Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement

TRYSTAN S. GOETZE, Harvard University, USA

When a computer system causes harm, who is responsible? This question has renewed significance given the proliferation of autonomous systems enabled by modern artificial intelligence techniques. At the root of this problem is a philosophical difficulty known in the literature as the responsibility gap. That is to say, because of the causal distance between the designers of autonomous systems and the eventual outcomes of those systems, the dilution of agency within the large and complex teams that design autonomous systems, and the impossibility of fully predicting how autonomous systems will behave once deployed, determining who is morally responsible for harms caused by autonomous systems is unclear at a conceptual level. I review past work on this topic, criticizing prior works for suggesting workarounds rather than philosophical answers to the conceptual problem presented by the responsibility gap. The view I develop, drawing on recent work on vicarious moral responsibility, explains why computing professionals are ethically required to take responsibility for the systems they design, despite not being blameworthy for the harms these systems may cause.

CCS Concepts: • **Social and professional topics** → *Socio-technical systems*; *Computing / technology policy*; **Computing profession**; *Codes of ethics*; • **Computing methodologies** → Artificial intelligence; Machine learning.

Additional Key Words and Phrases: moral responsibility, professional responsibility, autonomous systems, lethal autonomous weapons systems (LAWS), ethics of artificial intelligence, computer ethics, accountability

ACM Reference Format:

Trystan S. Goetze. 2022. Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement. In *FAccT '22: ACM Conference on Fairness, Accountability, and Transparency 2022, June 21–24, 2022, Seoul, South Korea, and Online.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/XXXXX

1 INTRODUCTION

Tech news in recent years has been marked by vacillation between, on the one hand, effusive praise for new innovations in artificial intelligence and the promise of these technologies for ushering in a new era of prosperity — or, at least, profit — and, on the other hand, criticism of technologists for failing to anticipate, mitigate, and properly respond to the harms caused by these same technologies. More than just a matter of perspective, this mix of hope and criticism is a concrete illustration of the fact that responsibility is a double-edged sword. That is to say, the same capacities that enable one to be praiseworthy for the *good* one brings about in the world also open one to being blameworthy for the *harms* one brings about.

However, despite early attempts in computer ethics to resolve the issue, it remains genuinely unclear whether computing professionals are morally responsible for the behaviour of the systems they design, deploy, and monitor. The long and complex causal chain between the computing professional and the actual behaviour of an autonomous system, the dilution of responsibility within large and complicated organizations, the absence of direct human control over how an autonomous system behaves once deployed, and the difficulty — perhaps impossibility — of predicting how an

Manuscript submitted to ACM

autonomous system will behave once deployed combine to create what is known as the *responsibility gap* between computing professionals and autonomous computer systems. Unless this gap can be bridged at the level of our concept of moral responsibility, it will be unclear whether technologists deserve any of the praise *or* blame for their innovations. And as long as this unclarity remains, computing professionals may exploit this conceptual ambiguity to accept the praise while deflecting calls for them to make right what their technologies have wrought.

In this paper, I revisit the responsibility gap in computing. While my central examples are of autonomous, AI-enabled systems, I take my arguments to apply to many other computer systems as well. In §2, I illustrate the conceptual problem posed by the responsibility gap, using the example of lethal autonomous weapons systems. Fundamentally, the issue is that it is unclear who, if anyone, is responsible when autonomous systems cause harm. In §3, I develop a more detailed theoretical account of the responsibility gap, explaining the relevant senses of the term *responsibility* and how autonomous systems disrupt our usual process of determining who to hold accountable for harms. In §4, I review and criticize four proposed solutions to the responsibility gap: develop a new theory of moral responsibility; blame the autonomous systems themselves; make the systems capable of moral deliberation; or establish professional and legal frameworks for liability. Each of these proposals, I argue, is merely a workaround; none actually solves the conceptual problem of the responsibility gap, and all smack of arbitrariness.

I turn next to my positive proposal. Using a recent account of vicarious moral responsibility — that is, of cases where one agent is responsible, in some sense, for the behaviour of another — I argue that we can accept that no human beings are *blameworthy* for the harms caused by autonomous systems, while at the same time affirming that computing professionals have distinctive moral obligations to *take responsibility* for these harms, by dint of the special connection between their agency and the pseudo-agency of the autonomous systems they design, deploy, and monitor. In particular, computing professionals have obligations to make amends for the harms caused by autonomous systems, and to help those affected make sense of these events. This solution has the advantage of being bottom-up rather than top-down: it derives from a set of existing — albeit unclear — moral norms, instead of a new framework of norms imposed by a professional or legal authority. §5 develops the account of vicarious responsibility; §6 applies it to the case of autonomous systems, again illustrating with lethal autonomous weapons. §7 concludes.

2 THE RESPONSIBILITY GAP AND THE CASE OF LETHAL AUTONOMOUS WEAPONS

In recent short documentary, *A.I. is Making it Easier to Kill (You)*, journalists for the *The New York Times* present some of the worries surrounding the development of lethal autonomous weapons systems (LAWS) by militaries across the world [19]. In the documentary, security policy analyst and former soldier Paul Scharre describes a harrowing event when he served as a sniper team leader in Afghanistan. He and his team had tracked a group of Taliban fighters to a compound near the border with Pakistan, and were monitoring their movements. During their stake-out, they noticed that the Taliban had sent out a small girl as a scout — and she had spotted their position. They watched as she radioed the Taliban commander, and were forced to flee when the militants opened fire.

In the interview, Scharre remarks that no one in his team, either in the moment or at the mission debriefing, ever suggested shooting the girl to prevent her from giving away their position. To have done so, Scharre thinks, would have been immoral. And yet, he says, it would have been perfectly *legal*: international law would have defined her as an enemy combatant, and a thus a legitimate target. Scharre worries that an autonomous weapon wouldn't make a distinction between killing that is *legally* permissible versus killing that is *morally* permissible. LAWS, typically conceived, would be designed to obey the laws of war, not the vague and difficult to apply *ethical* norms that also guide human soldiers' decision-making.

Suppose we agree with Scharre that killing the girl would have been immoral. And suppose also that instead of Scharre and his team, the coalition presence on that day had been a squadron of LAWS — a group of autonomous armed aerial drones, for example — which had determined that the girl was a threat and fired on her. Who should answer for this legal but immoral killing? Or, suppose instead that the drones determine, erroneously, that a mountain village is an insurgent compound, and fire on an innocent girl talking to a friend on a walkie-talkie. Who should answer for this war crime?

According to Michael Horowitz, what is unique about LAWS is that "the weapon system, not a person, selects and engages targets" [13, p. 26]. This creates a problem for determining who to hold responsible when LAWS cause morally unjustifiable harm. Horowitz describes the issue in terms of what the human beings involved might reasonably predict:

> The responsible party could be the programmer, but what if the programmer never imagined that particular situation? The responsible party could be the commander who ordered the activation of the weapon, but what if the weapon behaved in a way that the commander could not have reasonably predicted? [13, p. 30]

Here, Horowitz is pointing to an intuitive account of moral responsibility: namely, that in order to be responsible for an outcome, you must have *intended* to bring about that outcome, or, if not, you *should have known* that the outcome was a reasonably predictable (even if not highly probable) result of your actions.

This condition creates a *gap* in responsibility with regard to LAWS. That is to say, it isn't clear who should be held responsible for the immoral harms caused by LAWS. Could we hold the autonomous system itself responsible? Possibly, but we may think that unless the system has at least the same level of intelligence, self-awareness, and moral sensibility as an adult human being, holding a computer responsible for harm is pointless theatre — like Xerxes whipping the sea as punishment for inclement weather [12, book VII, chs. 34–35]. Could we hold the designers of the system responsible? Certainly we could, but the system designers might protest that they did not intend to cause immoral harm, passing the buck to the politicians who authorized the purchase of the system, the officers who ordered its use, or the soldiers who activated it — and any of *them* could invoke the same argument. We are left with a lacuna where a responsible party should be: this is the responsibility gap.

3 THE NATURE OF THE GAP

The responsibility gap is not an issue unique to LAWS; indeed, it is common to all autonomous systems and to many computer systems that do not depend on AI. As such, it has been a perennial topic in computer ethics. In this and the following section, I clarify the nature of the problem and review these prior works. My overall purpose here is to show that prior work has concentrated on finding ways to work *around* the responsibility gap, rather than bridging it in with a philosophical solution. The first subsection clarifies the theoretical details regarding moral responsibility itself; the second subsection provides more detail on the nature of the responsibility gap.

3.1 Moral Responsibility

First, let's get a bit more precise about the nature of the responsibility gap. To do so, we need to have an account of moral responsibility in view. Unhelpfully, much of the literature on the responsibility gap conflates, equivocates over, or runs together multiple senses of the term *responsibility* and derived terms, so our first task is to disambiguate the senses of the term that are at issue.

In the philosophical literature, there are at least ten different concepts of responsibility that have been identified (see [4, 7, 9, 10, 32, 35] for details on some of these distinctions):

(1) **Causal Responsibility (What caused this?)**: The agent's behaviour caused some morally significant outcome, either directly (their actions led immediately to the event) or proximally (their actions were among the most significant factors that produced the event).
(2) **Attributability (Who did this?)**: An event caused by the agent's behaviour can be attributed to their actions, which must be connected to their psychology in an appropriate way (e.g., via their intentions, motives, desires, or values).
(3) **Accountability (Who's to blame?)**: The agent may be held responsible via blame or punishment for some morally bad event, typically because the event is attributable to the agent.
(4) **Answerability (How do you justify this?)**: The agent may be asked to provide a justification of why they did something, by explaining their moral reasons for doing so, typically because the action or its outcomes are attributable to the agent.
(5) **Duty of Performance (Whose role is it to do this?)**: In virtue of the agent's professional or social role, they have a duty to perform certain actions.
(6) **Duty of Compensation (Who's gonna make things right?)**: The agent has an ethical duty to make amends for some harm, typically because the action is attributable to the agent.
(7) **Duty of Prevention (Who should have stopped this from happening?)**: The agent has an ethical duty to anticipate negative consequences and to take action to prevent them.
(8) **Virtuous Responsibility (What does it mean to be a *responsible person* or a *responsible professional*?)**: The agent has a robust ethical disposition to act with due care and a willingness to accept responsibility for harms that occur in their purview.
(9) **Hermeneutic Responsibility (How do we make sense of this?)**: The agent must produce a narrative that helps them or someone else to understand an event in terms not just of its causes, but also its social context and the significance the event holds in their life.
(10) **Vicarious Responsibility (Who's responsible for what that other person did?)**: The agent is responsible, in some sense, for the behaviour or actions of another.

Each of these senses is important to consider when it comes to autonomous systems. However, when it comes to the responsibility gap, it is senses (1), (2), and (3) that seem most relevant (though we shall see in §§5–6 that other senses, particularly (6), (9), and (10), are key to closing the responsibility gap). For in ordinary human action, we tend to say that a person A is morally responsible for a harmful event X only if A caused X, X was properly attributable to A, and A is a proper target of blame for X. And it is the satisfaction of these conditions that is made difficult to determine when autonomous systems replace human action or decision-making. Let's discuss each condition in turn.

When determining causal responsibility, we usually look to the agent or process whose actions or behaviour caused the event. For example, suppose X is someone's death. When we investigate, we find that the cause of death was a gunshot, and A fired the gun. Determining causal responsibility can be complicated, however: while the most immediate step in the causal chain might be that the bullet entered the victim's body, from an ethical point of view we aren't interested in these fine-grained details of the story. Rather, we look for the agent(s) whose actions precipitated the causal chain leading to the event in question. Because A's action — pulling the trigger — is the most significant part of the causal chain leading to the death, we say that A is causally responsible for X.

Causal responsibility is typically considered a necessary but not sufficient condition for an event to be attributable to the agent. The remaining conditions for attribution require some unpacking, and are subject to much debate. The following is representative of a wide range of views in contemporary analytic moral philosophy. In general, there are two conditions for responsibility attribution.

The first condition is the *control* condition: A must have been, in some sense, in control of whether or not X happened. Specifying the psychology of control is difficult. As mentioned above, one intuitive approach is to require that A *intended* to cause X. Another, developed by John Fischer and Mark Ravizza, holds that A's behaviour must have been caused by psychological mechanisms that are A's own, and which are responsive to moral reasons [5]. On yet another approach, propounded by Harry Frankfurt, A must have caused X through actions that A desired to take, and those desires must align with the sort of person A wishes to be — in other words, A must *want* to have had the desires that issued in the actions that caused X [6].

The second condition is the *epistemic* condition: A must have *known* that X would (or could) result from the actions that A took — or, if A did not know that X was a potential outcome of their actions, it must be true that A *should have known* this. The latter half of the epistemic condition — covering cases wherein A is said to be *culpably ignorant* — is the more challenging to theorize. Michael Zimmerman argues that in order to be culpably ignorant in causing X, A must have committed a prior wrongful act, knowing that it was wrong, to produce their ignorance that causing X would be wrong [36]. In other words, on this view, all responsibility for wrongdoing must trace to a knowingly wrongful act. By contrast, George Sher argues that the agent could be morally responsible despite never being aware of acting wrongfully, provided that their ignorance was caused by the combination of psychological traits (e.g. beliefs, desires, and dispositions) that constitute the person the agent is [31].

If the above conditions are satisfied — A caused X, A was in control of whether X occurred, and A knew or should have known that causing X would be wrong — then X is properly attributable to A. Furthermore, most theories of moral responsibility hold that when these conditions hold, A is also blameworthy for X. This means that A is accountable for X: it becomes appropriate to hold A responsible for X by feeling resentment towards A, reprimanding A, or imposing other sanctions on A. While feeling blame and imposing other penalties serve the purposes of emotional catharsis and retribution, they also have two further, arguably more important functions. The first is to communicate to the wrongdoer that they behaved badly, with the aim of bringing them to acknowledge the moral reasons that they ignored or flouted. The second is to spur the wrongdoer to do better in similar circumstances in the future. Typically, this is accomplished by making the wrongdoer feel bad for doing something wrong — it is unpleasant to be resented, rebuked, and punished — though there are other, less harsh ways of holding people accountable for their actions.

To summarize: When we attempt to determine who is responsible for some harmful event, we look for the coincidence of causal responsibility and attributability. When these conditions are met for some agent, we may hold that agent accountable for the harms caused.

3.2 The Gap

As we saw in the case of LAWS, the responsibility gap arises where autonomous systems take the place of human action or decision-making. When we attempt to determine who is responsible for harms that result, our usual process, described in the previous subsection, runs into difficulties. Let's run through those steps again, in the case of an autonomous system that causes harm.

With regard to causal responsibility, the most significant cause of the harmful event is the autonomous system itself. Through some automated process, the system determines what course of action to take without human intervention. For example, the drone determines that the girl is a threat, and opens fire.

But when we turn to attributability, things begin to break down. Consider the control condition. While the autonomous system is in some sense in control of the outcome, the way in which it exercises this control is quite different from the human case. In particular, the autonomous system does not have any of the psychological structures that moral philosophers take to be necessary for the relevant conception of control. Autonomous systems do not have intentions. They do not have psychological mechanisms that are responsive to moral reasons. They do not have desires, much less higher-order desires about the kind of person they wish to be. Similar remarks apply to the epistemic condition: while the autonomous system processes information about its environment, it would be controversial to say that it is capable of knowing what it is doing, or knowing that to take some action would be wrong.

Taken together, these observations tell us that — as the technology currently exists — an autonomous system is *incapable* of responsibility in the sense required for attributability. It simply doesn't have the right kind of mind, if it has a mind at all. As John Ladd puts the point, "The special responsibility problems raised by computers are due...to the fact that they are used to replace minds, or brains, which...are the source of human responsibility" [21, p. 219].

Finally, even if we decide to ignore these problems with responsibility attribution, we run into another stumbling block with accountability. For, what use would it be to blame the autonomous system for the harm it caused? While we might get some emotional catharsis from this — as when we yell at a printer for jamming or whip the ocean for bad weather — that is all that such blame would accomplish. The autonomous system has no sensitivity to moral reasons, and no capacity to feel bad for wrongdoing. Blame without a responsible subject is merely shouting into the void.

At this point, we might naturally wish to bring human responsibility back into the picture, by finding *someone* to take the blame for the autonomous system's behaviour. Moving back to the step of determining causal responsibility, we find several potential candidates: the programmers and data scientists who created the system, the managers who ordered the system's use, the lower-level employees who activated it, and so on. But because of the complexity of this causal chain, it would be controversial, in many cases, to identify any single individual or group as causally responsible for the specific harm caused by the autonomous system. At every step, decisions are likely made not by single individuals, but by teams or members of group agents. The causal responsibility thus is diluted across many different agents, such that assigning it to any subset of them is difficult to justify. And while we may be tempted to pin the causal responsibility on the origin of the autonomous system, the programmers and data scientists who created it are so far up the chain that it would be just as controversial to pin the responsibility on them.

But suppose we overcome this difficulty and identify a causally responsible party, be they the programmers, the managers, the employees, or someone else. Can we make a judgement of moral responsibility attribution? Again, we run into trouble. Take the control condition. To the extent that human beings have control over autonomous systems, much of it is exercised at a level that is, again, causally distant from the system's actual behaviour, and difficult to trace. As Andreas Matthias argues, because many machine learning techniques have the computer do much of the programming, control often leaves human hands well before the system is deployed [22]. Similarly, while the managers or employees who set up and activate the system have control over *these* actions, they may not have have control over how the system behaves thereafter. While it is true that they could pull the plug to prevent the system from causing harm, it is likely that they would become aware of the situation too late. Furthermore, the complex, collaborative nature of both modern computer systems design and modern organizational structures complicates any attempt to trace the harmful behaviour of an autonomous system to the controlled actions of any one human individual or group.

Next, consider the epistemic condition. We have already seen that Horowitz raises the possibility that the human beings who design, authorize, or activate autonomous systems might not be in a position to know what specific harms may come from their behaviour [13]. Likewise, Matthias argues that because of the "black box" effect of artificial neural networks, reinforcement learning techniques, and genetic programming methods, the designers of an autonomous system may not be in a position to predict how the system might respond to any particular situation [22]. Often, the only way to know how an autonomous system will behave in some situation is to perform rigorous testing. Even then, since real-world circumstances often introduce new complications and end users often configure or deploy systems in ways not anticipated by the designers, the system may act in unexpected ways once deployed.

Given these difficulties in making causal responsibility or responsibility attribution stick to a human being when an autonomous system causes harm, we reach the conclusion that there is no one to take the blame. No human subject is (uncontroversially) blameworthy for the harm — there is no one to hold accountable. As Helen Nissenbaum observes, "If we apply standard conceptions of accountability to identify who should step forward and answer for the injuries, we see an intricate web of causes and decisions," with no clear way to identify a single, overall responsible individual at any particular node within that web [27, p. 76]. Thus, we are left with a gap where a responsible party should be.

Before moving on, it's worth briefly acknowledging the other edge of the sword of responsibility, namely, praiseworthiness. While I am mainly concerned with blame and harm in this paper, as mentioned earlier, I am also interested in explaining why it might be appropriate for computing professionals to accept *praise* for the successes of autonomous systems and the *goods* they produce. But, if the foregoing is correct, then we face exactly the same problem legitimating such praise as we do in determining who to blame. For it is a common assumption in moral philosophy that praise and blame are two ways of expressing the judgement that someone is responsible for some outcome — the difference lies in the moral evaluation that comes along with that judgement, namely, moral approval or disapproval. If computing professionals are appropriate targets of praise for when autonomous systems do good and appropriate targets of blame for when autonomous systems cause harm, then they must be responsible for those good and bad outcomes equally. But it is precisely this latter judgement that the responsibility gap puts in question.

4 PROPOSED SOLUTIONS TO THE RESPONSIBILITY GAP

In the last section, I provided some theoretical detail to substantiate the responsibility gap. In this section, I review four different solutions that have been considered in the existing literature: develop a new theory of moral responsibility; blame the autonomous systems themselves; make the systems capable of moral deliberation by advancing the state of machine ethics; and establish professional and legal frameworks for liability.

Below, I criticize each of these solutions in turn. But a general criticism applies to them all, namely, that every one of these solutions in some way introduces new ethical practices, instead of working from within our pre-existing moral intuitions. On the one hand, we might think that the responsibility gap is such a new and unique problem that it requires a degree of arbitariness in its solution, so long as the solution can be justified. But, as I will argue in the next section, there is another solution that is preferable precisely because it avoids this sense of arbitariness.

4.1 Rethink Moral Responsibility

One potential solution would be to devise a new theory of moral responsibility that avoids the difficulties posed by the causal, control, and epistemic conditions. For example, one attempt to do away with the control condition in order to bridge the responsibility gap is proposed by Matteo Santoro, Dante Marino, and Guglielmo Tamburrini [29]. They

argue that there are many situations where responsibility attribution becomes fraught, and control over the outcome falls away as a necessary condition.

However, in making this move, Santoro et al. explicitly shift from moral responsibility to legal liability: "The crucial move here is to acknowledge the distinction between liability or objective responsibility on the hand, and moral responsibility on the other hand" [29, p. 310]. There are two problems with this approach. Firstly, it simply isn't clear, despite decades of discussion in jurisprudence, whether software developers are or should be held liable for harms caused by their products [20, 25, 34]. This approach would thus require a significant change in legislation and in judicial practice. Secondly, this strategy substitutes a legal solution where a conceptual solution was called for. The responsibility gap is, fundamentally, about *moral* responsibility, and a legal solution, whether or not it is morally desirable, fails to address the root of the problem. While legal norms should align with moral norms, they are not the same domain.

More to the point, before considering a theory of moral responsibility that does away with the control condition, it is important to note that unless such a revisionist theory can be independently motivated, this attempt to bridge the responsibility gap will seem suspicious. Why change how we think about responsibility *in general* for the sake of solving one specific problem? We need good reason to think that a new theory would still make sense of typical cases of moral responsibility, and, ideally, that it also helps to solve other problems that traditional theories cannot.

Accounts of moral responsibility that eschew the control condition do exist in the philosophical literature. One, developed by Robert Adams [1], is motivated by the need to explain why we sometimes consider people to be blameworthy for their involuntary attitudes and emotional states. For example, sometimes we blame others or ourselves for getting angry when they ought not to, even though one cannot choose to be angry the same way one can choose to, say, raise one's hands. On Adams's view, causal responsibility for moral badness suffices for us to judge the agent blameworthy.

We might be able to develop Adams's account to explain why computing professionals or some other human agent is responsible for harms caused by autonomous systems. However, making such a case would still require substantial work. For one, Adams's view does not yet address the causal complexities discussed above with regard to autonomous systems. Moreover, accounts such as Adams's are controversial, as they conflict with widely held intuitions about the connection between control and responsibility attribution. It remains to be seen whether a non-voluntarist account of moral responsibility can overcome these difficulties.

4.2 Blame the Computer

Because the responsibility gap is fundamentally a problem of whom to hold responsible for the harms caused by the behaviour of autonomous systems, it would be convenient if we could simply hold the systems *themselves* responsible for their harmful behaviour. Above, I suggested that this would be pointless, as computer systems lack the required psychology for blame to make any difference to how they behave. But perhaps this was too quick.

Thomas Hellström, for example, argues that LAWS have sufficient autonomy that people are inclined to think of them as morally responsible for their behaviour. Additionally, Hellström suggests that systems that are (re)trained using reinforcement learning techniques might be sensitive to something like praise or blame for their behaviour, meaning that there would be a sense in which these systems could be held responsible for what they do, in a way that is functionally similar to how we hold human beings responsible [11].

Again, however, this solution amounts to changing the subject. When one retrains a machine learning model by associating a penalty with the harmful actions it took the last time it was deployed, one isn't blaming anything or holding anyone responsible. Rather, retraining a model is more akin to retraining an animal with dangerous impulses, such as a poorly raised dog, to resist or to lose these impulses. In the words of Peter Strawson, this way of responding

to an entity that causes harm involves taking the *objective* attitude towards the autonomous system — and this attitude is fundamentally incompatible with the practice of holding someone or something morally responsible, which requires that we treat them as a member of the moral community [33]. We would still be left with a responsibility gap when it comes to the actual participants in the moral community — one feels there is some ethical residue of unclaimed responsibility, that someone is dodging blame. As as Deborah Johnson puts it, "computer systems cannot *by themselves* be moral agents" [16, p. 203].

4.3 Machine Ethics

Perhaps we could overcome the problem just raised, that autonomous systems are not the sort of entity that can be held responsible for their behaviour, by designing the system such that it has sufficient moral understanding. After all, being *human* is neither necessary nor sufficient to be a participant in the moral community. Rather, what is necessary is that the entity have the same capacities for moral agency that most human beings possess.

The effort to include ethical decision-making into autonomous systems themselves is known as *machine ethics*. There are several suggested approaches to accomplishing this. James Moor distinguishes between *implicit* ethical agents, *explicit* ethical agents, and *full* ethical agents [26].

Implicitly ethical computer systems are those which "constrain the machine's actions to avoid unethical outcomes" as a matter of the system's design [26, p. 19]. The "choice" to do something unethical is eliminated from the autonomous system's decision space, because it simply does not have the ability to take those actions. For example, Horowitz describes a possible type of LAWS that would devise tactical plans to advise field officers and issue orders to unit commanders [13]. Such a system might be set up so that it always provides factual and complete answers to a human operator when queried about its tactical plans, even if leaving out some details might be better for its long-term strategic goals. A system like this would simply not have the ability to "lie," and as such the ethical decision of whether or not to lie never comes up.

Explicitly ethical computer systems "would be able to make plausible ethical judgments and justify them" [8, p. 20]. These systems have an internal representation of ethical rules that enable them to evaluate the moral worth of one behaviour over another. They contrast with implicit ethical agents in that unethical behaviour is within their power to do, but in practice the behaviour they choose is subject to a moral processing stage where they evaluate potential courses of action against a formally defined moral framework. The system's ethical rules might take inspiration from a number of different ethical theories, such as consequentialism [23], deontology [18], contractualism [30], or some combination of these.

A full ethical agent would be a computer system possessed of all the same moral capacities as a human agent, replicated in silicon. It would not only have the capacity to evaluate its behaviour against a hard-coded moral framework; it would also be able to evaluate the framework itself. It would have autonomy in the Kantian sense, the ability to reason its own way to general principles of right and wrong, and to choose to act in accordance with these moral principles for their own sake. Moreover, it would have sufficient self-awareness to understand that it is the source of its own actions.

It is possible that implicitly or explicitly ethical autonomous systems would produce more morally desirable outcomes. But these approaches do little to address the responsibility gap. An implicitly ethical system is simply following the rules of its programming. From the perspective of moral responsibility, it is no different from an autonomous system that has no specific rules designed around ethical values. Explicit ethical agents are not much different, except that some of their programming is intended to produce ethically desirable results. What is missing in each case is the reflective understanding that is necessary to have the right sort of moral psychology to be a moral agent. Only a full ethical

agent can achieve this standard, and AI systems are a long way off from having the required critical self-awareness and meta-cognitive abilities. The only way to become a full ethical agent is to become a fully autonomous, generally intelligent agent.

4.4 Professional Frameworks

We saw above that Santoro et al. shifted from thinking about moral responsibility to legal liability. There, I criticized this strategy for substituting legal norms where we needed a solution in terms of our concepts of moral responsibility. But could we use some other formal mechanism to address the responsibility gap, by justifying why a computing professional ought to receive the blame for harmful autonomous systems? Several authors have suggested that we might do so by way of professional codes of conduct in computing.

One attempt is outlined by Donald Gotterbarn. He finds fault with all attempts to hold computers responsible for harmful outcomes rather than human beings, calling this a strategy for "dodging" or "side-stepping" responsibility [8]. On his view, computing professionals should assume responsibility for the harmful behaviour of the systems they design, deploy, maintain, and monitor. Following Ladd [21], he calls this set of professional duties *positive responsibilities*, to distinguish them from the "negative" responsibility of blameworthiness.

Similar proposals include the "five rules" of moral responsibility for computing artefacts propounded by Keith Miller [24], Johnson's argument that every computing professional in the complex causal chain between the creation and deployment of a harmful computer system should share some blame for the harm [16], and Nissenbaum's argument that legal and professional frameworks should hold computing professionals morally accountable for harms caused by their technologies [27].

Each of these works suggests that professional and legal standards must bridge the responsibility gap by creating clear rules for determining who is responsible for the behaviour of computer systems. The focus is on holding human beings to account when autonomous systems cause harm, as well as imposing professional duties on computing professionals to design, deploy, maintain, and monitor such systems with care. By inculcating the expectation of ethical design into the work of the computing professions, and enforcing this expectation through professional censure or legal penalties, the aim is to ensure that the buck always stops at a human being.

Why pass the responsibility to a computing professional, if, as we saw, the causal chain and organizational structures involved are complex enough to dilute their moral responsibility for harms caused by autonomous systems? The thinking is that, of all the people involved in the creation, deployment, and maintenance of autonomous systems, computing professionals are those in the best position to ensure that these systems are designed with ethically desirable outcomes in mind, to evaluate whether systems are fit-for-purpose from an ethical standpoint, and to monitor their performance for unethical outcomes, putting a stop to their use if need be. While in some circumstances there might be human agents who are better candidates for those who should be held responsible for the harms of autonomous systems — such as a senior leader in an organization who insists on deploying an untested and unreliable AI system despite the warnings of technical employees — as a general rule, it is reasonable to assign the responsibility to those with the computing expertise.

One potential challenge to the professional standards approach is the fact that the professionalization of the various computing specializations remains a work in progress. As Johnson and Miller observe, well-established professions, such as medicine, law, or education, are *strongly differentiated* from other sectors of society [17]. A strongly differentiated profession is marked by stricter social, legal, and ethical requirements, typically because of the heightened moral risk of entrusting ourselves to the services of these professionals. For our purposes, the most relevant aspects of these

professions is that they have fundamental values that are shared across the profession, which are expressed in a code of ethics, adherence to which can be enforced by expulsion from the profession. While the computing professions have several influential codes of conduct [2, 3, 14], the enforcement of these codes is less effective than in strongly differentiated professions. It is rare for a professional association in computing to censure one of their members, and when they do, while the censured party may be barred from membership in the association, this punishment does not prevent them from practising in the field — as a finding of malpractice in medicine, law, or education would.

Furthermore, while I think the above motivation for holding computing professionals accountable for harmful autonomous systems is on the right track, and while it is surely to the good to create a culture of professional duty and responsibility in computing, this approach still fails to bridge the responsibility gap. It remains unclear whether computing professionals actually *are* morally responsible for the behaviour of autonomous systems they have a hand in creating or maintaining. Adding professional standards would help insofar as they would ensure that *someone* is held accountable when these systems cause harm, but it might yet seem unfair that computing professionals should be the ones who bear the brunt of this regulatory apparatus, given that we lack an account of why they should be held responsible. Indeed, we might consider the noticeable lack of regulation and the toothlessness of professional standards in this area, despite decades of campaigning by scholars in computer ethics, to be a sign that computing professionals have *not* accepted that they should be held accountable for these outcomes. One contributing factor may be that these proposals go *around* the responsibility gap, acknowledging but not solving the problem at the conceptual level. What we need, rather, is something to bridge it.

5 VICARIOUS RESPONSIBILITY AND MORAL ENTANGLEMENT

How can we bridge the responsibility gap, given the difficulties discussed? I suggest that we should turn to a *different* sense of responsibility that we already acknowledge in our everyday life, even if it remains under-discussed in the philosophical literature. In particular, I contend that the notion of *vicarious responsibility* can help explain why we have the intuition that computing professionals and other agents who have some significant connection to the autonomous system that causes harm ought to take responsibility for that harm. At the same time, this account can explain why it is legitimate for computing professionals to take some of the credit for the *goods* produced by autonomous systems.

As described above, *vicarious responsibility* concerns cases where one agent is responsible, in some sense, for the actions or behaviour of another. There are some cases where one might be vicariously responsible in the accountability sense for another's actions, as when one issues a direct order to a subordinate, which, when carried out, cause harm. But this is not the situation we find with regard to autonomous systems. As already discussed, the causal chain is much more complex. We must, therefore, look to a different sense of moral responsibility to fill in the details of vicarious responsibility.

Picking up on the thread introduced by the discussion of professional standards in computing, we can notice that that discussion at times shifts the focus from moral accountability — who to blame — to ethical duties demanded of computing professionals. Recall, in the list of responsibility-concepts introduced in §3.1, there were three forms of duty that we identified as attached to the notion of someone's "having a responsibility." Gotterbarn seems most concerned with duties of performance — "In accordance with the malpractice model, a [computing practitioner] has a responsibility to conform to good standards and operating procedures of the profession" — and duties of prevention — "The professional commits to a 'higher degree of care' for those affected by the computing product" [8, p. 229].

By contrast, the kinds of responsibility that I am interested in are what I called the duty of compensation — given that some harm has occurred, who has the responsibility (that is to say, the duty) to make things right? — and what

Coeckelbergh calls hermeneutic responsibility [4] — given that something awful has happened, who do we turn to for help making sense of the event and its place in our lives? My reason for this focus is that it brings us closer to the accountability sense of responsibility. One of the things we are doing when we hold a wrongdoer to account is to look for a way to redress the harms caused. Another thing that we do when we hold a wrongdoer accountable is also captured by the answerability sense of responsibility: we ask the wrongdoer to explain the reasons why they took the actions that produced the harmful outcome. My question is: can these duties — to make amends for and to make sense of the harmful event — arise in cases of vicarious responsibility?

Cases where we expect someone to offer some form of amends or to offer an explanation for wrongdoing caused by another sometimes arise with respect to the actions of our family members or friends. Consider a case of family drama: imagine a holiday gathering with your in-laws, and an elderly member of your spouse's family says something offensive about a social group to which you belong. It would be reasonable to anticipate that your spouse — or at least, someone else in their family — would offer an apology on behalf of the offensive elder, and perhaps to explain why the elder is the way they are — e.g., they were raised badly or in a different time, or their religion has fringe beliefs. While these explanations are often taken as if they are aimed at excusing the relative from accountability for their offensive utterances, this interpretation is open to criticism. While an explanation goes some way to helping us understand why they would say such a thing, they do not give a reason to think that the elder is not in a position to know better.

A better way to understand these awkward apologies and explanations is the following. The spouse who offers the apology and explanation isn't offering excuses on behalf of their relative. Rather, they are personally *taking responsibility* for the harm caused by the offensive elder. In particular, they are taking on the duties of compensation and hermeneutic responsibility that the elder would presumably refuse to accept were they held accountable for their behaviour. The spouse is trying to make things right and to make sense of the event for their harmed partner — not to question whether the offensive elder is blameworthy for their bigoted remarks. In other words, they are making themselves vicariously responsible for the harm caused by their relative, by fulfilling unmet duties.

Similar cases can be recognized when our friends, coworkers, and others with whom we share some close and morally significant association cause harm. In a recent paper on vicarious responsibility, Trystan Goetze recognizes these relationships as forms of *moral entanglement*: these relationships are "circumstantially morally salient aspects of our identities" which "generate obligations to apologize, explain, or regret" harms caused by another person with whom we share such connections [7, pp. 219–20]. In addition to these standing aspects of our identities that create moral obligations to assume vicarious responsibility for those to whom we have close relations, Goetze also acknowledges a kind of moral entanglement that can occur because of a connection between our agency and that of another, "particularly when our own activities contribute to those others' behaviour" [7, p. 220] — consider, here, the case of ordering a subordinate to cause some harm.

The concepts of moral entanglement and of vicarious responsibility, when connected to the senses of responsibility as duty of compensation and hermeneutic responsibility, thus explain why we sometimes look to people other than the wrongdoer for actions that can make amends for the harm caused, and explanations that can help make sense of why the harms occurred. We may judge that, due to their close connection to the wrongdoer, they have a duty to offer such a response. At minimum, it would be appropriate for them to take on such a duty, but not for someone without that relationship to the wrongdoer. In the family drama case, if a stranger overheard the offensive remarks and butted in to apologize and to try to explain the offensive elder's behaviour, we would see this as inappropriate and nosy.

Finally, moral entanglement can also account for why we sometimes *praise* people for the actions and accomplishments of others. Just as we often look to those who have a close relationship to wrongdoers for amends, apology, or explanation,

so too do we sometimes look to those who have close relationships with moral exemplars to offer congratulations and to ask for their take on how this praiseworthy individual became so admirable. Consider cases where some heroic behaviour is reported on in the news, and journalists interview not just the hero in question, but also their close friends and family members. These conversations frequently look to give the audience a satisfying narrative explaining the life of the hero of the hour and how it led to their praiseworthy act, as well as offering the hero's close relations a kind of admiration of their own.

6 CLOSING THE GAP: TAKING RESPONSIBILITY FOR AUTONOMOUS SYSTEMS

Let's now apply the framework of vicarious responsibility just sketched to the case of autonomous systems that cause harm.

We have seen in previous sections that to make a straightforward judgement that a computing professional is morally responsible for the harmful behaviour of an autonomous system, even one that they designed, is fraught. The responsibility gap ensures that it is difficult to make claims of blameworthiness stick in connection with these events. Yet, many still have a sense that because of their relationship to these morally undesirable outcomes, computing professionals ought to do something in response. It is this intuition that the notions of vicarious responsibility and moral entanglement can satisfy.

There are two significant ways in which computing professionals are morally entangled with autonomous systems (and other computer systems) that they design, deploy, and monitor. Firstly, as Johnson argues [16], computer systems may be incapable of forming intentions or any of the other kinds of mental states taken to be necessary for attributability, but they still contain *intentionality*. That is to say, the ways in which computer systems are "poised to behave in certain ways in response to input" [16, p. 201] is no accident: how computer systems behave in response to inputs is a result of how they have been designed and how they are used. The agency of the designers and that of the users of computer systems are mixed with the autonomous behaviour of the system itself. This remains true even for systems created using machine learning techniques that have the system program itself on the basis of training data or reward functions. For the computing professional is the judge of when the training has been successfully completed, on the basis of their goals as a designer. The intentionality of the computing professional is still embedded in even these highly complex systems.

While the embedding of intentionality into a computer system muddies the waters with regard to the classic conditions for responsibility attribution, it makes for a clear case of vicarious responsibility. The computing professionals and others who design and use autonomous systems are morally entangled with these systems and their behaviour because of how they have set up the system to respond to inputs, and because of the ways in which they have deployed the system, which determine the kinds of inputs that the system receives. The agency of the computing professional is thus entangled with the behaviour of the system. When the system causes harm, then, the professionals who designed or used the system ought to take action to make things right, and to help those who have been harmed to make sense of what happened.

Secondly, the specific role of being a computing professional is a morally salient aspect of one's identity when an autonomous system that one designed or deployed causes some harm. As Gotterbarn [8] and Ladd [21] point out, because a computing professional is in the best position to anticipate potential harms when designing a computer system, and to take steps to correct for the harmful behaviour of a computer system once it is in use, the duty to take responsibility for an autonomous system falls most clearly on the computing professional. Other kinds of professional roles may also become relevant when an autonomous system causes harm. For example, if the system was employed in a medical or

educational context, perhaps for aiding in hospital resource allocation [cf. 28] or evaluation of teacher effectiveness [cf. 15], the professional roles of medical or educational practitioners and administrators become morally salient. These other professional roles generate similar duties associated with the entanglement of the relevant individuals' professional identities with the behaviour of the autonomous system.

To illustrate, let's return to the case of a LAWS-powered drone that kills a child soldier or an innocent child. For reasons already given, it is hard to argue definitively that the computing professionals who created the LAWS, or the military personnel or politicians who authorized the use of the LAWS are to blame for this immoral killing. However, each of them is morally entangled with the behaviour of the LAWS that they have created and deployed. Their choices, values, and intentions — in a word, their *agency* — is significantly connected with the behaviour of the LAWS, even if the wrongful death of the child cannot clearly be attributed to their agency. Furthermore, as a matter of professional duty, the computing professionals, military personnel, and policy-makers who are connected to this incident have obligations to respond it in a distinctive way. As such, they owe it to those harmed to take responsibility for this killing. Making amends might be difficult in wartime, but the possibility that the military who used the LAWS, or its government, should apologize for the killing, offer other forms of compensation, or give some satisfying account of the event, could emerge in parley and diplomatic negotiation during the conflict, and in truth and reconciliation inquiries afterward. As for the computing professionals involved, in addition to apologies and other forms of amends, one way they could make things right is to make adjustments to the system to prevent similar incidents in the future — or, perhaps, to repudiate that use of LAWS, dropping their military contracts, and advocating against the development and use of LAWS in the future.

As this example shows, exactly which actions are morally appropriate or required responses to a tragedy with which one is morally entangled are not easy to specify. Centrally, they involve acknowledging one's connection to the harm and making an apology. In many cases, they will include additional forms of making amends and taking action to prevent similar harms in the future. Many of the specifics will depend on the nature of the harm, the needs or desires of the victims, and one's own moral framework.

Finally, let's consider the advantages that the framework of vicarious responsibility and moral entanglement has over other proposed solutions to the responsibility gap. Firstly, by exploiting a pre-existing aspect of our everyday moral practices, my proposed framework avoids the arbitariness of other potential solutions, such as the top-down imposition of new duties and regulations. Instead, my approach recognizes and makes clearer duties that we already feel, however inchoately, should apply in cases of harm caused by autonomous systems. In a way, my approach is bottom-up instead, building a theoretical framework on the basis of live practices.

Secondly, by bringing out the importance of vicarious responsibility, we can shift away from the problems introduced by a focus on causal responsibility, attributability, and accountability. My account allows us to preserve the useful functions of these classical approaches to moral responsibility, without shifting to a different normative domain — and without making fundamental changes to how we understand typical cases of blameworthiness. While I take my proposal to be novel, it is not radical.

Thirdly, my account keeps moral responsibility firmly in the human realm. No computer systems, autonomous, intelligent, or otherwise, are implicated as the responsible parties or the holders of duties. We thus needn't wait for machine ethics to produce a full artificial ethical agent, nor need we fall back to the theatrics of blaming the computer. Even if no human beings are, strictly speaking, to blame for harms caused by autonomous systems, they are the ones who must take responsibility for the consequences.

Fourthly, and finally, as mentioned in the last section, vicarious responsibility also helps to make sense of instances where we accept some form of praise for the actions of others to whom we have some close relation. This can be

extended to computing professionals as well. Where autonomous systems genuinely do good, computing professionals, like proud parents of exemplary children, are entitled to some credit. This would not be justifiable were the classical forms of responsibility — causal responsibility, attributability, and accountability — the only conceptual options available. Vicarious responsibility and moral entanglement thus vindicate both aspects of recent popular writing on autonomous systems: computing professionals deserve a great deal of congratulations *and* criticism for their accomplishments.

7 CONCLUSION

In this paper, I have offered a new solution to the responsibility gap with regard to harmful autonomous systems. I outlined that we should understand the responsibility gap as mainly concerned with how to trace causal responsibility, attribute moral responsibility, and hold someone accountable for the harms caused by autonomous systems. I then explained that given the complex causal chain involved and the qualities of autonomous systems themselves, standard conditions for these senses of responsibility cannot clearly be met in typical cases. After rejecting four kinds of solution to the responsibility gap — rethinking moral responsibility, blaming the computer, advancing machine ethics, and enforcing professional frameworks and regulations — I offered vicarious responsibility and moral entanglement as a theoretical framework that overcomes the deficiencies of these earlier proposals. On my view, because the agency and professional identity of a computing professional are closely connected to the behaviour of autonomous systems, they have or should take on moral obligations to make amends for harms caused by these systems, and to help those affected to understand the events these systems have caused. Computing professionals might not be to blame for the harms caused by their inventions, but they must nevertheless take responsibility for these consequences.

REFERENCES

[1] Robert Merrihew Adams. 1985. Involuntary Sins. *The Philosophical Review* 94, 1 (1985), 3–31. https://doi.org/10.2307/2184713

[2] Association for Computing Machinery. 2018. *ACM Code of Ethics and Professional Conduct.* Retrieved 2022-01-19 from https://ethics.acm.org/

[3] Canadian Information Processing Society. 2018. *CIPS Code of Ethics.* Retrieved 2022-01-19 from https://cips.ca/ethics/

[4] Mark Coeckelbergh. 2021. Narrative responsibility and artificial intelligence: How AI challenges human responsibility and sense-making. *AI and Society* Online first (2021), 14 pages. https://doi.org/10.1007/s00146-021-01375-x

[5] John M. Fischer and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility.* Cambridge University Press, Cambridge, UK.

[6] Harry G. Frankfurt. 1971. Freedom of the Will and the Concept of a Person. *The Journal of Philosophy* 68, 1 (1971), 5–20. https://doi.org/10.2307/2024717

[7] Trystan S. Goetze. 2021. Moral Entanglement: Taking Responsibility and Vicarious Responsibility. *The Monist* 104 (2021), 210–223. https://doi.org/10.1093/monist/onaa033

[8] Donald Gotterbarn. 2001. Informatics and Professional Responsibility. *Science and Engineering Ethics* 7 (2001), 221–230. https://doi.org/10.1007/s11948-001-0043-5

[9] H. L. A. Hart. 1968. *Punishment and Responsibility.* Oxford University Press, Oxford, UK.

[10] Graham Haydon. 1978. On Being Responsible. *The Philosophical Quarterly* 28, 110 (1978), 46–57. https://doi.org/10.2307/2219043

[11] Thomas Hellström. 2013. On the moral responsibility of military robots. *Ethics and Information Technology* 15 (2013), 99–107. https://doi.org/10.1007/s10676-012-9301-2

[12] Herodotus. 1920. In *The Histories*, A. D. Godley (Ed.). Harvard Unversity Press, Cambridge, MA.

[13] Michael C. Horowitz. 2016. The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons. *Daedalus* 145, 4 (2016), 25–36. https://doi.org/10.1162/DAED_a_00409

[14] Institute of Electrical and Electronics Engineers. 2020. *IEEE Code of Ethics.* Retrieved 2022-01-19 from https://www.ieee.org/about/corporate/governance/p7-8.html

[15] Eric Isenberg and Heinrich Hock. 2012. *Measuring School and Teacher Value Added in DC, 2011-2012 School Year. Final Report.* Technical Report 06860.501. Mathematica Policy Research. 30 pages. https://eric.ed.gov/?id=ED565712

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

[17] Deborah G. Johnson and Keith W. Miller. 2009. *Computer Ethics* (4th ed.). Prentice Hall, Upper Saddle River, NJ.

[18] Immanuel Kant. 2012. In *Groundwork of the Metaphysics of Morals*, Mary Gregor and Jens Timmermann (Eds.). Cambridge University Press, Cambridge, UK.

[19] Jonah M. Kessel, Melissa Chan, and Natalie Reneau. 2019. *A.I. Is Making it Easier to Kill (You). Here's How.* The New York Times. Retrieved 2021-12-13 from https://youtu.be/GFD_Cgr2zho

[20] Sunghyo Kim. 2017. Crashed Software: Assessing Product Liability for Software Defects in Automated Vehicles. *Duke Law and Technology Review* 16, 1 (2017), 300–317. https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=1322&context=dltr

[21] John Ladd. 1990. Computers and Moral Responsibility: A Framework for an Ethical Analysis. In *The Information Web: Ethical and Social Implications of Computer Networking*, Carol C. Gould (Ed.). Westview Press, Boulder, CO, San Francisco, CA, and London, UK, 207–227.

[22] Andreas Matthias. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6, 3 (2004), 175–183. https://doi.org/10.1007/s10676-004-3422-1

[23] John Stuart Mill. 1906. *Utilitarianism.* University of Chicago Press, Chicago, IL. https://books.google.ca/books?id=nhERAAAAYAAJ

[24] Keith W. Miller. 2011. Moral Responsibility for Computing Artifacts: "The Rules". *IT Professional* 13, 3 (2011), 57–59. https://doi.org/10.1109/MITP.2011.46

[25] Patrick T. Miyaki. 1992. Computer Software Defects: Should Computer Software Manufacturers Be Held Strictly Liable for Computer Software Defects? *Computer and High Technology Law Journal* 8 (1992), 121–144. Issue 1. https://digitalcommons.law.scu.edu/chtlj/vol8/iss1/4

[26] James Moor. 2006. The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems* 21, 4 (2006), 18–21. https://doi.org/10.1109/MIS.2006.80

[27] Helen Nissenbaum. 1994. Computing and Accountability. *Commun. ACM* 37, 1 (1994), 72–80. https://doi.org/10.1145/175222.175228

[28] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. https://doi.org/10.1126/science.aax2342

[29] Matteo Santoro, Dante Marino, and Guglielmo Tamburrini. 2008. Learning robots interacting with humans: from epistemic risk to responsibility. *AI and Society* 22 (2008), 301–314. https://doi.org/10.1007/s00146-007-0155-9

[30] T. M. Scanlon. 1998. *What We Owe to Each Other.* The Belknap Press of Harvard University Press, Cambridge, MA.

[31] George Sher. 2009. *Who Knew? Responsibility without Awareness.* Oxford University Press, New York, NY. https://doi.org/10.1093/acprof:oso/9780195389197.001.0001

[32] David Shoemaker. 2011. Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics* 121, 3 (2011), 602–632. https://doi.org/10.1086/659003

[33] P. F. Strawson. 1962. Freedom and Resentment. *Proceedings of the British Academy* 48 (1962), 1–25.

[34] Gerhard Wagner. 2019. Robot, Inc.: Personhood for Autonomous Systems. *Fordham Law Review* 88, 2 (2019), 592–612. https://fordhamlawreview.org/wp-content/uploads/2019/11/Wagner_November_S_8.pdf

[35] Gary Watson. 1996. Two Faces of Responsibility. *Philosophical Topics* 24, 2 (1996), 227–248. https://doi.org/10.5840/philtopics199624222

[36] Michael J. Zimmerman. 1997. Moral Responsibility and Ignorance. *Ethics* 107, 3 (1997), 410–426. https://doi.org/10.1086/233742