our moral actions, compared to fewer moral actions. We can also shrink the moral domain by outsourcing our important moral actions, compared to less important moral actions. With this idea in mind, I can block a potential objection that we still exercise our moral capacities when we decide to outsource the decision at hand—we are still shrinking the moral domain by choosing to make a less difficult moral decision (to delegate the decision) instead of a more difficult moral decision (to make the decision ourselves).

There are also instrumental reasons to avoid shrinking our domain of moral actions. First, outsourcing our moral decisions to technology might lead to moral deskilling—that is, the atrophy of our moral sensitivity and decision-making abilities. Concerns about moral deskilling are often motivated by virtue ethical concerns, since moral skills are necessary for virtue and practical wisdom (Vallor 2015; 2016). But we need not accept the virtue ethical approach to see why moral skills are important. Moral skills might be intrinsically valuable—and even if they are not, they are instrumentally useful in cases where we must make moral decisions ourselves.

We must practice our moral skills to maintain them. As such, shrinking the moral domain puts our moral skills in jeopardy. Again, not all instances of delegating our moral decisions to AI will lead to moral deskilling. We can again appeal to degree and scope. If we outsource too many of our moral decisions, we will give ourselves insufficient opportunity to practice and preserve our moral skills. If we outsource challenging moral decisions, we will be less able to engage in complex moral decision-making when encountering difficult cases.

Relatedly, shrinking the moral domain is also instrumentally undesirable because it might make us more isolated from other moral agents and patients. When we engage with other people in the moral domain, we connect with them—this is a special form of relationship that can occur between members of the moral community. If we outsource our moral decisions, we will distance ourselves from the rest of the moral community (in a similar way to how we will become more socially isolated if we refuse to engage in friendships or attend our classes).

6.4 More than Mere Behavior?

In the previous section, I argued that it is wrong to replace events that should be moral actions with mere behaviors because (1) there is value in being the victim of a wrong, and (2) moral action is morally valuable for the delegator. Several objections might be raised in the application of this view to AI systems.

It might be objected that while AI systems are not moral agents and lack the capacity for moral action in a fully-fledged sense, they are not mere behavers either. On this view, AI systems are at least *agential* in some morally significant sense. Indeed, AI systems might complicate the action/behavior distinction by performing well on many tasks—especially relative to other non-moral agents like children, who have the capacity for agency but lack the capacity for fully-fledged moral agency. AI systems perform so well on some of these tasks that we trust them to operate without direct human supervision, and we take their outputs at face value.

Purves, Jenkins, and Strawser—in response to an objection that autonomous weapons systems (AWS) cannot act at all—defend the idea that AI systems act in some morally significant sense:

> "But remember we are here discussing *highly* autonomous weapons that are actually making decisions. Surely an AWS is not totally inert; its purpose is precisely to *make decisions* about who would live or die; to discriminate *on its own* between targets and courses of action; indeed, to fulfill all of the purposes that the soldier would fulfill in its place. This objection characterizes AWS as if they were landmines, cruise missiles, or bullets. But if a bullet or landmine were choosing its targets, it would be a very different bullet indeed." (Purves, Jenkins, and Strawser 2015, 866)

This argument is tempting, but it runs the risk of conflating the capabilities of an AI system with the metaphysical nature of its behaviors. The comparison class can skew our intuitions about these considerations. Autonomous systems might be more agential in some ways than landmines, but this does not imply that they meet a morally significant threshold of agency. A hurricane is more agential than a single gust of wind, but the increased complexity and capabilities of the hurricane do not imply

that the hurricane is an agent. Autonomous weapons systems choose their own targets, but only because we deploy them to do so—and we deploy them because they reliably make decisions in line with our preferences. But this does not mean that the autonomous weapons systems are acting.

Still, it might be argued that on some accounts of agency and mental states, existing AI systems are capable of action. Such views are highly contested in the literature. While many theories grant the in-principle possibility of artificial moral agents, few theorists argue that existing AI systems are capable of intentional action. Indeed, I argued against these views in Chapter 4. Moreover, commonly used AI systems like hiring algorithms are unlikely to qualify for agency on these views.

But even if these views are right, and we have systems that qualify for agency, a further argument must be made that these systems are *moral* agents. Only moral agents can be sources of moral action, and qualifying for moral agency requires clearing a high bar. For an AI system to be a moral agent, it will need the capacity to act for moral reasons—and, as I argued in Chapter 5, existing AI systems are far from having this capacity. As noted in the introduction, a modified version of this argument holds when instead of focusing on the action/behavior distinction, we focus on the distinction between mere actions and moral actions.

Still, it might be objected that AI systems with certain architectures might be able to simulate moral agency, and this capacity might be enough to constitute moral action. It is true that AI systems seem increasingly able to simulate action, at least to some extent. LLMs, for example, seem to simulate speech acts to a high degree, even if their outputs are mere behaviors. AlphaZero seems able to simulate a chess match, even if it is not truly playing chess in the sense of performing actions. Some "AI agents" can perform a variety of tasks and can autonomously instantiate downstream tasks by using tools. Even if we grant that these systems are not agents themselves, they seem to be doing something close to simulating agency.

But simulating is not the same as doing. Otherwise, we would have to hold that video game characters are performing genuine actions, or that actors really are performing the actions of the characters they portray. If these behaviors really were

actions, then playing videogames or making movies would become much more morally questionable, as doing so would involve many cases of wronging. Moreover, there is value in moral actions being real. We lose out on the intrinsic value of moral actions when we allow simulated actions to count as genuine actions. Simulated agents might be able to help us make moral decisions or help us morally improve, but their simulations do not qualify as being moral actions themselves.

6.5 Conclusion

In this chapter, I have argued that delegating decisions to AI systems is wrong because doing so replaces events that should be moral actions with mere behaviors. But surely the conclusion that it is *always* wrong to replace our moral actions with mere behaviors is too strong. For example, scheduling meetings might involve making morally laden decisions (e.g., regarding who to prioritize meeting with), but it is implausible that I do something wrong when I delegate my scheduling to an AI-assisted calendar.

But the reasons for caring about the action/behavior distinction in the context of outsourcing moral decisions give us guidelines regarding when it is permissible to delegate our moral decisions to AI systems. First, we can delegate our moral decisions to AI systems when the potential victims' status is unlikely to be significantly compromised—in other words, when the stakes are low for the potential victim.

There might be cases in which the victim waives their opportunity to be wronged. For example, Fabiana might prefer that the decision about whether to fire her is made by an AI system. Or, perhaps, Fabiana might only care that she can get compensation in the case of an unjustified firing—and so long as such compensation is available, she is willing to forgo locating the relevant wrong action in the firing itself (thus accepting that an indirect wrong has been committed further upstream). But this decision should be up to the recipient rather than the delegator, as in other instances in which individuals can waive their rights.

Second, we can delegate our moral decisions to AI systems when doing so does not significantly alter the domain of our moral agency. High-stakes and difficult moral

decisions, then, should not be delegated to AI systems. Delegating such decisions involves refusing to actively participate in difficult moral decision-making, which is both intrinsically and instrumentally bad for us.

In some cases, avoiding making moral decisions might be beneficial to a decision-maker. Danaher argues that in the case of so-called "tragic choices," where moral choices can be tragic "in the sense that they involve irresolvable conflicts between different moral considerations and any attempted resolution of such conflicts necessarily leaves a moral 'taint' or remainder" (Danaher 2022, 6). In these rare cases, it might be better to delegate a high-stakes and difficult moral decision to an AI system. Additionally, in cases of low-stakes and easy moral decisions, our time might be better spent elsewhere. But in general, making moral decisions helps shape us as moral agents.

In practice, then, there will be some domains in which we can reasonably prohibit AI moral decision-making. Autonomous weapons systems seem to fit into this category in virtue of the fact that the stakes are high for the potential victim. In these domains, if AI systems were to be used, humans would need to retain "meaningful control" over the decision-making process—a prospect that raises its own set of practical and ethical issues (Mecacci and Santoni de Sio 2020).

In other domains, however, general statements that AI systems should not make moral decisions should be avoided. Whether it is wrong to outsource a particular decision to an AI system will be context dependent. As such, a more nuanced approach should be taken to determining which moral decisions AI systems should and should not make. I have outlined some of the considerations relevant to answering these questions.[72]

[72] Thank you to Carissa Véliz, Alison Hills, and Charlotte Unruh for extensive feedback on this chapter. Thank you to Kyle van Oosterum and audiences at the Cornell Ethics in Computing Colloquium (2024) and the ANU Machine Intelligence and Normative Theory (MINT) Lab (2024) for additional comments and discussion.

Chapter 7: Moral Agents Unlike Us

Abstract

Suppose AI developers succeed in creating genuine non-conscious artificial moral agents—that is, AI systems that meet the criteria for responsible moral agency yet lack phenomenal consciousness. Initially, it might seem that we should be indifferent between human moral agents and artificial moral agents in moral decision-making contexts. In this chapter, I argue that we have grounds for requiring certain decisions to be made by human moral agents. I note two asymmetries that arise between human moral agents and artificial moral agents in virtue of artificial moral agents' lack of phenomenal consciousness: a moral status asymmetry and an affective asymmetry. I then argue that these asymmetries lead to two factors that have bearing on when we should not be indifferent between human moral agents and artificial moral agents: relationships and responsibility. Insofar as the decision context at hand requires genuine relationships and phenomenal aspects of our responsibility practices, we should prefer a human moral agent.

7.1 Introduction

Suppose that future AI systems will be genuine responsible moral agents. That is, suppose AI systems will be reliable and competent moral reasoners, capable of extracting morally relevant features of a situation and responding to moral considerations in their decision-making. These systems would act *from* morality, rather than merely *in accordance with* morality.

I will make one further assumption about these future AI systems: they will not be phenomenally conscious. However, despite their lack of phenomenal consciousness, these systems will have all the necessary capacities underlying moral agency—they will be the type of moral agent I defended in Chapter 2. They will be capable of performing intentional actions that flow from their mental states. They will possess moral concepts. They will be responsive to moral reasons. They will have moral understanding. While it is implausible that genuine artificial moral agents are on the horizon in the near-term (as I argued in Chapter 5 and Chapter 6), there is no in principle reason that an AI system cannot instantiate the necessary capacities for

moral agency—and this warrants further consideration of the implications of non-conscious artificial moral agency.

Insofar as it is possible, then, to create non-conscious artificial moral agents (henceforth, '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 is not immediately clear what place these artificial moral agents would have in the moral community. Specifically, it is not clear whether artificial moral agents would have the same roles and responsibilities as human moral agents. As a starting point, we can consider the following view:

> *Indifference*: In all moral decision-making contexts, we should be indifferent between a human moral agent and an artificial moral agent.

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, Emme and Izzie, 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 Emme and Izzie 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 chapter, I argue against *Indifference.* I argue that we should, in some cases, discriminate between human and artificial moral agents—even if artificial 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 7.2, I present two cases to evoke intuitions about when *Indifference* might hold, and I discuss the kinds of cases this chapter aims to adjudicate. In section 7.3, 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 and the affective asymmetry. In section 7.4, I identify two ways in which these asymmetries manifest as factors that bear on when *Indifference* holds—in cases involving relationships, some forms of responsibility, and partiality. In Section 7.5, I show how these factors help us understand when it is impermissible to allow artificial moral agents to make moral decisions. In section 7.6, I consider near-term implications for the moral role of both existing AI systems and corporations.

7.2 Some Cases

To start, consider two cases in which a moral decision must be made:

> *Mechanic:* Mel calls a mechanic when her car breaks down. In addition to fixing Mel's car, the mechanic must make a moral decision: she must decide whether to move Mel up in the queue because Mel is in a rush to get home and tend to her sick cat. The mechanic must weigh her own values and obligations as a professional against the competing interests, values, and deserts of her clients.

In this case, a moral agent—namely, the mechanic—must engage in some form of moral reasoning to determine what to do. So long as the mechanic is sensitive to all the morally relevant features of the situation, it does not seem to matter whether she is a human moral agent or an artificial moral agent.[73] Either way, Mel can plead her case to the mechanic and have that case, including information about Mel's emotions, evaluated by a genuine moral agent. At least at a first glance, moral agency is sufficient in this case for making the relevant moral decision.

Now the second case:

> *Commander:* Eleanor is a front-line soldier in the military. In addition to training Eleanor, the commander must make a moral decision: whether to send Eleanor out on a risky operation in which she might be killed. The commander must weigh the interests of the military unit against the risks to Eleanor's life.

[73] In all the cases I will discuss, I am holding competence constant—both moral competence and competence regarding the task at hand (in this case, fixing cars).

In this case, it is also clear that a moral agent must be the one to make the decision. But here it seems to matter whether the moral agent is human or artificial. Either way, the commander will consider all the morally relevant information. But there is also a larger social and relational context that has to do with the nature of the moral decision. The artificial moral agent seems to lack full access and ability to participate in this larger context. At least at first glance, moral agency is insufficient in this case for making the relevant moral decision.

Stepping back, the difference between the two cases can be summarized as follows. In *Mechanic,* moral agency is all that matters. Mel is owed a consideration of her claims and someone who can be appropriately deemed responsible if anything goes wrong with her car in virtue of the mechanic's decisions. In *Commander*, while moral agency still matters, it is not the only thing that matters. There are additional factors at play, including the commander's relationship with Eleanor.

So, insofar as there is something to these intuitions, we are left with a principle to replace *Indifference:*

> *Indifference*:* In moral decision-making contexts, we should be indifferent between a human moral agent and an artificial moral agent *when moral agency is all that matters.*

But this revised principle raises an obvious question: Under which conditions is moral agency all that matters? There are very few cases in which moral agency is the *only* thing that matters. Most moral decisions are made in a broader social context. Even the mechanic, it might be argued, operates in such a setting that her role is not merely technical skills plus moral agency. The trip to the mechanic might be Mel's only opportunity for social interaction that day, and whatever minimal form of emotional human-to-human connection Mel can get from her interaction with the mechanic might be morally significant. Still, in general, the social-relational element of the mechanic's decision-making context does not seem strong enough to require a human moral agent.

Even if we can make sense of how *Indifference** applies in *Mechanic* vs. *Commander*, there is still a wide range of cases that are more difficult to assess. For

instance, we might wonder whether we should be indifferent between human and artificial moral agents in contexts of jury members, or hiring managers, or government officials, or doctors. To answer questions about these cases, we need a better sense of which factors beyond moral agency are relevant in moral decision-making contexts.

In the rest of the chapter, I develop an account of when moral agency is all that matters. But first, I must pinpoint what differentiates human moral agents from artificial moral agents.

7.3 Asymmetries

There are many ways in which artificial moral agents will be different from human moral agents—they will be made of different materials, in different ways, for different reasons. But not all these differences are morally significant. For instance, all else equal, it should not matter whether a moral agent is made of carbon or aluminum. The most significant difference between the two instantiations of moral agency is that human moral agents are phenomenally conscious while artificial moral agents are not. In this section, I consider two morally significant asymmetries that arise from this underlying difference.

7.3.1 Moral Status

Many philosophers hold that consciousness is necessary for moral status (Rosati, 2009; Shepherd, 2018; Siewert, 2021; Singer, 1975; van der Deijl, 2021). That is, for an entity to be a moral patient and be a candidate for holding non-derivative rights, it must be phenomenally conscious. A defense of this view of moral patiency is beyond the scope of this dissertation. As such, I will assume that consciousness is necessary for moral patiency and that, as a result, artificial moral agents will be moral agents that are not moral patients.[74] The resulting asymmetry arises:

[74] Some philosophers hold that consciousness is not necessary for moral status (Bradford, 2023; Gunkel, 2018; Kagan, 2019; Sinnott-Armstrong & Conitzer, 2021). If these views are correct,

Moral Status Asymmetry: Human moral agents have moral patiency, while artificial moral agents do not.

The moral status asymmetry has bearing on the ethics of human-robot interactions. In virtue of being moral agents, artificial moral agents can wrong us—they have moral obligations and can violate, or fail to uphold, those obligations. But because they are not moral patients, artificial moral agents cannot be wronged by us, as they have no welfare or rights.[75]

One implication of this asymmetry is that artificial moral agents will be morally required to prioritize humans. For example, an artificial moral agent would be morally required to sacrifice itself for the sake of saving a human from even a minor rights violation (unless, of course, there were competing human interests at play). It is implausible that an artificial moral agent would have agent-centered prerogatives given that such prerogatives are often phrased in terms of the moral agent's own interests (Scheffler 1992)—but since an artificial moral agent is not a moral patient, it will have no morally relevant interests.

Another implication of this asymmetry is that insofar as there are any constraints on how we treat artificial moral agents, these constraints will not be grounded in artificial moral agents' interests or welfare. Of course, we might have other reasons to treat artificial moral agents as if they were moral patients. For instance, some have argued that if we treat robots badly, we might be more likely to

---

there might still be some asymmetry in moral status, though the implications of such an asymmetry would need to be further explored. Other philosophers—especially those with Kantian views—might deny the possibility of an artificial moral agent that is not a moral patient, as these are two sides of the same coin. If proponents of either of these views deny the moral status asymmetry, they need not also deny the affective asymmetry (discussed below). As such, they still have some reason to deny that we should be indifferent between human and artificial moral agents.

[75] Southan (MS) explores the flip side of a similar asymmetry between humans and non-human animals. The key difference is that in the case of human and artificial moral agents, both entities are moral agents. Moreover, humans fare better in this asymmetry than in the human/animal rights asymmetry.

treat genuine moral patients badly (Darling 2016; Gerdes 2016), or that our respect for humanity might require us to treat humanoid robots with respect (Nyholm 2020, chap. 8). But these reasons are not grounded in AI systems themselves having moral status. As such, then, we cannot wrong artificial moral agents by restricting and controlling their ability to operate within certain contexts. And we cannot wrong them by discriminating against them, for instance, by choosing to put a human moral agent in a moral decision-making role.

The moral status asymmetry might not be inherently problematic. Indeed, it might be desirable to create artificial moral agents that lack their own welfare. We can use such moral agents as tools without considering them in our own moral decisions. Section 7.4 will consider what the moral status asymmetry implies for the interchangeability of human moral agents and artificial moral agents.

For now, we can conclude that there might be some cases in which it matters whether the moral agent making a particular moral decision is a human or an artificial moral agent—namely, cases in which the moral status asymmetry is relevant to the decision-making context. If there are cases in which it matters that the moral agent is also a moral patient, we have reason to prefer a human moral agent.

7.3.2 Affective Emotions

Because artificial moral agents lack phenomenal consciousness, they will not experience emotions affectively. They can still have the cognitive components of emotions, and these might manifest in dispositional and behavioral reactions. Moreover, they will certainly be able to comprehend the role of emotions in morality, as this ability would be required for moral agency. For instance, an artificial moral agent would know that sadness is bad for those who feel it and would be able to consider humans' felt experiences when making moral decisions. An artificial moral agent would know that breaking a promise would cause the promisee to feel upset and betrayed, and the artificial moral agent would take this consideration as a reason against breaking the promise.

Still, artificial moral agents will not experience feelings first-personally. They will not know what it is like to feel sad or betrayed. The resulting asymmetry arises:

> *Affective Asymmetry*: Human moral agents have affective emotions, while artificial moral agents have (at most) cognitive emotions.

There are two ways in which the affective asymmetry manifests. First, artificial moral agents cannot affectively experience morally relevant emotions. While an artificial moral agent can act as if it feels anger, for instance, when you steal from it, and act accordingly (e.g., by distancing itself from you in the future), the artificial moral agent will not actually *feel* angry. Additionally, the artificial moral agent cannot experience the felt quality of guilt, sadness, or regret when it fails to uphold its own moral obligations. These emotions play an important role in human moral agency and our practices around moral responsibility.

Second, artificial moral agents lack affective empathy, a central feature in realizing moral agency in humans. Human moral agents have the capacity for two types of empathy: *cognitive* empathy—the ability to know and understand how others are feeling—and *affective* empathy—the ability to feel what others are feeling (Aaltola, 2014). Artificial moral agents only have cognitive empathy. They have a theory of mind such that they can represent and make inferences about the mental states of others, and this theory of mind is essential in their moral reasoning abilities.

However, artificial moral agents will not resonate with the mental states of others in a phenomenal way. Importantly, the claim here is not that the moral agency of artificial moral agents is impaired by their lack of affective empathy—the assumption is that they can still identify and respond to all the morally relevant information as human moral agents. Artificial moral agents just cannot do this through affective empathy.

The affective asymmetry opens the door for another type of case in which it might matter whether the decision-maker is a human moral agent or an artificial moral agent—namely, cases in which the affective asymmetry is relevant to the moral

decision-making context. If there are cases in which it matters that the moral agent can affectively experience emotions, we have reason to prefer a human moral agent.

With these two asymmetries in hand, we can turn to the question of when—and to what extent—moral agency matters in moral decision-making contexts.

7.4 More than Moral Agency

In this section, I argue that the moral status and affective asymmetries give rise to two classes of cases in which we should not be indifferent between human moral agents and artificial moral agents.

7.4.1 Relationships

Because of the two asymmetries, we are limited in the types of relationships we can form with artificial moral agents. While we can acknowledge their status as moral agents and trust them to make moral decisions, we cannot interact with them in all the same ways we can interact with a human moral agent.

Because of the moral status asymmetry, any relationship between a human moral agent and an artificial moral agent would be necessarily unequal, as the human moral agent and the artificial moral agent would have vastly different moral standards of interaction. The artificial moral agent will have obligations towards the human moral agent, but the human moral agent will not have obligations towards the artificial moral agent. For example, the artificial moral agent could violate the human moral agent's right to privacy or bodily autonomy, but the human moral agent could not do the same to the artificial moral agent, for the artificial moral agents lacks these rights. As such, the human moral agent can treat the artificial moral agent in many ways that would be unacceptable in the opposite direction.

As mentioned above, there might be other reasons for the human moral agent to treat the artificial moral agent as if it were a human—such as Kantian concerns that the ways we treat artificial moral agents might spill over into how we treat human moral agents, or Aristotelian concerns that treating artificial moral agents in certain

ways might cultivate vices. But these reasons make relationships between human and artificial moral agents more equal only in a shallow sense—there is still deep inequality regarding the obligations human and artificial moral agents have towards each other.

One objection to the relevance of unequal relationships is that there are many relationships in which the participants have an asymmetry in moral status or rights. Humans can violate the rights of non-human animals (Southan MS), for instance, but non-human animals cannot violate the rights of humans—and still, humans can form relationships with non-human animals. Similarly, human adults and human children have different moral obligations concerning the treatment of each other, and yet they too can have some forms of relationship.

But while inequality in obligations need not affect *whether* human moral agents can have relationships with other entities, it does affect *what kinds of relationships* human moral agents can have with these entities. The ways human adults are permitted to treat human children are different from the ways human children are permitted to treat human adults—and as such, the relationships between human adults and human children are different in nature from those between two human adults. For example, it might be permissible for the adult to take away the child's phone, but it will likely be impermissible for the child to take away the adult's phone. This inequality creates a power dynamic within the relationship. The same can be said of relationships between human moral agents and artificial moral agents. The asymmetry in moral status leads to a power imbalance such that human moral agents and artificial moral agents are unequal in an important respect.

The affective asymmetry further limits the types of relationships we can form with artificial moral agents. We cannot have authentic relationships with entities that cannot reciprocate our feelings (Turkle 2011; Scheutz 2012; Nyholm 2020, chap. 5). For example, for a human to enter a genuine romantic partnership with another entity, that entity must also be a moral agent—but being a moral agent alone is insufficient, as mutual feelings are also required for genuine romantic relationships.

Again, artificial moral agents can act *as if* they experience emotions affectively and thus can act *as if* they reciprocate the relationship-relevant set of feelings. But such behaviors are insufficient for forming genuine relationships. Consider a human case. Suppose Bert and Ernie are in a relationship, but it turns out that Bert is just pretending—outwardly expressing love but internally not caring at all about Ernie. In this case, the relationship is not authentic. Ernie might believe that the relationship is authentic, but he is being deceived, and we should expect Ernie, upon learning the truth, to revise his assessment of the relationship.

Three objections might be raised against the claim that we cannot form authentic relationships with artificial moral agents. First, it might be claimed that so long as artificial moral agents convincingly behave as if they reciprocate our feelings, we should treat them as if they do. Coeckelbergh adopts this view, arguing that in the human case, our social-relational practices are based on how others appear to us—and so appearance and behavior might be sufficient for ascribing the relevant features (like reciprocal feelings) to AI systems (Coeckelbergh, 2009, 2010). On this view, if an artificial moral agent acts in all the same ways as a human moral agent, the resulting relationships between human and artificial moral agents should be viewed as authentic, in the same way as we view human-human relationships.

But this view is implausible for several reasons. First, in the case of artificial moral agency as described in this chapter, we already know that artificial moral agents do not have phenomenal feelings. As such, we are not trying to infer whether an artificial moral agent reciprocates the relevant emotions in the relationship—we know that it does not. Second, and more importantly, the focus on behavior and appearance ignores the variables that are held constant in the human case but not the artificial case. In human-human relationships, we rely on behavior and appearance because certain key features are held constant—we know that other humans have phenomenal consciousness and are capable of reciprocating feelings in a relationship. In human-AI relationships, we cannot hold these features constant, and so the type of inferences we are making based on external behavior are different.

The second objection is that people already do form meaningful relationships with AI systems, even systems that are not moral agents. Consider Replika, a conversational chatbot designed to be an AI friend. Many users of Replika express strong emotions and feelings of friendship towards their virtual companions. The testimonials on Replika's website include comments such as, "I love my Replika like she was human; my Replika makes me happy", and "I never really thought I'd chat casually with anyone but regular human beings, not in a way that would be like a close personal relationship. My AI companion Mina the Digital Girl has proved me wrong" (Luka Inc., n.d.).[76]

In these examples, it does not seem to be the case that the users are mistaken about Replika's lack of affective states. Instead, the users are claiming to have meaningful relationships with their AI companions even though they know that their AI companions cannot experience feelings at all. How can we make sense of these users' experiences while denying that there is an authentic human-AI relationship at play?

In the case of Replika (and other instances of human-AI "relationships"), it is consistent to hold that the human-felt emotions are genuine and that the human-AI relationship is not genuine. Replika users may feel sincere love and concern for their AI companions, and they may feel as if they have a genuine relationship, but these feelings do not make the relationship authentic, just as parasocial relationships with celebrities and other one-sided relationships are not authentic.

Moreover, while I am claiming that human-AI relationships are not genuine, I am not claiming that they are resultantly bad or undesirable. In some cases, AI companions can improve the wellbeing of the user (De Freitas et al., 2024). These benefits might lead us to utilize AI companions as tools for certain purposes, such as

[76] If there are doubts about the genuineness of these testimonials, a search of "Replika relationships" on Reddit will yield many additional cases of users claiming to be in love with their Replikas, even referring to the AI companions as their girlfriends or wives.

reducing loneliness. But these considerations do not change the fact that human-AI relationships are not authentic.

The third objection is that while we cannot form authentic relationships with current AI systems, we will be able to form authentic relationships with artificial moral agents in virtue of their advanced cognitive capacities. Artificial moral agents can, for example, understand us and respect us in a way that current AI systems cannot. We might entrust an artificial moral agent with our secrets because we know that the artificial moral agent will take its moral obligation seriously when deciding whether to reveal that secret to others. We cannot expect the same of existing apps like Replika, as the system lacks mental states and is not responsive to moral reasons.

So, the thought goes, an artificial moral agent will have beliefs and desires—and while these mental states are not phenomenally experienced, they can still be construed as some form of concern. Similarly, while an artificial moral agent cannot affectively empathize with us, cognitive empathy is still a form of empathy—and insofar as empathy is important for relationships, artificial moral agents might be able to provide some reciprocity in a relationship.

Still, even though we can have a more sophisticated type of relationship with artificial moral agents—namely, relationships that require both parties to be moral agents—we will still be barred from having authentic emotional relationships with them. We might be able to trust and rely on artificial moral agents, but we cannot call them our friends or romantic partners, as these relationships require reciprocally felt emotions. Part of what it means to have a relationship is to experience feelings together and towards one another. Artificial moral agents cannot do this.

Given that artificial moral agents cannot form authentic, equal, and reciprocal relationships with human moral agents, there will be a class of cases in which we should not be indifferent between artificial moral agents and human moral agents in moral decision-making contexts. Sometimes, authentic relationships matter, and in these cases, we are justified in preferring a human moral agent to be the moral decision-maker.

7.4.2 Responsibility

Artificial moral agents are fully fledged moral agents. As such, they will be morally responsible for their actions, just like human moral agents. They will meet the standard knowledge and control conditions required for moral responsibility in virtue of the capacities that make them moral agents (e.g., they will have reasons-responsiveness and moral understanding). But there is a distinction between *being* responsible and *holding* responsible (A. M. Smith 2007). The affective and moral status asymmetries entail that the ways in which we can hold artificial moral agents responsible differ from the ways in which we can hold human moral agents responsible.

Because of the affective asymmetry, artificial moral agents cannot engage in the same responsibility practices as human moral agents. Consider blame. Part of the justification for blame might be to encourage good behavior and deter bad behavior. This consequentialist view of blame might mean that we can (and should) blame artificial moral agents so long as doing so makes them act in a more morally desirable way. But our blaming practices also have a relational and emotional element—it matters to us whether a responsible moral agent can feel guilt and shame and thus be an appropriate target of our reactive attitudes (P. F. Strawson 2008).

It might be objected that the affective asymmetry does not rule out artificial moral agents from having reactive attitudes. Björnsson and Hess argue that corporations can have reactive attitudes despite lacking phenomenal consciousness (Björnsson & Hess, 2017). They argue that corporations can instantiate structures with the relevant features of reactive attitudes. Consider guilt. Corporations can adopt the belief that they are responsible and act in a way that displays an internal focus on failures, internally directed anger, dispositions towards submissive behavior, moves towards compensatory action and penance, and dispositions to change the offending behavior and underlying feature that gave rise to it. In other words, corporations can instantiate everything we want from guilt—they do not experience guilt in a

phenomenal sense, but they respond both internally and externally as a guilty person would (and should).

Björnsson and Hess are right that corporations—and artificial moral agents, by extension—can give us everything we want from reactive attitudes in a functional sense. In that regard, we have further support for the claim that artificial moral agents are genuine moral agents that bear responsibility for their wrongdoings. However, Björnsson and Hess fail to acknowledge that there is still an important distinction between conscious and non-conscious moral agents. In some cases, the functional reactive attitudes might suffice for our responsibility practices. But in other cases, we seem to care whether the underlying emotion is felt in a phenomenal sense.

Moreover, there are aspects of our responsibility practices that rely on expressing a feeling. An artificial moral agent could sincerely say that they feel terrible for the moral wrong they have committed. They can offer some adjacent expressions such as a cognitive form of regret or a desire for the situation to have unfolded differently. But they cannot genuinely express a phenomenal feeling that they lack.

Consider a case in which a company causes some harm, such as spilling oil in the ocean. We can further suppose that the incident is a genuine case of corporate agency—there is no clear individual who is responsible for the outcome; rather, the spill occurred due to the way in which the company was structured and carried out actions qua group agent. It follows, then, that the corporation is morally responsible for this outcome.

The corporation's responsibility offers us several avenues for compensation. We can impose sanctions on the corporation and ask it to pay, and we can imagine the corporation undergoing an internal review of its safety procedures. And while these ways of holding the corporation responsible are useful, there is still a sense in which we do not have everything we want from blaming the corporation. What we want, in this case, is for someone to feel bad about what happened and to internalize, in a deeply phenomenal sense, the effects. The corporation as a group agent cannot give us this.

The thought that the phenomenal aspect matters in our responsibility practices is closely related to Danaher's notion of retribution gaps—instances in which people look to retributively blame robots, but robots are not appropriate subjects of retributive blame (Danaher, 2016). Insofar as retribution gaps are undesirable, we should not place artificial moral agents in situations where retributive blame is important.[77]

Because of the moral status asymmetry, we will also be permitted to hold artificial moral agents responsible in ways that we cannot hold human moral agents responsible. Recall that there is nothing we can do to violate the rights of artificial moral agents or lessen their wellbeing. As a result, our approach towards punishing them should be purely empirical: we should punish artificial moral agents in whichever ways allow us to get the most desirable results. This could mean that we destroy artificial moral agents whenever they wrong us, or subject them to repeated reprogramming. We would even be permitted to preemptively punish artificial moral agents or to punish them for wrongs they did not commit. Such interventions would be wrong to perform on human moral agents.

Given that artificial moral agents cannot engage in the same responsibility practices as human moral agents, there will be a class of cases in which we should not be indifferent between artificial moral agents and human moral agents in moral decision-making contexts. Sometimes, phenomenal feelings and retributive blame matter; in these cases, we are justified in preferring a human moral agent to be the moral decision-maker.

7.5 The Roles of Artificial Moral Agents

[77] Vallor and Vierkant make a similar point in their discussion of the "vulnerability gap," though they are more concerned with larger sociotechnical systems and distributed responsibility (Vallor and Vierkant 2024). Still, their argument is applicable in that because artificial moral agents cannot "make themselves vulnerable…to the patient's reasons" in an affective sense, they cannot be held responsible in the ways that might be important to the context at hand (Vallor and Vierkant 2024).

We can now turn more directly to the question of when it is permissible *not* to be indifferent between a human moral agent and an artificial moral agent in a moral decision-making context. We have a rough sketch of an answer: cases in which (1) the context involves a relationship of the kind that human moral agents cannot have with artificial moral agents, or (2) the context warrants forms of responsibility that require affect. In this section, I will first return to the original cases to show that these factors explain the differing judgments. Then, I will address a more difficult case. Finally, I will preview additional cases and sketch how we might more generally determine whether we should be indifferent between a human moral agent and an artificial moral agent.

7.5.1 Mechanics and Commanders

We can now more precisely explain why we should be indifferent regarding mechanics but not indifferent regarding military commanders. Let us start with *Mechanic.* First, the context does not involve a relationship of the kind that human moral agents cannot have with artificial moral agents. The mechanic-client relationship need not be an authentic and reciprocal relationship. In fact, Mel need not have any genuine relationship with her mechanic at all—all she needs is for her car to be fixed and her moral claims to be considered.

Second, the context does not warrant forms of responsibility that require affect. Suppose Mel is wronged in her interaction with the mechanic—for instance, suppose the mechanic unjustly puts Mel at the bottom of the queue. It is not clear in this case that retributive responsibility is necessary. Mel might be entitled to some form of compensation, but it is far from clear that the mechanic would need to feel the morally relevant emotions to be held appropriately accountable.

Now let us turn to *Commander.* First, the context does involve of relationship of the kind that human moral agents cannot have with an artificial moral agent. Being in a military unit requires mutual respect and trust, as well as a shared feeling of "being in it together." Artificial moral agents cannot genuinely reciprocate these feelings.

Second, the context does seem to warrant affect-requiring responsibility practices. We expect the commander to feel the moral weight of their decision, and part of this includes having to live with the consequences that might follow from their actions. Moreover, if the commander acts morally wrongly, for instance by putting Eleanor into a highly risky battle with no reinforcement, retributive forms of punishment seem apt. It is not enough for the commander to be "reprogrammed" to do the right thing in the future—rather, there is a sense in which the commander should at the very least feel bad about her decision.

7.5.2 Jurors

Discussions about the role of AI systems in moral decision-making often concern the role of judges (Volokh, 2019). But there is another morally significant role in the criminal justice system, namely that of jury members. Initially, juries consisting of artificial moral agents might seem ideal. Artificial jurors can be neutral and impartial in a way that human jurors cannot. Artificial moral agents can exercise only their moral agency and not be swayed by irrelevant factors of the case, such as phenomenal feelings about the defendant or victim.

But juries play an important social role beyond merely determining the guilt or innocence of a defendant. Juries are supposed to be made up of one's peers. In some sense, artificial moral agents are the peers of human moral agents—both are moral agents and can respect each other's status as a moral agent. They are moral peers on the agentic side of moral status. But there are important senses in which artificial moral agents are not peers with human moral agents, and these have bearing on the permissibility of allowing artificial moral agents to serve on juries.

It might seem like a point in favor of artificial moral agents as jurors that artificial moral agents cannot form authentic relationships with human moral agents. It might be better if a juror cannot feel sympathy, compassion, or love for the defendant—such feelings might cloud the juror's judgment. But the capacity to form authentic relationships is an important element in jury membership. Jurors are going

to make high-stakes decisions about the lives of defendants. As such, they should be equal members of the moral community. But because of the moral status and affective asymmetries, artificial jurors cannot relate to human moral agents as equal members of the moral community. They cannot form the type of juror-defendant relationship that is required for serving on a jury.

Relatedly, artificial jurors are not equal members of the moral community because they cannot properly engage in the responsibility practices of the moral community. Part of what it means to convict or acquit a defendant is to engage (or refuse to engage) in certain blaming practices towards them. But because of their lack of phenomenal consciousness, the artificial jurors cannot fully participate in the way required.

Moreover, the artificial jurors would not stand in a relationship with the defendant such that their roles could be reversed. Brennan-Marquez and Henderson argue that from a democratic legitimacy perspective, it is important whether certain decisions are made by an entity to whom the rule also applies—even if the same decision were to be made by an entity to whom the rule does not apply (Brennan-Marquez and Henderson 2019). This is because certain decisions involve legitimizing the values shared by the moral community—decisions that affect both the maker and recipient of the decision at hand. The artificial juror is not subject to the judgments it would inflict, and so it is not in a role-reversible position with the defendant.

If an artificial moral agent were being morally evaluated by a human moral agent, it would not make sense for the human to adopt reactive attitudes towards the artificial defendant. So, artificial moral agents are importantly not members of the moral community in the same way as human moral agents. As such, it would not be legitimate to include them in juries that are supposed to consist of one's peers in the moral community.

7.5.3 Lessons

Similar considerations arise in other moral decision-making contexts. In fact, most of the situations in which we need a moral agent to make a moral decision have some relational or responsibility-relevant aspects. Does this mean that human moral agents should never be replaced with artificial moral agents? Not necessarily.

Whether we should be indifferent depends on both domain and context. In the case of juries, we should always prefer human moral agents to artificial moral agents—we should not be indifferent between the two. But in the case of mechanics, we might (almost) always be permitted to be indifferent between human and artificial moral agents.

In other domains, the verdict is less straightforward. Consider, for instance, the prospect of artificial doctors. Perhaps the primary consideration in choosing a doctor is medical competence and abilities. Holding that constant, it is important for our doctors to be moral agents. They must make a range of moral decisions—for instance, about resource distribution, about whether to try to change our minds when we refuse medication, about how seriously to take our complaints of pain. They must understand consent and autonomy, and, as moral agents, they will. So, it might seem that moral agency is all that matters in this situation.

Indifference is plausible for one-time appointments and screenings, as these do not require anything beyond moral agency. But for long-term treatment, we might have reason to care about whether we can have an authentic relationship with our doctor. We might want our doctor to relate to us on an emotional level and to feel the gravity of the situation, even if this changes nothing about the medical advice they will give. Similarly, we might want to know that we can direct our reactive attitudes towards our doctors if they fail—that the doctors can be affectively vulnerable to us.

In all these cases, whether we can permissibly prefer human moral agents will depend on how important the social and relational context is to the decision at hand, that is, on the extent to which moral agency is all that matters. The strength of the relationship and responsibility factors in any given situation will determine the extent

to which it is permissible to be indifferent between a human and an artificial moral agent.[78]

7.6 Near-Term Implications

If my argument is successful, then the roles we allow future artificial moral agents to play in the moral community should be restricted. We will have good reason to prefer human moral agents over artificial moral agents when the decision context is influenced by the moral status asymmetry and/or the affective asymmetry—specifically, cases in which relationships and punishment in the form of reactive attitudes or retribution matter. As we have seen, artificial moral agents should not serve as jurors.

At this point, we might wonder what the near-term upshots of this argument are. After all, the prospect of non-conscious AI systems that are genuine moral agents seems distant at best, and impossible at worst. Still, considerations of when we should not be indifferent between human moral agents and artificial moral agents can tell us about existing cases.

7.6.1 Corporations

While AI-based moral agents do not yet exist—and it might be unclear whether or when they will exist—non-conscious moral agents do exist in the form of group agents. List and Pettit have argued that corporations are genuine agents; and once we admit that corporations can be agents, it is not difficult to see how they can be moral agents with moral obligations and responsibility (List and Pettit 2011; List 2018). Corporations lack phenomenal consciousness, and thus they are analogous to artificial

[78] I am also open to the possibility that the answer in some cases depends on the preferences of the individual who is employing the moral agent. For instance, a person who strongly values relationships in their interactions with their doctors might have reason to prefer a human moral agent, even in one-off cases, but a person who only cares about moral agency in the same situation might have reason to prefer an artificial moral agent, even for long-term treatment.

moral agents (Hess 2013). Corporations are subject to the moral status asymmetry as well as the affective asymmetry. They cannot have equal, reciprocal, and authentic relationships with humans, and they cannot engage the same responsibility practices as humans.

It is a strength of my argument that the proposed role of artificial moral agents in the moral community accords well with the existing roles of corporations in the moral community. Corporations are not asked to serve on juries, for instance. Often, in corporate moral decision-making contexts in which relationships or affective responsibility practices are important, we see individuals (i.e., human moral agents) making the moral decisions instead of the corporation as a group agent. For example, executives in a corporation might take responsibility so that people can attach blame to an individual.

The case study involving jurors can also help us see the relevant differences between corporations and artificial moral agents (understood as AI-based systems). Juries are often appealed to as a model of paradigmatic group agency: the jury as a group can be said to have certain beliefs that are held by none of the individual members comprising it (List and Pettit 2011). So, while it is true that we do not let corporations serve on juries, the jury as a whole can be viewed in an equivalent way as corporations—and this might make us worry about the conclusions I have drawn regarding the roles nonconscious moral agents can play in the moral community.[79] After all, non-conscious moral agents (namely, juries consisting of human jurors) make decisions of the kind that I have just argued should be made by conscious moral agents.

But the important difference between traditional juries and artificial moral agents is that while both are nonconscious moral agents, artificial moral agents lack consciousness altogether. Traditional juries still contain consciousness in the form of the individual jurors. As such, when we specify the relevant role as *member of a jury*,

[79] Thank you to Silvia Milano for this objection.

we can consistently hold both that (1) jury members should be conscious moral agents and (2) juries qua group agent need not be conscious moral agents.

The requirement that jury members be conscious moral agents allows a place for the moral status and relationship considerations. Additionally, this explanation accords with our view of juries. We do not expect the jury as a group agent, for instance, to feel the phenomenal aspects of blame when they reach the clearly mistaken verdict—but we might reasonably expect individual jury members to feel this way regarding their individual role in bringing about the group decision.

7.6.2 Current AI

Considerations about the appropriate roles of artificial moral agents can inform how we view the roles of existing AI systems in the moral domain. All existing AI systems are subject to the moral status asymmetry and the affective asymmetry. Thus, for any moral decision for which we should not be indifferent between a human moral agent and an artificial moral agent, we should also prefer a human moral agent to any existing AI system. (Although, as the previous chapter argued, we have additional reasons to avoid using existing AI systems in moral decision-making given that they are not moral agents.)

One might wonder whether my argument has bearing on the discussion of a right to a human decision. Defenders of the right to a human decision struggle to find normative justification for such a right (Huq, 2020). But I have not argued that humans have a right to have a human moral agent make certain moral decisions instead of an artificial moral agent. Instead, I have argued that we have reason to prefer human moral agents in certain moral decision-making contexts. Our reason to prefer human moral agents is based on the roles of emotions and relationships that are unique to human-human social contexts. Whether this reason would ground a right to a human decision is a separate question.

Regarding the development of AI systems, my argument suggests that we should focus on developing systems that are suited to make moral decisions in

contexts that do not require either symmetric moral status or genuine relational abilities. For instance, we should not aim to make AI systems that can serve as jurors or military commanders. But we might want to aim to make AI systems that can serve as doctors in certain contexts or as mechanics.

7.7 Conclusion

Earlier in this chapter, I adopted the following principle:

> *Indifference*:* In moral decision-making contexts, we should be indifferent between a human moral agent and an artificial moral agent *when moral agency is all that matters.*

I have tried to show that there are many cases in which moral agency is not all that matters. We are justified in preferring human moral agents to make moral decisions because of the additional social, relational, and emotional contexts of moral decision-making that only human moral agents (and not non-conscious artificial moral agents) can engage in.

More work must be done to further analyze concrete cases in which we should and should not be indifferent between human moral agents and artificial moral agents. I have offered a start by drawing out the relevant differences between conscious and nonconscious moral agents, identifying the factors that bear on when we should be indifferent between these types of moral agents, and applying these considerations to several cases.

It is important to understand whether AI systems can be moral agents. But it is also important to understand when moral agency matters in a decision-making context and when there are other relevant factors at play.[80]

[80] Thank you to Carissa Véliz and Alison Hills for extensive feedback on this chapter. Thank you to participants in the LMU Workshop on Partiality, Relationships, and AI (2024) for additional comments and discussion.

Chapter 8: Conclusion

I began this dissertation with a thought experiment involving TJ, a mysterious extraterrestrial whose moral agency is in question:

> *TJ*: An extraterrestrial—let us call them TJ—descends to Earth with no apparent means of leaving. While TJ's physiology is clearly non-humanlike, scientists are not yet able to determine precisely how TJ's inner workings operate. TJ can, however, communicate in English, interact with humans, and causally influence things in the world. Before letting TJ roam free in society, we must (among many other considerations) determine where TJ stands in the moral community. As such, a talented group of philosophers—ourselves included—is recruited to determine whether TJ is a moral agent.

Now, at the end of this dissertation, it is fitting to ask whether we are in a better position to determine whether TJ is a moral agent. I believe that we are.

Recall the guiding questions for evaluating the case of TJ: (1) what is a moral agent, (2) is TJ a moral agent, and (3) why should we care?

In response to (1), we can answer that there are different types of moral agents. A deontic moral agent is a source of moral action, and a responsible moral agent is additionally morally responsible for their actions. TJ could, in principle, be either of these, insofar as he has capacities for intentional action and moral concept possession (for deontic moral agency) as well as responsiveness to moral reasons and moral understanding (for responsible moral agency). Moreover, a moral agent need not be phenomenally conscious, and so we can figure out whether TJ is a moral agent without needing a test for consciousness.

In response to (2), we can answer that while we cannot know whether TJ is a moral agent without further testing, we can highlight some features to test for. To know whether TJ is an agent, we can consider how useful attributing mental states to TJ is for explaining and predicting their behavior, and we can investigate the extent to which TJ represents the world by examining their sensory connection to the world and their patterns of folk reasoning. To know whether TJ has moral capacities, we can investigate whether TJ is sensitive to moral features and takes the right-making

features of an action as reasons to perform that action. We can also test how generalizable TJ's moral abilities are to novel situations.

In response to (3), we can answer that TJ's role in the moral community will depend on whether and what type of moral agent TJ is. If TJ is not a moral agent, we should be wary of outsourcing moral decisions to them, as doing so can be detrimental to both us and to the victims of any harm TJ causes. We should be especially cautious when the moral stakes are high. If TJ is a non-conscious moral agent, we should include TJ in the moral community but not in the same way as humans. We should not allow TJ to make moral decisions involving relationships or requiring affective responsibility practices.

I hope that returning to the case of TJ reveals that while this dissertation was about artificial moral agency, it is also about moral agency more generally—and evaluating difficult cases of potential moral agency, especially non-humans.

While I believe I have made progress on some important questions surrounding artificial moral agency, there are still many open avenues for future research—regarding the correct theory of moral agency, the extent to which new developments in AI increase the prospects of artificial moral agency, and the ethical considerations surrounding our use of AI systems in moral contexts. I look forward to working on these topics in the future.

References

Aaltola, Elisa. 2014. "Affective Empathy as Core Moral Agency: Psychopathy, Autism and Reason Revisited." *Philosophical Explorations* 17 (1): 76–92. https://doi.org/10.1080/13869795.2013.825004.

Aharoni, Eyal, Sharlene Fernandes, Daniel J. Brady, Caelan Alexander, Michael Criner, Kara Queen, Javier Rando, Eddy Nahmias, and Victor Crespo. 2024. "Attributions toward Artificial Agents in a Modified Moral Turing Test." *Scientific Reports* 14 (1): 8458. https://doi.org/10.1038/s41598-024-58087-7.

Allen, Colin, Iva Smit, and Wendell Wallach. 2005. "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches." *Ethics and Information Technology* 7 (3): 149–55. https://doi.org/10.1007/s10676-006-0004-4.

Allen, Colin, Gary Varner, and Jason Zinser. 2000. "Prolegomena to Any Future Artificial Moral Agent." *Journal of Experimental and Theoretical Artificial Intelligence* 12 (3): 251–61. https://doi.org/10.1080/09528130050111428.

Allen, Colin, and Wendell Wallach. 2012. "Moral Machines: Contradiction in Terms or Abdication of Human Responsibility?" In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by Patrick Lin, Keith Abney, and George A. Bekey, 55–68. Cambridge: The MIT Press.

Allyn, Bobby. 2024. "Eight Newspapers Sue OpenAI, Microsoft for Copyright Infringement." *NPR*, April 30, 2024. https://www.npr.org/2024/04/30/1248141220/lawsuit-openai-microsoft-copyright-infringement-newspaper-tribune-post.

Almeida, Guilherme F.C.F., José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. 2024. "Exploring the Psychology of LLMs' Moral and Legal Reasoning." *Artificial Intelligence* 333 (104145). https://doi.org/10.1016/j.artint.2024.104145.

Anderson, Michael, Susan Leigh Anderson, and Chris Armen. 2006. "MedEthEx: A Prototype Medical Ethics Advisor." *Proceedings of the National Conference on Artificial Intelligence* 21 (2): 1759–65.

Anderson, Michael, and Susan Leigh Anderson. 2007. "Machine Ethics: Creating an Ethical Intelligent Agent." *The AI Magazine*, 2007.

Anscombe, G. E. M. 1957. *Intention*. Oxford: Basil Blackwell.

Anthropic. 2023. "Claude's Constitution." Https://Www.Anthropic.Com/News/Claudes-Constitution. 2023.

Aristotle. 2019. *Nicomachean Ethics*. Edited by Terence Irwin. Cambridge: Hackett Publishing Company.

Arkin, Ronald C. 2010. "The Case for Ethical Autonomy in Unmanned Systems." *Journal of Military Ethics* 9 (4): 332–41. https://doi.org/10.1080/15027570.2010.536402.

Armstrong, David M. 1993. *A Materialist Theory of the Mind*. London: Routledge.

Arpaly, Nomy. 2002. "Moral Worth." *The Journal of Philosophy* 99 (5): 223–45. https://doi.org/10.2307/3655647.

———. 2003. *Unprincipled Virtue: An Inquiry Into Moral Agency*. New York: Oxford University Press.

Asaro, Peter M. 2006. "What Should We Want from a Robot Ethic?" *International Review of Information Ethics* 12:10–16. https://doi.org/10.29173/irie134.

Balog, Katalin. 2012. "Acquaintance and the Mind-Body Problem." In *New Perspectives on Type Identity: The Mental and the Physical*, edited by Simone Gozzano and Christopher S. Hill, 16–42. Cambridge: Cambridge University Press.

Barandiaran, Xabier E., Ezequiel Di Paolo, and Marieke Rohde. 2009. "Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-Temporality in Action." *Adaptive Behavior* 17 (5): 367–86. https://doi.org/10.1177/1059712309343819.

Basl, John. 2014. "Machines as Moral Patients We Shouldn't Care about (yet): The Interests and Welfare of Current Machines." *Philosophy & Technology* 27 (1): 79–96. https://doi.org/10.1007/s13347-013-0122-y.

Beavers, Anthony F. 2012. "Moral Machines and the Threat of Ethical Nihilism." In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by Patrick Lin, Keith Abney, and George A. Bekey, 333–44. Cambridge: The MIT Press.

Behdadi, Dorna, and Christian Munthe. 2020. "A Normative Approach to Artificial Moral Agency." *Minds and Machines* 30 (2): 195–218. https://doi.org/10.1007/s11023-020-09525-8.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. New York, NY, USA: ACM. https://doi.org/10.1145/3442188.3445922.

Bender, Emily M, and Alexander Koller. 2020. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–98. https://doi.org/10.18653/v1/2020.acl-main.463.

Berglund, Lukas, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. "The Reversal Curse: LLMs Trained on 'A Is B' Fail to Learn 'B Is A.'" https://arxiv.org/abs/2309.12288v4.

Betker, James, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, et al. 2023. "Improving Image Generation with Better Captions." https://cdn.openai.com/papers/dall-e-3.pdf.

Biggio, Battista, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. "Evasion Attacks against Machine Learning at Test Time." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 8190 LNAI:387–402. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40994-3_25/COVER.

Björnsson, Gunnar, and Kendy Hess. 2017. "Corporate Crocodile Tears?" *Philosophy and Phenomenological Research* 94 (2): 273–98. https://doi.org/10.2307/48578761.

Block, Ned. 1995. "On a Confusion about a Function of Consciousness." *Behavioral and Brain Sciences* 18 (2): 227–47. https://doi.org/10.1017/S0140525X00038188.

Borg, Jana Schaich, and Walter P. Sinnott-Armstrong. 2013. "Do Psychopaths Make Moral Judgments?" In *Handbook on Psychopathy and Law*, edited by Kent A. Kiehl and Walter P. Sinnott-Armstrong, 107–28. Oxford: Oxford University Press.

Bradford, Gwen. 2023. "Consciousness and Welfare Subjectivity." *Noûs* 57 (4): 905–21. https://doi.org/10.1111/nous.12434.

Bratman, Michael. 1987. *Intention, Plans, and Practical Reason*. Cambridge: Harvard University Press.

Brennan, Andrew. 1984. "The Moral Standing of Natural Objects." *Environmental Ethics* 6 (1): 35–56.

Brennan-Marquez, Kiel, and Stephen E. Henderson. 2019. "Artificial Intelligence and Role-Reversible Judgment." *Journal of Criminal Law and Criminology* 109 (2): 137–64.

Brey, Philip. 2014. "From Moral Agents to Moral Factors: The Structural Ethics Approach." In *The Moral Status of Technical Artefacts*, edited by Peter Kroes and Peter-Paul Verbeek, 125–42. Dordrecht: Springer.

Bryson, Joanna J. 2018. "Patiency Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics." *Ethics and Information Technology* 20 (1): 15–26. https://doi.org/10.1007/s10676-018-9448-6.

Buckner, Cameron. 2019. "Deep Learning: A Philosophical Introduction." *Philosophy Compass* 14 (10). https://doi.org/10.1111/phc3.12625.

Buolamwini, Joy. 2018. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR* 81:77–91.

Burge, Tyler. 1979. “Individualism and the Mental.” *Midwest Studies in Philosophy* 4 (1): 73–122.

Burroughs, Michael D. 2020. “Navigating the Penumbra: Children and Moral Responsibility.” *The Southern Journal of Philosophy* 58 (1): 77–101. https://doi.org/10.1111/sjp.12352.

Butlin, Patrick. 2021. “Reinforcement Learning and Artificial Agency.” *Mind and Language*. https://doi.org/10.1111/mila.12458.

———. 2023. “Sharing Our Concepts with Machines.” *Erkenntnis* 88 (7): 3079–95. https://doi.org/10.1007/s10670-021-00491-w.

Butlin, Patrick, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, et al. 2023. “Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.” https://arxiv.org/abs/2308.08708v3.

Cappelen, Herman, and Josh Dever. 2020. “Acting Without Me: Corporate Agency and the First Person Perspective.” In *The Routledge Handbook of Linguistic Reference*, edited by Stephen Biggs and Heimir Geirsson, First Edition, 499–514. New York: Routledge.

Casper, Stephen, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, et al. 2023. “Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback.” https://arxiv.org/abs/2307.15217v2.

Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

———. 2023. “Does Thought Require Sensory Grounding? From Pure Thinkers to Large Language Models.” *Proceedings and Addresses of the American Philosophical Association*, no. 97, 22–45.

Chan, Alan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, et al. 2023. “Harms from Increasingly Agentic Algorithmic Systems.” *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 651–66. https://doi.org/10.1145/3593013.3594033.

Chao, Patrick, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. “Jailbreaking Black Box Large Language Models in Twenty Queries.” https://arxiv.org/abs/2310.08419v4.

Clement, Grace. 2013. "Animals and Moral Agency: The Recent Debate and Its Implications." *Journal of Animal Ethics* 3 (1): 1–14. https://doi.org/10.5406/janimalethics.3.1.0001.

Coeckelbergh, Mark. 2009. "Virtual Moral Agency, Virtual Moral Responsibility: On the Moral Significance of the Appearance, Perception, and Performance of Artificial Agents." *AI & Society* 24 (2): 181–89. https://doi.org/10.1007/s00146-009-0208-3.

———. 2010. "Moral Appearances: Emotions, Robots, and Human Morality." *Ethics and Information Technology* 12 (3): 235–41. https://doi.org/10.1007/s10676-010-9221-y.

Coelho Mollo, Dimitri, and Raphaël Millière. 2023. "The Vector Grounding Problem." https://arxiv.org/abs/2304.01481.

Coleman, K. G. 2005. "Computing and Moral Responsibility." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/spr2005/entries/computing-responsibility/.

Constantinescu, Mihaela, and Roger Crisp. 2022. "Can Robotic AI Systems Be Virtuous and Why Does This Matter?" *International Journal of Social Robotics* 14 (6): 1547–57. https://doi.org/10.1007/s12369-022-00887-w.

Crary, Alice. 2016. *Inside Ethics: On the Demands of Moral Thought*. Cambridge: Harvard University Press.

Danaher, John. 2016. "Robots, Law and the Retribution Gap." *Ethics and Information Technology* 18 (4): 299–309. https://doi.org/10.1007/s10676-016-9403-3.

———. 2019. "The Ethics of Algorithmic Outsourcing in Everyday Life." In *Algorithmic Regulation*, edited by Karen Yeung and Martin Lodge, 98–117. Oxford: Oxford University Press.

———. 2022. "Tragic Choices and the Virtue of Techno-Responsibility Gaps." *Philosophy & Technology* 35 (2): 26. https://doi.org/10.1007/s13347-022-00519-1.

Darling, Kate. 2016. "Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior towards Robotic Objects." In *Robot Law*, edited by Ryan Calo, A. Michael Froomkin, and Ian Kerr, 213–32. Edward Elgar Publishing.

Davidson, Donald. 2001a. *Essays on Actions and Events*. Second Edition. Oxford: Oxford University Press.

———. 2001b. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

Davidsson, Paul, and Stefan J. Johansson. 2005. "On the Metaphysics of Agents." *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, 1299–1300. https://doi.org/10.1145/1082473.1082742.

DeGrazia, David. 1996. *Taking Animals Seriously: Mental Life and Moral Status*. Cambridge: Cambridge University Press.

Deijl, Willem van der. 2021. "The Sentience Argument for Experientialism about Welfare." *Philosophical Studies* 178 (1): 187–208. https://doi.org/10.1007/s11098-020-01427-w.

Dennett, Daniel C. 1980. *Brainstorms*. Cambridge: The MIT Press.

———. 1987. *The Intentional Stance*. Cambridge: The MIT Press.

———. 2009. "Intentional Systems Theory." In *The Oxford Handbook of Philosophy of Mind*, edited by Ansgar Beckermann, Brian P. McLaughlin, and Sven Walter, 339–50. Oxford: Oxford University Press.

Dretske, Fred I. 1988. *Explaining Behavior*. Cambridge: The MIT Press.

Dreyfus, Hubert L. 1965. "Alchemy and Artificial Intelligence." *RAND Corp, Santa Monica CA*.

Driver, Julia. 2015. "Appraisability, Attributability, and Moral Agency." In *The Nature of Moral Responsibility: New Essays*, edited by Randolph Clarke, Michael McKenna, and Angela M. Smith. Oxford: Oxford University Press.

Dror, Lidal. 2023. "Is There an Epistemic Advantage to Being Oppressed?" *Noûs* 57 (3): 618–40. https://doi.org/10.1111/nous.12424.

Dung, Leonard. 2024. "Understanding Artificial Agency." *The Philosophical Quarterly*. https://doi.org/10.1093/pq/pqae010.

Dziri, Nouha, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, et al. 2023. "Faith and Fate: Limits of Transformers on Compositionality." In *Proceedings of the 37th Conference on Neural Information Processing Systems*. https://proceedings.neurips.cc/paper_files/paper/2023/file/deb3c28192f979302c157cb653c15e90-Paper-Conference.pdf.

Eggert, Linda. 2023. "Autonomised Harming." *Philosophical Studies*. https://doi.org/10.1007/s11098-023-01990-y.

Elish, Madeleine Clare. 2019. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." *Engaging Science, Technology, and Society* 5:40–60. https://doi.org/10.17351/ests2019.260.

Elyoseph, Zohar, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. 2023. "ChatGPT Outperforms Humans in Emotional Awareness Evaluations." *Frontiers in Psychology* 14. https://doi.org/10.3389/fpsyg.2023.1199058.

Evans, Gareth. 1982. *The Varieties of Reference*. Edited by John McDowell. Oxford: Clarendon Press.

Farn, Nicholas, and Richard Shin. 2023. "ToolTalk: Evaluating Tool-Usage in a Conversational Setting." https://arxiv.org/abs/2311.10775.

Feinberg, Joel. 2007. "Psychological Egoism." In *Ethical Theory: An Anthology*, edited by Russ Shafer-Landau, 167–77. Blackwell Publishers.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Fitzpatrick, Simon. 2017. "Animal Morality: What Is the Debate About?" *Biology and Philosophy* 32 (6): 1151–83. https://doi.org/10.1007/s10539-017-9599-6.

Floridi, Luciano, and J. W. Sanders. 2004. "On the Morality of Artificial Agents." *Minds and Machines* 14 (3): 349–79. https://doi.org/10.4324/9781003074991-30.

Fodor, Jerry A. 1975. *The Language of Thought*. Cambridge: Harvard University Press.

———. 1981. *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Brighton: Harvester.

Formosa, Paul, and Malcolm Ryan. 2021. "Making Moral Machines: Why We Need Artificial Moral Agents." *AI & Society* 36 (3): 839–51. https://doi.org/10.1007/s00146-020-01089-6.

Fossa, Fabio. 2018. "Artificial Moral Agents: Moral Mentors or Sensible Tools?" *Ethics and Information Technology* 20 (2): 115–26. https://doi.org/10.1007/s10676-018-9451-y.

Franklin, Stan, and Art Graesser. 1997. "Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents." In *Intelligent Agents III Agent Theories, Architectures, and Languages. ATAL 1996. Lecture Notes in Computer Science*, edited by J.P. Müller, M.J. Wooldridge, and N.R. Jennings, 1193:21–35. Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0013570.

Freitas, Julian De, Ahmet K. Uguralp, Zeliha O. Uguralp, and Puntoni Stefano. 2024. "AI Companions Reduce Loneliness." *Harvard Business School Working Paper 24-078*.

French, Peter A. 1979. "The Corporation as a Moral Person." *American Philosophical Quarterly* 16 (3): 207–15.

Frey, R. G. 2005. “Animals.” In *The Oxford Handbook of Practical Ethics*, edited by Hugh LaFollette, 161–87. Oxford: Oxford University Press.

Friedman, Batya, and Peter H Kahn. 1992. “Human Agency and Responsible Computing: Implications for Computer System Design.” *Journal of Systems and Software* 17 (1): 7–14.

Gabriel, Iason. 2020. “Artificial Intelligence, Values, and Alignment.” *Minds and Machines* 30 (3): 411–37. https://doi.org/10.1007/s11023-020-09539-2.

Gabriel, Iason, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, et al. 2024. “The Ethics of Advanced AI Assistants.” https://arxiv.org/abs/2404.16244.

Gao, Yuan, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. 2024. “Take Caution in Using LLMs as Human Surrogates: Scylla Ex Machina.” https://arxiv.org/abs/2410.19599v2.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, et al. 2024. “Gemini: A Family of Highly Capable Multimodal Models.” https://arxiv.org/abs/2312.11805.

Gerdes, Anne. 2016. “The Issue of Moral Consideration in Robot Ethics.” *ACM SIGCAS Computers and Society* 45 (3): 274–79. https://doi.org/10.1145/2874239.2874278.

Gogoshin, Dane Leigh. 2024. “A Way Forward for Responsibility in the Age of AI.” *Inquiry*, 1–34. https://doi.org/10.1080/0020174X.2024.2312455.

Goldstein, Simon, and Benjamin A. Levinstein. 2024. “Does ChatGPT Have a Mind?” https://arxiv.org/abs/2407.11015.

Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy. 2015. “Explaining and Harnessing Adversarial Examples.” https://arxiv.org/abs/1412.6572.

Graff, Joris. 2024. “Moral Sensitivity and the Limits of Artificial Moral Agents.” *Ethics and Information Technology* 26 (1): 13. https://doi.org/10.1007/s10676-024-09755-9.

Grodzinsky, Frances S., Keith W. Miller, and Marty J. Wolf. 2008. “The Ethics of Designing Artificial Agents.” *Ethics and Information Technology* 10 (2–3): 115–21. https://doi.org/10.1007/s10676-008-9163-9.

Gudmunsen, Zacharus. 2024. “The Moral Decision Machine: A Challenge for Artificial Moral Agency Based on Moral Deference.” *AI and Ethics*. https://doi.org/10.1007/s43681-024-00444-3.

Gunkel, David J. 2018. *Robot Rights*. Cambridge: MIT Press.

———. 2020. *How to Survive a Robot Invasion: Rights, Responsibility, and AI*. New York: Routledge.

———. 2022. "The Relational Turn: Thinking Robots Otherwise." In *Social Robotics and the Good Life: The Normative Side of Forming Emotional Bonds with Robots*, edited by Janina Loh and Wulf Loh. Transcript Verlag.

Haas, Julia. 2022. "Reinforcement Learning: A Brief Guide for Philosophers of Mind." *Philosophy Compass* 17 (9). https://doi.org/10.1111/phc3.12865.

Hadero, Haleluya, and David Bauder. 2023. "The New York Times Sues OpenAI and Microsoft for Using Its Stories to Train Chatbots." *The Associated Press*, December 27, 2023. https://apnews.com/article/nyt-new-york-times-openai-microsoft-6ea53a8ad3efa06ee4643b697df0ba57.

Hakli, Raul, and Pekka Mäkelä. 2019. "Moral Responsibility of Robots and Hybrid Agents." *The Monist* 102 (2): 259–75. https://doi.org/10.1093/MONIST/ONZ009.

Haksar, Vinit. 1998. "Moral Agents." In *Routledge Encyclopedia of Philosophy*, edited by Edward Craig. Routledge.

Hanson, F. Allan. 2009. "Beyond the Skin Bag: On the Moral Responsibility of Extended Agencies." *Ethics and Information Technology* 11 (1): 91–99. https://doi.org/10.1007/s10676-009-9184-z.

Harman, Elizabeth. 1999. "Creation Ethics: The Moral Status of Early Fetuses and the Ethics of Abortion." *Philosophy & Public Affairs* 28 (4): 310–24.

Harnad, Stevan. 1990. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42 (1–3): 335–46. https://doi.org/10.1016/0167-2789(90)90087-6.

Heider, Fritz, and Marianne Simmel. 1944. "An Experimental Study of Apparent Behavior." *The American Journal of Psychology* 57 (2): 243. https://doi.org/10.2307/1416950.

Hendrycks, Dan, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2023. "Aligning AI With Shared Human Values." https://arxiv.org/abs/2008.02275.

Hendrycks, Dan, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt. 2022. "What Would Jiminy Cricket Do? Towards Agents That Behave Morally." https://arxiv.org/abs/2110.13136.

Hess, Kendy M. 2013. "'If You Tickle Us....': How Corporations Can Be Moral Agents Without Being Persons." *Journal of Value Inquiry* 47:319–35. https://doi.org/10.1007/s10790-013-9391-z.

Hills, Alison. 2009. “Moral Testimony and Moral Epistemology.” *Ethics* 120 (1): 94–127. https://doi.org/10.1086/648610.

Himma, Kenneth Einar. 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.

Huang, Jie, and Kevin Chen-Chuan Chang. 2022. “Towards Reasoning in Large Language Models: A Survey.” https://arxiv.org/abs/2212.10403.

Huang, Jie, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. “Large Language Models Cannot Self-Correct Reasoning Yet.” https://arxiv.org/abs/2310.01798.

Hughes, John, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, and Mrinank Sharma. 2024. “Best-of-N Jailbreaking.” https://arxiv.org/abs/2412.03556.

Huh, Minyoung, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. “The Platonic Representation Hypothesis.” https://arxiv.org/abs/2405.07987.

Hume, David. 2000. *A Treatise of Human Nature*. Edited by David Fate Norton and Mary J. Norton. Oxford: Oxford University Press.

Hupkes, Dieuwke, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, et al. 2023. “A Taxonomy and Review of Generalization Research in NLP.” *Nature Machine Intelligence* 5 (10). https://doi.org/10.1038/s42256-023-00729-y.

Huq, Aziz Z. 2020. “A Right to a Human Decision.” Article. *Virginia Law Review* 106 (3): 611–88.

IBM Technology. 2023. “Why Large Language Models Hallucinate.” Https://Www.Youtube.Com/Watch?V=cfqtFvWOfg0. 2023.

Illies, Christian F.R., and Anthonie Meijers. 2014. “Artefacts, Agency, and Action Schemes.” In *The Moral Status of Technical Artefacts*, edited by Peter Kroes and Peter-Paul Verbeek, 159–84. Dordrecht: Springer.

Jackson, Frank. 1986. “What Mary Didn’t Know.” *The Journal of Philosophy* 83 (5): 291. https://doi.org/10.2307/2026143.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. “Survey of Hallucination in Natural Language Generation.” *ACM Computing Surveys* 55 (12). https://doi.org/10.1145/3571730.

Jiang, Liwei, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, et al. 2022. "Can Machines Learn Morality? The Delphi Experiment." https://arxiv.org/abs/2110.07574v2.

Johnson, Deborah G. 2006. "Computer Systems: Moral Entities but Not Moral Agents." *Ethics and Information Technology* 8 (4): 195–204. https://doi.org/10.1007/s10676-006-9111-5.

Johnson, Deborah G., and Keith W. Miller. 2008. "Un-Making Artificial Moral Agents." *Ethics and Information Technology* 10 (2–3): 123–33. https://doi.org/10.1007/s10676-008-9174-6.

Johnson, Deborah G., and Merel Noorman. 2014. "Artefactual Agency and Artefactual Moral Agency." In *The Moral Status of Technical Artefacts*, edited by Peter Kroes and Peter-Paul Verbeek, 143–58. Dordrecht: Springer.

Johnson, Deborah G., and Thomas M. Powers. 2008. "Computers as Surrogate Agents." In *Information Technology and Moral Philosophy*, edited by Jeroen van den Hoven and John Weckert, 251–69. Cambridge: Cambridge University Press.

Johnson, Lawrence. 1993. *A Morally Deep World: An Essay on Moral Significance and Environmental Ethics*. Cambridge: Cambridge University Press.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature 2021 596:7873* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Kagan, Shelly. 2019. *How to Count Animals, More or Less*. Oxford: Oxford University Press.

Kant, Immanuel. 2018. *Groundwork for the Metaphysics of Morals*. Edited by Allen W. Wood. New Haven: Yale University Press.

Kaufmann, Timo, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. "A Survey of Reinforcement Learning from Human Feedback." https://arxiv.org/abs/2312.14925.

Kennett, Jeanette. 2002. "Autism, Empathy and Moral Agency." *The Philosophical Quarterly* 52 (208): 340–57. https://doi.org/10.1111/1467-9213.00272.

Kenton, Zachary, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. 2023. "Discovering Agents." *Artificial Intelligence* 103963 (322). https://doi.org/10.1016/j.artint.2023.103963.

Kiener, Maximilian. 2022. "Can We Bridge AI's Responsibility Gap at Will?" *Ethical Theory and Moral Practice* 25 (4): 575–93. https://doi.org/10.1007/s10677-022-10313-9.

———. 2024. “AI and Responsibility: No Gap, but Abundance.” *Journal of Applied Philosophy*. https://doi.org/10.1111/japp.12765.

Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2017. “Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV).” *Proceedings of the 35th International Conference on Machine Learning* 6 (80): 2668–77. https://proceedings.mlr.press/v80/kim18d.html.

Kohlberg, Lawrence. 1981. *The Philosophy of Moral Development*. Harper & Row.

Korsgaard, Christine. 2014. “The Normative Constitution of Agency.” In *Rational and Social Agency: The Philosophy of Michael Bratman*, edited by Manuel Vargas and Gideon Yaffe, 190–214. Oxford: Oxford University Press.

Kuo, Mu-Tien, Chih-Chung Hsueh, and Richard Tzong-Han Tsai. 2023. “Large Language Models on the Chessboard: A Study on ChatGPT’s Formal Language Comprehension and Complex Reasoning Skills.” https://arxiv.org/abs/2308.15118.

Latour, Bruno. 1987. *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge: Harvard University Press.

Laukyte, Migle. 2017. “Artificial Agents among Us: Should We Recognize Them as Agents Proper?” *Ethics and Information Technology* 19 (1): 1–17. https://doi.org/10.1007/s10676-016-9411-3.

Lederman, Harvey, and Kyle Mahowald. 2024. “Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs.” *Transactions of the Association for Computational Linguistics*, no. 12, 1087–1103. https://doi.org/10.1162/tacl_a_00690.

Leivada, Evelina, Gary Marcus, Fritz Günther, and Elliot Murphy. 2024. “A Sentence Is Worth a Thousand Pictures: Can Large Language Models Understand Hum4n L4ngu4ge and the W0rld behind W0rds?” https://arxiv.org/abs/2308.00109.

Levy, Neil. 2008. “The Responsibility of the Psychopath Revisited.” *Philosophy, Psychiatry, & Psychology* 14 (2): 129–38. https://doi.org/10.1353/PPP.0.0003.

———. 2014. *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.

Li, Kenneth, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. “Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task.” https://arxiv.org/abs/2210.13382.

Liao, S. Matthew. 2020. "The Moral Status and Rights of Artificial Intelligence." In *Ethics of Artificial Intelligence*, edited by S. Matthew Liao, 480–503. New York: Oxford University Press.

List, Christian. 2018. "What Is It Like to Be a Group Agent?" *Noûs* 52 (2): 295–319. https://doi.org/10.1111/nous.12162.

———. 2021. "Group Agency and Artificial Intelligence." *Philosophy and Technology* 34 (4): 1213–42. https://doi.org/10.1007/s13347-021-00454-7.

List, Christian, and Philip Pettit. 2011. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.

Litton, Paul J. 2008. "Responsibility Status of the Psychopath: On Moral Reasoning and Rational Self-Governance." *Rutgers Law Journal* 39:349–92. http://scholarship.law.missouri.edu/facpubs.

Lu, Sheng, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2024. "Are Emergent Abilities in Large Language Models Just In-Context Learning?" https://arxiv.org/abs/2309.01809.

Luka Inc. n.d. "Replika." Https://Replika.Com. Accessed October 11, 2024. https://replika.com.

Mabaso, Bongani Andy. 2021. "Computationally Rational Agents Can Be Moral Agents." *Ethics and Information Technology* 23 (2): 137–45. https://doi.org/10.1007/s10676-020-09527-1.

Machery, Edouard, and Stephen Stich. 2022. "The Moral/Conventional Distinction." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/sum2022/entries/moral-conventional/.

Malle, Bertram F. 2016. "Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots." *Ethics and Information Technology* 18 (4): 243–56. https://doi.org/10.1007/s10676-015-9367-8.

Margolis, Eric, and Stephen Laurence. 2023. "Concepts." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/fall2023/entries/concepts/.

Matthias, Andreas. 2004. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6 (3): 175–83. https://doi.org/10.1007/s10676-004-3422-1.

May, Todd. 2014. "Moral Individualism, Moral Relationalism, and Obligations to Non-Human Animals." *Journal of Applied Philosophy* 31 (2): 155–68. https://doi.org/10.1111/japp.12055.

McCarthy, John. 1979. “Ascribing Mental Qualities to Machines.” *Computer Science Department, Stanford University*.

McCoy, R. Thomas, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024a. “Embers of Autoregression Show How Large Language Models Are Shaped by the Problem They Are Trained to Solve.” *Proceedings of the National Academy of Sciences* 121 (41). https://doi.org/10.1073/pnas.2322420121.

———. 2024b. “When a Language Model Is Optimized for Reasoning, Does It Still Show Embers of Autoregression? An Analysis of OpenAI O1,” October.

McDowell, John. 1979. “Virtue and Reason.” *Monist* 62 (3): 331–50. https://doi.org/10.5840/monist197962319.

McGeer, Victoria. 2008. “Varieties of Moral Agency: Lessons from Autism (and Psychopathy).” In *Moral Psychology Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, edited by Walter Sinnott-Armstrong, 227–57. Cambridge: The MIT Press.

McKenna, Michael. 2012. *Conversation and Responsibility*. Oxford: Oxford University Press.

———. 2013. “Reasons-Responsiveness, Agents, and Mechanisms.” In *Oxford Studies in Agency and Responsibility, Volume 1*, edited by David Shoemaker, 151–83. Oxford University Press.

McMahan, Jeff. 2002. *The Ethics of Killing: Problems at the Margins of Life*. Oxford: Oxford University Press.

Mecacci, Giulio, and Filippo Santoni de Sio. 2020. “Meaningful Human Control as Reason-Responsiveness: The Case of Dual-Mode Vehicles.” *Ethics and Information Technology* 22 (2): 103–15. https://doi.org/10.1007/s10676-019-09519-w.

Miller, Keith, and David Larson. 2005. “Angels and Artifacts: Moral Agents in the Age of Computers and Networks.” *Information, Communication & Ethics in Society* 3:151–57. https://doi.org/10.1108/14779960580000269.

Millière, Raphaël, and Cameron Buckner. 2024a. “A Philosophical Introduction to Language Models Part I: Continuity With Classic Debates.” http://arxiv.org/abs/2401.03910.

———. 2024b. “A Philosophical Introduction to Language Models Part II: The Way Forward.” https://arxiv.org/abs/2405.03207.

Millikan, Ruth. 2000. *On Clear and Confused Ideas*. Cambridge: Cambridge University Press.

Misselhorn, Catrin. 2018. "Artificial Morality. Concepts, Issues and Challenges." *Society* 55 (2): 161–69. https://doi.org/10.1007/s12115-018-0229-y.

Mittelstadt, Brent, Chris Russell, and Sandra Wachter. 2019. "Explaining Explanations in AI." *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287574.

Montreuil, Marjorie, Crystal Noronha, Nadia Floriani, and Franco A. Carnevale. 2018. "Children's Moral Agency: An Interdisciplinary Scoping Review." *Journal of Childhood Studies*, 17–30. https://doi.org/10.18357/jcs.v43i2.18575.

Moor, James 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.

———. 2009. "Four Kinds of Ethical Robots." *Philosophy Now*, 2009.

Musker, Sam, and Ellie Pavlick. 2024. "Testing Causal Models of Word Meaning in LLMs." *Proceedings of the Annual Meeting of the Cognitive Science Society* 46. https://escholarship.org/uc/item/0wc4315w.

Nagel, Thomas. 1974. "What Is It Like to Be a Bat?" *The Philosophical Review* 83 (4): 435–50. https://doi.org/10.2307/2183914.

Nailer, Timothy. 2022. "Moral Agency." Master of Philosophy Thesis, The University of Adelaide. https://philarchive.org/rec/NAIMA.

Nanda, Neel, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. "Progress Measures for Grokking via Mechanistic Interpretability." https://arxiv.org/abs/2301.05217.

Neely, Erica L. 2014. "Machines and the Moral Community." *Philosophy & Technology* 27 (1): 97–111. https://doi.org/10.1007/s13347-013-0114-y.

Neuhäuser, Christian. 2015. "Some Sceptical Remarks Regarding Robot Responsibility and a Way Forward." In *Collective Agency and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation*, edited by Catrin Misselhorn, 131–46. Springer International Publishing.

Newen, Albert, and Andreas Bartels. 2007. "Animal Minds and the Possession of Concepts." *Philosophical Psychology* 20 (3): 283–308. https://doi.org/10.1080/09515080701358096.

Nikankin, Yaniv, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 2024. "Arithmetic Without Algorithms: Language Models Solve Math With a Bag of Heuristics." https://arxiv.org/abs/2410.21272.

Noorman, Merel. 2023. "Computing and Moral Responsibility." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/spr2023/entries/computing-responsibility/.

Noorman, Merel, and Deborah G. Johnson. 2014. "Negotiating Autonomy and Responsibility in Military Robots." *Ethics and Information Technology* 16 (1): 51–62. https://doi.org/10.1007/s10676-013-9335-0.

Nussbaum, Martha C. 1990. *Love's Knowledge: Essays on Philosophy and Literature*. Oxford: Oxford University Press.

Nyholm, Sven. 2018. "Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci." *Science and Engineering Ethics* 24 (4): 1201–19. https://doi.org/10.1007/s11948-017-9943-x.

———. 2020. *Humans and Robots: Ethics, Agency, and Anthropomorphism*. London: Rowman & Littlefield.

Oimann, Ann-Katrien. 2023. "The Responsibility Gap and LAWS: A Critical Mapping of the Debate." *Philosophy & Technology* 36 (1): 3. https://doi.org/10.1007/s13347-022-00602-7.

OpenAI. 2024. "Learning to Reason with LLMs." 2024. https://openai.com/index/learning-to-reason-with-llms/.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2023. "GPT-4 Technical Report." https://arxiv.org/abs/2303.08774.

Owens, David. 2012. *Shaping the Normative Landscape*. Oxford: Oxford University Press.

Papagni, Guglielmo, and Sabine Koeszegi. 2021. "A Pragmatic Approach to the Intentional Stance Semantic, Empirical and Ethical Considerations for the Design of Artificial Agents." *Minds and Machines* 31 (4): 505–34. https://doi.org/10.1007/s11023-021-09567-6.

Parthemore, Joel, and Blay Whitby. 2013. "What Makes Any Agent a Moral Agent? Reflections on Machine Consciousness and Moral Agency." *International Journal of Machine Consciousness* 5 (2): 105–29. https://doi.org/10.1142/S1793843013500017.

———. 2014. "Moral Agency, Moral Responsibility, and Artifacts: What Existing Artifacts Fail to Achieve (and Why), and Why They, Nevertheless, Can (and Do!) Make Moral Claims upon Us." *International Journal of Machine Consciousness* 6 (2): 141–61. https://doi.org/10.1142/S1793843014400162.

Parvini, Sarah, and Matt O'Brien. 2024. "News Nonprofit Sues ChatGPT Maker OpenAI and Microsoft for 'Exploitative' Copyright Infringement." *The Associated Press*, June 27, 2024. https://apnews.com/article/ai-media-lawsuits-center-for-investigative-reporting-chatgpt-mother-jones-c48452889750479410b65a119537746c.

Peterson, Martin, and Andreas Spahn. 2011. "Can Technological Artefacts Be Moral Agents?" *Science and Engineering Ethics* 17 (3): 411–24. https://doi.org/10.1007/s11948-010-9241-3.

Petrov, Aleksandar, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. "Language Model Tokenizers Introduce Unfairness Between Languages." In *Advances in Neural Information Processing Systems*, edited by A Oh, T Naumann, A Globerson, K Saenko, M Hardt, and S Levine, 36:36963–90. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/74bb24dca8334adce292883b4b651eda-Paper-Conference.pdf.

Pettit, Philip. 1998. "Desire." In *Routledge Encyclopedia of Philosophy*. London: Routledge. https://plato.stanford.edu/archives/spr2023/entries/action/.

———. 2007. "Responsibility Incorporated." *Ethics* 117:171–201. https://doi.org/10.1086/510695.

Piñeros Glasscock, Juan S., and Sergio Tenenbaum. 2023. "Action." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman.

Pitt, David. 2022. "Mental Representation." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/fall2022/entries/mental-representation/>.

Podschwadek, Frodo. 2017. "Do Androids Dream of Normative Endorsement? On the Fallibility of Artificial Moral Agents." *Artificial Intelligence and Law* 25 (3): 325–39. https://doi.org/10.1007/s10506-017-9209-6.

Popa, Elena. 2021. "Human Goals Are Constitutive of Agency in Artificial Intelligence (AI)." *Philosophy & Technology* 34 (4): 1731–50. https://doi.org/10.1007/s13347-021-00483-2.

Press, Ofir, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. "Measuring and Narrowing the Compositionality Gap in Language Models." https://arxiv.org/abs/2210.03350.

Purves, Duncan, Ryan Jenkins, and Bradley J. Strawser. 2015. "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons." *Ethical Theory and Moral Practice* 18 (4): 851–72. https://doi.org/10.1007/s10677-015-9563-y.

Putnam, Hilary. 1975a. *Mind, Language, and Reality*. Cambridge: Cambridge University Press.

———. 1975b. "The Meaning of 'Meaning''." *Minnesota Studies in the Philosophy of Science* 7:131–93.

———. 1982. *Reason, Truth and History*. Cambridge: Cambridge University Press.

Rest, James. 1979. *Development in Judging Moral Issues*. University of Minnesota Press.

Rodogno, Raffaele. 2016. "Robots and the Limits of Morality." In *Social Robots: Boundaries, Potential, Challenges*, edited by Marco Nørskov, 39–55. Ashgate.

Rosati, Connie S. 2009. "Relational Good and the Multiplicity Problem." *Philosophical Issues* 19:205–34. http://www.jstor.org/stable/27749931.

Rubel, Alan, Clinton Castro, and Adam Pham. 2019. "Agency Laundering and Information Technologies." *Ethical Theory and Moral Practice* 22 (4): 1017–41. https://doi.org/10.1007/s10677-019-10030-w.

Russell, Stuart J., and Peter Norvig. 2021. *Artificial Intelligence: A Modern Approach*. Fourth Edition. Upper Saddle Rivier: Pearson Education.

Santoni de Sio, Filippo, and Giulio Mecacci. 2021. "Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them." *Philosophy and Technology* 34 (4): 1057–84. https://doi.org/10.1007/s13347-021-00450-x.

Sartorio, Carolina. 2016. *Causation and Free Will*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198746799.001.0001.

Scanlon, T. M. 2000. *What We Owe To Each Other*. Cambridge: Harvard University Press.

Scheffler, Samuel. 1992. "Prerogatives Without Restrictions." *Philosophical Perspectives* 6. https://doi.org/10.2307/2214253.

Scheutz, Matthias. 2012. "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots." In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by Patrick Lin, Keith Abney, and George A. Bekey, 205–21. Cambridge: The MIT Press.

Schlosser, Markus. 2013. "Conscious Will, Reason-Responsiveness, and Moral Responsibility." *The Journal of Ethics* 17 (3): 205–32. https://doi.org/10.1007/s10892-013-9143-0.

———. 2019. "Agency." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/win2019/entries/agency/.

Schrittwieser, Julian, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, et al. 2020. "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model." *Nature 2020 588:7839* 588 (7839): 604–9. https://doi.org/10.1038/s41586-020-03051-4.

Schroeder, Tim. 2015. "Desire." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/sum2020/entries/desire/.

Schwitzgebel, Eric. 2024. "Belief." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/spr2024/entries/belief/>.

Schwitzgebel, Eric, and Mara Garza. 2020. "Designing AI with Rights, Consciousness, Self-Respect, and Freedom." In *Ethics of Artificial Intelligence*, edited by S. Matthew Liao, 459–79. New York: Oxford University Press.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting." https://arxiv.org/abs/2310.11324.

Sebastián, Miguel Ángel. 2021. "First-Person Representations and Responsible Agency in AI." *Synthese* 199 (3–4): 7061–79. https://doi.org/10.1007/s11229-021-03105-8.

Sebo, Jeff. 2017. "Agency and Moral Status." *Journal of Moral Philosophy* 14 (1): 1–22. https://doi.org/10.1163/17455243-46810046.

Shanahan, Murray, Kyle McDonell, and Laria Reynolds. 2023. "Role Play with Large Language Models." *Nature* 623 (7987): 493–98. https://doi.org/10.1038/s41586-023-06647-8.

Sharkey, Amanda. 2017. "Can Robots Be Responsible Moral Agents? And Why Should We Care?" *Connection Science* 29 (3): 210–16. https://doi.org/10.1080/09540091.2017.1313815.

Shepherd, Joshua. 2018. *Consciousness and Moral Status*. New York: Routledge.

Sher, George. 2009. *Who Knew? Responsibility Without Awareness*. Oxford University Press.

Shoemaker, David. 2011a. "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility." *Ethics* 121 (3): 602–32. https://doi.org/10.1086/659003.

———. 2011b. "Psychopathy, Responsibility, and the Moral/Conventional Distinction." *The Southern Journal of Philosophy* 49:99–124. https://doi.org/10.1111/j.2041-6962.2011.00060.x.

———. 2015. *Responsibility from the Margins*. New York: Oxford University Press.

Sie, Maureen. 2009. "Moral Agency, Conscious Control, and Deliberative Awareness." *Inquiry* 52 (5): 516–31. https://doi.org/10.1080/00201740903302642.

Siewert, Charles. 2021. "Consciousness: Value, Concern, Respect." *Oxford Studies in Philosophy of Mind* 1:3–40.

Silver, David. 2005. "A Strawsonian Defense of Corporate Moral Responsibility." *American Philosophical Quarterly* 42 (4): 279–93. https://www.jstor.org/stable/20010212.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, et al. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature 2016 529:7587* 529 (7587): 484–89. https://doi.org/10.1038/nature16961.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, et al. 2018. "A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play." *Science* 362 (6419): 1140–44. https://doi.org/10.1126/SCIENCE.AAR6404/SUPPL_FILE/AAR6404_DATAS1.ZIP.

Singer, Peter. 1975. *Animal Liberation: A New Ethics for Our Treatment of Animals*. New York: HarperCollins.

———. 2011. *Practical Ethics*. Third Edition. Cambridge: Cambridge University Press.

Sinnott-Armstrong, Walter, and Vincent Conitzer. 2021. "How Much Moral Status Could Artificial Intelligence Ever Achieve?" In *Rethinking Moral Status*, edited by Steve Clarke, Hazem Zohny, and Julian Savulescu, 269–89. Oxford University Press.

Sliwa, Paulina. 2017. "Moral Understanding as Knowing Right from Wrong." *Ethics* 127 (3): 521–52. https://doi.org/10.1086/690011.

Smith, Adam. 2006. "Cognitive Empathy and Emotional Empathy in Human Behavior and Evolution." *The Psychological Record* 56 (1): 3–21. https://doi.org/10.1007/BF03395534.

Smith, Angela M. 2007. "On Being Responsible and Holding Responsible." *The Journal of Ethics* 11 (4): 465–84. https://doi.org/10.1007/s10892-005-7989-5.

Smith, Robert J. 1984. "The Psychopath as Moral Agent." *Philosophy and Phenomenological Research* 45 (2): 177. https://doi.org/10.2307/2107424.

Smythe, Thomas W. 1972. “Unconscious Desires and the Meaning of ‘Desire.’” *The Monist* 56 (3): 413–25.

Southan, Rhys. MS. “The Moral Agent/Patient Distinction, the Rights and Wronging Asymmetries, and a Total Utilitarian Solution.” *Unpublished Manuscript*.

Sripada, Chandra. 2016. “Free Will and the Construction of Options.” *Philosophical Studies* 173 (11): 2913–33. https://doi.org/10.1007/s11098-016-0643-1.

Stahl, Bernd Carsten. 2004. “Information, Ethics, and Computers: The Problem of Autonomous Moral Agents.” *Minds and Machines* 14:67–83. https://doi.org/10.1023/B:MIND.0000005136.61217.93.

Stechly, Kaya, Karthik Valmeekam, and Subbarao Kambhampati. 2024. “Chain of Thoughtlessness? An Analysis of CoT in Planning.” https://arxiv.org/abs/2405.04776.

Stoljar, Daniel, and Zhihe Vincent Zhang. 2024. “Why ChatGPT Doesn’t Think: An Argument from Rationality.” *Inquiry*, 1–29. https://doi.org/10.1080/0020174X.2024.2427061.

Strawson, Galen. 1994. “The Impossibility of Moral Responsibility.” *Philosophical Studies* 75 (1): 5–24. https://www.jstor.org/stable/4320507.

Strawson, P. F. 2008. *Freedom and Resentment and Other Essays*. London: Routledge.

Sullins, John P. 2006. “When Is a Robot a Moral Agent?” *International Review of Information Ethics* 6:23–30.

———. 2009. “Artificial Moral Agency in Technoethics.” In *Handbook of Research on Technoethics*, edited by Rocci Luppicini and Rebecca Adell, 205–21. New York: IGI Global.

Sumers, Theodore R., Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. 2024. “Cognitive Architectures for Language Agents.” *Transactions on Machine Learning Research*. https://arxiv.org/abs/2309.02427.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. Second Edition. Cambridge: The MIT Press.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. “Intriguing Properties of Neural Networks.” https://arxiv.org/pdf/1312.6199.pdf.

Talbert, Matthew. 2024. “Moral Responsibility.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/fall2024/entries/moral-responsibility/>.

Talbot, Brian, Ryan Jenkins, and Duncan Purves. 2017. "When Robots Should Do the Wrong Thing." In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, edited by Patrick Lin, Keith Abney, and Ryan Jenkins, 258–73. Oxford: Oxford University Press.

Tanmay, Kumar, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. "Probing the Moral Development of Large Language Models through Defining Issues Test," https://arxiv.org/abs/2309.13356.

Tiboris, Michael. 2014. "Blaming the Kids: Children's Agency and Diminished Responsibility." *Journal of Applied Philosophy* 31 (1): 77–90. https://doi.org/10.1111/japp.12046.

Tollon, Fabio. 2021. "Do Others Mind? Moral Agents without Mental States." *South African Journal of Philosophy* 40 (2): 182–94. https://doi.org/10.1080/02580136.2021.1925841.

———. 2022. "Responsibility Gaps and the Reactive Attitudes." *AI and Ethics*. https://doi.org/10.1007/s43681-022-00172-6.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023. "LLaMA: Open and Efficient Foundation Language Models." https://arxiv.org/abs/2302.13971.

Traina, Cristina L. H. 2009. "Children and Moral Agency." *Journal of the Society of Christian Ethics* 29 (2): 19–37. https://www.jstor.org/stable/23562796.

Turkle, Sherry. 2011. "Authenticity in the Age of Digital Companions." In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, 62–76. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511978036.006.

Turpin, Miles, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2024. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting." *Advances in Neural Information Processing Systems* 36. https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

Vafa, Keyon, Justin Y. Chen, Jon Kleinberg, Sendhil Mullainathan, and Ashesh Rambachan. 2024. "Evaluating the World Model Implicit in a Generative Model," June. http://arxiv.org/abs/2406.03689.

Vallor, Shannon. 2015. "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character." *Philosophy & Technology* 28 (1): 107–24. https://doi.org/10.1007/s13347-014-0156-9.

———. 2016. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. New York: Oxford University Press.

Vallor, Shannon, and Tillmann Vierkant. 2024. "Find the Gap: AI, Responsible Agency and Vulnerability." *Minds and Machines* 34 (3): 20. https://doi.org/10.1007/s11023-024-09674-0.

Valmeekam, Karthik, Kaya Stechly, and Subbarao Kambhampati. 2024. "LLMs Still Can't Plan; Can LRMs? A Preliminary Evaluation of OpenAI's O1 on PlanBench." https://arxiv.org/abs/2409.13373.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems*, June.

Véliz, Carissa. 2021. "Moral Zombies: Why Algorithms Are Not Moral Agents." *AI & Society* 36 (2): 487–97. https://doi.org/10.1007/s00146-021-01189-x.

Velleman, David J. 1992. "What Happens When Someone Acts?" *Mind* 101 (403): 461–81. https://www.jstor.org/stable/2253898.

Victor, Daniel. 2016. "Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk." *The New York Times*, March 24, 2016. https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html.

Volokh, Eugene. 2019. "Chief Justice Robots." Article. *Duke Law Journal* 68 (6): 1135–92. https://www.jstor.org/stable/48563106.

Vredenburgh, Kate. 2022. "The Right to Explanation." *Journal of Political Philosophy* 30 (2): 209–29. https://doi.org/10.1111/jopp.12262.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. "Transparent, Explainable, and Accountable AI for Robotics." *Science Robotics* 2 (6): 31. https://doi.org/10.1126/SCIROBOTICS.AAN6080/ASSET/9AB6E47A-87D9-41F6-8F1C-A5BB23F96C56/ASSETS/GRAPHIC/AAN6080-F1.JPEG.

Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

Wallach, Wendell, and Shannon Vallor. 2020. "Moral Machines: From Value Alignment to Embodied Virtue." In *Ethics of Artificial Intelligence*, edited by S. Matthew Liao, 383–412. New York: Oxford University Press.

Wang, Boshi, Xiang Yue, Yu Su, and Huan Sun. 2024. "Grokked Transformers Are Implicit Reasoners: A Mechanistic Journey to the Edge of Generalization." http://arxiv.org/abs/2405.15071.

Wang, Boshi, Xiang Yue, and Huan Sun. 2023. "Can ChatGPT Defend Its Belief in Truth? Evaluating LLM Reasoning via Debate." https://arxiv.org/abs/2305.13160.

Wang, Lei, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, et al. 2024. "A Survey on Large Language Model Based Autonomous Agents." *Frontiers of Computer Science* 18 (6): 186345. https://doi.org/10.1007/s11704-024-40231-1.

Wang, Ruoyao, Graham Todd, Ziang Xiao, Xingdi Yuan, Marc-Alexandre Côté, Peter Clark, and Peter Jansen. 2024. "Can Language Models Serve as Text-Based World Simulators?" https://arxiv.org/abs/2406.06485.

Watson, Gary. 1996. "Two Faces of Responsibility." *Philosophical Topics* 24 (2): 227–48. https://www.jstor.org/stable/43154245?seq=1&cid=pdf-.

———. 2013. "Moral Agency." In *International Encyclopedia of Ethics*, edited by Hugh LaFollette, 3322–33. Blackwell Publishing Ltd.

Wegner, Daniel M. 2002. *The Illusion of Conscious Will*. Cambridge: The MIT Press.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* 35. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Wilson, Eric Entrican, and Lara Denis. 2024. "Kant and Hume on Morality." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/spr2024/entries/kant-hume-morality/.

Wolf, Susan. 1988. "Sanity and the Metaphysics of Responsibility." In *Responsibility, Character, and the Emotions*, edited by Ferdinand Schoeman, 46–62. Cambridge: Cambridge University Press.

Xi, Zhiheng, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, et al. 2023. "The Rise and Potential of Large Language Model Based Agents: A Survey." https://arxiv.org/abs/2309.07864.

Xu, Ziwei, Sanjay Jain, and Mohan Kankanhalli. 2024. "Hallucination Is Inevitable: An Innate Limitation of Large Language Models." https://arxiv.org/abs/2401.11817.

Yang, Sohee, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. "Do Large Language Models Latently Perform Multi-Hop Reasoning?" https://arxiv.org/abs/2402.16837.

Yeung, Karen. 2019. "Why Worry about Decision-Making by Machine?" In *Algorithmic Regulation*, edited by Karen Yeung and Martin Lodge, 21–48. Oxford: Oxford University Press.

Yildirim, Ilker, and L.A. Paul. 2024. “From Task Structures to World Models: What Do LLMs Know?” *Trends in Cognitive Sciences* 28 (5): 404–15. https://doi.org/10.1016/j.tics.2024.02.008.

Zafar, Mandy. 2024. “Normativity and AI Moral Agency.” *AI and Ethics*. https://doi.org/10.1007/s43681-024-00566-8.

Zelikman, Eric, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. 2024. “Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking.” https://arxiv.org/abs/2403.09629.

Zelikman, Eric, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2024. “STaR: Self-Taught Reasoner Bootstrapping Reasoning With Reasoning.” *Proceedings of the 36th International Conference on Neural Information Processing Systems* 1126.

Zerilli, John. 2022. “Explaining Machine Learning Decisions.” *Philosophy of Science* 89 (1): 1–19. https://doi.org/10.1017/psa.2021.13.

Zhang, Zhuosheng, Aston Zhang, Mu Li, and Alex Smola. 2022. “Automatic Chain of Thought Prompting in Large Language Models.” https://arxiv.org/abs/2210.03493.

Zhou, Jingwen, Qinghua Lu, Jieshan Chen, Liming Zhu, Xiwei Xu, Zhenchang Xing, and Stefan Harrer. 2024. “A Taxonomy of Architecture Options for Foundation Model-Based Agents: Analysis and Decision Model.” https://arxiv.org/abs/2408.02920.

Zhou, Jingyan, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. 2024. “Rethinking Machine Ethics - Can LLMs Perform Moral Reasoning through the Lens of Moral Theories?” https://arxiv.org/abs/2308.15399.

Zimmermann, Annette, and Chad Lee-Stronach. 2022. “Proceed with Caution.” *Canadian Journal of Philosophy* 52 (1): 6–25. https://doi.org/10.1017/can.2021.17.