Canadian Journal of Philosophy (2025), 1–20
doi:10.1017/can.2025.10008

ARTICLE

Moral Agency without Consciousness

Jen Semler

Faculty of Philosophy, University of Oxford, Oxford, UK
Email: jen.semler@philosophy.ox.ac.uk

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. This paper casts doubt on the consciousness requirement. I argue that consciousness is not necessary for instantiating four key capacities necessary for moral agency: action, moral concept possession, responsiveness to moral reasons, and moral understanding. I defend my picture of nonconscious moral agency as a plausible account of an entity that can act for moral reasons and can be morally responsible. Lastly, I discuss broader implications of my argument, especially on the possibility of artificial moral agency.

Keywords: moral agency; artificial moral agency; phenomenal consciousness; moral responsibility; moral understanding

1. Introduction

Suppose a talented group of philosophers—ourselves included—is recruited by a team of computer scientists to help develop a highly sophisticated robot. Specifically, our goal is to build a robot that is a *moral agent*. We can imagine this project occurring in the future, such that the technology is sufficiently advanced that we can equip our robot with various sophisticated capacities. But suppose further that there is one key limitation to our project: we are unable to provide our robot with consciousness.

The notion that we could create such an entity—a nonconscious moral agent—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.[1] The consciousness requirement has a strong intuitive appeal. But in this paper, I cast doubt on the consciousness requirement. I argue that the core capacities necessary for moral agency can be instantiated without consciousness. The question at hand is a question about the nature of moral agency—and we must confront the possibility that beings like us might not be the only entities that qualify for moral agency.

The question of whether consciousness is necessary for moral agency features prominently in debates about artificial moral agency, that is, debates about whether artificial entities (particularly

[1] The consciousness requirement appears in various forms. On one version, the consciousness requirement is part of the epistemic, or knowledge, condition of moral responsibility, such that being morally responsible for an action requires being consciously aware of certain features of the scenario—and the conscious awareness is what enables the agent to have the relevant knowledge to be morally responsible. Sher, before arguing against it, 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 maintains that consciousness is necessary for moral agency (Sie, 2009).

© The Author(s), 2025. Published by Cambridge University Press on behalf of The Canadian Journal of Philosophy, Inc. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

autonomous systems like AI and robots) can be moral agents (Behdadi and Munthe, 2020). Artificial moral agency skeptics often claim that AI systems cannot be moral agents because they lack consciousness.

Some authors explicitly highlight a lack of consciousness in their arguments. Such views include claims that intentionality requires experiencing psychological states (Friedman & Kahn, 1992); that deliberation and understanding require consciousness (Himma, 2009); that making moral judgments requires phenomenal quality (Purves et al., 2015); that robots cannot have the necessary mental capacities for moral agency in virtue of their lack of phenomenal consciousness (Talbot et al., 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 (Brey, 2014; Fossa, 2018; Johnson, 2006; Johnson & Noorman, 2014; Johnson & Powers, 2008; Parthemore & Whitby, 2013, 2014; Peterson & Spahn, 2011; Stahl, 2004).[2]

Some views of moral agency already deny the necessity of consciousness (Arpaly, 2003; Sher, 2009; Sie, 2009; Wegner, 2002). Often, however, these views hold that conscious awareness of certain things (e.g., one’s reasons for actions or morally salient features of a situation), or that consciousness in particular cases, is not necessary for exercising moral agency; they tend not to make the more controversial claim that consciousness is not at all necessary for moral agency. In other words, these views can be seen as addressing the question of whether consciousness is necessary for the *exercise* of moral agency, whereas I am concerned with whether consciousness is necessary for the *capacity* for moral agency.

Additionally, theories of group moral agency often hold that corporations are moral agents without holding that corporations are conscious (Björnsson & Hess, 2017; List, 2018; List & Pettit, 2011; Pettit, 2007; Silver, 2005). 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 necessarily plays some role in moral agency, by giving rise to group agency.[3] This paper generalizes the phenomenon of nonconscious moral agency, leaving room for moral agents that contain no consciousness at all. Moreover, my argument does not rely on the claim that group agents are genuine agents *qua* group (rather than a collection of individual agents), or any view about the kind of moral agency groups might have.

Some views of artificial moral agency also deny the necessity of consciousness. But these views tend to offer highly revisionary accounts of moral agency by offering new, more inclusive, criteria for moral agency (Floridi & Sanders, 2004; Sullins, 2009). My argument considers the criteria invoked by more standard—and stringent—accounts of moral agency.

My argument, then, is not the first argument against the consciousness requirement. 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 moral agency in particular cases, my view denies that consciousness is necessary to be a moral agent in general.

Establishing 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 consciousness might be required. My approach is to go through the strongest candidates for consciousness-requiring

[2] 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.

[3] 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 & Dever, 2020).

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

capacities and show that consciousness is not, in fact, necessary for instantiating those capacities in the way moral agency demands. Another way to think about this approach is to return to our thought experiment. I argue that we can build a nonconscious moral agent: we can, conceptually, create an entity that has the necessary capacities for moral agency but lacks the capacity for consciousness.

As mentioned earlier, the consciousness requirement can be cashed out in two ways. First, consciousness might be necessary for the *capacity* for moral agency; that is, it might be the case that only conscious entities can be moral agents. Second, consciousness might be necessary for the *exercise* of moral agency; that is, consciousness might play a role in an agent performing morally evaluable actions. In this paper, I address the former. I argue that consciousness is not necessary for the capacities that are constitutive of moral agency, such that an entity can be a moral agent without being conscious.

In Section 2, I define the two key terms in this paper: moral agency and consciousness. In Sections 3–6, I argue that consciousness is not necessary for four candidate necessary conditions for moral agency: action, moral concept possession, moral reasons-responsiveness, and moral understanding. For each capacity, I describe how the capacity can be instantiated without consciousness—and I argue that this instantiation fulfills a requirement for moral agency. In Section 7, I respond to two objections: that moral motivation requires consciousness and that my account fails to capture the true sense of moral agency. In Section 8, I conclude by considering implications for artificial moral agency.

2. Definitions

Before I can assess the role of consciousness in moral agency, I must specify what these terms mean.

2.1. Moral Agency

"Moral agency," despite its pervasive use, is a slippery concept. Moral agency is often discussed in close connection with moral responsibility. But moral agency is not always straightforwardly defined in line with moral responsibility.

Haksar defines moral agents as "those 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, an ambiguity arises, as these three features might come apart. Haksar does not clarify whether moral agency involves all 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). Watson, however, does not specify whether 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 "moral worth"—that is, moral praiseworthiness and blameworthiness. Arpaly is concerned with what makes it the case that the same action can prompt different degrees of praise and blame in different agents (Arpaly, 2003). Of course, 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). But Arpaly does not directly focus 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 moral agency can mean one, or some combination, of various related but distinct concepts: entities that are expected to meet moral standards, that have moral obligations, that are accountable, that are subject to moral praise and blame, that have moral worth, that

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

act competently regarding morality, that are autonomous, or that are answerable to others. At least some of these concepts seem to come apart (Shoemaker, 2011; Watson, 1996), and so we must be clear about what we are talking about when we are evaluating whether consciousness is necessary for moral agency.

I will define moral agency as follows: a moral agent is a genuine source of moral action. Roughly, a moral agent is an entity that can act from moral reasons and can be morally responsible for their actions. Different theories of moral agency will have different ways of cashing out precisely what moral agency is and which capacities are necessary for it. My aim is not to adjudicate between these different theories. Rather, I argue that on a plausible conception of the key capacities associated with moral agency, consciousness is not necessary.

2.2. Consciousness

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). If an entity is phenomenally conscious, it will feel the experiential quality of pain—the pain will hurt.

Sometimes, phenomenal consciousness is used interchangeably with *sentience.* Some accounts differentiate the two, viewing sentience as a subcategory of phenomenal consciousness: sentience is the capacity to feel valenced phenomenal states (e.g., pain and pleasure). Phenomenal consciousness is taken to be broader and to include non-valenced experiences that do not feel good or bad for the experiencer, such as the perceptual experience of seeing a square. I use the broader term to capture all aspects of first-personal experience, though proponents of the consciousness requirement seem to focus on felt emotions and other valenced phenomenal states.

Phenomenal consciousness is conceptually distinct from access consciousness (Block, 1995), which I take as straightforwardly necessary for moral agency (Levy, 2014; Schlosser, 2013). 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. A philosophical zombie—a functional duplicate of a human without phenomenal consciousness—would have only access consciousness (Chalmers, 1996).[4] Such an entity could utilize all the same information as a phenomenally conscious human (and thus act in all the same ways) but would not experience anything first-personally. They might scream and retract their hand when it touches the stovetop, but they will not feel any pain or sadness.

With these definitions in hand, I turn to my argument against the consciousness requirement.

3. Consciousness and 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.

But the argument needs more precision. The philosophy of agency includes a wide range of views about which features enable an entity to act. On some minimal views of agency that focus on goal- or norm-directed behavior, entities like bacteria (Barandiaran et al., 2009) and simple reinforcement learning computer systems (Butlin, 2023) are agents. In discussions of artificial agency, AI systems are often referred to as agential in nature, and agency is linked to an ability to perform an increasing range of sophisticated tasks. Indeed, moral agents likely must have at least this type of agency.

[4] See Véliz (2021) for an argument that philosophical zombies are not moral agents.

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

But the kind of agency relevant to *moral* agency is more sophisticated—it is the type of agency that renders us able to act for moral reasons and to act in a way that undergirds our moral responsibility. The notion of *intentional action* is closely connected to the notion of acting for reasons, and theories of intentional action can be viewed as providing accounts of what it means for an agent to act for reasons.

The standard theory of action is an event-causal view, according to which an event is an intentional action if it has the right kind of causal connection[5] to certain mental states (Glasscock et al., 2023). While there is some disagreement about which mental states are required for intentional action, contemporary views tend to highlight beliefs, desires, and intentions. Again, because we are focused on moral agency, a focus on mental states is apt—it is difficult to see how an entity could be subject to moral obligations, for instance, without having any beliefs.[6]

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.

Beliefs are generally taken to be non-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 be explicitly aware that she holds that belief, even though she uses it, for instance, when she travels 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 (Dretske, 1988; Fodor, 1975, 1981)—there need not be any phenomenal experiences associated with such representations. Interpretationism requires exhibiting appropriate patterns of behavior (Davidson, 2001; Dennett, 1987, 1980)—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 (Armstrong, 1993; Putnam, 1975)—and belief states need not be connected to any phenomenal states.

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). 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 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, aren't 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.

Even if noninstrumental desires are important to action (or 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,

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

[6] There are some views of "mind-less morality" that deny the role of mental states in moral agency (Floridi & Sanders, 2004). But if I'm wrong about the kind of agency required for moral agency—if it is the case that some simpler, mind-less form of agency is sufficient for moral agency—then all the better for my argument, as these more minimal types of agency do not require consciousness either.

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

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 phenomenally experience fulfillment when she desires the marker.

Intentions similarly do not require consciousness. When Ambre intends to raise her arm and does so, the act might involve some phenomenal feeling (perhaps her arm feels heavy), but the intention does not have a phenomenal character. There is not something it is like for Ambre to intend to raise her arm; 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 moral agency through action. A nonconscious entity can have the capacity for intentional action—the kind of agency necessary for moral agency.

4. Consciousness and Moral Concepts

Moral agency requires the possession of moral concepts. Toddlers, for instance, are agents in that they have the capacity for action, but they are not 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 concepts are required for moral agency is difficult to determine, but obvious candidates include the concept of *moral wrongness* and concepts that factor into moral reasoning, such as *pain* or *equality*.

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 *moral* concepts. A moral agent must be able to "grasp or apply moral predicates" (McKenna, 2012, 11). Initially, this gloss of concept possession might seem too minimal. After all, some existing AI systems can accurately and appropriately use concepts, even moral concepts. Even if we are willing to admit that such systems possess concepts, we seem to mean something different than when we are talking about human concept possession.

While I am committed to a functional account of concept possession—if "functional" means excluding phenomenal consciousness—I am not committed to the claim that existing AI systems have concepts or that simplistic concept usage is sufficient for concept possession. Indeed, it is plausible that concept possession requires meeting some difficult criteria. For example, possessing concepts might require exhibiting 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" (Butlin, 2021; Evans, 1982). Moral concept possession plausibly involves, at the very least, the ability to distinguish between moral and conventional norms. But possessing any particular moral concept will require the ability to apply the concept in sophisticated ways—like the ways in which humans use them.

Is consciousness required for moral concept possession? Some concepts straightforwardly do not require consciousness. For instance, it is not clear how phenomenal states would be relevant to concepts of *subtraction* and *atom*. Included in this category of concepts are some abstract concepts relevant to morality, such as *democracy* 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 sandpaper. Knowing that sandpaper is rough in texture might be an important part of 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 the experience of having a promise

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

broken. While these phenomenal experiences might add more content to the concept of *promise*, they 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 inherently phenomenal concepts—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 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 abstractly in a way that does not require the first-personal concept.

But can we truly possess the concept of *pain* without the first-personal aspect? Of course, we cannot *fully* possess the concept without the first-personal component. But no one possesses any concept in its entirety. More importantly, we need not fully possess the concept to possess the concept in the way that is necessary for moral agency. Suppose an entity possesses all aspects of the concept of *pain* except for the first-personal aspect—all that it is missing is the knowledge of what pain feels like from the inside. Such an entity will know a great deal about pain—that those who feel pain desire not to feel pain, that causing pain is bad, that (and how) pain influences the way individuals act. This information is sufficient for the moral agent to know what their obligations are and how to fulfill them.

It might be objected that we can only fully understand the meaning of the concept *morally wrong* by experiencing moral emotions. Rodogno, 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 *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 could not grasp the meaning of certain inferences, such as the connections between colors and mental states (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. The argument is reminiscent of the Mary the color scientist thought experiment. Mary knows everything about the physical world but lives in a black-and-white room—the key question 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. The blind person can still learn how different colors relate to different mental states, even if experiencing those colors does not cause the mental states in the blind agent herself. She can grasp, for instance, that blue makes people feel calm. She can also learn about why blue makes people feel calm 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 might be an open empirical question whether we can design systems that can correctly use the concept without consciousness. But there is no reason to

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

rule out the possibility of such an entity.[7] Return again to the philosophical zombie. Such an entity is functionally identical to a human—as such, it will use concepts in the same ways as humans use concepts.

I have argued that Rodogno sets the bar too high for moral concept possession. But the opposite objection might be raised against my argument: perhaps I have set the bar too low. Perhaps possessing a concept without the first-personal aspect is not the right kind of concept possession for moral agency. The phenomenal component of a concept might add something to the agent's concept possession—the conscious agent can have a sense of what it is like to be wronged or feel pain.

But it is not clear why this first-personal aspect would be required for an agent to possess the morally relevant form of the concept. Let us compare the conscious concept possessor and the nonconscious concept possessor. Why should we think, for instance, that the nonconscious concept possessor does not possess enough of the concept to be subject to moral obligations? Given that the agent knows what the morally relevant concepts are and how they are used, it seems to have what is required to be bound by moral obligations—it can form beliefs about the moral domain, and it can have moral knowledge. While moral concept possession is not sufficient for moral agency, the defender of the consciousness requirement needs an argument for why lacking the conscious aspects of moral concepts renders one ineligible for moral agency.

I suspect that the resistance to my argument might stem from the idea that the agent cannot *grasp* the concept in the right way. But this objection has more to do with understanding than mere concept possession—as such, I will address the objection more thoroughly in Section 5.

Overall, then, consciousness is not necessary for moral agency through the possession of moral concepts. A nonconscious entity can possess both nonmoral and moral concepts—while it will lack the first-personal component of such concepts, it can accurately use the concepts in sophisticated ways.

5. Consciousness and Moral Reasons-Responsiveness

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.

First, moral reasons-responsiveness requires sensitivity to ethical considerations. A moral agent must be able to identify morally relevant features of a situation as morally relevant (Wallach & Allen, 2009). 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. 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, there are other ways to do so. We might invoke definitions of morally relevant features and apply 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

[7] 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 it is possible that moral concept acquisition can happen in other ways—either through examples or explicit definitions.

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

pain, and thus we can identify an instance of pain as occurring (and as being morally relevant) without having any associated phenomenal states.

I am not claiming that consciousness plays no role in moral sensitivity—I am just claiming that it does not play a *necessary* role. My account is pluralist regarding the ways in which an agent can be sensitive to ethical considerations. It might be the case that nonconscious entities must be sensitive to a greater extent in these alternative ways to compensate for their lack of consciousness. But given that there are different ways to acquire the relevant information, what matters is *whether* this information is acquired—not how.

Second, responsiveness to moral reasons requires recognizing moral reasons *qua* reasons. Scanlon offers a widely accepted definition of a normative reason as "a consideration that counts in favor" of some action (Scanlon, 2000, 17). A *moral* reason is 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 moral evaluation. An example can help further distinguish these capacities. Suppose Aleks must choose a path home: taking path A would get Aleks home quickly but injure a bystander; 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 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 consciousness. The argument for this claim is an extension of the argument that identifying morally salient features does not require consciousness. Recognizing something as a reason might be more complex than 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 reason in evaluating potential actions—she will be able to recognize moral reasons as reasons.

I have already established that a nonconscious entity can act for reasons—it has the capacity for intentional action and can possess beliefs, desires, and intentions. As such, the nonconscious entity has the capacity to take considerations as reasons for action. The nonconscious agent can pick up the phone for the reason of talking to its friend—its desire to chat plus its belief that its friend is calling constitute a reason for action. So, the proponent of the consciousness requirement must argue that a nonconscious agent cannot act for *moral* reasons. But this claim is implausible. Suppose the nonconscious agent has the belief that helping a stranger is morally right and the desire to do what is morally right—when the nonconscious agent helps the stranger, it is acting for a moral reason.

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. Moral reasons-responsiveness involves adaptability. A moral agent must be able to evaluate and weigh competing reasons—and allow such reasons to guide their actions.

Regulative control does not require 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,

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

this description coincides with how we make some moral decisions. We do not just 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 we are making the right decision.

Consciousness often guides the accuracy and efficiency of moral decision-making 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) guides us toward morally right actions.

However, we must 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 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 someone 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 and intuitions that guide our moral reasoning can lead us astray, and the point of reflective equilibrium is to scrutinize these intuitions in light of moral reasoning. Sometimes we must make a moral decision despite the phenomenal weight of the reason pulling us in another direction. 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 relevant to moral agency.

Arkin claims that unmanned weapons systems 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 also guides us toward 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 consciousness adds aside from making this inductive process more salient to the decider. Intuition is a form of inference, and while the associated intuitive feelings might help guide us, 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.

Purves, Jenkins, and Strawser argue that robots cannot, in principle, replicate moral judgment because morality is not codifiable—and so moral deliberation cannot be programmed (Purves et al., 2015). While these authors do not claim that consciousness is necessary for moral judgment, they do highlight "phenomenal quality" as a plausible requirement of moral judgment. Talbot, Jenkins, and Purves later claim that robots' lack of consciousness renders them unable to engage in moral decision-making and act for reasons (Talbot et al., 2017).

But the authors do not explain why a lack of consciousness precludes an entity from moral agency. If the problem is about codifiability, then it is a problem with how machines are programmed, not a problem with consciousness. The nonconscious moral agent need not follow a moral rulebook or algorithm—it can engage in moral deliberation in a similar way to humans, minus the phenomenal aspects.

Given that reasons-responsiveness is often taken as a basis for moral responsibility, it is important to ask whether my account of nonconscious reasons-responsiveness captures the kind of reasons-responsiveness that underlies moral agency. Let us return to the example of the

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

nonconscious agent that helps the stranger because it is the right thing to do. It is difficult to see why such an agent would not be morally responsible: it took up a moral consideration as a reason and acted from that reason. Now, certain responsibility *practices* might be unjustified—if the purpose of certain praising behaviors is aimed at causing the moral agent to feel happy, then such behaviors will be out of place for the nonconscious moral agent. But the nonconscious moral agent can still be praiseworthy for the action in the sense described above.

Relatedly, because the nonconscious agent can identify and respond to moral reasons, it makes sense to say that the agent is subject to moral obligations. We expect it to uphold the requirements of morality because it can take those requirements as reasons for action.

Overall, then, consciousness is not necessary for moral agency through moral reasons-responsiveness.

6. Consciousness and Moral Understanding

A closely related capacity to moral reasons-responsiveness is moral understanding. There is a difference between taking up moral reasons in determining how to act (being responsive to moral reasons) and understanding why and how those reasons are 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.[8]

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 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 & 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 understand well. Wallach and Vallor do, however, highlight a key feature of moral understanding: the role of connectivity between and among moral reasons and actions.

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)

None of these abilities requires consciousness. Rather, they involve abilities to reason and apply moral concepts in novel situations. Grasping the relationships between reasons, explanations, and propositions does not require phenomenal capacities.

[8] We might think of a moral version of the Chinese Room—the man inside the room is responsive to moral reasons but lacks moral understanding.

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

It might be objected that moral understanding is not purely cognitivist, as the above description characterizes it. There are several reasons to think that moral understanding might not be cognitivist. First, 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 folks 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 (Smith, 2006, 2009). Aaltola takes this as evidence that affective empathy, rather than cognitive empathy, is necessary for moral agency. Affective empathy, of course, involves consciousness because it requires feeling the same emotional states as those one is empathizing with.

But Aaltola's argument does not show that affective empathy is necessary for moral agency. Consider the psychopath. There are other explanations of psychopaths' behavior that are consistent with the claim that affective empathy is not necessary for moral agency. Psychopaths might, despite struggling to act morally, be moral agents. They might understand moral concepts and apply them correctly but choose not to act in accordance with them, or they might lack the relevant motivation to do so (Borg & Sinnott-Armstrong, 2013). Put differently, psychopaths might just be bad moral agents. Alternatively, even if psychopaths are not moral agents, their lack of moral agency might not stem from their lack of affective empathy. Psychopaths exhibit a wide range of deficits in rational self-governance that might impair their reasons-responsiveness (Litton, 2008).

Imagine an agent that only possesses cognitive empathy. The agent can accurately represent and reason about how others are feeling—the agent simply does not feel those emotions itself. Given that there are no such entities in existence, there is no empirical evidence that demonstrates that an entity entirely absent of affective empathy can have moral agency. However, given that cognitive empathy enables the agent to reason through all the morally relevant reasons pertaining to a decision, the fact that the agent cannot affectively empathize seems irrelevant. Cognitive empathy can provide all the resources required to be a moral agent.

Another reason to reject the cognitivist account of moral understanding is that it "conflates having moral understanding and having the ability to articulate it" (Sliwa, 2017, 541). On Sliwa's account, knowing right from wrong is what constitutes moral understanding—sometimes, we just know that something is wrong even if we cannot fully express or explain it. Sliwa's account also emphasizes the role of first-personal experiences in moral understanding—she claims that affective experiences give us a more complete conception of the wrong-making features of a situation. While Sliwa argues that there is an important phenomenal aspect to moral understanding, she also notes that her account is pluralist in nature: moral understanding can be realized through various mechanisms, one of which is our affective responses.

Insofar as moral understanding is multiply realizable in this way, a nonconscious agent can have moral understanding despite lacking one method of acquiring moral understanding. The nonconscious agent will simply need other ways to grasp the wrong-making features of a situation. It might be the case that the nonconscious moral agent cannot *fully* grasp these wrong-making features, or that it cannot have a *rich* conception of these features. But a lack of richness need not preclude the nonconscious agent from the level of moral understanding required for moral agency.

Still, it seems difficult to let go of the intuition 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, some kernel of phenomenal experience, 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 requires us to extrapolate beyond our own experiences. Clearly, moral understanding cannot require us to have gone through the exact

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

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, Mel's experience of sadness in her life does not imply that she can understand the experience of someone with depression. In fact, the depressed person seems to have grounds for criticizing Mel 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 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 consciousness 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 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 knowledge about how marginalization operates. But the lack of firsthand experience of being oppressed need not be a barrier to understanding how 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 won't 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.

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 this experience, nor does it imply that attaining understanding is impossible without firsthand experience. For instance, Cillian might develop a deeper and fuller understanding of disloyalty when he is betrayed by a friend. But this admission does not mean that Cillian had no understanding of disloyalty before experiencing it. If Cillian had never experienced disloyalty, he could still have sufficient understanding of the phenomenon to engage in moral reasoning about it—and, for instance, to avoid being disloyal himself.

Consider one more example. Many people think that nonhuman animals have moral status in virtue of their sentience—if an animal can feel pain, then it deserves moral consideration. We might seem to reach this conclusion from our own first-personal experience of pain. But we have an inherent inability to understand, phenomenally, what it is like to be another species, especially species very different from us. Our consciousness is not doing the work in helping us understand why it might be wrong to harm octopuses. We understand that octopus suffering is wrong, and we understand why it is wrong despite our inability to fathom what it is like for the octopus to suffer.

Still, it might be objected that I have not established that moral understanding is possible without any consciousness whatsoever. The people in my examples are generally phenomenally conscious —they just lack specific phenomenal experiences. I have not shown that moral understanding is

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

possible without some baseline of consciousness. Véliz argues, "We do not need to experience every kind of pain to empathise with others' pain…. But [it] is enough to have a sense of what pleasure and pain are to act like competent moral agents" (Véliz, 2021, 493). On this view, some amount of consciousness is required to truly understand moral wrongness; it is this kernel of consciousness that renders one able to access the wider range of considerations relevant to moral reasons-responsiveness and understanding.[9]

There are no existing entities that have moral understanding without consciousness. As such, my claim that such an entity is possible is, to some extent, speculative. This is a limitation of my argument. However, I have offered two reasons to think that moral understanding does not require consciousness: the information required for moral understanding can be obtained in ways that do not involve consciousness, and the kernel of consciousness view involves problematic claims about extrapolation to link consciousness to moral understanding.

The proponent of the consciousness requirement must provide an account of why phenomenal experience—and only phenomenal experience—enables genuine understanding. Such an account would require two parts. First, it would need to explain why having non-phenomenal moral knowledge, for instance, the knowledge that pain is bad, is an insufficient basis for understanding wrongness. Second, it would need to explain why some phenomenal acquaintance with negatively valenced experience—some sense of what pain feels like—provides an agent with a relevant resource for understanding morality more broadly.

Again, return to our philosophical zombie. Suppose our zombie has harmed someone. Even worse, our zombie knows that what it did was wrong—it can offer an account of the reasons for which it acted, and it can explain why its action was wrong. The zombie can also recognize that certain responses are appropriate—for instance, it might offer to compensate the victim for their injuries. The zombie can do all of this but does not affectively empathize with the victim—the zombie does not know what it feels like to be harmed or wronged. It is not clear why this component, given all the other aspects of understanding the zombie possesses, should stop us from deeming the zombie morally responsible.

Overall, then, consciousness is not straightforwardly necessary for moral agency through moral understanding. While my argument is not definitive, at the very least, the burden is on the proponent of the consciousness requirement to explain why some kernel of phenomenal consciousness—some minimal firsthand phenomenal experience—is required to enable genuine moral understanding.

7. Further Objections

In this section, I consider two further objections: that consciousness is necessary for moral motivation and that my account fails to capture the true sense of moral agency.

7.1. Motivation

To be a moral agent, it might not be enough for an entity to have the capacities I outlined above. Moral agents must be *motivated* to act morally.

Consciousness plays a strong motivational role for humans. We have desires that are associated with positive phenomenal states, and we are 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 motivated to act to avoid them. For example, we often feel bad when we see others in pain —and we feel guilty when we refrain from intervening. On a higher level, the desire to be morally good might also be associated with phenomenal states. It might feel fulfilling to view oneself as

[9] Thank you to an anonymous reviewer and to David Shoemaker for helping me make this objection precise.

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

morally virtuous. This feeling can motivate us to put significant weight on moral reasons in our decision-making.

The necessity of consciousness for 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 pleasure. On these views, phenomenal states are the only—or the main—motivator of moral actions. But most philosophers reject such views (Feinberg, 2007), acknowledging that some moral actions are performed because they are the right thing to do, or for the sake of others, 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. 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. Consciousness often makes it easier to do the right thing, and it is likely no evolutionary surprise that humans have developed phenomenal states in line with prosocial behavior. A diminished capacity for consciousness might make morality more difficult for humans, and 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.

Still, the skeptic might push back and claim that consciousness renders us morally motivated in a broader sense. Véliz writes, "When we think about doing something, we imagine the possible consequences we might cause, and consider the kind of pleasure of pain we might create, which motivates us to act one way or another" (Véliz, 2021, 493). On this account, consciousness enables moral agents to *care* about morality—and without such care, we cannot genuinely act for moral reasons.

But the nonconscious moral agent can have desires, and these desires can drive its actions. While it will not care about others in the sense of imagining *how* others might feel and being moved to act on that basis, the nonconscious moral agent can still desire that others are well-off. A lack of consciousness does not imply that an entity can only have self-regarding desires. In fact, a nonconscious agent might have more other-regarding desires than self-regarding desires, given that it does not experience any of the phenomenal benefits associated with fulfilling its self-regarding desires.

7.2. The Wrong Sense of Moral Agency

I have presented a picture of a nonconscious moral agent—one that acts intentionally from nonconscious mental states, possesses moral concepts through nonconscious processes, identifies and responds to moral reasons without consciously appreciating those reasons, and possesses moral understanding through cognitive empathy and grasping the relations between moral reasons. But it might be objected that this picture is not one of a genuine moral agent. The concern is that I have, perhaps tacitly, diluted the concept of moral agency.[10]

Recall that I have defined a moral agent as a genuine source of moral action—an entity that can act for moral reasons and can be morally responsible for its actions. Does the entity I have described fit this definition? I believe that it does.

Consider first the notion of a genuine source of moral action. I have argued that a nonconscious entity can have the capacity for intentional action because it can have the mental states that underlie intentional action. So, at the very least, a nonconscious entity can be a source of action. The proponent of the consciousness requirement might interject with the following objection: *You have shown that intentional action does not require consciousness, but there is a certain class of intentional*

[10] Thank you to an anonymous reviewer for raising this objection.

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

actions, namely, conscious intentional actions, that render an agent a genuine source of action. But this objector has a difficult task ahead: they must explain why there is this special class of intentional actions, and they must provide an account of why only these actions count towards moral agency. Moreover, if the objector is successful, they might inadvertently render all actions that conscious moral agents perform nonconsciously as failures of moral agency.[11]

I have also provided reason to think that a nonconscious entity can do more than act—it can act morally. I argued that such an entity can have moral concepts. While consciousness might be necessary to *fully* grasp certain concepts (i.e., to grasp the first-personal component of that concept), a nonconscious agent can still use moral concepts accurately in sophisticated ways. This capacity renders the agent able to have beliefs and desires about moral matters. Here, the proponent of the consciousness requirement might once again object: *The nonconscious agent is missing the most important component of concepts: the phenomenal aspect.* But the objector must explain why the first-personal component of concepts is needed to possess the concept, especially in cases where the agent can apply the concept correctly. Moreover, the objector must offer an account of why certain concepts cannot be possessed to the right degree without phenomenal experience, while other (e.g., abstract) concepts can.

The key remaining question is whether the cognitivist accounts of reasons-responsiveness and moral understanding can ground moral responsibility. Insofar as moral responsibility is directly linked to an agent's capacity to identify, respond to, and understand moral reasons, I have offered a preliminary account on which a nonconscious agent can be morally responsible. Holders of certain views of moral responsibility will not be convinced by my argument. Some theories of responsibility, for instance, forefront the role of affective emotions (Shoemaker, 2015; Strawson, 2008). Perhaps proponents of these views will embrace my argument as a reason to support emotion-centric theories of responsibility. But proponents of views of moral responsibility that do not forefront emotions must identify where my nonconscious moral agent falls short in instantiating reasons-responsiveness and moral understanding.

There are two ways to view my conclusion, then. Here is the first way. Recall the original thought experiment: we are trying to build a robotic moral agent without consciousness. I have argued that if we go as far as we can in building the relevant capacities into the robot without consciousness, we will end up with a robot that meets the standards for moral agency.

The second way to view my conclusion is this. While I have not shown that consciousness is not necessary for moral agency, I have shifted the burden to the defender of the consciousness requirement. Those who maintain the consciousness requirement must explain either (a) why a lack of consciousness precludes an entity from "properly" meeting the criteria I have outlined, or (b) which additional necessary capacity for moral agency requires consciousness.

8. Conclusion

This paper has put pressure on the consciousness requirement, according to which consciousness is necessary for moral agency. I have argued that phenomenal consciousness is not required for the key capacities required for moral agency.

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 simple chatbots. Corporations may or may not qualify for moral agency in my view—but if they fail to qualify, it will not be because they are not conscious.

[11] A parallel problem has been raised in discussions of the role of consciousness in moral patiency. The challenge is, for those who claim that consciousness is necessary for moral patiency but acknowledge that there are non-experiential welfare goods, to explain why nonconscious entities cannot be welfare subjects (Bradford, 2023).

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

The most significant implications of my argument lie in discussions of artificial moral agency. My argument opens the door for the possibility of artificial nonconscious moral agents. In some ways, the prospect of AI-based moral agents is 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. 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 must reassess their reasons for believing that artificial systems cannot be moral agents. But some existing views that deny the consciousness requirement for artificial moral agency are not vindicated by my argument. Views that do not include the necessary capacities for moral agency considered in this paper 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 must 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 nonconscious artificial moral agents 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 (c.f., Basl, 2014; Bryson, 2018; Gunkel, 2018, 2020; Liao, 2020; Neely, 2014). 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 nonconscious moral agents that are not moral patients).[12]

On the other hand, questions will arise about the contexts in which it is appropriate to deploy nonconscious 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 robot judges should not make sentencing decisions, the reason cannot simply be that nonconscious robot judges lack moral agency—the reason must appeal specifically to why a nonconscious 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, nonconscious moral agents will be unlike human 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. Future research should explore the moral significance of moral agents that are very different from us in these ways.

Acknowledgments. Thank you to Carissa Véliz, Alison Hills, Milo Phillips-Brown, Roger Crisp, David Shoemaker, and two anonymous reviewers at this journal for extensive feedback and encouragement. Thank you to Max Kiener, Kyle van Oosterum, Seth Lazar, 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 for helpful comments and discussion.

Funding statement. This work has been supported by the Oxford-Reuben Graduate Scholarship.

Jen Semler recently completed her doctorate in philosophy at the University of Oxford. In August 2025, she will join Cornell Tech as a Postdoctoral Associate at the Digital Life Initiative.

[12] Some views of moral patiency do not require consciousness (Sinnott-Armstrong & Conitzer, 2021). On such views, nonconscious artificial moral agents will have some degree of moral status and some rights.

References

Aaltola, E. (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.

Arkin, R. C. (2010). The case for ethical autonomy in unmanned systems. *Journal of Military Ethics*, 9(4), 332–341. https://doi.org/10.1080/15027570.2010.536402.

Armstrong, D. M. (1993). *A materialist theory of the mind*. Routledge.

Arpaly, N. (2003). *Unprincipled virtue: An inquiry into moral agency*. Oxford University Press.

Balog, K. (2012). Acquaintance and the mind-body problem. In S. Gozzano & S. C. Hill (Eds.), *New perspectives on type identity: The mental and the physical*. Cambridge University Press.

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and Spatio-temporality in action. *Adaptive Behavior*, 17(5), 367–386. https://doi.org/10.1177/1059712309343819.

Basl, J. (2014). Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. *Philosophy & Technology*, 27, 79–96. https://doi.org/10.1007/s13347-013-0122-y.

Behdadi, D., & Munthe, C. (2020). A Normative Approach to Artificial Moral Agency. *Minds and Machines*, 30(2), 195–218. https://doi.org/10.1007/s11023-020-09525-8.

Björnsson, G., & Hess, K. (2017). Corporate crocodile tears? *Philosophy and Phenomenological Research*, 94(2), 273–298. https://doi.org/10.2307/48578761.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227–247. https://doi.org/10.1017/S0140525X00038188.

Borg, J. S., & Sinnott-Armstrong, W. P. (2013). Do psychopaths make moral judgments? In K. A. Kiehl & P. Walter (Eds.), *Handbook on psychopathy and law*. Oxford University Press.

Bradford, G. (2023). Consciousness and welfare subjectivity. *Noûs*, 57(4), 905–921. https://doi.org/10.1111/nous.12434.

Bratman, M. (1987). *Intention, plans, and practical reason*. Harvard University Press.

Brey, P. (2014). From moral agents to moral factors: The structural ethics approach. In P. Kroes & P.-P. Verbeek (Eds.), *The moral status of technical artefacts*. Springer.

Bryson, J. 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.

Butlin, P. (2021). Sharing our concepts with machines. *Erkenntnis*, 88(7), 3079–3095. https://doi.org/10.1007/s10670-021-00491-w.

Butlin, P. (2023). Reinforcement learning and artificial agency. *Mind and Language*, 39(1), 22–38. https://doi.org/10.1111/mila.12458.

Cappelen, H., & Dever, J. (2020). Acting without me: Corporate agency and the first person perspective. In S. Biggs & H. Geirsson. (Eds.), *The Routledge handbook of linguistic reference*, Routledge.

Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Davidson, D. (2001). *Inquiries into truth and interpretation: Philosophical essays*. Oxford University Press.

Dennett, D. C. (1980). *Brainstorms*. MIT Press.

Dennett, D. C. (1987). *The intentional stance*. MIT Press.

Dretske, F. I. (1988). *Explaining behavior*. MIT Press.

Dror, L. (2023). Is there an epistemic advantage to being oppressed? *Noûs*, 57(3), 618–640. https://doi.org/10.1111/nous.12424.

Evans, G. (1982). *The varieties of reference*. Edited by J. McDowell. Clarendon Press.

Feinberg, J. (2007). Psychological egoism. In R. Shafer-Landau (Ed.), *Ethical theory: An anthology*. Blackwell Publishers.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379. https://doi.org/10.4324/9781003074991-30.

Fodor, J. A. (1975). *The language of thought*. Harvard University Press.

Fodor, J. A. (1981). *Representations: Philosophical essays on the foundations of cognitive science*. Harvester.

Fossa, F. (2018). Artificial moral agents: Moral mentors or sensible tools? *Ethics and Information Technology*, 20(2), 115–126. https://doi.org/10.1007/s10676-018-9451-y.

Friedman, B., & Kahn, P. H. (1992). Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software*, 17(1), 7–14.

Glasscock, P., Juan, S., & Tenenbaum, S.. (2023). Action. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of philosophy*. https://plato.stanford.edu/archives/spr2023/entries/action/.

Gunkel, D. J. (2018). *Robot rights*. MIT Press.

Gunkel, D. J. (2020). *How to survive a robot invasion: Rights, responsibility, and AI*. Routledge.

Haksar, V. (1998). Moral agents. In E. Craig (Ed.), *Routledge Encyclopedia of philosophy*. Routledge.

Himma, K. E. (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.

Hills, A. (2009). Moral Testimony and Moral Epistemology. *Ethics*, 120(1), 94–127. https://doi.org/10.1086/648610.

Jackson, F. (1986). What Mary didn't know. *The Journal of Philosophy*, 83(5), 291. https://doi.org/10.2307/2026143.

Johnson, D. 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, D. G., & Noorman, M. (2014). Artefactual agency and artefactual moral agency. In P. Kroes & P.-P. Verbeek (Eds.), *The moral status of technical artefacts*. Springer.

Johnson, D. G., & Powers, T. M. (2008). Computers as surrogate agents. In J. van den Hoven & J. Weckert (Eds.), *Information technology and moral philosophy*. Cambridge University Press.

Kennett, J. (2002). Autism, empathy and moral agency. *The Philosophical Quarterly*, 52(208), 340–357. https://doi.org/10.1111/1467-9213.00272.

Levy, N. (2014). *Consciousness and moral responsibility*. Oxford University Press.

Liao, S. M. (2020). The moral status and rights of artificial intelligence. In S. M. Liao (Ed.), *Ethics of artificial intelligence*. Oxford University Press.

List, C. (2018). What is it like to be a group agent? *Noûs*, 52(2), 295–319. https://doi.org/10.1111/nous.12162.

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

Litton, P. J. (2008). Responsibility status of the psychopath: On moral reasoning and rational self-governance. *Rutgers Law Journal*, 39, 349–392. http://scholarship.law.missouri.edu/facpubs.

McKenna, M. (2012). *Conversation and responsibility*. Oxford University Press.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450. https://doi.org/10.2307/2183914.

Nailer, T. (2022). *Moral agency*. Master of Philosophy Thesis, The University of Adelaide. https://philarchive.org/rec/NAIMA.

Neely, E. L. (2014). Machines and the moral community. *Philosophy & Technology*, 27(1), 97–111. https://doi.org/10.1007/s13347-013-0114-y.

Parthemore, J., & Whitby, B. (2013). What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness*, 5(2), 105–129. https://doi.org/10.1142/S1793843013500017.

Parthemore, J., & B. Whitby. (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), 141–161. https://doi.org/10.1142/S1793843014400162.

Peterson, M., & Spahn, A. (2011). Can technological artefacts be moral agents? *Science and Engineering Ethics*, 17(3), 411–424. https://doi.org/10.1007/s11948-010-9241-3.

Pettit, P. (1998). Desire. In *Routledge Encyclopedia of philosophy*. Taylor and Francis. https://www.rep.routledge.com/articles/thematic/desire/v-1.

Pettit, P. (2007). Responsibility incorporated. *Ethics*, 117, 171–201. https://doi.org/10.1086/510695.

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice*, 18(4), 851–872. https://doi.org/10.1007/s10677-015-9563-y.

Putnam, H. (1975). *Mind, language, and reality*. Cambridge University Press.

Rodogno, R. (2016). Robots and the limits of morality. In M. Nørskov (Ed.), *Social robots: Boundaries, potential, challenges*. Ashgate.

Scanlon, T. M. (2000). *What we owe to each other*. Harvard University Press.

Schlosser, M. (2013). Conscious will, reason-responsiveness, and moral responsibility. *The Journal of Ethics*, 17(3), 205–232. https://doi.org/10.1007/s10892-013-9143-0.

Schroeder, T. (2015). Desire. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of philosophy*. https://plato.stanford.edu/archives/sum2020/entries/desire/.

Schwitzgebel, E. (2024). Belief. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of philosophy*. https://plato.stanford.edu/archives/spr2024/entries/belief/>.

Sebastián, M. Á. (2021). First-person representations and responsible agency in AI. *Synthese*, 199, 7061–7079. https://doi.org/10.1007/s11229-021-03105-8.

Sher, G. (2009). *Who knew? Responsibility without awareness*. Oxford University Press.

Shoemaker, D. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics*, 121(3), 602–632. https://doi.org/10.1086/659003.

Shoemaker, D. (2015). *Responsibility from the margins*. Oxford University Press.

Sie, M. (2009). Moral agency, conscious control, and deliberative awareness. *Inquiry*, 52(5), 516–531. https://doi.org/10.1080/00201740903302642.

Silver, D. (2005). A Strawsonian Defense of corporate moral responsibility. *American Philosophical Quarterly*, 42(4), 279–293. https://www.jstor.org/stable/20010212.

Sinnott-Armstrong, W., & Conitzer, V. (2021). How much moral status could artificial intelligence ever achieve? In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking moral status*. Oxford University Press.

Sliwa, P. (2017). Moral understanding as knowing right from wrong. *Ethics*, 127(3), 521–552. https://doi.org/10.1086/690011.

Smith, A. (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, A. (2009). The Empathy Imbalance Hypothesis of Autism: A Theoretical Approach to Cognitive and Emotional Empathy in Autistic Development. *The Psychological Record*, 59(3), 489–510. https://doi.org/10.1007/bf03395675.

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press

Stahl, B. C. (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.

Strawson, P. F. (2008). *Freedom and resentment and other essays*. Routledge.

Sullins, J. P. (2009). Artificial moral agency in technoethics. In R. Luppicini & R. Adell (Eds.), *Handbook of research on technoethics*. IGI Global.

Talbot, B., Jenkins, R., & Purves, D. (2017). When robots should do the wrong thing. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence*. Oxford University Press.

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

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wallach, W., & Vallor, S. (2020). Moral machines: From value alignment to embodied virtue. In S. M. Liao (Ed.), *Ethics of artificial intelligence*. Oxford University Press.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, 24(2), 227–248. http://www.jstor.org/stable/43154245.

Watson, G. (2013). Moral agency. In H. LaFollette (Ed.), *International Encyclopedia of ethics*. Blackwell Publishing Ltd.

Wegner, D. M. (2002). *The Illusion of Conscious Will*. The MIT Press.

Cite this article: Semler, J. 2025. Moral Agency without Consciousness. *Canadian Journal of Philosophy*: 1–20, doi:10.1017/can.2025.10008

https://doi.org/10.1017/can.2025.10008 Published online by Cambridge University Press