Moral Agents Unlike Us[1]

Jen Semler // Cornell Tech

Draft for AIES 2025

Abstract

Suppose AI developers succeed in creating advanced non-conscious artificial moral agents—AI systems that can act in sophisticated ways in the moral domain, systems that meet all the standard criteria for moral agency yet lack phenomenal consciousness. We can call these systems "cognitive moral agents" (short for "merely cognitive moral agents"), and we can contrast them to "affective moral agents" (short for "cognitive and affective moral agents") like humans. Initially, it might seem that we should be indifferent between cognitive moral agents and affective moral agents as moral decision-makers. In this paper, I argue that we have reason to prefer affective moral agents to make moral decisions in certain contexts. I outline two asymmetries that arise between cognitive moral agents and affective moral agents in virtue of cognitive moral agents' lack of phenomenal consciousness: a moral status asymmetry and a valance asymmetry. I then argue that these asymmetries lead to two factors that have bearing on when we should not be indifferent between cognitive moral agents and affective moral agents: relationships and responsibility. Insofar as the decision context at hand requires genuine relationships and phenomenal aspects of our responsibility practices, have reason to prefer an affective moral agent. This conclusion sheds light on the role of consciousness in moral agency as well as the roles of existing AI systems and corporations in moral decision-making.

1 Introduction

Suppose AI developers succeed in creating genuine artificial moral agents. That is, suppose AI systems will be able to act *from* morality, rather than merely *in accordance with* morality. To qualify as genuine moral agents, these systems will need to have certain capacities. For instance, they will need to be capable of performing intentional actions that flow from their mental states, and they will need to be responsive to moral reasons. By "moral agent," I mean that an entity is a source of moral action: it is the type of entity that can act for moral reasons, have moral obligations, wrong moral patients, and be morally responsible (at least in some sense) for its actions (Haksar 1998; Watson 2013).

[1] Thank you to Carissa Véliz, Alison Hills, Roger Crisp, David Shoemaker, Milo Phillips-Brown, and participants in the LMU Workshop on Partiality, Relationships, and AI for helpful comments and discussion. Thank you to the Digital Life Initiative at Cornell Tech for support and feedback.

We can envision these artificial moral agents in various ways. For the purposes of this paper, I will focus on the implications of moral agents that lack phenomenal consciousness; that is, moral agents that have no first-personal experiences or affectively felt emotions. We can call such entities *merely cognitive moral agents* (henceforth *cognitive moral agents*), or *non-conscious moral agents*. Cognitive moral agents are like philosophical zombies: they meet all the standard criteria for moral agency, but there is nothing going on inside—there is nothing it is like to be them (Chalmers 1996; Véliz 2021).

Contrast such moral agents to *cognitive and affective moral agents* (henceforth *affective moral agents*) or *conscious moral agents*. Affective moral agents are moral agents like us: in addition to having the purely cognitive aspects of moral agency, they are also phenomenally conscious—they can feel and experience things, including the badness of pain. In this paper, I consider some morally relevant differences between cognitive moral agents and affective moral agents in terms of their moral decision-making roles.

Whether cognitive moral agency is conceptually possible—that is, whether an entity can meet all the necessary conditions of moral agency, and thus be a genuine moral agent, without phenomenal consciousness—is open to debate, especially within the context of AI (Behdadi and Munthe 2020; Floridi and Sanders 2004; Semler 2025; Véliz 2021).

The aim of this paper is not to argue that cognitive moral agency is possible. Regardless of whether the reader accepts the assumption, this paper offers something of use. Those who accept the assumption that cognitive moral agency is possible can be viewed as claiming that cognitive moral agency really is moral agency. For them, this paper is an exploration of the limits of cognitive moral agency. Those who reject the assumption that cognitive moral agency is possible can be viewed as claiming that cognitive moral agency is not truly moral agency—the affective component is required to be a genuine moral agent. For them, this paper is an explanation of what cognitive moral agency is missing. In sum, the features I am outlining as differentiating cognitive moral agents from affective moral agents can be viewed as either differences between two types of moral agents or as differences between non-moral-agents (cognitive "moral agents") and moral agents (affective moral agents).

Still, for the purposes of this paper, I will proceed in line with the former interpretation. Plausibly, cognitive moral agents (whether they are genuine moral agents or not) are easier to create than affective moral agents, and as such, it is worth considering their role in the moral community.

Given that cognitive moral agents are, in many ways, unlike the paradigm case of moral agency (because human moral agents are affective moral agents), it's not immediately clear what place these moral agents would have in the moral community. Specifically, it's not clear whether cognitive moral agents would have the same roles and responsibilities as affective moral agents. As a starting point, we can consider the following view:

> *Indifference:* For a given moral decision-making context, there is no reason to prefer an affective moral agent to a cognitive moral agent as the decision-maker.

Indifference is motivated by the thought that, put simply, a moral agent is a moral agent, full stop—all moral agents, in virtue of being moral agents, should occupy the same moral roles. For instance, suppose there are two human doctors, 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 paper, 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 2, I present two cases to evoke intuitions about when *Indifference* might hold, and I discuss the kinds of cases this paper aims to adjudicate. In section 3, I explain two underlying asymmetries between human moral agents and artificial moral agents that stem from their asymmetry in phenomenal consciousness: the moral status asymmetry and the valance asymmetry. In section 4, I identify two ways in which these asymmetries manifest as factors that bear on when *Indifference* holds—in cases involving relationships and some kinds of responsibility practices. In Section 5, I show how these factors help us understand when it is impermissible to allow artificial moral agents to make moral decisions. In section 6, I consider near-term implications for the moral role of both existing AI systems and corporations.

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 cede Mel's request to be moved to the front of the queue. The mechanic must weigh the interests, values, and deserts of herself and her clients in deciding whether to expedite Mel's service.

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 cognitive moral agent or an affective moral agent.[2] Either way, Mel can plead her case to the mechanic and have that case,

[2] 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).

including any relevant information about Mel's reasons, 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:

> *Date:* Eleanor is scheduled to go out to dinner with a romantic interest. An hour before the dinner, her date gets a message from his friend asking for help. The date must make a moral decision: whether to cancel the date with Eleanor or instead turn down his friend's request for help. The date must consider his various obligations as well as the possible effects on both his friend and Eleanor.

In this case, it is also clear that a moral agent must be the one to make the decision. But here it matters whether the moral agent is a cognitive moral agent or an affective moral agent. In theory, the decision *can* be made by any moral agent, given that any moral agent will be able to evaluate the relevant information. But in this case, there is a larger social, emotional, and relational context that has to do with the nature of the moral decision. The cognitive moral agent seems to lack full access and ability to participate in this larger context. It's not that the cognitive moral agent cannot understand the context; rather, it's that the cognitive moral agent can't engage in the context in the right way. 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 *Date*, while moral agency still matters, it is not the only thing that matters. There are additional factors at play, including the date's particular relationship with Eleanor.

If it's true that there is reason to prefer an affective moral agent over a cognitive moral agent in *Date*, then *Indifference* is straightforwardly false. But the case of *Date* might be explained away by appealing to the date's special obligations to Eleanor. Suppose, for instance, that I promise to take a friend out to dinner to celebrate her promotion at work. The obligation to do so is *mine*—it doesn't count if I ask my sister to take my friend out instead. So, it might reasonably be claimed, *Indifference* was never meant to hold for *all* moral decisions. The kinds of cases I'm considering, then, are those in which a decision need not be made by any particular moral agent. I'm concerned with decisions in which it is *prima facie* plausible that any moral agent has standing to make a moral decision.

The question at hand is about the conditions under which *Indifference* holds. We've seen that it fails in one type of case: those that require a decision to be made by a particular moral agent. But I want to explore whether there are other reasons to prefer an affective moral agent over a cognitive moral agent. In the rest of the paper, I develop an account of the factors beyond moral agency that are relevant to which type of moral agent should make a given moral decision. I start, in the next section, by pinpointing what differentiates cognitive moral agents from affective moral agents.

3 Asymmetries

There are many ways in which cognitive moral agents will be different from affective 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 affective moral agents are phenomenally conscious while cognitive moral agents are not. In this section, I consider two morally significant asymmetries that arise from this underlying difference.

3.1 Moral Status

Many philosophers hold that consciousness is necessary for moral status (van der Deijl 2021; Rosati 2009; Shepherd 2018; Siewert 2021; Singer 1975). 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, cognitive moral agents will be moral agents that are not moral patients.[3] The resulting asymmetry arises:

> *Moral Status Asymmetry:* Affective moral agents have moral patiency, while cognitive moral agents do not.

The moral status asymmetry has bearing on the ethics of human-robot interactions. In virtue of being moral agents, cognitive 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, cognitive moral agents cannot be wronged by us, as they have no welfare or rights.[4]

[3] Some philosophers hold that consciousness is not necessary for moral status (Bradford 2023; Gunkel 2018; Kagan 2019; Sinnott-Armstrong and Conitzer 2021). If these views are correct, there might still be some asymmetry in moral status arising from differences in degree of moral status or particular rights, 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 valance asymmetry (discussed below). As such, they still have some reason to deny that we should be indifferent between human and artificial moral agents.

[4] Southan explores the flip side of a similar asymmetry between humans and non-human animals. In that case, both humans and non-humans are moral patients, while only humans are moral agents (Southan MS). The key difference is that in the case of cognitive and affective moral agents, both entities are moral agents, while only affective moral agents are moral patients. Moreover, humans fare better in this asymmetry than in the human/animal rights asymmetry.

One implication of this asymmetry is that cognitive moral agents will be morally required to prioritize affective moral agents. For example, a cognitive moral agent would be morally required to sacrifice itself for the sake of saving an affective moral agent from even a minor rights violation (unless, of course, there were competing interests at play from other moral patients). It is implausible that a cognitive 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 a cognitive 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 cognitive moral agents, these constraints will not be grounded in cognitive moral agents' interests or welfare. Of course, we might have other reasons to treat some, or all, cognitive 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 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 reveals a class of cases in which it matters whether the moral agent making a particular moral decision is a cognitive or an affective moral agent. These will be cases in which the moral agent should have some moral skin in the game, some welfare interests that are also at stake for the moral agent. Consider the following case:

> *Taxi:* Hunter is deciding between taking a human-operated taxi or a driverless taxi. Both the human driver and the autonomous vehicle meet the conditions for moral agency, but the human driver is an affective moral agent while the autonomous vehicle is a cognitive moral agent. They are equally skilled at driving, and both options on equally convenient for Hunter.

Does Hunter have reason to prefer one driver over the other? Intuitively, it might seem not to make a difference; indeed, plausibly many people in Hunter's situation would be indifferent. But to Hunter, it might reasonably matter whether the driver is putting their own welfare at stake. When a human driver makes a moral decision about, say, speeding on the highway, she is putting herself at risk too. Because she is also a moral patient, she has a phenomenal stake in what happens to Hunter. And this matters because she is making moral decisions that can risk Hunter's life.

Generally, then, the instances in which the moral status asymmetry is relevant are those in which the moral patient and the moral decision-maker are significantly in the decision, or bearing the consequences of the decision, together. Section 4 will further explore consider what the moral status asymmetry implies for the interchangeability of human moral agents and artificial moral agents. For now, we can conclude that in cases in which it matters that the moral agent is also a moral patient, we have reason to prefer an affective moral agent over a cognitive moral agent as the decision-maker.

3.2 Valanced Experience

Because cognitive moral agents lack phenomenal consciousness, they will not have any valanced experience; they will not experience anything as positive or negative in terms of affect. The cognitive moral agent will not, for instance, feel the badness of pain, even if it acts the same as those who do (e.g., by crying in response to bodily damage).

Cognitive moral agents 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, a cognitive 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. A cognitive moral agent would know that breaking a promise would cause the promisee to feel upset and betrayed, and the cognitive moral agent would take this consideration as a reason against breaking the promise.

Moreover, the cognitive moral agent would have desires, albeit ones that are not associated with any phenomenal sense of wanting. So, the cognitive moral agent might desire to make others feel happy, or to avoid making others suffer.

Still, cognitive 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:

> *Valance Asymmetry*: Affective moral agents have valanced experience, while cognitive moral agents have (at most) functional or behavioral equivalents of valanced states.

There are two ways in which the valance asymmetry manifests. First, cognitive moral agents cannot affectively experience morally relevant emotions. While an artificial moral agent can behave 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 cognitive moral agent will not actually *feel* angry. Additionally, the cognitive moral agent cannot experience the felt quality of guilt, sadness, or regret when it fails to uphold its own moral obligations—even though it might have non-conscious versions of these states. The non-conscious versions of these states and the corresponding actions might be enough for some purposes (Björnsson and Hess 2017), but, as Section 4 will outline in more detail, our responsibility practices will plausibly look different for our interactions with cognitive moral agents.

Second, cognitive moral agents lack affective empathy, a central feature in realizing moral agency in humans. Affective 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). Cognitive 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, cognitive 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 cognitive 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 affective moral agents. Cognitive moral agents just cannot do this through affective empathy.

The valance asymmetry reveals another class of cases in which it matters whether the decision-maker is a cognitive moral agent or an affective moral agent, namely cases in which felt experience is relevant. Consider the following case:

> *Life support:* Marv is on life support. Marv has not listed a medical proxy, nor does he have any family or friends to serve as his decision surrogate. As such, Marv's physician must act as the surrogate decision-maker. Marv has two equally competent physicians: one cognitive moral agent and one affective moral agent.

Assuming only one physician can serve as Marv's surrogate decision-maker, is there reason to prefer one over the other? On the one hand, perhaps it doesn't matter: both options are moral agents and can thus decide what's best for Marv. I am not claiming that the affective moral agent will make a better decision in virtue of its ability to experience valence.

On the other hand, the affective moral agent can do something the cognitive moral agent cannot: she can *feel for* Marv in a phenomenal sense. In this regard, the affective moral agent can empathize with Marv in an experiential way and can thus engage with the decision on an affective level. Given how important it is for Marv to be treated as a unique human and given that the decision will include tradeoffs involving pain, quality of life, and death, the ability to have valanced experience seems morally relevant.

Generally, then, the instances in which the valance asymmetry is relevant are those in which the moral decision warrants affective engagement. In cases in which it matters that the moral agent can have valanced experience, we have reason to prefer an affective moral agent over a cognitive moral agent as the decision-maker.

With these two asymmetries in hand, we can now turn to the question of when—and to what extent—moral agency is all that matters in moral decision-making contexts.

4 More than Moral Agency

The previous section revealed two asymmetries between cognitive and affective moral agents that give us reasons to prefer an affective moral agent in certain contexts. Already, then, for any moral decision, evaluating whether the two asymmetries are relevant to the decision context will provide guidance on whether there is reason to reserve the decision for an affective moral agent. In this section, I argue that the moral status and valance asymmetries give rise to two broader factors that bear on whether we should be indifferent between cognitive moral agents and affective moral agents as moral decision-makers.

4.1 Relationships

Because of the two asymmetries, we are limited in the types of relationships we can form with cognitive 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 an affective moral agent.

Because of the moral status asymmetry, any relationship between a cognitive moral agent and an affective moral agent would be necessarily unequal, as the agents would have vastly different moral standards of interaction. Because of the moral status asymmetry, the cognitive moral agent will have obligations towards the affective moral agent, but the affective moral agent will not have obligations towards the cognitive moral agent. For example, the cognitive moral agent could violate the affective moral agent's right to privacy or bodily autonomy, but the affective moral agent could not do the same to the cognitive moral agent, for the cognitive moral agents lacks these rights. As such, the affective moral agent can treat the cognitive moral agent in many ways that would be unacceptable in the opposite direction.[5]

One objection to the relevance of unequal relationships is that we already engage in many relationships in which participants have an asymmetry in moral status or rights (and thus an asymmetry in obligations owed towards each other). Humans can violate the rights of non-human animals, 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. Indeed, there are even more cases in which the exact suite of rights each participant has varies within a relationship; bosses and employees have many rights in common but also share an asymmetric set of rights.

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

[5] As mentioned in section 3.1, there might be other reasons for the affective moral agent to treat the cognitive moral agent as if it were a moral patient, such as Kantian concerns that the ways we treat cognitive moral agents might spill over into how we treat moral patients, or Aristotelian concerns that treating cognitive moral agents in certain ways might cultivate vices. But these reasons make relationships between affective and cognitive moral agents more equal only in a shallow sense—there is still a deep inequality regarding the obligations that cognitive and affective moral agents have towards each other.

between two human adults. This inequality creates a power dynamic within the relationship. The same can be said of relationships between affective moral agents and cognitive 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 valance asymmetry further limits the types of relationships we can form with artificial moral agents. Because of the valance asymmetry, cognitive moral agents cannot reciprocate our feelings. In light of this fact, several arguments have been made that we cannot form authentic relationships with technological entities (Nyholm 2020, chap. 5; Scheutz 2012; Turkle 2011). 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, cognitive moral agents can act *as if* they experience emotions affectively and thus can act *as if* they reciprocate the relationship-relevant set of feelings. Additionally, they can have the cognitive and behavioral equivalents of the relevant emotions. So, a cognitive moral agent could care about someone in an emotionless way and could have desires and beliefs about the person. But the cognitive moral agent couldn't *feel* love or care towards a partner. The cognitive moral agent could also not affectively empathize with a partner—it could *know* how someone feels but not *feel* how they feel. These phenomenal aspects are important features of many relationships.

It might be objected that while the kinds of relationships we can have with cognitive moral agents differ from those we can have with affective moral agents, the difference is not morally relevant. There are three versions of this objection. First, it might be claimed that so long as cognitive moral agents convincingly behave as if they experience valanced states, 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 us to interact with cognitive moral agents as if they were affective moral agents (Coeckelbergh 2009; 2010). On this view, if a cognitive moral agent acts in all the same ways as an affective moral agent (e.g., by outwardly expressing what looks to us like reciprocal emotions), the resulting relationships are the same in all the relevant ways.

But this view is implausible. In the case of cognitive moral agency as described in this paper, we already know that the cognitive moral agent does not have valanced experience—this is posited by the very notion of a cognitive moral agent. We also know that the cognitive moral agent lacks moral status, under the assumption that consciousness is necessary for moral status. As such, there is a discrepancy between the behavior exhibited by cognitive moral agents and what we know to be true under the surface. It's not the case that we're unclear about whether cognitive moral agents really do experience emotion, sensation, and other valanced states. We're not trying to infer whether they feel the same way about us as we feel about them. We know that they do not. Regardless of appearance and behavior, how cognitive moral agents seem to us, or how we perceive cognitive moral agents, the fact remains that they do not *feel* love for us, affectively empathize with us, or care in a phenomenal sense about us. Without these ingredients, the relationship is deeply unequal.

The second version of the 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.).[6]

In these examples, it does not seem to be the case that the users are mistaken about Replika's lack of first-personal experience. 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 importantly different from a human-human relationship. 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 equal, just as parasocial relationships with celebrities and other one-sided relationships are not equal in morally significant ways—they are not reciprocal.

While I have claimed that our relationships with cognitive moral agents (and other non-conscious entities) are not reciprocal in morally important ways, 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. 2025). These benefits might lead us to utilize cognitive moral agents for various purposes, such as reducing loneliness and helping us work through our moral dilemmas, and to have certain types of relationships with them. But these considerations do not change the fact that there is a morally significant difference between the relationships we can have with affective moral agents and those we can have with cognitive moral agents.

The third version of the objection is that because cognitive moral agents will have advanced capacities, we will be able to form relationships with them in a way that we cannot with existing non-conscious technological artifacts. Cognitive moral agents can, for example, understand us and respect us in a way that current AI systems cannot. We might entrust a cognitive moral agent with our secrets because we know that the cognitive moral agent will take its moral obligation

[6] 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.

seriously when deciding whether to reveal that secret to others. We cannot expect the same of existing technologies that lack mental states and are not responsive to moral reasons. So, the thought goes, a cognitive 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 a cognitive 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 equal 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. Cognitive moral agents cannot do this.

Given that cognitive moral agents cannot form equal and reciprocal relationships with human moral agents, there will be a class of cases in which we should not be indifferent between cognitive moral agents and affective moral agents in moral decision-making contexts. Sometimes, reciprocal relationships matter, and in these cases, we are justified in preferring an affective moral agent to be the moral decision-maker.

4.2 Responsibility

Cognitive moral agents are fully fledged moral agents. As such, they will be morally responsible for their actions, just like affective 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 (Smith 2007). The moral status and valance asymmetries entail that the ways in which we can hold cognitive moral agents responsible differ from the ways in which we can hold affective moral agents responsible.[7]

Because of the valance asymmetry, cognitive moral agents cannot engage in the same responsibility practices as affective 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 cognitive moral agents so long as doing so makes them act in a more morally desirable way. But our blaming practices often additionally include 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 (Strawson

[7] The moral status and valance asymmetries might also affect the ways in which cognitive moral agents can hold us responsible. For example, a cognitive moral agent cannot blame us when blame involves felt emotions. This consideration might challenge and reshape our responsibility practices and their effectiveness. Further discussion is beyond the scope of this paper.

2008). Directing resentment towards one’s car for breaking down is inapt not only because the car isn’t a moral agent, but also because the car cannot phenomenally receive resentment.

It might be objected that the valance asymmetry does not rule out cognitive moral agents from having reactive attitudes. Björnsson and Hess argue that corporations can have reactive attitudes despite lacking phenomenal consciousness (Björnsson and 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). All cognitive moral agents could give us these same features.

Björnsson and Hess are right that corporations—and cognitive 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 cognitive 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 because of the larger social-relational practices around responsibility. 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. A cognitive moral agent could not 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 cognitive moral agents in situations where retributive blame is important.[8] We must, then, think ahead of time whether retributive blame will be important for a given moral decision. It might not always be important; sometimes all we care about is compensation or behavioral changes. But sometimes retributive blame will matter, and in those cases, we have reason to prefer an affective moral agent to make the decision, so that no retribution gap subsequently arises.

Because of the moral status asymmetry, we will also be permitted to hold cognitive moral agents responsible in ways that we cannot hold affective moral agents responsible. Recall that there is nothing we can do to violate the rights of cognitive moral agents or lessen their wellbeing. As a result, our approach towards punishing them should be purely empirical: we should punish cognitive moral agents in whichever ways allow us to get the most desirable results. This could mean that we destroy cognitive moral agents whenever they wrong us, or subject them to repeated reprogramming. We would even be permitted to preemptively punish cognitive 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 cognitive moral agents cannot engage in the same responsibility practices as affective moral agents, there will be a class of cases in which we should not be indifferent between cognitive moral agents and affective moral agents in moral decision-making contexts. Sometimes, phenomenal feelings and retributive blame matter; in these cases, we are justified in preferring an affective moral agent to be the moral decision-maker.

5 The Roles of Artificial Moral Agents

This paper isn't about settling whether any particular role is one that ought not be filled by a cognitive moral agent. Rather, the aim of this paper is to help guide our thinking about when we have reason to prefer an affective moral agent as a moral decision-maker. Whether we should be indifferent between a cognitive moral agent and an affective moral agent will depend on the context. namely on the applicability and importance of the asymmetries and factors described above for the decision at hand.

The arguments so far offer a rough sketch of a framework: we should not be indifferent between cognitive and affective moral agents as moral decision-makers in cases in which (1) the context involves a relationship of the kind that we cannot have with cognitive moral agents, or (2) the

[8] 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).

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 cognitive moral agent and an affective moral agent.

5.1 Mechanics and Dates

We can now more precisely explain why we should be indifferent regarding mechanics but not indifferent regarding dates. Let's start with *Mechanic.* First, the context does not involve a relationship of the kind that human moral agents cannot have with cognitive moral agents. The mechanic-client relationship need not be an equal and reciprocal relationship. In fact, Mel need not have any 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 *Date.* First, the context does involve of relationship of the kind that human moral agents cannot have with a cognitive moral agent. Being a romantic interest involves having some degree of care for the date. Cognitive moral agents cannot genuinely experience these feelings and form the relevant socio-relational connections with their dates.

Second, the context does seem to warrant affect-requiring responsibility practices. We expect the date to feel the moral weight of their decisions, and part of this includes feeling some responsibility and concern for Eleanor's comfort and interests. Moreover, if the date acts morally wrongly, for instance by leaving the date early to go meet another woman, retributive forms of punishment seem apt. It is not enough for the date to be "reprogrammed" to do the right thing in the future—rather, there is a sense in which the date should at the very least feel bad about his decision.

5.2 Jurors

Discussions about the role of AI 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 cognitive moral agents might seem ideal. Cognitive jurors can be neutral and impartial in a way that affective jurors cannot. Cognitive 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, victim, or lawyers.

But juries play a social role beyond merely determining the guilt or innocence of a defendant. Juries are supposed to be made up of one's peers, and participation in a jury is a civic duty that arises from membership in a political community. In some sense, cognitive moral agents are the peers of human moral agents. Both are moral agents and can respect each other's status as a moral agent; both understand their obligations and responsibilities; both can engage in moral reasoning and deliberation. They are peers in the moral community on the agentic side of moral status. But there are important senses in which cognitive moral agents are not peers with human moral agents in the context of jury participation.

Cognitive moral agents, in their capacity as jurors, cannot form equal relationships with human moral agents—whether this be the defendant or other jurors. Again, this inability might seem like a point in favor of preferring cognitive moral agents—it might be better than a cognitive juror's judgment cannot be clouded by feelings of sympathy, compassion, attraction, or any other phenomenal experiences of the relevant parties. But the capacity to form reciprocal relationships is an important element of jury membership. Jurors make high-stakes decisions about the lives of defendants. As such, they must be able to relate to the defendant as an equal in the moral community. And because of the moral status and valance asymmetries, cognitive moral agents cannot do this. They cannot form the type of juror-defendant relationship that is required for serving on a jury.

Relatedly, cognitive 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 cognitive jurors cannot fully participate in the way required. They can also not be the recipients of certain blaming practices if they, for instance, fail to take their role seriously or make an error in judgment. They are unable to feel the moral weight of their decision, an important feature in the responsibility landscape of criminal trials.

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 cognitive juror is not subject to the judgments it would inflict, and so it is not in a role-reversible position with the defendant. For example, if a cognitive moral agent were being morally evaluated by a moral agent, it would not make sense for the human to adopt reactive attitudes towards the cognitive moral agent.

So, cognitive 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.

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 affective moral agents should never be replaced with cognitive moral agents? Not necessarily.

Whether we have reason to prefer an affective moral agent depends on both domain and context. In the case of juries, we always have strong reason to prefer affective moral agents over cognitive moral agents—we should not be indifferent between the two. But in the case of mechanics, we do not have strong reasons to prefer an affective moral agent.

In other domains, the verdict is less straightforward. Consider, for instance, the prospect of cognitive moral agents as 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 a reciprocal 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 have reason to prefer affective 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 cognitive and an affective moral agent.[9]

6 Near-Term Implications

If my argument is successful, then the roles we allow future cognitive (i.e., artificial) moral agents to play in the moral community should be restricted. We will have good reason to prefer affective

[9] 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 an affective 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 a cognitive moral agent, even for long-term treatment.

moral agents (i.e., humans) over cognitive moral agents when the decision context is influenced by the moral status asymmetry and/or the valance asymmetry—specifically, cases in which relationships and punishment in the form of reactive attitudes or retribution matter.

At this point, one 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 have reason to prefer affective moral agents over cognitive moral agents can tell us about existing cases.

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 cognitive moral agents (Hess 2013). Corporations are subject to the moral status asymmetry as well as the valance asymmetry. They cannot have equal and reciprocal 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 cognitive 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., affective 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—or, at the very least, if no individual affective moral agent takes responsibility, the public often demands that someone does, or believes that there is injustice in the lack of retributive blame.

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 non-conscious moral agents can play in the moral community.[10] After all, non-conscious moral agents (in this instance, 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 non-conscious moral agents, artificial moral agents lack consciousness altogether.

[10] Thank you to Silvia Milano for this objection.

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*, 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. Similar considerations apply within corporations: there are certain roles that individual members of a corporation play that the corporation as a whole should not play.

The requirement that jury members be affective 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.

6.2 Current AI

Considerations about the appropriate roles of cognitive 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 valance asymmetry. Thus, for any moral decision in which we have reason to prefer an affective moral agent over a cognitive moral agent, we should also prefer a human moral agent to any existing AI system. (Of course, we likely 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. Regardless of whether there is a *right* to a human decision, my argument shows that there are cases in which we have good reason to give people the option to have a human decision-maker.

These considerations can also inform the trajectory of AI development. My argument suggests that we should focus on creating systems that are suited to make moral decisions in contexts that do not require either symmetric moral status or symmetric relational abilities. For instance, we should not aim to make AI systems that can serve as jurors or dates. But we might want to aim to make AI systems that can serve as doctors in certain contexts or as mechanics.

7 Conclusion

In this paper, I have considered the following principle:

> *Indifference:* For a given moral decision-making context, there is no reason to prefer an affective moral agent to a cognitive moral agent as the decision-maker.

I have tried to show that there are many cases in which moral agency is not all that matters. We are justified in preferring affective moral agents to make moral decisions because of the additional social, relational, and emotional contexts of moral decision-making that only affective moral agents (and not cognitive moral agents) can engage in. In particular, I have identified two asymmetries—a moral status asymmetry and a valance asymmetry—that give us reason to prefer affective moral agents in contexts where those asymmetries are relevant. I have also identified two broader factors—relationships and affectively laden responsibility practices—for which these asymmetries are particularly relevant.

More work must be done to further analyze concrete cases in which we should and should not be indifferent between affective moral agents and cognitive moral agents, and to offer a framework for determining when the factors I have identified are relevant to the decision context. Moreover, future work should consider cases in which these asymmetries give us reason to prefer cognitive moral agents over artificial agents—and how our reasons for preferring cognitive moral agents trade off with our reasons for preferring affective moral agents.

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.

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.

Behdadi, Dorna, and Christian Munthe. 2020. "A Normative Approach to Artificial Moral Agency." *Minds and Machines* 30: 195–218. https://doi.org/10.1007/s11023-020-09525-8.

Björnsson, Gunnar, and Kendy Hess. 2017. "Corporate Crocodile Tears?: On the Reactive Attitudes of Corporate Agents." *Philosophy and Phenomenological Research* 94 (2): 273–98.

Bradford, Gwen. 2023. "Consciousness and Welfare Subjectivity." *Noûs* 57 (4): 905–21. https://doi.org/10.1111/nous.12434.

Brennan-Marquez, Kiel, and Stephen E. Henderson. 2019. "Artificial Intelligence and Role-Reversible Judgment." *Journal of Criminal Law and Criminology* 109 (2): 137–64.

Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

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: 181–89. https://doi.org/10.1007/s00146-009-0208-3.

Coeckelbergh, Mark. 2010. “Moral Appearances: Emotions, Robots, and Human Morality.” *Ethics and Information Technology* 12: 235–41. https://doi.org/10.1007/s10676-010-9221-y.

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.

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. Edward Elgar Publishing.

De Freitas, Julian, Zeliha Oğuz-Uğuralp, Ahmet Kaan Uğuralp, and Stefano Puntoni. 2025. “AI Companions Reduce Loneliness.” *Journal of Consumer Research*, June 25, ucaf040. https://doi.org/10.1093/jcr/ucaf040.

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.

Floridi, Luciano, and J.W. Sanders. 2004. “On the Morality of Artificial Agents.” *Minds and Machines* 14: 349–79. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

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.

Gunkel, David J. 2018. *Robot Rights*. MIT Press.

Haksar, Vinit. 1998. “Moral Agents.” In *Routledge Encyclopedia of Philosophy*, edited by Edward Craig. Routledge.

Hess, Kendy. 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.

Huq, Aziz Z. 2020. “A Right to a Human Decision.” *Virginia Law Review* 106 (3): 611–88.

Kagan, Shelly. 2019. *How to Count Animals, More or Less*. 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.

List, Christian, and Philip Pettit. 2011. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press.

Luka Inc. n.d. “Replika.” Replika. Accessed October 11, 2024. https://replika.com.

Nyholm, Sven. 2020. *Humans and Robots: Ethics, Agency, and Anthropomorphism*. Rowman & Littlefield.

Rosati, Connie S. 2009. “Relational Good and the Multiplicity Problem.” *Philosophical Issues* 19: 205–34.

Scheffler, Samuel. 1992. “Prerogatives Without Restrictions.” *Philosophical Perspectives* 6: 377–97. 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. MIT Press.

Semler, Jen. 2025. “Moral Agency without Consciousness.” *Canadian Journal of Philosophy*, 1–20. https://doi.org/10.1017/can.2025.10008.

Shepherd, Joshua. 2018. *Consciousness and Moral Status*. Routledge.

Siewert, Charles. 2021. “Consciousness: Value, Concern, Respect.” In *Oxford Studies in Philosophy of Mind*, vol. 1.

Singer, Peter. 1975. *Animal Liberation: A New Ethics for Our Treatment of Animals*. HarperCollins.

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. Oxford University Press.

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.

Southan, Rhys. MS. “The Moral Agent/Patient Distinction, the Rights and Wronging Asymmetries, and a Total Utilitarian Solution.” Unpublished manuscript.

Strawson, P.F. 2008. *Freedom and Resentment and Other Essays*. Routledge.

Turkle, Sherry. 2011. “Authenticity in the Age of Digital Companions.” In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson. Cambridge 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.

Véliz, Carissa. 2021. “Moral Zombies: Why Algorithms Are Not Moral Agents.” *AI & Society* 36: 487–97. https://doi.org/10.1007/s00146-021-01189-x.

Volokh, Eugene. 2019. “Chief Justice Robots.” *Duke Law Journal* 68 (6): 1135–92.

Watson, Gary. 2013. “Moral Agency.” In *The International Encyclopedia of Ethics*, edited by Hugh LaFollette. Blackwell Publishing Ltd.