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

Trystan S. Goetze
tgoetze@fas.harvard.edu
Harvard University
Department of Philosophy > Embedded EthiCS
Cambridge, Massachusetts, USA

ABSTRACT

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 my earlier 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.

KEYWORDS

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 *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3531146.3533106

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533106

1 INTRODUCTION

Tech news in recent years has been marked by vacillation between, on the one hand, effusive praise for innovations in artificial intelligence and other emergent technologies for their promise to usher 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 develop. 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 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 morally responsible when autonomous systems cause harm. Answering this question is important for reasons beyond legal liability. 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: enact a legal framework of liability; develop a new theory of moral responsibility; blame the autonomous systems themselves; or establish enforceable professional ethical standards. 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. Drawing on my previous work on vicarious moral responsibility — i.e., 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 autonomous systems that they develop. This solution has the advantage of being bottom-up rather than top-down: it derives from a set of existing (albeit vague) moral norms, instead of an *ad hoc* 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 a recent short documentary, *A.I. is Making it Easier to Kill (You)*, journalists for *The New York Times* present some of the worries surrounding the development of lethal autonomous weapons systems (LAWS) by militaries across the world [21]. In the documentary, security policy expert 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 claims 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 — perhaps *couldn'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" [16, 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? [16, 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 improbable) result of your actions.

This kind of 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 [15, book VII, chs. 34–35]. Could we hold the developers of the system responsible? Certainly we could, but the system developers 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. (And arguably, similar gaps occur whenever a machine or creature that is not a moral agent — such as an automobile or a pet dog — causes harm.) As such, it has been a perennial topic in computer ethics. The proliferation of AI-enabled systems that increasingly make decisions instead of human beings, and the continued lack of clarity around who is responsible for their behaviour, makes it worth attending to this literature to see if it can offer us any guidance.

In this section and the next, 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 with a philosophical solution. This conclusion motivates the turn, in the following two sections, to another sense of moral responsibility. Below, the first subsection clarifies the theoretical details regarding moral responsibility, and 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 clarify the sense of the term that is at issue.

In the philosophical literature, there are many different concepts of responsibility that are potentially relevant to the responsibility gap (see [12, 13, 31, 34] for details on some of these distinctions). Sometimes, the focus of the specific conception of responsibility that is invoked is to attribute a specific event or its outcomes to a particular human agent, often for the purposes of determining blameworthiness. Other times, we speak of a responsibility as a duty or set of duties arising either as a matter of general moral obligation or due to one's specific social role.

The most relevant sense of "responsible" for the responsibility gap seems to be that of *personal responsibility*, where an agent is accountable for something that they have done. To be precise, we typically think that a person A is morally responsible for an event X only if A caused X, X was properly attributable to A, and A is a proper target of moral evaluation (such as praise or 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.

When determining causal responsibility, we usually look to the agent or process whose actions or behaviour caused the event. For example, suppose that X is someone's death. When we investigate, we find that the cause of death was a gunshot wound, 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 personal responsibility. The remaining conditions for personal responsibility require some unpacking, and are subject to much debate. However, there is widespread agreement in analytic moral philosophy that there are at least two.

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, and there are several popular approaches. As mentioned earlier, 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 [8]. 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 and acted on the desires that issued in the actions that caused X [9].

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, which produced their ignorance that causing X would be wrong [37]. 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 [30].

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 A is personally responsible for X. Furthermore, most theories of moral responsibility hold that when these conditions obtain, A is also blameworthy for X.

Blameworthiness is more than a merely conceptual category. If A is blameworthy for X, it becomes appropriate to hold A responsible for X by, for example, feeling resentment towards A, reprimanding A, or imposing other sanctions on A; just how harsh these punitive responses may be depends on additional factors, such as the severity of the wrongdoing or the blamer's relationship to the blamee. 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.

Note that these conditions are similar to, but distinct from, the legal conditions for criminal liability. The control condition is comparable to satisfying the *actus reus*, and the epistemic condition is comparable to satisfying the *mens rea*. But these conditions bear on whether one is liable for violating *legal* norms. While legal norms are connected to moral norms in important ways, ethics and the law are two distinct sources of practical reasons. Some actions that may be permitted by law are immoral (e.g. forced labour), and some actions that are morally required may be illegal (e.g. protest against an unjust ruler). Furthermore, our interpersonal practices of holding people accountable for immoral actions (e.g. rebuking, shaming, censuring, or avoiding the wrongdoer) are different from our judicial practices of holding people accountable for illegal actions (e.g. mandatory community service, fines, imprisonment, or death). This distinction between moral and legal responsibility will be important to bear in mind as we go forward.

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 (though it is not unique to these cases). When we attempt to determine who is responsible for any harms that result, our usual process, described in the previous subsection, runs into difficulties. Let's run through those steps 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 the control condition, 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 being morally responsible for anything. 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" [22, p. 219].

Finally, even if we decide to ignore these problems with personal responsibility, we run into another stumbling block with blame. 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 one chastises 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 developed 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 individuals or groups as causally responsible for the specific harm caused by the autonomous system. At every step, decisions are likely made not by individuals, but by teams or group agents. The causal responsibility thus is diluted across many different people, such that assigning it to any subset of them may be difficult to justify. And while we may be tempted to pin the causal responsibility on the developers of the autonomous system, the programmers and data scientists are so far up the causal chain that it would be just as controversial to pin the responsibility on them.

But suppose we simply chose a causally responsible party, be they the developers, 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 [23]. 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 organizations 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 [16]. Likewise, Matthias argues that because of the "black box" effect of multiple popular machine learning methods, the developers of an autonomous system that uses these techniques may not be in a position to predict how the system might respond to any particular situation [23]. 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 developers, the system may act in unexpected ways once deployed.

Given these difficulties in making personal responsibility 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 clearly blameworthy for the harm — and so, 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 responsible party at any particular node within that web [27, p. 76].

Before moving on, it's worth briefly acknowledging the other edge of the sword of responsibility, namely, praiseworthiness. While I am mainly concerned with harm in this paper, as mentioned earlier, I am also interested in explaining why it might be appropriate for computing professionals to accept responsibility for the *good* that autonomous systems 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, at root, two ways of expressing the judgement that someone is morally responsible for some outcome — the difference lies in the moral evaluation

that comes along with that judgement, namely, moral approval or disapproval. Responsibility for harm and responsibility for good stand or fall together.

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: deploy a legal framework of liability; develop a new theory of moral responsibility; blame the autonomous systems themselves; and establish professional accountability frameworks in computing.

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 arbitrariness in its solution, so long as the solution can be justified. But, as I will argue in the next section, not only is the responsibility gap in computing *not* so unique, there is another solution that is preferable precisely because it avoids this sense of arbitrariness.

4.1 Use a Legal Framework

One potential solution would be to deploy a legal accountability framework that avoids the problems posed by the responsibility gap. For example, Matteo Santoro, Dante Marino, and Guglielmo Tamburrini suggest that we use the legal notion of *strict liability* [28]. The framework enables the law to hold people accountable for certain kinds of harm or risk even when they are not at fault. Applications of strict liability often concern dangerous products (e.g. storage of hazardous materials) and dangerous activities (e.g. using explosives), but there are cases where people are held strictly liable for damage caused by their livestock or pets — that is to say, for the behaviour of autonomous non-human beings in their care.

This latter case seems especially relevant when we think about autonomous computer systems — and the former cases seem relevant to LAWS in particular. For autonomous systems act on their own initiative within their domain of operation: the drone moves through its patrol zone, analyzing and engaging targets, perhaps attacking targets that it should not, without regular human input. Similarly, livestock navigate their environment and choose how to behave in what they take to be their territories: the cows walk around the pasture without human direction, perhaps wandering into areas they should not, where they might cause damage to neighbouring properties. We might think that the owners and developers of autonomous systems should be held strictly liable just as the owners of livestock are.

While making this shift to legal liability may be helpful in terms of public policy, it makes a significant mistake by conflating moral norms and legal norms. The responsibility gap, as I described it, is fundamentally about *moral* responsibility. As explained above, there are important differences between legal liability and moral responsibility. A legal solution such as strict liability may well be desirable, since it would provide a method for holding *someone* accountable for the harms produced by autonomous systems. But it has several shortcomings. For one, there is no analogue of praise in the framework of strict liability, so it would still be unclear whether developers are responsible for the good their systems may produce. For another, if we adopt strict liability as our solution to the responsibility gap, it's not clear how this should inform our moral evaluation of those held strictly liable. It would be a mistake to hold our moral judgements hostage to specific legal frameworks, which are contingent on the specific legal system and legislative history of particular jurisdictions that could themselves be morally unjust. Furthermore, this account is silent on how we should respond interpersonally to someone who is found strictly liable. Should they be blamed? (And why? Strict liability explicitly disavows judgements of blame.) Using strict liability here simply solves the wrong problem.

4.2 Rethink Moral Responsibility

If we continue the line of thought that we should rethink the relevant conception of responsibility such that we eschew the control condition, but restrict our attention to accounts of *moral* responsibility, we find several options in the philosophical literature. For example, Robert Adams [3] and Angela Smith [32] have argued that we often hold people responsible for their involuntary attitudes and emotional reactions. Sometimes we blame others or ourselves for getting angry when we 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. On Smith's account, we are responsible for any action or reaction that is rationally connected to our evaluative judgements — that is to say, on her view, one is morally responsible for any of one's behaviour, voluntary or involuntary, if, in principle, one could offer a rational justification for that behaviour on the basis of a stable judgement of right and wrong that one holds.

We might be able to develop Adams's or Smith's account to explain why some human agent is responsible for harms caused by autonomous systems. However, making such a case would still require substantial work. Even setting aside the philosophical debate around these theories — both of which are controversial — neither view addresses the causal complexities discussed above. If we accept Adams's view, for instance, we would still have to explain why some *particular* agent(s) or group(s) within the complex causal chain that leads to an autonomous system producing harm are responsible.

Moreover, if we accept Smith's view, we would have to explain how the behaviour of an autonomous system is rationally related to the evaluative judgements of some agent — because, once again, the system itself is incapable of such judgements and so cannot be responsible for anything. For example, one might suggest that the developer of a drone that kills an underage combatant judges such killings to be morally permissible. But there are many cases where this approach is unlikely to work. The developer of the drone might well hold the opposite judgement — that killing underage combatants is *not* morally permissible — and they might be just as horrified that their creation behaved as it did. They might even have taken steps to try to prevent this sort of thing from happening. It would be implausible and unfair to think that the drone's behaviour

reflects this developer's evaluative judgements. Yet, we may still think that they should bear some kind of responsibility for what happened.

4.3 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 already inclined to think of them as morally responsible for their behaviour. Additionally, he suggests that systems which 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 [14].

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 has caused harm involves taking the *objective* attitude towards the autonomous system, and this attitude is fundamentally incompatible with the practice of holding its target 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 who are involved, such as the system's developers. As as Deborah Johnson puts it, "computer systems cannot *by themselves* be moral agents" [18, p. 203].

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 some kind of 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, as James Moor describes [26]. Some of the most common suggestions in this vein are either to encode an ethical theory — such as utilitarianism [cf. 24], Kantianism [cf. 20], or contractualism [cf. 29], or some hybrid of these — or the results of a robust survey of folk morality [such as the results of 4], and use this representation of moral understanding to constrain the behaviour of the system.

While it is possible that machine ethics could produce more morally desirable outcomes, the technology that is currently available or might be developed in the foreseeable future does little to address the responsibility gap. An ethical system could be developed by introducing constraints on its behaviour, whether these are hard-coded to simply eliminate the possibility of harmful behaviour (what Moor calls an "implicit" ethical agent) or contained within a stage of computation that compares probable outcomes against encoded moral rules (what Moor calls an "explicit" ethical agent). In both cases, the system simply follows the rules of its programming, regardless of their complexity. It has no actual moral understanding, which requires a critical reflective capacity. And it is hard to see how an autonomous system could be given such a capacity without furnishing it with artificial general intelligence (what Moor calls a "full" ethical agent) — an achievement that has proved elusive, and may be unethical to create on independent grounds.

4.4 Professional Frameworks

We saw above that Santoro et al. shifted from thinking about moral responsibility to legal liability [28]. 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 — some examples of which we saw in the last section — calling this a strategy for "dodging" or "side-stepping" responsibility [11]. On his view, computing professionals should assume responsibility for the harmful behaviour of the systems they design, deploy, maintain, and monitor. Following Ladd [22], 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 [25], 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 [18], and Nissenbaum's argument that professional frameworks should hold computing professionals morally accountable for harms caused by their technologies [27].

Each of these works suggests that professional 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 penalties for malpractice, 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? Why not hold computing professionals' bosses, or their organization as a whole, responsible?

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 [19]. 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 aspect of these professions is that they have fundamental values that are shared across the profession and 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 [1, 5, 17], the enforcement of these codes is less effective than in strongly differentiated professions. It is rare for a computing association 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. The relative weakness of ethics enforcement in computing remains despite recent updates to the ACM Code to clarify enforcement procedures [2].

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 the autonomous systems that they have a hand in creating or maintaining — and, if they are, in what sense. Adding professional standards would help insofar as they would ensure that *someone* is held accountable when these systems cause harm, but it might still 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, beyond the fact that they are in a good position to do something. 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 away from personal responsibility and focus instead on a *different* sense of responsibility that we already acknowledge in our everyday life, though it remains under-discussed in the philosophical literature. Namely, I contend that the notion of *vicarious responsibility* can explain and justify the intuition that computing professionals ought to take responsibility for harms caused by systems they create. At the same time, this account can explain why it is legitimate for computing professionals to take some of the credit for the *good* produced by autonomous systems.

Unlike the sense of responsibility that we have been considering so far, which concerns personal responsibility for one's own actions, *vicarious responsibility* concerns cases where one agent is responsible for the actions or behaviour of another entity. While some instances of vicarious responsibility can be assimilated into the framework of personal responsibility — as when one's own will is carried out by another, such as a subordinate to whom one issues orders — other cases are more complex. Autonomous systems are one of the complex cases, whence the responsibility gap: as explained in §3, it is not straightforward, on accounts of personal responsibility, to explain why we think that computing professionals (or anyone else) are in some way responsible for the behaviour of these systems.

But these are not the only cases where vicarious responsibility appears — and many are quite familiar. For example, parents are often taken to be in some way responsible for the behaviour of their young children. Suppose that, at a family dinner, a child angrily pours hot gravy onto the lap of their cousin. We would expect the child's parents to respond in a distinctive way, offering apologies, making sure that the cousin is unharmed, helping to clean up the mess, and so on, in addition to whatever corrective action they mete out to their child. Furthermore, were the parents *not* to respond in this way, we would think there was something wrong with them: suppose that they expressed blame toward their child but left things at that, much as a bystander at the next table in the restaurant might. In such a case, we might think that there is something about being a parent that these people have failed to grasp. In fact, the behaviour that we expect is notably similar to what we might expect of an adult whose blameworthily reckless actions lead to a similar result — say, wild tipsy gesticulations that knock the gravy boat into another person's lap. But it is also clear that the parents are not personally responsible for the harm caused by their child: this case is not as straightforward as that of a superior issuing an order.

We can think of many other cases of vicarious responsibility: A relative of a bigot might apologize for what the bigot says, despite having done nothing wrong themselves. The parents of an adult might be expected to respond in a distinctive way when their offspring commits a violent crime — whether it is an expression of something like remorse, or an apology, or some explanation of their offspring's behaviour or character, or something else; in any case, it seems that it would not be enough for them to blame their adult offspring and leave it at that, unless they were quite estranged. The owner of a pet might be held morally responsible for its bad

behaviour, such when their dog bites a passing pedestrian — or they might praised for the pet's heroism, such as when their dog, unprompted, rescues a stranger from an attacking wild animal. A citizen of a country that commits war crimes might feel shame despite not having any role in the atrocities. An employee of a large corporation that causes environmental damage might apologize for the damage even if they had no direct role in causing it; and, they might offer some explanation of how the company's internal processes contributed to the problem. And so on.

What explains these commonplace instances of vicarious responsibility? In previous work, I suggested that what unifies these kinds of cases is a relationship of *moral entanglement* between the vicariously responsible agent and the behaviour of the other entity [10]. This relationship obtains in instances where there is some uncertainty or vagueness around the extent to which one's own agency is implicated in the behaviour of someone or something else. In the cases of parents and their offspring or their pets, the parents' past actions are heavily implicated in how their offspring or pets behave and the sort of person or animal they have become, even though their offspring or pets act autonomously from them, and may become more autonomous as time goes on. In the cases of citizens and states or employees and employers, even if one is not the decision-maker responsible for some harm done by the group, one's membership in the group agent makes it difficult to draw a clean line between one's own agency and that of the group. The uncertainty of where one's own agency ends and where another's begins is what generates the entanglement.

We can even identify self-reflexive examples: for instance, Bernard Williams describes a case of a truck driver who, through no fault of his own (suppose the driver is following all the rules of the road, is alert and unintoxicated, is well-trained and generally a good driver, etc.) runs over a child who happens to run into the street at the wrong time. While it is clear that the driver is not to blame for this tragic accident, Williams tells us that "We feel sorry for the driver, but that sentiment co-exists with, indeed presupposes, that there is something special about his relation to this happening, something which cannot merely be eliminated by the consideration that it was not his fault" [35, p. 28]. In this case, the source of the uncertainty is around the driver's own agency: what might he have done differently to prevent this from happening, despite the fact that, at the time, he had no reason to think he should? Here, the driver is morally entangled with a past version of himself.

How should we respond to instances of moral entanglement? In my previous work, I suggested (drawing on David Enoch's discussion of similar scenarios [7]) that in these cases, the aspects of one's identity that connect one to the entity with which one is morally entangled become morally salient, generating moral reasons for *taking responsibility* for the behaviour of the other entity. The parents' identity *as* parents, for example, gives them reason to take responsibility when their child acts wrongfully, even though they are not personally responsible for what the child has done. The employee's identity *as* a member of a group agent gives them reason to take responsibility for the harms caused by their employer, despite not being at fault.

Taking responsibility, in this sense, can take many forms, depending on the context. When harm has been done, taking responsibility involves taking on (some of) the moral duties or obligations that someone who *was* personally responsible for the harm would have had. Someone responsible for recklessly knocking gravy into a dinner guest's lap might have a duty to apologize to the victim; they also have reason to feel bad for what they've done. In the parallel case of vicarious responsibility, where the child angrily pours gravy into their cousin's lap, the child's parents should take on these same obligations — they should feel bad for what their child has done and offer apologies of their own.

In cases of employees or citizens of employers or states that cause harm, the individual should take responsibility by, perhaps, offering apologies, feeling shame, or explaining the behaviour of the group agent to which they belong. (This latter duty of offering explanations is what Mark Coekelbergh, in connection with AI systems, calls *hermeneutic* or *narrative responsibility* [6].) In the self-reflexive cases, such as the truck driver, Williams argues that we should feel a kind of regret that things did not turn out differently, as well as a willingness to try to make things right [35]. And in those cases where one is vicariously responsible for some *good*, such as the owner of a heroic animal, while there are no particular obligations to take on, one's moral entanglement does entitle one to something akin to praise.

Notice that all of these responsibilities taken correspond to what, as I noted above, we expect these agents to do. And if at least some of these responsibilities are obligations, then we are entitled to *demand* that the vicariously responsible agent take responsibility in these ways. In some cases, it may even constitute an additional wrong for the agent not to take responsibility for what the other entity has done — and this seems more plausible in some cases rather than others. For example, it seems worse for a parent to refuse to take responsibility for harms caused by their children than for an individual citizen to refuse to take responsibility for harms caused by their state. In previous work, I suggested that we can explain these differences by thinking of moral entanglement as a continuum, where stronger entanglements produce moral obligations to take responsibility, and weaker entanglements merely produce *pro-tanto* moral reasons to take responsibility [10]. Stronger entanglements arise when the aspect of one's identity connecting one to the other entity's behaviour is more central or important to who one is. One's role as a parent, for example, is often more important to one's identity than one's nationality — though there can be exceptions.

Why do we have this practice of taking responsibility for things that are outside the scope of our personal agency? One reason is that human beings have a material and psychological need for someone to step up and take responsibility where harm has been caused. When harm has been done, someone is often left in need of care and support. And beyond that, it makes us uneasy to think that harm will go unanswered, as it seems to leave something important unfinished. (Perhaps this is also one reason why some of us believe that a higher power will make all things right in the end.) Often, it would be appropriate to call upon the wrongdoer to take responsibility for their actions. But there are many cases where either there is not clear wrongdoer, or where the victim has good reason to avoid interacting with the wrongdoer again. These are the cases where our moral entanglements call on us to take responsibility more widely.

Another reason is that the borders of what we are personally responsible for are often genuinely unclear. Philosophers pretend

that judgements of responsibility are clear-cut, but this is usually an artefact of the abstraction used to arrive at clearer intuitions. This vagueness or uncertainty around the scope of our agency is, as Susan Wolf argues, simply a fact about human life [36]. As such, Wolf claims that there is a virtue of appropriately taking responsibility for things beyond what is clearly one's fault.

To summarize: When one's connection to some other autonomous entity — be it an individual agent, a group agent, or an entity capable of autonomous behaviour but not of full agency, such as a child, an animal, or certain computer systems — makes the edges of one's agency unclear with respect to that entity's behaviour, one is morally entangled with that entity. When one is morally entangled with someone or something that causes harm, one acquires reasons to take responsibility for that harm, by taking actions similar to what we would expect one to do if one were personally responsible for that harm. The stronger the entanglement — i.e., the more important the connection between oneself and the other entity is to one's personal identity — the stronger the reasons one has to so take responsibility.

6 CLOSING THE GAP: TAKING RESPONSIBILITY FOR AUTONOMOUS SYSTEMS

Let's now apply the framework of vicarious responsibility and moral entanglement 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 renders it 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 develop. Firstly, as Johnson argues [18], computer systems may be incapable of forming intentions or any of the other kinds of mental states taken to be necessary for personal responsibility, 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" [18, 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 developers is thus mixed with the autonomous behaviour of the system itself, making it unclear just where the developer's agency ends and the autonomous behaviour of the system begins. 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, and the computing professional is the one who sets things in motion by choosing the training dataset, creating the reward function, tuning the hyperparameters, and so on. 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 conditions of personal responsibility, it makes for a clear case of moral entanglement. 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 developed 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 [11] and Ladd [22] 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 often falls most clearly on the computing professional.

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* — all contributed to the behaviour of the LAWS, even if the wrongful death of the child cannot clearly be attributed to any of them. Furthermore, as a matter of professional moral 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 some of the relevant actions may include: apologizing for the killing; offering forms of compensation to the family or community, if possible; or giving some account of the event, explaining why things happened the way they did.

One distinctive thing that the computing professionals involved might do is to make adjustments to the system to prevent similar incidents in the future, and to review their processes for creating systems such as the LAWS that is implicated in the tragedy. Or, perhaps they should repudiate that use of LAWS, drop their military contracts, and advocate against the development and use of LAWS in the future.

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, 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.

Thirdly, 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. 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 offered a new solution to the responsibility gap. I outlined that the responsibility gap arises when the complexities of creating a computing system render it unclear just who is responsible for its behaviour. After rejecting four solutions to the responsibility gap, I offered vicarious responsibility and moral entanglement as a theoretical framework that overcomes the deficiencies of these 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. Computing professionals might not be to blame for the harms caused by their inventions, but they must nevertheless take responsibility for them.

ACKNOWLEDGMENTS

Thanks to Alison Simmons, Jeff Behrends, Jenna Donohue, William Cochran, Kevin Mills, Kiran Bhardwaj, Brian Chance, Stacy Doore, Tony Steinbock, Bob Crease, the Cyberethics Forum Works-in-Progress Group, the attendees of a colloquium I gave at Stony Brook University, and three anonymous referees for discussion of these ideas. I also acknowledge that Harvard University is situated in the traditional and ancestral territory of the Massachusett people.

REFERENCES

[1] ACM. 2018. *ACM Code of Ethics and Professional Conduct.* Association for Computing Machinery. Retrieved 2022-05-02 from https://ethics.acm.org/

[2] ACM. 2018. *Association for Computing Machinery Code of Ethics Enforcement Policy.* Association for Computing Machinery. Retrieved 2022-05-09 from https://ethics.acm.org/wp-content/uploads/2018/07/2018-ACM-Code-of-Ethics-Enforcement-Procedure.pdf

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

[4] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan. 2018. The Moral Machine experiment. *Nature* 563 (2018), 59–64. https://doi.org/10.1038/s41586-018-0637-6

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

[6] 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

[7] David Enoch. 2012. Being Responsible, Taking Responsibility, and Penumbral Agency. In *Luck, Value, and Commitment: Themes From the Ethics of Bernard Williams*, Ulrike Heuer and Gerald Lang (Eds.). Oxford University Press, Oxford, UK, 95–132. https://doi.org/10.1093/acprof:oso/9780199599325.001.0001

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

[9] 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

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

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

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

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

[14] 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

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

[16] 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

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

[18] 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

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

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

[21] 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

[22] 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.

[23] 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

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

[25] 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

[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] 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

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

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

[31] 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

[32] Angela M. Smith. 2005. Responsibility for Attitudes: Activity and Passivity in Mental Life. *Ethics* 115 (2005), 236–271. https://doi.org/10.1086/426957

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

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

[35] Bernard Williams. 1981. Moral Luck. In *Moral Luck: Philosophical Papers, 1973–1980.* Cambridge University Press, Cambridge, UK, 20–39.

[36] Susan Wolf. 2001. The Moral of Moral Luck. *Philosophic Exchange* 31, 1 (2001), 15 pages. http://hdl.handle.net/20.500.12648/3203

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