Cambridge Quarterly of Healthcare Ethics (2021), 30: 3, 435–447
doi:10.1017/S0963180120000985

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible

Daniel W. Tigard*

Institute for History and Ethics of Medicine, Technical University of Munich, Ismaninger Str. 22 81675 Munich, Germany
***Corresponding author.** Email: daniel.tigard@tum.de

(Received 26 March 2019; revised 04 December 2019; accepted 06 May 2020)

Abstract

Our ability to locate moral responsibility is often thought to be a necessary condition for conducting morally permissible medical practice, engaging in a just war, and other high-stakes endeavors. Yet, with increasing reliance upon artificially intelligent systems, we may be facing a widening *responsibility gap*, which, some argue, cannot be bridged by traditional concepts of responsibility. How then, if at all, can we make use of crucial emerging technologies? According to Colin Allen and Wendell Wallach, the advent of so-called 'artificial moral agents' (AMAs) is inevitable. Still, this notion may seem to push back the problem, leaving those who have an interest in developing autonomous technology with a dilemma. We may need to scale-back our efforts at deploying AMAs (or at least maintain human oversight); otherwise, we must rapidly and drastically update our moral and legal norms in a way that ensures responsibility for potentially avoidable harms. This paper invokes contemporary accounts of responsibility in order to show how artificially intelligent systems might be held responsible. Although many theorists are concerned enough to develop artificial conceptions of *agency* or to exploit our present inability to regulate valuable innovations, the proposal here highlights the importance of—and outlines a plausible foundation for—a workable notion of *artificial moral responsibility*.

Keywords: moral responsibility; moral agency; blame; machine ethics; AI ethics; artificial intelligence; human-robot interaction

Introduction

January of 2019 marked the 40th anniversary of what is said to be the first human fatality caused by a robot. Robert Williams, a factory worker for the Ford Motor Company, reportedly climbed atop a faulty casting machine when a mechanical arm of the 1-ton unit struck him in the head.[1] Although the incident generated some degree of attention and a significant legal settlement for the family, today we see widespread public concern over who is to blame when machines cause us harm. No doubt, with the increasing prevalence of artificially intelligent (AI) systems comes a growing urgency to address the question of who, if anyone, can be held responsible for the harms resulting from AI.

As the rapidly emerging body of literature attests, the question is not only *who* can be responsible, but *how* we can coherently locate responsibility when the source of harm is an autonomous AI system. According to Andreas Matthias, our use of autonomous machines creates a "*responsibility gap*, which cannot be bridged by traditional concepts of responsibility."[2] Indeed, it is commonly accepted that moral responsibility is a necessary condition for such high-stakes endeavors as warfare and medical practice, among others. If AI systems in fact create a gap in responsibility, our use of AI in these domains appears to be morally indefensible. In this way, it seems we should either scale-back our deployment of AI systems, or else find novel ways to bridge the gap.

Naturally, there are many advocates for the more conservative approach, with some supporting a complete moratorium on autonomous machines, namely those designed to harm human lives.[3] Robert Sparrow, for example, argues that responsibility for harms in warfare cannot fall upon the programmers or users of truly autonomous AI systems.[4] Importantly, the machines themselves are also implausible loci of responsibility; reasons for this claim will be clarified and challenged below. For Sparrow and others, because our ability to identify moral responsibility is necessary for engaging in a just war, it appears that our use of AI in warfare is unjustified. To be sure, comparable arguments are being put forward to show the problems and the need for swift regulation of AI in healthcare, self-driving cars, and more.[5]

According to Colin Allen and Wendell Wallach, the advent of so-called 'artificial moral agents' (AMAs) is both necessary and inevitable.[6] As I will discuss, this notion helps us to understand what it means for AI technologies to operate in morally significant circumstances, like warfare and medicine. Still, it remains unclear how we might hold AI responsible. Despite the frequent overlap in philosophical buzzwords, on most accounts, moral agency is not coextensive with moral responsibility.[7] That is, one might fulfill whatever conditions for moral agency are on the table—knowledge and free will are usual candidates—but for some reason not be morally responsible. Picture, for example, those who face genuine moral dilemmas; they may well know what they are doing and act freely in choosing the lesser of two evils, yet the practice of blaming them for the consequences is often inappropriate. In short, determinations of moral agency, whether natural or artificial, do not yet tell us about moral responsibility. As such, the idea of artificial moral *agency* seems to push back the problem of locating responsibility for harms resulting from AI systems. Although the notion of AMAs is conceptually useful, what we need is an account of moral *responsibility* that can be applied to emerging technologies and novel beings, in a way that allows us to locate moral responsibility in AI. In other words, we need a workable notion of *artificial moral responsibility*.

In this paper, I look to contemporary accounts of responsibility in order to show how AI systems—such as AMAs—might be held responsible. I begin by expanding upon the ideas of moral agency, both natural and artificial. As I have suggested, addressing questions of agency cannot fully satisfy the search for mechanisms by which we hold others responsible. Accordingly, I then explain how agency matters and how it does not, and I challenge those who find our use of AI unjustified on the grounds that we lack responsibility. I then turn to an analysis of moral responsibility as advanced by Gary Watson and David Shoemaker. What these notable developments provide, as I will make clear, is a rich conceptual and practical foundation for managing our interactions with a host of questionable subjects, from children and psychopaths to non-human animals and AI systems. Building upon this foundation, I suggest, offers an understanding of how we can and cannot hold machines responsible.

Natural and Artificial Moral Agency

The main focus of my attention in this work will be moral *responsibility*, in both its natural and artificial manifestations. However, I wish to draw upon and distinguish my account of artificial moral responsibility from the notion of AMAs. Thus, a fruitful point of departure will be an examination of moral agency, again, in both the natural and artificial manifestations. So, what exactly does it mean, naturally speaking, to be a moral agent?

As theorists since Aristotle have established, when we speak of agency we typically have in mind the ultimate source of an action.[8] Beyond being a random movement, actions can be intentional and performed on the basis of reasons. Intentional actions follow from decisions, or might be said to include decision-making as a sort of mental action.[9] Importantly, for those with the capacity to act intentionally, it is often thought, we see an ability to form a judgment based upon consideration of a multitude of reasons—for and against bringing about some state of affairs, or for bringing about some state of affairs rather than another.[10] In this way, agents have the ability to display desires, at least very loosely, along with beliefs and plans for how to fulfill them.[11] In sum, to be the source of an action entails a number of crucial capacities, many of which appear to require something like consciousness.[12] It will be helpful to note, then, what it is about these capacities that makes an agent a *moral* agent.

Crudely, if an agent is one who is capable of being the source of an action, along with the coinciding capacities for acting intentionally (and so on), moral agents can be thought of as those who maintain capacities for initiating action bearing moral significance. As Michael McKenna aptly states, a "moral agent is a person who is capable of action that can be morally evaluated as good or bad, right or wrong, virtuous or vicious."[13] What this conception provides is a line with which we can draw several fundamental distinctions. First, we see the beginnings of a separation between moral agency and moral responsibility; I will discuss this further in the following section. At present, what we also begin to see is that moral agency typically describes one's membership in a special subset of a wider class, known as the moral 'patients' or, simply, moral subjects. We commonly think, for example, that children or the environment are proper subjects of moral concern; while, normally, they do not (yet) have the capacity to *be* morally concerned. That is, the broader moral community is made up of a great diversity of natural specimens, only some of which maintain and display moral concern.

I suggested above that two key capacities are widely associated with moral agency, and often held as necessary conditions, namely a certain sort of knowledge and the ability to act freely. While agency *simpliciter* can be fruitfully seen as involving intentional action on the basis of reasons, when it comes to moral agency it seems that one must be able to act intentionally on *moral* reasons. Indeed, for those who can act intentionally but cannot grasp or feel the force of moral reasons, we would be hesitant to ascribe moral agency.[14] One must know *right from wrong*, as it were, along with a host of moral and non-moral considerations pertaining to the immediate situation.[15] Here we see why children and persons with cognitive or emotional impairments are not thought to possess full moral agency. They *don't know any better*. Likewise, even where one does understand and is motivated by moral reasons, moral agency is not ascribed to those who cannot freely control their behavior. In cases of coercion—whether facing a gunman or a psychological compulsion—coerced subjects simply cannot act in ways they otherwise would or in ways they know to be right. Without a sufficient degree of control, again, we see that moral agency is compromised or entirely absent. The question, then, is who or what else, if anyone or anything other than fully functioning adult human beings, can truly possess moral agency?

On the natural picture of moral agency, as I have in mind here, the short answer is no one and nothing. Although the moral community is broad, encompassing all kinds of creatures properly demanding moral concern, the class of moral agents is a rather exclusive one.[16] Granted, it may be that the especially intelligent non-human animals—dolphins or chimpanzees, for instance—are capable of the sort of empathy we associate with moral concern.[17] Such creatures are also natural, biological beings, of course. But what appears non-natural, I am suggesting, is the ascription of full moral agency to anything other than creatures like ourselves. No doubt, we occasionally consider non-human animals as possessors of something *like* moral agency. But when we think, for example, of dolphins as non-human persons, or of the family dog as happy to see us, we are artificially ascribing a sense of moral agency. Such practices are not necessarily incorrect. The claim here is simply that we employ non-natural ascriptions of moral agency when the concept is used to denote creatures other than those who can grasp and act freely upon moral reasons.

Turning to the notion of AMAs, it should be stated upfront that Allen and Wallach are concerned with ascriptions of moral agency specifically to machines, from robots with human-like structures to automated software. Still, with the account offered here, we gain a broader understanding of what it means, generally, to ascribe moral agency artificially. Allen and Wallach state that the complexity of today's systems will require "the systems themselves to make moral decisions…[and] will expand the circle of moral agents beyond humans to artificially intelligent systems."[18] On this picture, AMAs are a fairly straightforward concept. They can be known simply as AI systems, whether robots or software, within the exclusive class of moral agents. Note that my initial framing of this class, namely as a subset of the wider circle of subjects deserving moral concern, might imply that some machines are to be considered capable of both exhibiting moral concern as well as being the target of our moral concern. It is here we see room for reasonable resistance to this expansion of moral agency.

Some thinkers encourage us to be open to the idea of being cared for *and caring for* AI. Consider the tales of Isaac Asimov, creator of the famed Laws of Robotics.[19] More recently, the work of David Gunkel persuades us to think of machines as possessing capacities for action and as deserving protections with

rights of their own.[20] For Gunkel, when considering the social and legal status of machines, moral agency should indeed remain a subset of the wider class of moral subjects. Yet, it is not clear why we must retain this standard conceptual map when navigating our emerging regulatory mechanisms for novel technologies. Undeniably, it would be difficult to justify showing absolutely no moral concern for a creature like ourselves, that is, for a *natural* moral agent. But the notion of AMAs, I take it, accounts precisely for the fact that non-natural ascriptions of moral agency to AI systems are just that—*artificial*—and nothing more. In other words, there appear to be few reasons to retain our conceptual map of natural moral agency and patiency when dealing with artificially intelligent beings. Admittedly, some of those reasons will find support. It might be argued, for example, that exhibiting disrespectful behavior toward AI systems, treating them as if they lack moral patiency, might globalize by negatively affecting our behavior toward fellow human beings.[21]

My main objective, as stated, is to examine moral responsibility as it might be applied to AI. Thus, without firmly committing myself to what we owe (or do not owe) to machines, I want to suppose that AMAs represent an outlying specimen, a deviation from our standard conceptual map of moral agency and patiency. When it comes to *artificial* moral agents, moral agency is no longer a proper subset; instead, it is likely a largely overlapping Venn diagram, with some AMAs occupying the space beyond the class of moral subjects. Indeed, this picture appears to resemble what Allen and Wallach have in mind with their deployment of the term AMAs.

An AMA is not yet capable of consciousness, Allen and Wallach admit. Still, "if its architecture and mechanism allow it to do many of the same tasks" as human consciousness, it can be considered *functionally conscious*.[22] Analogously, even if AI is not capable of full moral agency, the architecture and mechanisms of many emerging systems allow for the same sorts of behaviors as natural moral agents. Accordingly, many AI systems can be considered *functionally moral*, occupying a space somewhere between "operational morality"—where decisions concerning values remain in the hands of the designers and users—and full moral agency.[23] In short, while AI cannot be said to possess capacities for understanding moral reasons and for freely acting upon decisions in the ways we enjoy, today's technologies are capable of behaving in ways that appear morally significant.[24]

Picture, for example, personal AI assistants programmed to encourage courteous interactions from their users.[25] Clearly, this level of technological sophistication is not enough to warrant full moral agency, nor does it call for such devices being included within the sphere of our moral concern. The reasons to encourage courteous interactions are, of course, to support a certain developmental pattern in the users, particularly with children.[26] Likewise, for a host of AI systems, the concerns for potentially negative interactions globalizing to fellow humans does not show that we should treat AI with moral concern *for the sake of AI*. It is because we wish to foster healthy moral development *in humans* and maintain moral concern for *each other* that we see reasons to treat some AI systems as if they were moral subjects and perhaps as something like moral agents.[27]

In what follows, I want to continue supposing that many AI systems already possess—and will increasingly exhibit—the sort of "functional morality" that Allen and Wallach outline. AMAs are surely not natural moral agents. Yet, they are capable of behaving in ways that are more significant, morally speaking, than machines that simply execute the commands of their designers and users. And although the notion of AMAs allows us to redraw our standard conceptual map of moral agency and patiency, we are nonetheless left without a firm grasp on how we might hold machines responsible. In order to reveal these processes, we must, to some degree, put aside questions of natural and artificial moral agency.

How Agency Does and Does Not Matter

I claimed above that moral agency is not coextensive with moral responsibility. In many cases, we might easily grant that one is a moral agent but that she is not morally responsible for some action performed, even one bearing great moral significance. Genuine moral dilemmas and cases of moral luck, for instance, begin to elicit intuitions supporting this division. But how do we draw the line in theory? What exactly is moral responsibility *apart from* moral agency? Addressing these inquiries will help to clarify the extent of agency's significance in our attempt to locate responsibility in AI.

With his widely influential essay, "Freedom and Resentment," P.F. Strawson moved the debate over the nature of moral responsibility beyond a strict focus on the conditions for moral agency.[28] According to Strawson's observations, there is something much more fundamental to what it means to be morally responsible than simply assuring that one is sufficiently informed and capable of acting freely. The moral community, he pointed out, is built upon relationships with one another. As members of the community and within our personal relationships, we hold expectations and we demand, at least implicitly, that we are treated by others with good will or, at minimum, without ill will. Accordingly, our concerns for the conduct of others are focused both on the actions themselves and on how they are performed. To illustrate, consider that we respond quite differently to the fellow subway passenger who steps on our toe to avoid tripping onto an elderly person, versus the fellow passenger who steps on our toe deliberately and with a mischievous smirk. Clearly, in the latter case, we are likely to respond with something like contempt or resentment. These sorts of responses—along with more positive sentiments, like admiration and gratitude—are commonly known as the *reactive attitudes*. For Strawson and numerous followers, being morally responsible is a function of our "natural human reactions to the good or ill will or indifference of others towards us."[29] We resent others for insulting us. We are grateful when others go out of their way to help us. In these ways, our attitudes as well as our corresponding practices of blaming and praising, punishing and rewarding, make up our *responsibility responses*.[30]

On its surface, Strawson's development may appear rather commonplace, for indeed, the theory is constructed consistently with our everyday reflections and interactions. Yet, for many, its significance is hard to overstate. On an extreme reading, moral responsibility turns out to be a highly subjective property of its object, if it can even be called a property. On this view, moral responsibility is a process—we hold others responsible and, importantly, *holding* is conceptually prior to *being* responsible. Here we see what looks like a reversal of traditional concepts of responsibility. That is, the facts of responsibility are determined primarily by our responses, our attitudes and practices, and not the other way around.[31] In other words, moral agency does not *really* matter for our determinations of responsibility. If anything, considerations of agency are secondary, serving to affirm or call into question our existing attitudes and practices. For example, when I respond to the subway toe-stepper with anger, I might do so before knowing exactly who or what stepped on my toe. Upon learning that it was the smirking misfit, my anger is surely vindicated and I may well act accordingly, say, by communicating the offense. By contrast, were I to learn that it was purely an accident, that the toe-stepper was a child who knew no better, or that it was a lost bowling ball rolling down the aisle, I would likely work to quell my anger—or at least shift it from the immediate object to, say, the Fates. Thus, considerations of agency still matter, but are supplemental. We do not continue to hold ignorant children or bowling balls responsible, at least not appropriately. In any case, a key takeaway is that being morally responsible, on this view, is clearly something more than being a moral agent. It is to *be held* morally responsible by being targeted with a moral agent's attitudes and practices. I'll refer to this as the 'process view' of responsibility.

On the opposite extreme, we must—and we often do—account for the agential status of the objects to which we respond. On this line of thought, responsibility is an objective property, and our determinations of it will be based largely upon our awareness of fundamental features of the object. Is the source of harm human? Did he or she (or they) know what they were doing, and did they do so freely? Notice that the conditions of moral agency play a more prominent role here. Our attitudes and practices can certainly *aid* us when it comes to holding a given target responsible. For example, blame is often seen paradigmatically as a form of communication, and as we can imagine, communicating an offense may well be assisted by expressions of anger.[32] Still, on this picture, our responses are merely "epistemic markers" of responsibility, where *being* responsible remains conceptually prior to being *held* responsible.[33] Again, on this view as well, we see that moral responsibility is something other than moral agency. It is to *be identified* as the appropriate target of our attitudes and practices. I'll refer to this as the 'property view'.

No doubt, the question of responsibility as a process or a property contains layers of complexity that cannot be fully addressed here. But again, I want to avoid coming down definitively in favor of either extreme. This is in part due to spatial limitations, but also because I can grant that both views have their merits. Specifically, the process view holds that moral responsibility depends upon our actual attitudes

and practices, which seems highly plausible. As Strawson and many followers helped us to understand, without the interactions and responses that occur within our interpersonal relationships, it would be far from clear what we mean when we ascribe responsibility. At the same time, the property view illuminates the fact that we naturally modify and refine our responses according to observable features of the target.[34] For these reasons, in what remains of this section, I will return to the issue of AMAs with the broadly Strawsonian framework, assessing where and why some theorists argue against locating moral responsibility in machines.

At the outset, I introduced Sparrow as a staunch advocate for the position that we lack responsibility when machines act autonomously. Considering that responsibility is implied by the *jus in bello* principles, our deployment of AI systems in warfare appears morally indefensible. For Sparrow, neither the programmers, the users, nor the machines themselves are plausible loci of responsibility. I take this line of argumentation to be generalizable to many public domains where AI is being developed or is already deployed.[35] And while locating responsibility in programmers, users, or machines may indeed present difficulties, I want to clarify and challenge the argument, and thereby call into question any generalizations to other domains. With this task, my goal is not to promote AI in military and public contexts, but only to demote those who find it absolutely unjustified.

Sparrow states of autonomous systems that "it is not possible to hold anyone else responsible for its actions."[36] Note both the descriptive, absolute nature of this claim and that already he is loading the dice by singling-out machines themselves as the only potential loci of responsibility. Shortly thereafter, it is claimed "the more these machines are held to be autonomous the less it seems that those who program or design them, or those who order them into action, *should* be held responsible."[37] Note here two points. First, this is a normative claim; it tells us not that it is impossible to hold others responsible for a machine's actions, only that doing so goes against some norm. Second, although it was not likely intended, with both claims cited here (among others) Sparrow displays support for a process view of responsibility. In other claims, however, the focus is less on *holding* and more on *being* responsible: "The question of who *is* responsible," he says, for example, "remains crucial."[38]

What is clear is that Sparrow does not consistently maintain a view on the nature of responsibility, which seems peculiar considering that the key to his overarching argument is to negate our chances of locating it. At times he deploys talk of responsibility as a process that can be subject to normative guidelines, and at other times as a property which simply can or cannot be found. Nevertheless, rather than settling on charges of inconsistency on a distinction that was likely unnoticed, I want to take his premises on the terms invoked. Doing so, I will show, serves to render the overarching argument unconvincing.

In his survey of possible sources of responsibility for autonomous weapons, Sparrow begins with the design and programming. Holding designers or programmers responsible, it is said, "will only be fair if the situation described [war crimes] occurred as a result of negligence."[39] Notice again the normativity at work, namely in the notion of fairness. The basic idea is not that we cannot hold designers and programmers responsible, but only that we should not. We have reasons not to, which are grounded in the broken chain of control or predictability in cases of fully autonomous systems. Even for the users, on Sparrow's picture, the claim is not that we cannot locate responsibility, but that once machines choose their own targets "it will no longer be *fair* to hold the Commanding Officer responsible."[40] Aside from showing his earlier claims regarding the *impossibility* of holding others responsible to be misleading, a potential solution is being overlooked.

If responsibility is a procedural matter, a fair or unfair process of *holding* responsible, it may still be that we can *and do* exercise our usual attitudes and practices. We can blame, say, Microsoft for the racist remarks generated by its autonomous AI bot, Tay.[41] We can praise doctors for their successful use of AI-enabled diagnostic tools.[42] Often, our responsibility responses will seem inappropriate, but largely because we maintain a natural conceptual link, namely moral responsibility falling directly upon those who exercised moral agency in the action being evaluated. I will have more to say on this in the following section. At present, the point to be made is simply that the direct link between agency and responsibility, however natural it may be, is not necessary. We can and often do assign responsibility to individuals

other than the agent who brought about the consequences in question. Importantly for weaponized AI systems and other high-stakes applications, responsibility can (and likely should) be *taken* by individuals associated with the effects.[43] This is precisely what Marc Champagne and Ryan Tonkens aptly argue with their notion of "blank check" responsibility.[44] Of course, blaming those who *take* responsibility does not appear as natural as blaming the agent identified as the causal source of an action. Still, for conceptual and practical purposes, when dealing with artificial moral agents, we must move beyond our reliance upon natural moral agency and responsibility.

When examining whether or not responsibility might be found in machines themselves, understandably, Sparrow adopts the more objective view, wherein responsibility is a property. Adopting this position here allows him to argue, on a purely conceptual basis, that autonomous weapons "could never be held morally responsible."[45] Yet, with the concept invoked—namely, punishment—we again see a plausible solution overlooked. For someone to be held morally responsible, Sparrow notes, they must be capable of receiving blame or praise, punishment or reward. Since AI systems cannot be punished or rewarded, they cannot be held responsible. But here it is explicitly assumed that the most plausible account of the nature and justification of punishment is retributivist or deontological in spirit. For punishment to be punishment, he says, its target "must be capable of suffering."[46] However, there are other highly plausible accounts of the nature and justification of punishment. In short, according to a consequentialist account, the reasons we punish are grounded in the goods to be achieved and not in any type of treatment supposedly deserved by the target. In everyday situations, and in many legal cases, the target's capacities for suffering are simply irrelevant to why we must, say, imprison those who cause great harm. Sparrow claims that "to serve as punishment" our treatments "must evoke the right sort of response in their object."[47] Yet, with a moment's consideration of a smiling psychopath behind bars, we know this to be false. In other words, some who are punished *suffer the consequences*, even if they cannot experience the suffering of a natural moral agent.[48]

To conclude the present section, several points are worth reiterating. First, we punish individuals, very often because of the goods to be achieved. Punishment, indeed, serves as a key mechanism by which we hold individuals responsible. At times, locating responsibility is a matter of identifying a property in the target—say, the capacity for grasping and acting freely upon moral reasons—which turns out to resemble the search for moral agency. At other times, responsibility is a process, whereby we respond with attitudes and practices, and where a direct link to moral agency is unnecessary.[49] What is certain, at least, is that moral responsibility is not a singular nor unified enterprise. Luckily, the complexities can help us to locate responsibility in novel ways and in novel beings.

How We Can and Cannot Hold Machines Responsible

With this final section, my suggestion is not that the various ways of understanding responsibility necessarily apply to AI systems. Nor is it my goal to show that the process view of responding naturally to one another can be neatly mapped onto AMAs. Instead, just as AMAs provide new conceptions of moral agency, I aim to help us move beyond natural conceptions of moral responsibility by showing that, often, our attitudes and practices are adaptable to AI systems.

In his renowned essay "Two Faces of Responsibility," Gary Watson articulated our ambivalence toward agents that appear responsible but are not appropriately *held* responsible.[50] We can, for example, attribute a harmful action to a person's character while stopping short of engaging in other, more overt blaming practices. This might be due to a recognition that the potential blamee was raised in a world (an abusive family, perhaps) where the harmful action (say, insulting others) was a normal part of life. As a result, he may be seen as a bad person in light of his propensity to insult others, but outwardly resenting him for it seems inappropriate. What these sorts of cases show is that locating responsibility comes apart into two "faces." On the one hand, we can assess an agent's character by thinking some conduct is *attributable*; on the other hand, we can hold the agent *accountable* by openly blaming, praising, or otherwise maintaining a "readiness to respond."[51]

Expanding this pluralistic approach, David Shoemaker shows, first, that one can be responsible in the *attributability* sense when an action stems from his character. Here one reveals the things that truly matter to him, what Shoemaker calls "care-commitment *clusters*."[52] Given this process of revealing one's cares and commitments, an agent can be the object of others' admiration or disdain, since these sorts of responses target one's character. Secondly, one is responsible in the *accountability* sense when he is appropriately targeted with blame as a result of exhibiting poor regard for others. Typically, he has 'slighted' someone, namely by being inconsiderate of others' perspectives or insensitive to others' fortunes, both representing failures of empathy. Finally, one is responsible in the *answerability* sense when he is called upon to defend or 'answer for' some exercise of judgment. Here one must be capable of maintaining and citing reasons. Where one has this capacity but fails to properly utilize it, those who demand answers will naturally disapprove.

Although I have given only a very rough depiction of contemporary, pluralistic views of responsibility, we can nonetheless see how such accounts allow us to make sense of a range of questionable cases, from psychological abnormalities to many commonplace interactions. We might, say, disapprove of a friend's decisions when she fails to show up for a weekly meeting; but if her failure is a rare occurrence, surely we do not hold her in contempt. Naturally, we might say 'This isn't like you!' as we demand to hear her reasons. In other words, we take her to be answerable for her conduct, yet we would not likely attribute the action to her in a deep sense, given that her failure to show up does not reflect her underlying cares or commitments. Still, considering that she has the capacity to maintain cares and commitments, our friend is seen as responsible in terms of attributability. Because she is able to entertain and act upon various reasons, she is responsible in terms of answerability. And given that she could be appropriately targeted with blame based on her regard, she is responsible in terms of accountability.

Where one is lacking in any of the usual natural capacities, we find exemptions from moral responsibility.[53] Depending upon the exemption in question, our responses will likely display a sense of ambivalence. Psychopathy, for example, is given a robust explanation on a pluralistic theory. According to Shoemaker, a person with disorders characterized under psychopathy is often exempted from our accountability responses (like anger and resentment) given that they are typically incapable of empathizing. At the same time, this person might effectively display cares (namely for themselves) and may be capable of considering reasons (at least in non-moral decisions); thus, they can be considered responsible in the attributability and answerability senses.[54]

Pluralistic views of responsibility make sense of our common practices while providing theoretically rich understandings of individuals at the *margins* of moral agency, in Shoemaker's terms. My suggestion, then, is that with this framework we might come to better understand moral responsibility in *artificial* moral agents. Note that my aim here is not only to locate responsibility *for* AMAs. This mechanism can be seen in the consideration of others—like the designers or users—*taking* responsibility for a machine's behavior, as discussed above. Instead, my focus is also on what it would look like to hold AMAs themselves responsible.

Recall Sparrow's hasty assumption that punishment, by its nature, aims at retribution. As I claimed, we can also punish those who harm—and reward those who help—in order to achieve certain results. Hence, incarceration is often referred to as *rehabilitation*, particularly when we want to emphasize a hope and belief that the individual in question can progress toward a more desirable condition. We want to improve upon the past, help others to learn from mistakes, and assure that similar behaviors are not repeated. In our more proactive efforts, we might educate early and encourage desirable behaviors before mistakes are made. The point here is that, in some sense, these practices are ways of holding others *to account*. We are ascribing a sort of responsibility when we punish (or reward) with the expectation that the target of our treatment will improve upon (or maintain) existing behavioral patterns.

Under the present description, the process of holding others to account is not the type of responsibility defended by Shoemaker, as his notion was grounded in the empathic capacities of the target. Indeed, his was a notion of accountability for natural moral agents. Nonetheless, the basic mechanism can be effectively adapted for our responses to AMAs. Contrary to Sparrow's arguments, we can punish AMAs. We can impose sanctions on an AMA's domain of application, restrict its previously authorized

behaviors, or work to rewrite any deviant or undesirable lines of code.[55] No doubt, these efforts will not cause it to suffer in anything like the ways we are capable of suffering. But, as I have argued, searching for the target's capacities for suffering or consciousness is a means by which we locate agency and not necessarily the means by which we hold others responsible. While AMAs cannot suffer like us, they can and should suffer the consequences of carrying out harmful behaviors. AI systems capable of functional morality might one day learn from and improve upon their unique mistakes, as a sort of reinforcement learning.[56] Importantly, it will be largely up to us—natural moral agents—to respond and signal when something like a moral transgression has occurred, and for this reason we cannot sit idly with the conviction that AI cannot be held responsible. Considering the sophistication of existing sensory and computing capacities in today's AI systems, there are reasons to think that machines soon to be among us will be capable of recognizing and learning from our moral attitudes and practices—our anger and blame, gratitude and praise—perhaps just as effectively as from our simpler, non-moral commands.[57]

To summarize, what I have suggested is not only that it is possible to locate a sort of responsibility in AMAs—albeit, a non-natural sort of responsibility—but that we already possess mechanisms by which we can resolve the search. In particular, looking to recent pluralistic accounts, we see a range of types of responsibility and unique ways of holding others responsible, even those at the outermost margins of moral agency. Granted, some sorts of responsibility apply much less coherently to AMAs than others. It might seem odd, for example, to think we could attribute some action to an AI system's underlying character. Still, our moral attitudes and practices are adaptable and will likely continue to evolve. For the time being, we can hold some machines to account, namely by engaging in consequential-based forms of punishment or reward. Likewise, we can demand that an AI system's associates—its programmers or users—*take* responsibility, even where these individuals could not have controlled or foreseen the machine's behavior.[58] Undoubtedly, as an objective property, moral responsibility is most aptly identified only in natural moral agents, namely ourselves. That is to say, on a strict property view of responsibility, it may be that we cannot hold machines responsible. Nonetheless, as a procedural matter, we can coherently interact with AI systems—particularly those that are being developed to respond accordingly—in ways that assign a sort of responsibility. And although these mechanisms surely do not indicate the presence of our usual, natural sense of moral responsibility, they can function in very similar ways. If we are willing to grant, as some now are, that certain AI systems are artificial moral agents in virtue of engaging in functionally moral behaviors, we must likewise recognize that our moral attitudes and practices are adaptable in ways that serve to locate a sort of artificial moral responsibility.

Conclusion

I began by stating a problem for the development and use of AI in high-stakes domains, such as warfare and medicine. The case of Robert Williams, victim of the Ford manufacturing tragedy, suggests that for over 40 years we have been blaming sophisticated machines for harm to humans. Yet, it is now said that some systems may create a "gap" in moral responsibility. Some authors use this supposed gap to exploit our difficulties in regulating emerging technologies. If locating responsibility is necessary but we cannot find it when harms result from AI, they argue, our use of AI is morally unjustified. To reiterate, my efforts here have not been an attempt to justify the increasing deployment of AI systems in such sensitive domains as warfare and medicine. Instead, I simply mean to call into question the arguments demanding that we cease, or drastically scale-back, our use of AI. For, indeed, it seems that at this early stage in our lives with AI, we might still find ways to harness the benefits while minimizing the risks.

As some authors have argued, our use of AI is already helping to decrease the harms resulting from our usual risky activities. Engaging in warfare, for example, might stand to harm fewer innocent civilians where weapons are autonomously deployed to target only combatants.[59] Driving cars might harm fewer pedestrians where AI takes the wheel than where we remain in control.[60] Of course, even if such beneficial trends can be reliably produced by our use of AI systems, problems are sure to be encountered. Lives will still be lost, and when the loss of human life appears to result from AI, we will nonetheless want

to know where responsibility lies. For some, it seems that the prospect of losing our grip on responsibility, even for fewer harms, is worse than knowing exactly who is responsible in masses of tragedies.[61] How can we have both the benefits of newfound technology while satisfying our common psychological need to blame the sources of harm?

Recent proposals, namely in the work of Allen and Wallach, have provided fruitful conceptual tools: AMAs, they claim, are necessary for ensuring our future wellbeing in a world increasingly run by AI systems. Indeed, implementing the AMA framework might help us to cope with the otherwise mysterious effects of technology in our lives. No doubt, AMAs encourage us to redraw and better understand our everyday conceptions of moral agency and patiency. However, as I argued, determining moral agency—whether natural or artificial—does not necessarily tell us enough to be able to locate responsibility. Thus, while conceptually useful, the notion of AMAs does not yet provide concrete mechanisms by which we can hold machines responsible. To help with this task, I suggested new applications of the rich pluralistic views of responsibility seen in contemporary literature. While I do not claim here to have resolved the alleged gap in responsibility, it seems to me that our existing mechanisms, if taken seriously in our increasing interactions with AI, help to show how machines might be held responsible.[62]

Acknowledgments. Earlier versions of this paper were presented at the 11th Postgraduate Bioethics Conference at the University of Oxford, the 2018 Symposium on 'Regulating Intelligence' at Newcastle University, and the 2019 Neuroethics Network in Paris. I thank the participants and organizers of these events. I'm also grateful for conversations with Niël Conradie, Olya Kudina, Saskia Nagel, Sven Nyholm, and Wendell Wallach, and for postdoctoral support from RWTH Aachen University. Finally, a great debt of gratitude is owed to David Lawrence and Sarah Morley. I've been truly inspired by and honored to contribute to their important project on *Novel Beings*.

Notes

1. For the archived story of the family's settlement, see: https://www.nytimes.com/1983/08/11/us/around-the-nation-jury-awards-10-million-in-killing-by-robot.html (last accessed 4 Dec 2019).
2. Matthias A. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 2004;**6**:175–183.
3. See Sharkey N. Saying "no!" to lethal autonomous targeting. *Journal of Military Ethics* 2010;**9**:369–383; Asaro P. On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross* 2012;**94**:687–709; Wagner M. Taking humans out of the loop: Implications for international humanitarian law. *Journal of Law, Information & Science* 2012;**21**:155–165.
4. Sparrow R. Killer robots. *Journal of Applied Philosophy* 2007;**24**:62–77.
5. See Char DS, Shah NH, Magnus D. Implementing machine learning in healthcare – addressing ethical challenges. *New England Journal of Medicine* 2018;**378**:981–983; Sharkey A. Should we welcome robot teachers? *Ethics and Information Technology* 2016;**18**:283–297; Van Wynsberghe A. Service robots, care ethics, and design. *Ethics and Information Technology* 2016;**18**:311–321; Himmelreich J. Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice* 2018;**21**:669–684.
6. Allen C, Wallach W. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press; 2009; Allen C, Wallach W. Moral machines: Contradiction in terms or abdication of human responsibility? In: Lin P, Abney K, Bekey GA, eds. *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press; 2011:55–68.
7. This distinction will be discussed below. For a helpful breakdown of moral subjects, moral agency, and morally responsible agency, see the opening chapter of McKenna M. *Conversation and Responsibility*. New York: Oxford University Press; 2012.
8. See Book III of Aristotle's *Nicomachean Ethics*.

9. See Mele A. Agency and mental action. *Philosophical Perspectives* 1997;**11**:231–249.
10. These have recently been dubbed contrastive or "instead of" reasons. See Dorsey D. Consequentialism, cognitive limitations, and moral theory. In: Timmons M, ed. *Oxford Studies in Normative Ethics***3**. Oxford: Oxford University Press; 2013:179–202; Shoemaker D. *Responsibility from the Margins*. New York: Oxford University Press; 2015.
11. That is, assuming one's desires are consistent and not overruled by second-order desires. See Frankfurt H. Freedom of the will and the concept of a person. *Journal of Philosophy* 1971;**68**:5–20.
12. See Himma K. 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* 2009;**11**:19–29.
13. See note 7, McKenna 2012, at 11.
14. This condition may be diagnosed as an antisocial personality disorder, such as psychopathy. See the American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5).* Washington, DC; 2013.
15. In this way, those who plead ignorance are attempting to eschew responsibility by dissolving their agency. The common reply—"you *should have known*"—is, then, a way of restoring agency and proceeding with blame. See Biebel N. Epistemic justification and the ignorance excuse. *Philosophical Studies* 2018;**175**:3005–3028.
16. According to recent work on implicit biases, it seems very few of us are moral agents in the robust sense outlined here. See, e.g., Doris J. *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford University Press; 2015; Levy N. Implicit bias and moral responsibility: Probing the data. *Philosophy and Phenomenological Research* 2017;**94**:3–26; Vargas M. Implicit bias, responsibility, and moral ecology. In: Shoemaker D, ed. *Oxford Studies in Agency and Responsibility* **4**. Oxford University Press; 2017.
17. Much of Frans de Waal's work supports this idea; e.g., Preston S, de Waal F. Empathy: its ultimate and proximate bases. *Behavioral and Brain Sciences* 2002;**25**:1–20.
18. See note 6, Allen and Wallach 2009, at 4.
19. In Asimov's "A Boy's Best Friend," for example, the child of a family settled on a future lunar colony cares more for his robotic canine companion than for a real-life dog. Thanks to Nathan Emmerich for the pointer.
20. Gunkel D. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. Cambridge, MA: MIT Press; 2012.
21. Here I have in mind the Kantian idea that we have indirect duties to non-human animals on the grounds that cruelty towards them translates to cruelty towards humans. See Kant's *Lectures on Ethics* 27:459. For related discussion, on our treatment of artifacts, See Parthemore J, Whitby B. 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* 2014;**6**:141–161.
22. See note 6, Allen and Wallach 2009, at 68.
23. Ibid., at 25–26. See also Nyholm S. *Humans and Robots: Ethics, Agency, and Anthropomorphism*. Rowman & Littlefield; 2020.
24. In some ways, I've so far echoed the expansion of agency seen in Floridi L, Sanders JW. On the morality of artificial agents. *Minds and Machines* 2004;**14**:349–379. Differences will emerge, however, as my focus turns to various ways of holding others *responsible*, rather than expanding *agency* to encompass artificial entities. Similarities can also be drawn to Coeckelbergh M. Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society* 2009;24:181–189. Still, my account will rely less on AMAs' appearance and more on human attitudes and interactions within the moral community.
25. See Bastone N. Google assistant now has a 'pretty please' feature to help everybody be more polite. *Business Insider* 2018 Dec 1. available at: https://www.businessinsider.co.za/google-assistant-pretty-please-now-available-2018-11 (last accessed 12 Mar 2019).

26. See, e.g., Coninx A. Towards long-term social child-robot interaction: Using multi-activity switching to engage young users. *Journal of Human-Robot Interaction* 2016;**5**:32–67.
27. For the same reasons, it may be beneficial to design some AI and robotic systems with a degree of 'social responsiveness.' See Tigard D, Conradie N, Nagel S. Socially responsive technologies: Toward a co-developmental path. *AI & Society* 2020;35:885–893.
28. Strawson PF. Freedom and resentment. *Proceedings of the British Academy* 1962;**48**:1–25.
29. Ibid., at 5.
30. For more on our "responsibility responses" and their various targets, see Shoemaker D. Qualities of will. *Social Philosophy and Policy* 2013;**30**:95–120.
31. See Tognazzini N. Blameworthiness and the affective account of blame. *Philosophia* 2013;**41**:1299–1312; also Shoemaker D. Response-dependent responsibility; or, a funny thing happened on the way to blame. *Philosophical Review* 2017;**126**:481–527.
32. See note 7, McKenna 2012; also Fricker M. What's the point of blame? A paradigm based explanation. *Noûs* 2016;**50**:165–183.
33. See note 10, Shoemaker 2015, at 19–20.
34. The advantages sketched here are persuasively harnessed by the notion of *rational sentimentalism*, notably in D'Arms J, Jacobson D. Sentiment and value. *Ethics* 2000;**110**:722–748; D'Arms J, Jacobson D. Anthropocentric constraints on human value. In: Shafer-Landau R, ed. *Oxford Studies in Metaethics* **1**. Oxford: Oxford University Press; 2006: 99–126.
35. See note 5 for similar arguments in healthcare, education, and transportation.
36. See note 4, Sparrow 2007, at 65.
37. Ibid., at 66; italics added.
38. Ibid., at 69; italics added. Comparable inconsistencies are seen in Floridi and Sanders 2004 (note 24).
39. Ibid.
40. Ibid., at 71; italics added.
41. See, e.g., Wolf M, Miller K, Grodzinsky F. Why we should have seen that coming: Comments on Microsoft's Tay "experiment" and wider implications. *ACM SIGCAS Computers and Society* 2017;**47**:54–64.
42. For discussion of promising medical uses, see Jiang F, et al. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology* 2017;**2**:230–243.
43. For medical errors, and even unavoidable harms, blame should often be *taken* by attending practitioners. See Tigard D. Taking the blame: Appropriate responses to medical error. *Journal of Medical Ethics* 2019;**45**:101–105.
44. Champagne M, Tonkens R. Bridging the responsibility gap in automated warfare. *Philosophy and Technology* 2015;**28**:125–137. See also Johnson DG. Technology with no human responsibility? *Journal of Business Ethics* 2015;**127**:707–715. My account will be consistent with Johnson's view that the "responsibility gap depends on human choices." However, while Johnson focuses on the design choices in technology itself, the choices that occupy my attention concern how and where we direct our responsibility practices. I'm grateful to an anonymous reviewer for comments here.
45. See note 4, Sparrow 2007, at 71.
46. Ibid., at 72.
47. Ibid.
48. Consider also that we punish corporations (e.g. by imposing fines) despite the implausibility of such entities displaying the right sort of response, an anonymous reviewer aptly notes. By contrast, consequential accounts of punishment can be seen as inadequate depictions of moral blame, since they don't fully explain our attitudes and might not properly distinguish wrongdoers from others. See Wallace RJ. Responsibility and the Moral Sentiments. Harvard University Press 1994; 52–62. I'm grateful to Sven Nyholm for discussion here.
49. Proponents of the 'process view' applied to technology can be said to include Johnson DG, Miller KW. Un-making artificial moral agents. *Ethics and Information Technology* 2008;**10**:123–133.

Despite some similarities to this work, my account does not fit neatly into Johnson and Miller's Computational Modelers or Computers-in-Society group.

50. See Watson G. *Agency and Answerability*. Oxford: Oxford University Press 2004; 260–288.
51. See note note 50, Watson 2004, at 274.
52. See note 10, Shoemaker 2015, at 57.
53. Exemptions are contrasted with excuses (and justifications). See, e.g., Watson 2004;224–225 (note 50).
54. See note 10, Shoemaker 2015, at 146–182.
55. However, these sorts of sanctioning mechanisms are less likely to succeed where the target AI system has surpassed humans in general intelligence. See the discussion of 'incentive methods' for controlling AI, in Bostrom N. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press 2014: 160–163.
56. Such 'bottom-up' moral development in AI is discussed in Allen and Wallach 2009 (note 6). Compare also Hellström, T. On the moral responsibility of military robots. *Ethics and Information Technology* 2013;**15**:99–107. Again, for some (e.g. Wallace 1994, in note 48), consequential accounts of responsibility will be unsatisfying. While a fuller discussion isn't possible here, in short, my goal has been to unearth general mechanisms for holding diverse objects responsible, which admittedly will deviate from the robust sorts of responsibility (and justifications) we ascribe to *natural* moral agents. Again, I'm here indebted to Sven Nyholm.
57. See, e.g., Ren F. Affective information processing and recognizing human emotion. *Electronic Notes in Theoretical Computer Science* 2009;**225**:39–50. Consider also recent work on Amazon's Alexa, e.g., in Knight W. Amazon working on making Alexa recognize your emotions. *MIT Technology Review* 2016.
58. Similarly, Helen Nissenbaum suggests that although accountability is often undermined by computing, we can and should restore it, namely by promoting an 'explicit standard of care' and imposing 'strict liability and producer responsibility.' Nissenbaum H. Computing and accountability. *Communications of the ACM* 1994;**37**:72–80; Nissenbaum H. Accountability in a computerized society. *Science and Engineering Ethics* 1996;**2**:25–42. I thank an anonymous reviewer for connecting my account with Nissenbaum's early work.
59. See Smith PT. Just research into killer robots. *Ethics and Information Technology* 2019;**21**:281–293.
60. See Combs TS, et al. Automated vehicles and pedestrian safety: exploring the promise and limits of pedestrian detection. *American Journal of Preventive Medicine* 2019;**56**:1–7.
61. John Danaher likewise frames the problem in terms of trade-offs, namely increases in efficiency and perhaps well-being, but at the cost of human participation and comprehension. See Danaher J. The threat of algocracy: reality, resistance and accommodation. *Philosophy and Technology* 2016;**29**:245–268; also Danaher J. Robots, law and the retribution gap. *Ethics and Information Technology* 2016;**18**:299–309.
62. In a follow-up paper, I explain further how pluralistic conceptions of responsibility can address the alleged gap created by emerging technologies. See Tigard D. There is no techno-responsibility gap. *Philosophy and Technology* 2020; available at: https://doi.org/10.1007/s13347-020-00414-7.

Cite this article: Tigard, D. W. 2021. Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible. *Cambridge Quarterly of Healthcare Ethics* 30: 435–447, doi:10.1017/S0963180120000985