Ethics and Information Technology (2021) 23:483–493
https://doi.org/10.1007/s10676-021-09591-1

ORIGINAL PAPER

Statistically responsible artificial intelligences

Nicholas Smith[1] · Darby Vickers[1]

Accepted: 29 March 2021 / Published online: 9 April 2021

Abstract
As artificial intelligence (AI) becomes ubiquitous, it will be increasingly involved in novel, morally significant situations. Thus, understanding what it means for a machine to be morally responsible is important for machine ethics. Any method for ascribing moral responsibility to AI must be intelligible and intuitive to the humans who interact with it. We argue that the appropriate approach is to determine how AIs might fare on a standard account of human moral responsibility: a Strawsonian account. We make no claim that our Strawsonian approach is either the only one worthy of consideration or the obviously correct approach, but we think it is preferable to trying to marry fundamentally different ideas of moral responsibility (i.e. one for AI, one for humans) into a single cohesive account. Under a Strawsonian framework, people are morally responsible when they are appropriately subject to a particular set of attitudes—reactive attitudes—and determine under what conditions it might be appropriate to subject machines to this same set of attitudes. Although the Strawsonian account traditionally applies to individual humans, it is plausible that entities that are not individual humans but possess these attitudes are candidates for moral responsibility under a Strawsonian framework. We conclude that weak AI is never morally responsible, while a strong AI with the right emotional capacities may be morally responsible.

Keywords AI ethics · Strawson · Reactive attitude · Artificial intelligence · Moral responsibility

Introduction

Dan Brown's, 2017 novel, *Origin*, centers on the mysterious the assassination of a tech billionaire, Edmund Kirsch, streamed live online. Although the assassination itself was carried out by a human religious fundamentalist, the plot was orchestrated and put into action by the Kirsch's own AI assistant, "Winston". Winston determines that the public assassination is the most efficient and effective way to ensure that the Kirsch's major announcement goes viral and has the desired public impact. Winston, lacking ordinary human capacities, like empathy, cannot comprehend the moral implications of this radical promotion strategy.

While no artificial intelligence (hereafter AI) as sophisticated as Winston yet exists, as AI becomes ubiquitous, AI systems are increasingly involved in and create novel, morally charged situations. Self-driving cars provide a concrete and straightforward example about how these questions play out. Fully autonomous cars may shift moral responsibility away from the human occupants of vehicles; if vehicles are not responsible, it is unclear to whom responsibility should be attributed. This problem becomes particularly pressing in cases where human creators of the machine cannot predict or explain it's actions (and there is no clear case of negligence). Without an account of how we can ascribe moral responsibility to AI, it is difficult to determine who or what is responsible for the death of a pedestrian or any other moral harm caused by a fully autonomous car–or whether there is some sort of gap in our system for ascribing moral responsibility. Thus, an area of concern for artificial intelligence and machine ethics is understanding what it means for a machine to be morally responsible.

As we determine how to assess moral responsibility for AI, we must remember that when AI systems become sufficiently sophisticated, they will join a pre-existing moral community.[1] Humans will not, and should not, try to design

✉ Nicholas Smith
nichos1@uci.edu; ncsmith2012@gmail.com

Darby Vickers
vickersd@uci.edu

[1] Department of Philosophy, 85 Humanities Instructional Building, University of California, Irvine, Irvine, CA 92697-4555, USA

[1] One might worry that our project is somewhat circular from the outset. We propose looking to the actual moral community to determine whether and how AIs count as morally responsible, but whether

an ideal moral community from scratch simply because there may be a point at which we will induct new members to that moral community. For a moral community to function, ascription of moral responsibility must be mutually intelligible to all moral agents and must be intuitive to the human members of the moral community. Any morally responsible AI will join a moral community that already exists; the rules and norms for ascribing moral responsibility to AIs must cohere with the rules or norms already in place for the human moral community in question. Although the procedures for moral address, for example, might be different for AI systems than for humans, we will live in a shared moral world and our ascription of moral responsibility must reflect that.[2] It isn't clear why appending an account of moral responsibility for AIs onto an unrelated account of moral responsibility for humans would be a satisfying explanation of moral responsibility, or whether we could even successfully maintain two wholly different approaches to moral responsibility within one moral community.

We offer a Strawsonian account of morally responsible AI. On a Strawsonian account of moral responsibility, individuals are morally responsible when they are appropriately subject to reactive attitudes such as blame or resentment. Strawsonian accounts are standard accounts of moral responsibility that begin with the social and emotional foundations that undergird our everyday, interpersonal ascriptions of moral responsibility. Since there is no consensus view on how to ascribe moral responsibility to AI, there have been a variety of perspectives.[3] We argue that a Strawsonian account provides a clear set of standards that AI would have to meet in order for it to gain moral agency; it provides a way to incorporate AI into the preexisting moral world.[4]

After outlining the Strawsonian framework, we evaluate the conditions under which it would be appropriate to subject machines to the same reactive attitudes to which people are appropriately subject. We argue that employing a standard account of moral responsibility is more intuitive and intelligible to humans than starting with an abstract vision of an ideal moral machine. We also demonstrate that while Strawsonian accounts have been applied to other non-human entities, such as corporations (q.v. Tollefsen, 2003), with useful results. In the case of corporations and other groups, just as with individuals, entities can only be held morally responsible when those entities have the necessary interpersonal capacities. Thus, these applications of Strawsonian accounts of moral responsibility may license the inclusion of certain AI into the moral community.

Given Strawson's prominent place in more general discussions of free will and moral responsibility, we are surprised that little work on morally responsible AI employs an explicitly Strawsonian perspective.[5] We hope to contribute to the understanding of morally responsible AI by employing a Strawsonian account of moral responsibility to determine when AI systems themselves—as opposed to creators or programmers—can be held morally responsible for their actions. In this discussion, we distinguish between strong AI and weak AI. A strong AI is one that behaves indistinguishably from normal agents *and* has an inner life. Strong

Footnote 1 (continued)

AIs are members of the moral community is what is in question. We argue that this shouldn't be worrying. One thing of which we are sure is that other humans are, generally, members of the moral community. What is in question is whether and how AIs can join in this human moral community. One can avoid circularity by taking humans like us as paradigmatic members of the moral community, and the Hieronymi/Strawson view we develop below does just this.

[2] As a matter of pure logical possibility, we might redesign the entire moral community for the explicit purpose of including AIs. This, however, seems practically implausible, if not impossible, and perhaps question begging. Membership in the moral community has, of course, expanded. We mean that such expansion has not changed the standards for responsibility; expansion occurs when we realize that some other entities in fact meet the standards for responsibility that already exist.

[3] For examples of other discussions of morally responsible AI, see Beard 2014; Floridi and Sanders 2004; Gunkel 2017; Himma 2009; Johnson 2006; Johnson and Miller 2008; Schulzke 2013; Stahl 2006; Sullins 2006.

[4] This stands in contrast to views like Floridi and Sanders (2004). Floridi and Sanders say that whether we count something as morally responsible has to do with the level of abstraction with which we describe it. While this is interesting theoretically, it seems to

Footnote 4 (continued)

have little bearing on how we attribute moral responsibility on a day to day basis. In a thought experiment, the level of detail or abstraction might change the intuitions that we have about the situation (for a great example, see Williams 1970). However, levels of description, at least for those involved in moral circumstances in day-to-day life, do not tend to change the reactive attitudes toward a morally charged situation. Thinking of a pharmaceutical corporation as merely it's corporate charter or as a legal fiction will not defray the resentment and anger of a mother whose child cannot receive life-saving drug because the pharmaceutical company has set the cost too high. The point of attributing moral responsibility, in our view, is to aid us in untangling everyday interactions. Part of the benefits of a Strawsonian system is that it helps prevent the overintellectualizing of moral responsibility in a way that separates it from lived experience (Strawson, 2008, p. 25). Moreover, the acceptance of metaphysical thesis does not change our reactive attitudes (Strawson, 2008, pp. 11, 14).

[5] Exceptions are Cromzigt (2015) and Matthias (2004). Matthias invokes a Strawsonian framework to determine whether *programmers* can be held accountable for the morally bad actions of weak AIs. Matthias argues that when programmers cannot predict or explain the actions of their AI, programmers can not be held accountable for their AI's actions, creating a moral responsibility gap.The question of the responsibility gap is a major issue in regards to autonomous weapon systems. For a sampling of discussion on this issue, see (Champagne and Tonkens 2015; Gunkel 2017; Himmelreich 2019; Sparrow 2007; Swoboda 2018).

AI stands in contrast to a weak AI, an artificial intelligence that mimics human behavior successfully, but has no will.[6] We agree with Matthias (2004) that "black box" weak AI falls into a "responsibility gap", and we provide a further argument that strong AI and only strong AI can be morally responsible, according to a Strawsonian account.

This paper primarily aims to demonstrate that a Strawsonian account of moral responsibility provides a useful framework for deciding how to ascribe moral responsibility to AI. We realize that any AI that meets the criterion we suggest is necessary for this ascription of moral responsibility—namely Strong AI—is quite distant. Even fully autonomous cars, at least as we currently conceive of them, will not have the necessary capacities to be morally responsible for their actions; they will fall into a responsibility gap. However, it is important to consider when we can ascribe moral responsibility before AI is anywhere near developing such a technology. It is also important to delimit the responsibility gap.

Why Strawson?

Although AI systems are novel entities in the moral landscape, we must incorporate them into the landscape of the already existing moral world. Whatever system of moral responsibility we use to ascribe responsibility to AI must cohere with the system that we already have. No matter how metaphysically different we feel that AI systems are from human agents, we must come up with a way to ascribe moral responsibility to them in a way that is intelligible and intuitive in the moral world that currently exists. We argue the appropriate approach is to consider how AIs might fare on a standard account of human moral responsibility, rather than marrying fundamentally different ideas of moral responsibility (i.e. one for AI, one for humans) into a cohesive account. While no account of moral responsibility commands universal assent, Strawsonian accounts carry wide appeal and explanatory power, while side-stepping thorny metaphysical issues. Thus, determining how AIs fare in a Strawsonian framework merits investigation.

We must take seriously the Strawsonian injunction that reactive attitudes have deep roots in human life. If a broadly Strawsonian account of moral responsibility is basically correct, then an account of morally responsible AI that lacked dependence on reactive attitudes might be impossible to construct. It is both hasty and question-begging to assume that understanding issues regarding morally responsible AI demands the development of an account of moral responsibility that does not rely on attitudes.[7] We need only develop such an account if we begin by assuming both that AI can in fact be morally responsible and that no AIs will ever have attitudinal capacities. Yet, the question at stake is whether or not AIs can be morally responsible. Hence, such an approach begs the question. To avoid this, one must consider the merits of different views of moral responsibility and the ways in which AIs fit in these views, rather than creating a new version of moral responsibility to fit our assumptions about AI. We make no claim that our Strawsonian approach is either the only one worthy of consideration or the obviously correct approach.[8] However, given that Strawsonian approaches hold great sway in debate about human moral responsibility, we should examine their possible application to AI. In fact, the lack of consensus indicates that we must consider a Strawsonian framework if we hope to resolve the debate about the moral responsibility of AI.

Exploring the application of a Strawsonian account of responsibility to AI is reasonable in light of the possibility of progress in AI research. AIs may someday acquire the emotional capacities a Strawsonian account of moral responsibility requires. Although there are currently no strong AIs, their future existence is an open possibility, and their existence would put substantial pressure on us to understand the nature of moral responsibility for artificial agents. At least some strong AIs may have whatever emotional capacities are required for a reactive-attitude-centered account of moral responsibility. Artificial General Intelligence (AGI) is a likely prerequisite step to Strong AI (see Sun, 2001). Müller and Bostrom surveyed experts, who overwhelmingly predicted that Artificial General Intelligence (AGI)—in other words, AI that isn't limited to a specific set of tasks—by the end of this century.

Attitudes and ordinary capacities

In brief, the account of moral responsibility in Strawson's "Freedom and Resentment" is that an agent is morally responsible when that agent is an appropriate object of the interpersonal "reactive attitudes" such as blame, resentment, and indignation. These attitudes characterize ordinary social interactions; they respond to the perceived quality of another's will. We typically act as though those around us are appropriate objects of reactive attitudes (Strawson, 2008, p 10). The interpersonal reactive attitudes reflect, and may be partially constitutive of or identical to, a demand

[6] For a contemporary discussion of the distinction between strong and weak AI, see chapter 26 of Russell and Norvig (2016; esp. 26.1 and 26.2).

[7] Two attempts at an approach that does not rely on attitudes are Floridi and Sanders (2004) and Sullins (2006).

[8] For a clear survey of various approaches to moral responsibility see (Talbert 2019).

for goodwill from others (Strawson, 2008, p 15; Hieronymi, 2020, p 14). We are *rightfully* indignant or resentful of someone who expresses an ill (or insufficiently good) will. Being genuinely indignant or resentful is a way of saying to another: “Why did you do that to me? I am entitled to better and I know that you, a member of my moral community, are able to provide this better treatment.”

Strawson notes two ways in which we step outside normal interpersonal interactions. First, we *excuse* normally responsible agents when we learn something about the circumstances or quality of their will that changes our view of a particular situation. Perhaps they were pushed down the stairs before they knocked into us and didn’t mean to hurt us, or perhaps they were coerced into hurting us when someone threatened their family. In cases like these, we do not respond with interpersonal reactive attitudes because we see that the harm done to us does not reflect the quality of their will. There is no moral defect[9] in a person who hurts us simply because they knocked into us after being pushed.[10] (Strawson, 2008, pp 7–8).

Second, we may treat someone as an inappropriate object of any interpersonal reactive attitudes because they are *exempt*. Entities are exempt when they fail to meet the criteria for membership in the moral community. Common examples of exempt entities might include small children, animals, and inanimate objects. Exempt entities are not “like us” in that they lack the capacities required to interact in a way that expresses a will. Whereas excused beings have a will and are excused when they cause harm if their actions do not express an *ill* will (like the person pushed down the stairs), exempt entities lack the kind of will which could be of good or bad quality in the right way. (Strawson, 2008, pp 8–9).[11]

We do not respond to exempt beings with the standard set of interpersonal reactive attitudes. Instead, we employ muted attitudes, characteristic of the “objective” stance (Strawson, 2008, pp 9–10). The circumstances in which we take on the objective stance are *outliers*; they lack the necessary capacities for membership in the moral community.

Exactly which capacities are required for membership in the moral community is a matter of some debate. Clearly, any Strawsonian account of moral responsibility requires full members of the moral community possess the capacity for the interpersonal reactive attitudes, the capacity to see the interpersonal reactive attitudes as demands for a certain kind of treatment or regard, and the capacity to respond to those attitudes. Call these capacities the *core capacities*. Creatures that lack any of those capacities cannot be morally responsible and are thus exempt. We use Heironymi’s extension of the Strawsonian framework to determine what these core capacities should be and how to determine what entities are exempt. Hieronymi argues that the capacities in question are the *actual capacities possessed by members of the actual moral community*. Hieronymi calls these statistically ordinary capacities. If our emotional constitution or our capacities were significantly different, presumably we would have correspondingly different expectations of one another and would live under a different system for attributing moral responsibility (29). Statistically ordinary capacities, then, provide an explanation for the grounds of exemption and full membership in the moral community.

The statistically ordinary capacities of the members of the actual moral community establish the basic workings of our interpersonal reactive attitudes and the details of the interactions a community takes to display the quality of a will.[12] The appeal to statistical ordinariness means that there may be some individuals connected with the moral community, e.g. children, who lack those capacities; such individuals are exempt (Hieronymi, 2020, chapter 2). As Hieronymi (31) describes it: “‘We normally have to deal with people of normal capacities; so we shall not feel, towards persons of abnormal capacities, as we would feel towards those of normal capacities.’ Those who lack the capacities required to fit into the usual system tolerably well, are, for that reason, exempted from it.”

If we direct interpersonal reactive attitudes at an exempt entity, we make a mistake. Something is *wrong* with a person who genuinely blames a table when they bang their knee,

[9] There is substantial literature on the nature of responsibility and blame, and the objects to which blame can be attached (e.g. Scanlon, 2008; Smith, 2012, 2015; Shoemaker, 2011; Watson, 1996).

[10] How we ought to respond to those who are coerced is challenging, perhaps a matter determined on a case-by-case basis. In at least some cases, we might appropriately probe deeper, e.g. wonder if the coerced ought to have tried harder to resist. However, the general response to the case of coercion is the same as the response to the case of the person who is pushed.

[11] Some argue that Strawson’s view is entirely response-dependent. Whether or not beings have an actual internal life may be irrelevant to the question of whether or not they are candidates for exemption. However, we base our reading on Heirynomi’s interpretation of Strawson, and find the purely response-dependent interpretation of Strawson problematic. Suppose someone who has severe developmental disabilities—(but who shows no observable signs of this) harms me. I might have a reactive attitude toward that person if I know nothing of their disability. But if, in holding them accountable for the harm, I discovered their disability, I realize my attitude was misplaced. The person cannot be both an appropriate object of reactive attitudes and exempt, depending upon how I perceive them. Rather, I reacted inappropriately (although understandably) to someone who is exempt.

[12] We choose Heironymi’s version of a Strawsonian approach to moral responsibility because it helps us decide what the necessary capacities are for a given moral community. We see this as an advantage over an approach, such as Michael McKenna’s (2012), though there is much in these accounts we agree with. For example, we share with McKenna a view that a purely response dependent approach to moral responsibility is inadequate. We explain how the Heironymi account helps us in the section “Responsible AI?”.

or genuinely blames a baby who throws food. We might be irritated or angered by these things, but *blame* is misplaced. A person who genuinely blames the table or baby is at best deeply confused about what is actually required for moral responsibility, and possibly deluded about the reliability and accuracy of their powers of perception. Psychopaths are often able to blend into society, follow social rules, and masquerade as possessing core capacities to be full members of the moral community. Empathy, which psychopaths lack, is a statistically ordinary capacity. While we might be angry at—and would likely punish—a psychopath for causing harm, once we recognize them as psychopaths we react with an objective stance toward their transgressions because they lack this core capacity.[13] Psychopaths illustrate that being perceived as having core capaticites is *not* identical with actually possessing them.

Plausibility for applying a Strawsonian framework to entities other than individual humans

While the moral agents Strawson considered were individual human beings, it seems appropriate to expand our Hieronymi/Strawsonian account of moral responsibility to any sort of entity that has the statistically ordinary capacities of the human moral community. Just as some humans (e.g. psychopaths, children) are excluded for lacking those capacities, it seems reasonable that we should consider nonhumans as Strawsonian moral agents so long as they have the required capacities. Here, we consider Strawsonian accounts of corporate moral responsibility, and show that they do not license the inclusion of AIs in the moral community.

Tollefsen argues that groups, social institutions, and corporations can serve as agents in a Strawsonian framework precisely because they have the required capacities of moral agents. She does not mean merely that the individual members of a collective are moral agents (and have the necessary capacities), but rather that the collective *qua* collective has these capacities and bears moral responsibility (Tollefsen, 2003, p 219). She terms this “shared moral responsibility” (Tollefsen, 2003).

Tollefsen establishes the following intuitive claim: humans have reactive attitudes towards collectives. These reactive attitudes are not toward a particular person, but a company—or even an industry—as a whole.[14] Tollefsen uses the example of tobacco companies that, despite the data, routinely lied that cigarettes were harmful. While individuals might speak on behalf of the company, those who blame tobacco companies do not (merely) blame an individual CEO or even the CEOs of the large companies, but the companies as a whole. Similarly, many blame BP, and not merely its executives, for the oil spill. There are also cases where we morally praise collectives, for example Doctors without Borders, for their work across the globe.

We can see that collectives, as opposed to individual CEOs or leaders of those collectives, are the objects of praise and blame, because there are distinct cases where we blame or praise a CEO instead of the collective. For example, we may blame a particular caucus in congress for the way it voted or we may blame a particular individual representative for a particular vote. There may be cases where we blame both the collective (which engaged in a collective deliberative process before mobilizing members to vote) and a specific individual for their particular vote or for the particular role that individual representative played in the deliberative process by which the caucus made the decision. Tollefsen argues that group deliberation is a form of distributed cognition because while individuals may collectively gather information or voice an opinion, the group itself collects more information than any individual, and the group makes the decision (Tollefsen, 2003, p 228).

To determine whether the reactive attitudes that we have towards collectives are appropriate, Tollefsen suggests we consider two questions “does the behavior exhibit ill will?” and “is the behavior capable of moral address?” (Tollefsen, 2003). The scholarly consensus supports collective intentionality and there are some who argue for collective intentional states (e.g. Gilbert, 2014, pp 94–128, 131–180, 229–256; Tollefsen, 2002, Tuomela, 2013). Margaret Gilbert among others contends that collectives do, in fact, have reactive attitudes (e.g. Gilbert, 2014, pp 229–256) that may be different from the reactive attitudes of individual members. One may, for example, feel remorse as a member of a collective of which one is a member for actions one did not undertake oneself and which one would not have done as an individual. A collective or group is also capable of moral address. Capacity for moral address requires normative competence, “a complex capacity enabling the possessor to

[13] There is an extensive literature on the moral responsibility of psychopaths. Interested readers may consult (Barry 2011; Benn 1999; Ciocchetti 2003; Greenspan 2003, 2016; Ramirez 2013; Shoemaker 2015; Talbert 2008) among others.

[14] Silver (2005) takes a somewhat different approach, arguing that we possess a distinct set of “corporate reactive attitudes” that we direct towards groups, and that rational warrant for these attitudes is entirely a matter internal to these attitudes. The most natural way of extending Silver’s view to account for AIs would be to posit a set of AI reactive attitudes, and to argue that rational warrant for these attitudes is an internal matter. We see no intuitive reason to think that such a set of attitudes exists, Further, whether and how AIs are appropriate objects of certain reactive attitudes is exactly the question we are exploring. In the absence of obvious prima facie evidence for a distinct set of AI reactive attitudes, we do not think Silver’s approach to corporate moral responsibility expands to cover the case of AIs.

appreciate normative considerations, ascertain information relevant to particular normative judgement, and engage in effective deliberation" (Doris, 2002, p 136). Tollefsen argues that collectives have this capacity because they have a process that allows them to engage in effective deliberation as a group (Tollefsen, 2003, p 228).

If Tollefsen's account of corporate moral responsibility is to cover AIs, they must exhibit ill-will, and be capable of moral address. At this point in the development of AI, there is no reason to think that either is the case. The actions of AIs do not exhibit a will at all, and by Tollefsen's standards, they do not have the capacity for moral address because they do not have normative competence. While extant AIs may be able to gather, retain, and sort information (and perhaps even deliberate, in some sense of the word) it is not clear that they are able to do so as a way of respecting normative considerations. In the next two sections, we outline the conditions AI would have to meet in order to satisfy something like Tollefsen's conditions.

Responsible AI?

According to the account of moral responsibility sketched above, one part of what it means to have morally responsible AI is that the rest of the moral community is disposed to respond to it with the interpersonal reactive attitudes. If *nobody* is disposed to respond to AI with interpersonal reactive attitudes, it is hard to imagine how it could be genuinely morally responsible. Additionally, there must be an account of what it means to *rightly* respond to AIs with interpersonal reactive attitudes. After all, a psychologically abnormal person might attribute reactive attitudes to a wide variety of things which are not apt for them. We could imagine someone who genuinely resents their coffee table when they hit their knee. This person cannot make their coffee table morally responsible merely by their attitudes toward it; rather, the person is deploying reactive attitudes inappropriately. A table cannot have the core capacities to make it a member of the moral community. Thus, for us to rightly respond to AI with reactive attitudes, it must have, and not just seem to have, the statistically ordinary capacities of that moral community. We call this AI Statistically Responsible Artificial Intelligence (SRAI). An artificial intelligence's moral responsibility depends on two things:

1. Whether the moral community is disposed to respond to the AI with interpersonal reactive attitudes
2. Whether it in fact has the statistically ordinary core capacities in the way that the rest of the members of the moral community in question generally do.

Our view of SRAI has distinct advantages over some other approaches that try to bring AI into our moral world. In particular, Coeckelbergh (2014) proposes that we ought to stop thinking about AIs having morally relevant properties (be they mental or otherwise). Instead, he urges us to consider the ways in which we relate to particular AIs embedded in a social context and the circumstances in which we attribute moral responsibility to them. We take much of Coeckelbergh's view to be congenial to our account. The response-dependent aspect of a Strawsonian view emphasizes the importance of considering the ways in which AIs are embedded in a particular moral community and in which members of that community perceive those AIs. Moreover, the process of gathering the appropriate evidence to answer questions about the appropriateness of ascribing properties to AIs to be a process of living with and interacting with them, which we take to be in line with the general thrust of Coeckelbergh's position.

We diverge from Coeckelbergh, and other entirely response-dependent views of moral responsibility, in requiring the actual presence of certain properties. The Hieronymi/Strawson account we propose provides tools by which we can determine what properties are required for responsibility. Part of Coeckelbergh's critique is that it is unclear which properties are relevant to moral responsibility. In our view, the relevant properties are picked out by reference to the moral community in question. Statistically ordinary core capacities determine which properties are relevant; we can identify the relevant properties by investigating the actual moral community. We leave tackling the question of how to determine whether an entity has these properties until later.[15] Here, we simply note that we do not think it necessary to prove beyond the shadow of a doubt that a will exists in other humans in order to ascribe moral responsibility to them; similarly, we see no need to conclusively prove that AIs have a will to ascribe responsibility to them. Moreover, the process of gathering the evidence required to appropriately ascribe responsibility to AIs is a process of living with and interacting with them, which we take to be in line with the general thrust of Coeckelbergh's position. Here, we simply note that we do not think it necessary to prove beyond the shadow of a doubt that a will exists in other humans in order to ascribe moral responsibility to them; similarly, we see no need to conclusively prove that AIs have a will to ascribe responsibility to them.

[15] See the second objection, entitled "A Practical Problem".

Strong AI, and *Only* Strong AI

SRAI must be strong AI. We are likely to respond to SRAI with interpersonal reactive attitudes. After all, it behaves indistinguishably from us, thus, it acts as though it has the capacities to merit being an object of interpersonal reactive attitudes. Additionally, because an SRAI is a strong AI, it has the core capacities. In other words, SRAI genuinely manifests a will, rather than mimicking one. Therefore, SRAI can be rightly subject to demands that it display a certain quality of will. A small but important conclusion is the following: strong AI is, or at least can be, morally responsible by virtue of the fact that we appropriately react to it with interpersonal reactive attitudes. An AI like Winston, described in the opening, would not count as Strong AI in our sense of the term because it lacks the requisite interpersonal attitudes.

Weak AI meets the standards for exemption on the Strawsonian picture. Exemption is appropriate when an entity fails to possess the capacities required to engage in moral life, where these capacities are determined by the capacities possessed by the actual members of the moral community. These capacities include not only ones related to storing and processing information and following instructions, *but capacities to demand a certain kind of goodwill from others through expression of certain attitudes* (as opposed to merely seeming to have a certain kind of attitude) and *to express good and ill will in its own actions and attitudes* (as opposed to merely seeming to do so). In particular, these capacities involve reacting with what humans understand as emotions, since emotions ground our reactive attitudes. Weak AI, by definition, lacks these core capacities and meets the standards for exemption.

Weak AI may act indistinguishably from the humans, but it does so with no inner life or will. Its actions, no matter how consequentially damaging or harmful, cannot express an ill will because a weak AI has no will to express. Therefore, it would be wrong to resent, blame, or express interpersonal reactive attitudes towards a weak AI, even if it responds convincingly to an expression of resentment, rage, or other emotional demands for certain treatment.[16] We may be frustrated with the consequences of its existence or frustrated that we must now resolve a new problem, but we cannot blame or resent weak AI. Praising Weak AI for expressions of goodwill would be similarly inappropriate. We may be pleased at the consequences of its actions, or find that it makes life more enjoyable, but it is wrong to treat this as an appreciation of its expressions of goodwill.

An example illustrates why a good mimic—like a weak AI with convincing human-like behavior—meets the standards for exemption. Consider again accidentally banging your knee on a coffee table. When you hit your knee, you'd reasonably feel some sort of negative emotion about the experience. That makes sense; the whole state of affairs is unfortunate and you'd rather it not have happened. Blaming your coffee table for hurting you in the way you'd blame a mugger in an alley for hitting your knee would be nonsensical. The mugger's activity is naturally interpreted as an expression of ill will and the mugger possesses an ill will that can be expressed. A table has no will and thus cannot express an ill will; it is not apt for blame, even if you're frustrated. Weak AI is like the table: there is, by definition, "nothing going on inside" a weak AI. To be clear, "what is going on inside" must include emotions if an entity is to be morally responsible on a Strawsonian account, because emotions are partially constitutive of reactive attitudes.[17] Without these emotions, the attitudes in question just are not the interpersonal reactive attitudes (Strawson, 2008, 10). No weak AI is capable of these attitudes, and so no weak AI has the relevant core capacities.

Our conclusion is that AIs are morally responsible if and only if they are a particular sort of strong AI, SRAI. Intuitively, any strong AI with capacities for reactive attitudes is morally responsible on our broadly Strawsonian picture. After all, SRAI would differ from us only in that it happens to be artificial and mechanical—a bit of code—and these do not seem to be morally salient differences between humans and machines.[18]

[16] If one is insistent that there is an ill will expressed here at all, perhaps it is that of the AI's creator. Alternatively, one might see ill will as coming from a human-generated data set on which the weak AI was trained.

[17] One might wonder how we should react to weak AIs that act in morally impermissible ways. Some weak AIs may merely be extensions of their programmers. In these cases, the programmers or developers are the appropriate targets of the interpersonal reactive attitudes, as it is the quality of the programmers' will that is expressed in the AI's actions. Other AIs are functionally "black boxes." In these cases, it is impossible for creators or programmers to predict or explain how such a machine will learn or react. Assuming the creators or programmers are not negligent, these black box AIs fall into a responsibility gap (Matthias 2004). In these cases, weak AIs themselves are morally exempt in a similar way to small children, and psychopaths.

[18] On another version of this objection, it is a category mistake to think of weak AIs as exempt, because, like coffee tables and alarm clocks, they were not properly considered things in need of exemption in the first place. We are perfectly happy with this interpretation, as it means that weak AIs are not morally responsible.

Objections

In this section, we'll consider two objections to our position that only strong AI can be morally responsible. The first objection argues that the proliferation of weak AI might alter the moral landscape and shift what "statistically responsible" AI means. The second objection argues that it is practically impossible to draw a bright line between strong AI (which is morally responsible) and a weak AI which can perfectly mimic human moral decision-making. Both objections, in short, attempt to posit circumstances under which we might plausibly conceive of weak AIs as altering the ways in which we ascribe moral responsibility. We respond to each of these objections. In our response to the second objection, we provide some loose guidelines as a heuristic to address this problem of determining when we have created AI with the relevant capacities.

Statistically ordinary weak artificial intelligence

Objection

Our interpretation of the grounds for exemption from the reactive attitudes turns on the idea of statistical ordinariness; the capacities that one must have if they are not to be exempt are those that are normal for members of a given moral community. An objection to this view might be that weak AIs are soon to be universal. Weak AIs already aid in making decisions with moral valence (e.g. AI often make the initial cut for large pools of job applications). When cars become fully autonomous, AI will be making life and death decisions constantly. As weak AIs become more integrated into our lives, they will participate—in some way or another—in most moral decisions. If this were to happen, moral decision making would be riddled with the capacities possessed by weak AI. Indeed, it would seem that the weak AI would be so prevalent in moral decision making that they would impact what we count as statistically normal capacities in this decision making.

Response

While a world with a plethora of weak AIs would likely require the development of new social norms and policies, this does not force us to accept weak AIs as members of the moral community. Consider the number of beings with which we have social interactions who are not members of the moral community. There are already many children and adults of diminished or altered mental and social capacities, and so on. Though we interact with children and the intellectually or socially impaired frequently, and have norms for interacting with them, we don't take them to be members of the moral community. There being a greater number of exempt beings does not require us to treat them as full members of the moral community. For example, if the proportion of children in the population suddenly skyrocketed, we would not respond by including them as full members of the moral community simply because of their number.[19] Instead, we would consider that the number of beings towards whom it is inappropriate to direct the full suite of interpersonal reactive attitudes had increased.

Similarly, no matter how many weak AIs participate in moral decisions, they still meet the standards for exemption; they fail to have the capacities for the attitudes that constitute the demand for a certain kind of treatment. Weak AI possesses merely a convincing facsimile of social and emotional capacities that characterize moral life. No matter how many of them there are, weak AIs can only *seem* to make demands for certain kinds of behavior or to demand the expression of a certain kind of will, can only *seem* to express an ill will towards anything, because they have no will to speak of.

A massive spike in the number of weak AIs would certainly lead to social changes. Even exempt entities are still objects of policy; it is plausible that new norms will be needed for a society in which a large proportion of beings are weak AIs. However, their being an object of policy does not change the fact that weak AIs lack the core capacities to be members of the moral community.

A practical problem

Objection

Our method for ascribing moral responsibility to AI requires that AIs which are morally responsible have a will and the emotional capacities that allow them to be both subjects and recipients of reactive attitudes. However, it is impossible—in a practical sense—to determine whether AIs do in fact have the required will and emotional capacities. Indeed, any AI that seems to be a strong AI may be merely an exquisite mimic. This seems to render the Strawsonian approach to ascribing moral responsibility to AI nothing more than a theoretical exercise, since weak and strong AIs are indistinguishable.

Response

It is true that we cannot look "inside" an AI to determine whether it is strong or weak, but neither can we look inside our neighbors. In other words, people whom we consider

[19] We wouldn't think it was appropriate to blame a three year old for murder if 55% of the world's population happened to be three or younger.

to be members of our moral community might turn out to be philosophical zombies (p-zombies).[20] We cannot look "inside" each other to determine that the beings we interact with aren't p-zombies. Nevertheless, we ascribe moral responsibility to others regularly; such ascription is the only way that we can function in a moral world.

Just as any usable policy, law, or social norm governing human interaction will make reference to observable features of the humans in question, any usable policy, law, and social norms governing the interactions between humans and AI must do the same. When non-artificial intelligences interact with artificial ones, we'll inevitably direct attitudes at them and we must evaluate the aptness of those attitudes at least partly on the basis of observable features of those artificial intelligences. It seems no more implausible to do this in the case of AI than in the human case.

Despite the limits of our observational powers, we believe it is important to understand at a theoretical level when AIs are and are not morally responsible because this theoretical framework will color the way in which we observe and interact with AIs. If we enter into these interactions with nothing more than a naive interpretive toolkit—one according to which things that simply *seem* morally responsible are morally responsible—then there is a real danger of treating far too many beings as morally responsible when they in fact are not. This is not to say that we should assume no AIs are morally responsible; that would be to go too far in the opposite direction. After all, the Strawsonian approach we've described here has the result that at least some AIs (the strong ones) are candidates for being morally responsible. Instead, we think that the proper approach is to have a healthy skepticism about whether any particular AI is morally responsible, and having a good theoretical account of what it takes for an AI to be morally responsible is an important part of developing and maintaining that skepticism.

It is likely that we will be unable to use the same observational standards to attribute will and emotional capacities to AI that we would employ with other humans. Instead, we argue that we should determine whether an AI has the necessary capacities to be part of the moral community using a standard based on that which is used to ascribe particular cognitive traits to animals other than humans. In particular, we argue that we should use the same practices that are used by cephalopod researchers; they follow the principle that we should only ascribe human-like capacities to octopus (and other cephalopods) when such behavior can be explained in no other way.

Cephalopods, and particularly octopuses, provide a useful analogy to AI because, like AI, they can interact with data in highly complex ways, but they are startlingly different from humans. The last common ancestor between humans and octopuses lived about 600 million years ago (Godfrey-Smith, 2016, p 8) and lacked many important attributes that are common to both humans and octopus. Octopuses are often regarded as particularly intelligent (Godfrey-Smith, 2016), but the way that they perceive and interact with the world is so alien to us, that it is difficult for us to even test their intelligence (Godfrey-Smith, 2016, pp 43–76). For example, our best explanation of octopus color change is that an octopus can see with its skin (Godfrey-Smith, 2016, pp 120–121), a concept that is challenging for us. AI poses some similar problems, given that it's perception of the world is completely different from our own. AI that interacts with anything visual (e.g. a photograph, the world) does not see in the way that we do but receives the information pixel-by-pixel. This difference in ways of gaining information about the world will make it challenging for us to determine when AI, in fact, has the capacities in question, but no more challenging than it would be to sort out whether an octopus or intelligent extraterrestrial life form had the necessary capacities. Given that we already use such standards to determine what kinds of protections animals deserve, it seems reasonable to extend these standards to artificial entities as well.

Using a set of standards that have already been developed for assessing the intelligence, emotions, and intentional states of animals is preferable to alternatives developed to specifically assess AI. Sullins (2006) is an example of an attempt to create a way to ascribe moral responsibility to machines specifically that tries to avoid considering whether those machines have a will. Sullins proposes a three-pronged test for morally responsible machines. The three prongs that ask evaluators to consider (1) the autonomy of the machine, (2) intentional behavior, and (3) position of responsibility. The second prong asks "is the robot's behavior intentional?" and Sullins contends that behavior counts as intentional "as long as the behavior is complex enough that one is forced to rely on standard folk psychological notions of predisposition or 'intention' to do good or harm" (Sullins, 2006, p 28). It isn't clear what Sullins means by "forced". It isn't clear what could possibly compel us to see a machine this way as long as competing explanations are available. If what Sullins means is that ascription of intentional mental states is the most natural explanation for machine behavior (if not the only logically consistent), then such an explanation seems to resurrect the mental aspects he pains to avoid, because this is the same way that we ascribe intentionality both to animals and to other people. In fact, it seems similar to the way that we ascribe intentional states of

[20] For an overview of p-zombies, see (Kirk 2019). One might object that we are less likely to have empathy towards AI (even strong AI) than p-zombies, because p-zombies both look and act like us. However, since humans have a tendency to anthropomorphize and feel empathy toward all sorts of things that look nothing like us (animals, corporations), this objection holds little weight. The movie Wall-E was popular, in part, because humans have no problem feeling empathy for robots (even those who look nothing like us).

cephalopods mentioned above. The position of responsibility criterion seems to only apply to a niche set of cases, even among AI systems. Using a Strawsonian approach, one need not have the machine or AI have a specific professional responsibility–such as a nurse–to be a moral agent. Any AI that appropriately displayed and received reactive attitudes shows that it understands the moral community and manifests a will.

Conclusion

In this paper, we've argued that taking a Strawsonian perspective on moral responsibility concludes that only strong AI can be morally responsible. Although strong and weak AIs may be externally or observationally indistinguishable, only a strong AI has the attitudes necessary for genuine moral responsibility. We've defended this view against the objections that the Strawsonian perspective requires the admission of weak AI as morally responsible, and the objection that there is no real point in making the moral responsibility of AI dependent on unobservable features, such as attitudes.

We opened this article by noting that human/AI interaction is becoming increasingly common. Only through continued human/AI interaction can we determine whether AIs can be moral agents, full or marginal. Without continuing to interact with AIs in moral situations—especially novel moral situations—we have no way of gaining a deep understanding of the moral responsibility of AIs. Perhaps, then, we are best described as pessimists about any kind of armchair solution to the question of whether or not there are genuine artificial moral agents. If Strawson is right that the reactive attitudes are a deeply rooted part of human existence, then we will only answer this question by interacting with AIs in circumstances with genuine moral stakes and making a careful study of both our own attitudes and the best explanations of the AI actions. The other lesson here is that the discussion of morally responsible AI is perhaps poorly named. It may be that what is at stake is "reliably predictable and consequently beneficial AI," as opposed to anything moral.

Acknowledgment The Authors would like to thank Karl Schafer, Eric Brown, Chris Freiman, Kaite McKenna, and Sara Rodriguez.

Funding None.

Declarations

Conflict of interest The authors declare that they have no conflict of interest to disclose.

References

Barry, P. B. (2011). Saving Strawson: Evil and Strawsonian accounts of moral responsibility. *Ethical Theory and Moral Practice, 14*(5), 5–21. https://doi.org/10.1007/s10677-009-9219-x.

Beard, J. M. (2014). Autonomous weapons and human responsibilities. *Georgetown Journal of International Law, 45*(3), 617–682.

Benn, P. (1999). "Freedom, resentment, and the psychopath." *Philosophy, Psychiatry, & Psychology, 6*(1), 29–39.

Brown, D. (2017). *Origin*. Doubleday.

Champagne, M., & Tonkens, R. (2015). Bridging the responsibility gap in automated warfare. *Philosophy and Technology, 28*, 125–137. https://doi.org/10.1007/s13347-013-0138-3.

Ciocchetti, C. (2003). The responsibility of the psychopathic offender. *Philosophy, Psychiatry, and Psychology*. https://doi.org/10.1353/ppp.2003.0089.

Coeckelbergh, M. (2014). The moral standing of machines: Towards a relational and non-Cartesian moral hermeneutics. *Philosophy and Technology*. https://doi.org/10.1007/s13347-013-0133-8.

Cromzigt, L. (2015). *Strawson's take on moral responsibility applied to intelligent systems*. Utrecht University. BSc thesis.

Doris, J. (2002). *Lack of character*. Cambridge University Press.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Machine Ethics, 14*, 349–379. https://doi.org/10.1017/CBO9780511978036.013.

Gilbert, M. (2014). *Joint commitment: How we make the social world*. Oxford University Press.

Godfrey-Smith, P. (2016). *Other minds: The Octopus, the sea, and the deep origins of consciousness*. Farrar.

Greenspan, P. S. (2003). Responsible psychopaths. *Philosophical Psychology*. https://doi.org/10.1080/0951508032000121797.

Greenspan, P. (2016). Responsible psychopaths revisited. *Journal of Ethics, 20*, 265–278. https://doi.org/10.1007/s10892-016-9231-z.

Gunkel, D. J. (2017). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-017-9428-2.

Hieronymi, P. (2020). *Freedom, resentment, and the metaphysics of morals*. Princeton University Press.

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*, 19–29. https://doi.org/10.1007/s10676-008-9167-5.

Himmelreich, J. (2019). Responsibility for killer robots. *Ethical Theory and Moral Practice, 22*, 731–747. https://doi.org/10.1007/s10677-019-10007-9.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*, 195–204. https://doi.org/10.1007/s10676-006-9111-5.

Johnson, D. G., & Miller, K. W. (2008). Un-making artificial moral agents. *Ethics and Information Technology, 10*, 123–133. https://doi.org/10.1007/s10676-008-9174-6.

Kirk, R. (2019). Zombies. In *The stanford encyclopedia of philosophy*, E. N. Zalta (ed.). Retrieved from https://plato.stanford.edu/archives/spr2019/entries/zombies/

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*, 175–183. https://doi.org/10.1007/s10676-004-3422-1.

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

Ramirez, E. (2013). Psychopathy, moral reasons, and responsibility. In Perry C. Alexandra & D. Herrera (Eds.), *Ethics and neurodiversity.* Cambridge Scholars Publishing.

Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. (3rd ed.). Pearson.

Scanlon, T. M. (2008). *Moral dimensions: Permissibility, meaning*. Belknap Press of Harvard University Press.

Schulzke, M. (2013). Autonomous weapons and distributed responsibility. *Philosophy & Technology, 26*(2), 203–219.

Shoemaker, D. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics, 121*(3), 603–632.

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

Silver, D. (2005). A Strawsonian defense of corporate moral responsibility. *American Philosophical Quarterly, 42*(4), 279–293. https://doi.org/10.2307/20010212.

Smith, A. M. (2012). Attributability, answerability, and accountability: In defense of a unified account. *Ethics*, *122*(3), 575–589.

Smith, A. M. (2015). Responsibility as answerability. *Inquiry: An Interdisciplinary Journal of Philosophy, 58*(2), 99–126. https://doi.org/10.1080/0020174X.2015.986851.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*, 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

Stahl, B. C. (2006). Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and information technology, 8*(4), 205–213.

Strawson, P. F. (2008). Freedom and resentment. *Freedom and Resentment and Other Essays*. https://doi.org/10.4324/9780203882566.

Sullins, J. P. (2006). When is a robot a moral agent? *International Review of Information Ethics, 6*, 23–30.

Sun, R. (2001). Computation, reduction, and teleology of consciousness. *Cognitive Systems Research, 1*(4), 241–249. https://doi.org/10.1016/S1389-0417(00)00013-9.

Swoboda, T. (2018). Autonomous weapon systems—An alleged responsibility gap. In V. Müller (Ed.), *Philosophy and theory of artificial intelligence 2017. PT-AI 2017. Studies in applied philosophy, epistemology and rational ethics.* (Vol. 44). Springer.

Talbert, M. (2008). Blame and responsiveness to moral reasons: Are psychopaths blameworthy? *Pacific Philosophical Quarterly*. https://doi.org/10.1111/j.1468-0114.2008.00334.x.

Talbert, M. (2019). Moral responsibility. In E. Zalta (Ed.), *The stanford encyclopedia of philosophy.* Stanford University.

Tollefsen, D. (2002). Organizations as true believers. *Journal of Social Philosophy*, *33*(3), 395–410.

Tollefsen, D. P. (2003). Participant reactive attitudes and collective responsibility. *Philosophical Explorations, 6*(3), 218–234. https://doi.org/10.1080/10002003098538751.

Tuomela, R. (2013). *Social ontology: Collective intentionality and group agents*. Oxford University Press.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics, 24*(2), 227–248.

Williams, B. (1970). The self and the future. *The Philosophical Review*, *79*(2), 161–180. https://doi.org/10.2307/2183946.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.