Responsibility Before Freedom: closing the responsibility gaps for autonomous machines

Shervin Mirzaeighazi and Jakob Stenseke, Lund University (Sweden)
International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: responsibility gap, freedom of action, rights, accountability, autonomous machines

Abstract

The introduction of autonomous machines (AMs) in human domains has raised challenging questions about the attribution of responsibility; referred to as the *responsibility gap*. In this paper, we address the gap by positing that entities cannot be granted the freedom of action unless they can also recognise the same right for others—and be subject to blame or punishment in cases of undermining the rights of others. Since AMs fail to meet this criterion, the users who utilize an AM to pursue their goals can instead grant the machine *their* (the user's) right to act autonomously on their behalf. Thus, an AM's right to act freely hinges on the user's duty to recognise others' right to be free. Since responsibility is attributed *before* an entity is given the freedom to act, the responsibility gap only arises when we ignore the fact that AMs have no right of acting freely on their own. We also discuss some attractive features of the approach, address some potential objections, and compare our theory to existing proposals. We conclude by arguing that holding users responsible for the behaviour of AMs promotes a responsible use of AI while it indirectly motivates companies to make safer machines.

Keywords: Responsibility gap, right forfeiture, freedom of action, accountability, autonomous machines

1. Introduction

In recent years, the introduction of autonomous machines (AMs)[1] in an ever-growing set of human domains have raised challenging questions about the designation of responsibility. When an AM brings about a bad outcome—and seemingly no recklessness, malice or negligence was involved—it appears as if no one can be held responsible in any meaningful sense of the term. Should policymakers, manufacturing companies, users, the machines themselves, or all of them be held responsible?

This problem—commonly known as the "responsibility gap" (RG)—has been subject to extensive scholarly debate, with contributions ranging from the fatalistic (believing that RG presents a genuine and unsolvable problem) to the optimistic (believing that the RG either can be solved *or* that it is not a genuine problem in the first place). However, while these contributions have enriched the RG debate

[1] [1] Throughout this paper, the term "autonomous machine" refers to computational systems of software and hardware that can carry out a task (or a range of tasks) without human supervision or control. As such, the autonomy of these systems should not be conflated with personal autonomy (to exercise control over one's life), political autonomy (self-governing with regards to other political entities), or moral autonomy (e.g., in the Kantian sense of acting according to one's self-imposed rules without external influence).

and disentangled its many facets, no resolution or consensus seems to be in sight. As it stands, there seems to be no way forward that would reconcile the pessimist's moral concern with the solutionist's optimism. And while several of the proposed solutions can address *some* aspect of the complex issue, none of them seems capable of cutting through the many dimensions and nuances the debate presents.

In this paper, we present a novel solution to the RG that not only targets the core of both moral responsibility and legal accountability concerns, but is applicable to AMs in a broad range of domains; and in particular, the two areas—driving and warfare—that remain at the centre of RG discussions.

The solution can be summarized as "responsibility before freedom" and is based on the idea that entities cannot be granted the freedom of action in our society unless they can also recognise the same right for others—and be subject to blame or punishment in cases of undermining the rights of others. Essentially, since AMs fail to meet this criterion, the user who utilizes an AM to pursue their goals can instead grant the machine *their* (the user's) right to act autonomously on their behalf. Thus, an AM's right to act freely hinges on the user's duty to recognise others' right to be free. Since responsibility is attributed *before* an entity is given the freedom to act, the responsibility gap only arises when we ignore that AMs have no right to act freely on their own. Ultimately, we argue that "responsibility before freedom" not only provides clarity to a convoluted scholarly dispute, but it also subsumes many previous endeavours to "close" the responsibility gap,[2] all while answering to the original concern raised by pessimists (de Jong, 2020; Matthias, 2004; Roff, 2013; Sparrow, 2007).

The rest of the paper is structured as follows. In the next section, we introduce the responsibility gap debate; its different flavours, and what is missing from previous contributions. In section 3, we present our own approach, which centres around the idea that an entity's right to act freely presupposes that it recognizes the same right in others. Finally, in section 4 we discuss some attractive features of the approach, address some critical objections, and situate our theory in relation to existing proposals.

2. The responsibility gap(s)

Who is responsible for the behaviour of autonomous machines? In a landmark paper from 2004, Andreas Matthias was the first to frame this concern as the "responsibility gap" (RG). Matthias presented us with an intriguing choice: either we (as a society) ban the use of autonomous machines whose behaviour neither manufacturers nor operators can predict, or we face a situation—a RG—that cannot be addressed by our traditional responsibility practices. Over the nearly two decades that have followed Matthias' paper, the scholarly contributions addressing the RG and its many flavours have experienced an exponential increase.[3] These contributions can be roughly divided into four camps: (i) pessimists, (ii) solutionists, (iii) trivialisers, and (iv) differentialists.

(i) *Pessimists* take Matthias' challenge at face value and view the responsibility gap as an inevitable consequence of a genuine problem that resists a solution (de Jong, 2020; Matthias, 2004; Roff, 2013; Sparrow, 2007). In short, the reasoning is that, on the one hand, we cannot blame the AM itself—as it lacks the relevant capacity to be a genuine subject of praise, blame, or punishment—nor can we, on the other hand, hold any humans responsible, as they are not in control of the AM (or able to predict its

[2] For instance: (Champagne & Tonkens, 2015; Hindriks & Veluwenkamp, 2023; List, 2021; Saxon, 2016; Simmler & Markwalder, 2019; Simpson & Müller, 2016; Taddeo & Blanchard, 2022; Tigard, 2021a).

[3] See Santoni de Sio and Mecacci (2021) and Oimann (2023) for two recent critical overviews of the RG debate.

behaviour). However, the fact that we haven't reached a knockdown solution for RGs yet does not imply that there is no solution at all; and we aim to provide one in this paper.

(ii) *Solutionists*, by contrast, are those authors who believe that the responsibility gap can be bridged. While solutionists are united in their opposition to the pessimists, they can be differentiated in terms of the strategies they employ to close the RG, which come in a variety of flavours (Oimann, 2023). We will briefly discuss four:

(ii.a) *Technical solutionists* approach the RG as an empirical problem that can be solved via technical means, e.g., by identifying the link between an acting agent and a bad outcome (Saxon, 2016), or by designing AMs that can display a degree of risk that is morally tolerable (Hindriks & Veluwenkamp, 2023). The problem with this strategy, as pointed out by Oimann (2023), is that it views the attribution of responsibility as a problem of causality, and thus fails to address the normative dimension of the RG that concerns the inability of identifying individuals that are responsible for outcomes. For example, Hidriks and Velunwenkamp argue that a certain amount of risk is acceptable in society (because it can never be reduced to zero), and if the risk associated with the use of AMs is below that tolerable threshold, then we can write off individual cases of RG as accidents for which no responsibility can be assigned. To this end, Hindriks and Velunkamp believe that the RG is in fact a 'control gap' in disguise, which occurs when AMs fail to exhibit a level of risk that is morally tolerable. But again, this answer fails to do justice to the concerns of pessimists. While reducing the risk can be considered a reason to use AMs, it cannot, by itself write off cases of RG as accidents. To illustrate, suppose we have an employee who makes way fewer mistakes than that of the average employee. This, however, does not give them any moral discount. If they commit a moral wrongdoing, they are responsible for it; and the fact that they commit fewer mistakes does not provide sufficient grounds to turn their moral wrongdoing into an accident.

(ii.b) *Human solutionists*, instead, focus on closing the gap in virtue of practical arrangements, where, e.g., a human agent willingly takes on a "blank check" liability (Champagne & Tonkens, 2015) or an *ex ante* "moral gambit" (Taddeo & Blanchard, 2022) for an AM's actions. According to the latter—the moral gambit solution—it is presupposed that manufacturers or users know that it is possible that something goes wrong, and yet if they decide to deploy/use the AM, they are responsible in the case of a RG. A major problem with this solution is that it has counterintuitive implications. For example, I know there is a possibility that if I go on vacation, I might be mugged or killed by some criminals, or that my plane crashes. If I nevertheless go ahead and plan my trip with due care and precautions, and any of these events happen, I will not be responsible for being mugged, killed, or for my own death. Essentially,
just knowing that if you do an action and there is a possibility that something goes wrong does not make you responsible for it. In turn, the problem with "blank check" liability (Champagne & Tonkens, 2015) is that it is basically a form of scapegoating: someone must be held responsible, and if no one is, then let's blame those who occupy the highest office. According to this theory, prestige comes with a price and one of the prices of a high office—for example, being a military general—is to be held accountable in cases when you are not responsible. However, this way of dealing with RGs completely ignores their normative side. While such arrangements might serve to correct undesirable outcomes (e.g., compensate a victim), the strategy has been criticized for not addressing the "retributionist gap"; i.e., the mismatch between our retributionist desires to hold someone responsible and the fact that no one seems to be responsible (Amoroso & Giordano, 2019; Chengeta, 2016; Danaher, 2016). Moreover, contrary to the cases of lethal autonomous weapons (LAWs), in many situations—e.g., self driving

cars—the user does not hold any prestigious or high office, and therefore, this solution cannot be applied to them.

(ii.c) *Collective solutionists*, by contrast, attempt to bridge the gap by distributing responsibility over the collaborative agency that constitutes human-machine interactions (Galliott, 2020; Nyholm, 2018), or, in the case of killer robots, the group of agents that makes up the military industrial complex (Taylor, 2021). Nevertheless, while such solutions are perfectly apt to answer utilitarian concerns about the group level (e.g., the tragedy of the commons present in the climate crisis), and attend to some specific nuances of human-machine collaborations, they fail to consider the conditions under which individual agents are (un)fairly or (un)deservedly held responsible. For instance, in this view, a member of a group may be held responsible for the wrongdoings of its group, all while the individual agent herself did not have any intention of doing so, nor did she act with recklessness, malice, or negligence.[4] Furthermore, distributing responsibility to multiple entities runs the risk of creating an ambivalent situation where no-one is held accountable because everyone is responsible.

(ii.d) Finally, *machine solutionists* are those who argue that it may, in various degrees, be possible to hold autonomous machines responsible (Lagioia & Sartor, 2020; List, 2021; Simmler & Markwalder, 2019; Tigard, 2021a).[5] The reason is that, although AI systems are developed to carry out human ends, their exhibited level of autonomy cannot be simply reduced to human responsibility. The main problem with the strategy, however, is that it does not represent the current state of technology: nothing, both now and in the foreseeable future, suggests that computational systems could be genuine moral agents, nor that they should be treated as such.[6] Furthermore, even if it is a theoretical possibility that some AI systems may reach human-like moral agency in the future, machine solutionists fail to address the RGs of today.

(iii) *Trivialisers* are those who, unlike pessimists or solutionists, dismiss the RG as reflecting a genuine concern in the first place (Köhler, 2020; Königs, 2022; Simpson & Müller, 2016; Tigard, 2021b). Köhler (2020), for instance, argues that the responsibility of AMs is analogous to the responsible use of non-human animals as instruments (e.g., trained dogs), and hard cases can be discarded as accidents. Königs (2022), by contrast, argues that it is unclear when and whether RGs occur; and even if we believe that they do in fact occur, it is unclear why they are morally important. In a similar vein, Tigard (2021b) claims that the moral RG does not exist; at least not in a way that is already addressed by the dynamic and flexible process that the moral responsibility of emerging technology entails. A related argument is that, since the societal benefits of AMs override the potential responsibility concerns it raises, the latter can be ignored (Simpson & Müller, 2016). As might be expected, since trivialisers "solve" the RG by simply ignoring or underplaying it, the strategy has been criticized for underestimating the challenges that advanced AI systems present for existing responsibility practices (Santoni de Sio & Mecacci, 2021). And even if we grant that some of these challenges may be exaggerated, it does not offer an excuse to ignore them.

(iv) Finally, the *differentialists* encompass the contributions that, either add another facet to the RG

[4] See Oiman (2023) for a more in-depth discussion of this issue.

[5] As Tigard puts it: "While artificial moral agents cannot suffer like us, they can and should suffer the consequences of carrying out harmful behaviors. AI systems capable of functional morality might one day learn from and improve upon their unique mistakes, as a sort of reinforcement learning" (Tigard, 2021a, pp. 442–443).

debate, or provide a meta-perspective on the debate itself: e.g., by clarifying the arguments and tensions between the camps (Oimann, 2023), differentiating between different kinds of responsibility gaps (Santoni de Sio & Mecacci, 2021), or exploring RG in relation to retribution (Danaher, 2016), accountability (Chengeta, 2016), and broader issues pertaining collective and distributed responsibility (Bovens, 1998; Nyholm, 2018; Taylor, 2021; Thompson, 1980). Nevertheless, although differentialists pinpoint many relevant aspects of RGs, they do not present us with a solution. In fact, one might even say that they make the prospect of reaching a solution more difficult by introducing differences that may or may not be necessary for reaching a solution. There is, for instance, disagreement regarding the extent to which RG is connected to the "problem of many hands" (Oimann, 2023), and about the role and importance of different forms of responsibility (e.g., accountability versus attributability).

While each contribution has enriched the RG debate and disentangled its many facets, no resolution or consensus has been reached. As it stands, there seems to be no way forward that would reconcile the pessimist's moral concern with the trivialiser's optimism. Furthermore, while several of the proposed solutions are able to address *some* aspect of the complex issue, as evident in the differentialist contributions, none seems capable of cutting through the many dimensions and nuances the debate presents.

Against this backdrop, we will present our approach. As a first move, we believe that the most important as well as urgent sense of responsibility with regard to RG concerns *accountability*; i.e., being blameworthy or punishable for wrongful actions. Even if we agree with Santoni de Sio & Mecacci (2021) that there are different nuances of responsibility to be considered, we believe that the most pressing aspects of RG find a common moral core and end in accountability.[6] That is, while it may be informative to examine responsibility in terms of conditions for aretaic appraisal—the special form of accountability an elected politician has in relation to the public, or the forward-looking responsibility of promoting certain values in one's life—in cases of RGs, we first and foremost want to blame someone for their wrongful doings. In essence, when an AM's behaviour has resulted in some undesirable outcome, we want a straightforward answer to the question: who is accountable? As such, introducing additional distinctions may even be counter-productive to the RG debate, as they lead to further confusion instead of clarification.[7] Thus, if sound, the move is naturally appealing for addressing the RG for a variety of reasons: e.g., it allows one to (i) cut through the tortuous layers of the RG debate, (ii) target the core of both moral responsibility and legal accountability, and (iii) address a broad range of domains where AMs are deployed (e.g., transportation, education, and warfare).

As a second move, we believe that a promising yet overlooked way to address RG is through the lens of *rights*, and in particular, what gives entities the right to act freely in a society (i.e., without external control or supervision). A natural candidate for this purpose is to turn to the right forfeiture theory of punishment. There are, prima facie, two reasons that motivates this approach. First, one of the main issues of AMs in the military section concerns the responsibility of using LAWS in war. In the ethics of war, it is claimed that killing enemy soldiers is only permissible in self-defence. In turn, self-defence is commonly explained based on a right forfeiture theory according to which enemy

[6] Of course, the nature of moral responsibility is itself a highly contentious question in the moral responsibility debate (Shoemaker, 2011; Smith, 2012; Watson, 1996).

[7] Thus, after reaching a solution for RGs in this distilled form, we can go on and refine our solution by introducing more distinctions and particularities.

soldiers forfeited their rights by imposing a threat on other people's life.[8]By using this theory to deal with RG in general, our solution enjoys a form of cohesiveness, as it can be applied to any domain where AMs are used. Secondly, until now, RG has mainly been considered as a moral and legal issue, but by using the right forfeiture theory, we can also pinpoint the political aspects of the problem—especially in virtue of the connection made between responsibility and civil rights.

3. Responsibility before freedom

Right forfeiture is a theory of punishment that primarily deals with the question: "is it permissible to impose harm on a wrongdoer as a response to their crime?". Punishment involves treatment that seems to be impermissible and in conflict with citizens' natural rights. For example, incarceration is a prevalent form of punishment that involves temporarily stripping wrongdoers from their right to freedom. Now the question is: what justifies treating wrongdoers in ways that are not normally permissible? The answer, as indicated by the theory's name, is that by performing the wrongful action, those citizens forfeited some of their rights to not be interfered with by the state for a period of time. Then, it would be permissible for the state to punish them as a way of securing other citizens' rights and compensating the victim.[9]

But in what sense can we say that wrongdoers forfeit their rights? To answer this question, we must consider the relationship between rights and duties. In a civil society, every right presupposes the duty to acknowledge the same rights for other citizens.[10] So, as my right to property puts a restriction on others' behaviour—e.g., not taking my belongings without permission or by force—I have a duty to recognise the same right for them by restricting my behaviour accordingly. In this way, every right presupposes certain duties, and conversely, enjoying certain rights is bound to respecting certain duties. Therefore, in a society in which my rights are secured by the way of other people being bound by certain duties, I cannot retain those rights without being bound by the same duties. This is the sense in which a wrongdoer forfeits their rights. As Robert Nozick (1974) puts it:

> […] those who themselves violate another's boundaries forfeit the right to have certain of their own boundaries respected. On this view, one is not morally prohibited from doing certain sorts of things to others who have already violated certain moral prohibitions (and gone unpunished for this). Certain wrongdoing gives others a liberty to cross certain boundaries (an absence of a duty not to do it) (1974, pp. 137-138)

Now, before using the theory to address the responsibility gap, let's delve a little bit more into one of our most fundamental rights, namely being free to act in our society. It is commonly accepted that every citizen in civil society has a right to carry out their own life plans and goals in a way that is free from the interference of others. This right, as mentioned, entails certain duties for community members: a duty to recognize the rights of others and a duty to avoid actions that restrict others' ability to pursue their goals.[11] In this context, wronging another agent would amount to a failure to recognize these rights for others and, in this manner, forfeiting one's own right to freedom. This forfeiture would,

---

[8] On this topic, see (McMahan, 2004; Otsuka, 1994; Thomson, 1991).

[9] For instance, Hegel (Wood, 1990) and Locke (1689) are two prominent defenders of different forms of right forfeiture view. For a more recent discussion on the topic see (Goldman, 1982; Morris, 1991; Ross, 1930; Simmons, 1991; Wellman, 2012).

[10] There are some special exceptions to this claim, e.g., the exclusive right of the Sámi to raise reindeer.

[11] [12] For a detailed exposition of the relationship between rights and duties, see Goldman (1979).

in turn, authorise others to blame or punish the person, as a way of holding the agent responsible for her wrongdoing.

Our freedom of action in society presupposes responsibility for those actions, and one cannot be granted this right if one cannot be responsible for those actions. This means that those who are not capable of recognising these rights and cannot be held responsible—e.g., small children and people with severe mental disorders—are not granted this right. In other words, we should put responsibility *before* freedom. As Simmons (1991) writes:

> Protection under the rules is contingent on our obeying them; any rights the rules may define are guaranteed only to those who refrain from violating them (independent, of course, of unanimous agreement to alternative arrangements). Surely no one could reasonably complain of being deprived of privileges under rules he refuses to live by. (1991, p. 335)

Considering the issue from this perspective, the problem of the responsibility gap turns to a question regarding the freedom of AMs in our moral community. In order to answer the question "*who* is responsible for an AM's actions?", we must first ask "*why* should we grant AMs freedom of action while they are not capable of being held responsible". After all, to adequately carry out their function, autonomous machines need to be granted the freedom of action. Others should not interfere with their actions, for example, by stopping them or cause
disturbance in their functioning.

This question can be answered in two different ways. First, we can say that autonomous machines cannot be granted the freedom of action until they are capable of being held responsible. According to this view, we can still develop AMs, but using them would be only possible after they acquired the capacity of recognising others' rights and being held responsible. This answer, we believe, is hasty and problematic. It is not clear how long it would take for these AMs to acquire these capacities—even if we accept that they can eventually become responsible. And given the enthusiasm for AMs and their extensive benefits, it is not clear that such a solution would even be seriously considered.

There is however a better way of dealing with this issue. Instead of granting the right of non interference to the machine itself, those who use a machine to pursue their plans can transfer to the machine *their* right to act freely on their behalf. Thus, AMs operate based on their user's right to freedom of action, e.g., to achieve the user's aims. In this way, an AM's right to act freely hinges on the user's duty to recognize others' right to be free and compensate them if anything goes wrong—the user, to a certain extent, will be responsible to compensate for the victim.

To illustrate, consider X, who is the manager of a company that needs to hire a person to do data analysis on some sensitive data. X hires Y, a data analyst, to do the job and is given access to the data. Unfortunately, Y makes a mistake, and the sensitive data is leaked on the internet. In this case, X is responsible for hiring Y and giving her access to the data. X had a right of non-interference in his decision to hire someone for the job and transferred this right to Y while she was working on the data. Other members of the board could not interfere with Y's actions because she had that right. Now, when things went wrong, X is responsible for the leakage.

Of course, in this situation, Y shares a part of the responsibility while in the case of RG, the machine cannot bear any responsibility. This, however, does not mean that the user must accept all the

responsibility themself. In the case of a RG, we must acknowledge the unfortunate nature of the harm. There is no bad intention, recklessness or ignorance involved. Therefore, the kind of responsibility involved is different from cases like the aforementioned example. Responsibility for the actions of AMs is a restricted form of our normal responsibility, in the sense that it is analogous to cases of unintentional wrongdoings. In our life, there are situations in which our actions bring harm to others when we didn't mean to harm them. In these cases, just as in RGs, it seems that no ill intention or ignorance was involved; but due to bad luck, things went bad and someone got hurt. Still, in these cases, we can hold the person responsible but in this restricted sense.

Imagine you see an old man who is trying to put a piece of furniture in his car, but it is not possible to do it single-handedly. He politely asks for your help, and as a good Samaritan you help him with all good intentions. As it turns out, the man was a thief and you helped him steal someone else's property. There was no bad intentions, recklessness, or ignorance involved,[12] but still, you helped a person to steal and can be held responsible—but not as if you committed the crime yourself. If you didn't help the old man, he could not steal the couch, but in helping him, you didn't act out of bad intentions, ignorance, or recklessness. Still, you are responsible for enabling him to execute his plans. As your free actions contributed to a theft, you are required to answer to what you have freely chosen to do.

A similar case can be made concerning the use of AMs: they could not act freely in society if you didn't grant them permission to act on your behalf in reaching your goals; but still, your action did not involve bad intention or ignorance.[13] Therefore, to a certain extent, you are responsible for the actions of the machine. Note that we, unlike Taddeo & Blanchard (2022) and Champagne & Tonkens (2015)**,** do not hinge responsibility on any foreknowledge or gambit. Being responsible for AMs was already a prerequisite for being able to use them for your plans by giving them freedom to act in our society.

In summary, based on right forfeiture, we argue that in the case of a responsibility gap, it is the user—who can be a person, a company, a state, or a combination of them—who is responsible for the action of the autonomous machine. We also claimed that the user is not responsible for everything the AM does per se, but rather, for using a machine whose actions can potentially cause unfortunate results. In other words, the user's responsibility for what happened due to the machine's action would be like cases in which someone's action caused unintended bad results. The user is responsible because she granted the machine the freedom to bring about her aims. In the next section, we consider the implications of our theory for the usage of AMs.

4. Discussion

In this section, we take a more applied approach and consider the implications our theory has for the development and use of AMs and LAWs. The first aspect of the presented theory that we want to pinpoint is that it is somewhat unique in the RG literature. Our account is the only one in the debate that solely holds users—who can be a human, a company, or a state—responsible in the case of RG. Other accounts either claim that only manufacturers should be responsible, or that responsibility should be divided between users and manufacturers (including managers, designers, etc.). This feature, in turn, may give rise to an objection against our theory. For instance, it can be argued that avoiding blame and punishment are important incentives that motivate manufacturers to do as much as they can

[12] Of course, someone might here say that you were ignorant or reckless since you didn't ask the old man to prove that the old couch was his; but we believe that such objections are unrealistic in the case at hand.

[13] Considering that you had good reason to use an AM and it was relatively safe.

to create safer machines. Holding the user responsible for the actions of an AM can therefore make manufacturers negligent in their development of AMs, as it removes their motivating incentives. To that end, some authors claim that the responsibility gap has exculpatory powers which can be exploited by manufacturing companies to avoid responsibility for their faulty products. [14] As a result, they suggest that we should always hold companies responsible in cases of responsibility gaps.

However, we believe that this objection is based on a conflation which treats the responsibility gap as an excuse. This conflation might arise because, in cases of both exculpation and responsibility gaps, we are faced with a wrongful action that seemingly cannot be blamed on anyone. This may create the false impression that a responsibility gap can work as an excuse. But there is an important difference between these two cases. Blame and punishment are warranted when someone fails to satisfy certain minimal moral expectations—e.g., unnecessarily causing harm to others. Excuses work by showing that, considering all the relevant facts about the context, the culprit did not undermine any moral expectation. For instance, if I trod on someone's foot and it turns out that I have been pushed by someone else, I have an excuse because it could not be expected of me not to trod on that person's foot while being pushed. But if I did it because I was reckless or negligent, it is perfectly reasonable for others to blame me. The same is true in the case at hand: if it turns out that some harmful action performed by an AM was due to recklessness or negligence in the designing or manufacturing process, the company is responsible, and no responsibility gap arises. In this way, holding users responsible for an AM's actions is sound in contexts in which responsibility gaps typically arise; namely, when a manufacturing company has discharged its responsibility by creating a purportedly "safe machine", and provided the potential user with the information needed to decide whether to use the machine.

Moreover, our theory gives a systematic answer to all RG cases. In the literature, questions about RG in the context of autonomous vehicles (AVs) are typically considered separately from questions regarding lethal autonomous weapon systems (LAWs). For this reason, proposed solutions for the two domains may be inconsistent or even contradictory.[15] As an example, consider Sparrow's treatment of the two (Sparrow, 2007; Sparrow & Howard, 2017). In the case of self-driving cars, he claims that after a certain point, they will be significantly safer than human drivers, and for this reason, the state has a duty to enforce the usage of AMs to save citizens' lives (Sparrow & Howard, 2017). According to him: human drivers look like drunk drivers when they are compared to AVs; and since we prohibited drunk driving due to its dangers, we should ban human driving as well. In this view, holding *users* responsible for the actions of AMs would be absurd, in the same way it is unreasonable to hold someone responsible for a choice you (state) imposed on them (user). Moreover, since a lot of people have the desire to drive themselves, the transition to fully AI-administered transportation would become harder to motivate, and any obstacle that demotivates this transition should be dealt with in some way. Therefore, we should hold developers responsible because this not only solves the above problem but also motivates the designers to build safer machines.

In the case of LAWs, Sparrow takes a completely different approach and claims that we should stop

---

[14] See for example: Johnson (2015, p. 174); Santoni de Sio and Mecacci (2021, p. 1063).

[15] One might argue that these questions must be considered separately because of the significant difference between the two cases. While this may – at least prima facie – be reasonable, we believe that the ones who argue this must also show that the mentioned differences are sufficient to make such an impact for our approach. In other words, they must show why considerations that led us to one answer in the case of AVs are not valid in the case of LAWs and *vice versa*.

developing them altogether (Sparrow, 2007). However, the reasons he presents for this standpoint conflict with his views on AMs. First, he claims that the use of LAWs undermines responsibility by creating unsolvable RGs, writing: "[…] it will be unethical to deploy autonomous systems involving sophisticated artificial intelligences in warfare unless someone can be held responsible for the decisions they make where these might threaten human life"
(Sparrow, 2007, p. 74). But in the case of self-driving cars, he rejects that RGs are a problem, and instead argues that we should hold manufacturing companies responsible. Why cannot designers be held responsible in both situations; or, alternatively, that nobody can be held responsible in both? On the other hand, if in the future human drivers look like drunk drivers compared to AMs self-driving cars, how is it that we cannot say the same thing about LAWs? For instance, why it is not appropriate to say that after a certain point, human soldiers compared to LAWs look like drunk soldiers? Or, based on the same reasoning, should we ban normal human soldiers and only use LAWs?

By contrast, our theory gives the same consistent answer in both cases: the user is responsible. With regards to self-driving cars, it is the user of the car, and in the case of LAWs, it is the military leaders, presidents, or countries that are using the AM in battle. Until the proponents of other accounts justify the inconsistency in their answers in one way or another, we believe our theory has the upper hand. To the extent that being systematic and consistent is a virtue, our theory has more merits than others.

Furthermore, we believe that holding customers responsible has the additional forward-looking benefit of encouraging customers to use AMs more responsibly. This is often ignored by scholars who exclusively focus on the responsibility of manufacturers. But as the use of AMs becomes more prevalent, the responsible use of AMs becomes more critical. AMs make our lives easier and provide us with a lot of freedom and possibilities. For this reason, many of us, understandably, are excited about new technologies and eager to let machines do many of the boring tasks of everyday life. Although this enthusiasm may have many direct or indirect benefits for developing better AMs, and in this way decrease the risk of harm we are exposed to every day, it can also encourage irresponsible behaviour. Since people might see AMs as a way of avoiding responsibility, this situation can be intensified if we ignore users in discussing RGs. Moreover, holding users responsible would also make users more cautious in choosing what products to use, which indirectly motivates manufacturers to produce safer AMs; and manufacturing companies that produce safer machines gets an edge over rivals. Thus, if users are held responsible, it promotes the responsible use of AI while strengthening the motivations for companies to make safer machines.

Finally, we want to discuss how our theory stands in relation to the four camps outlined in the first section: pessimists, solutionists, trivializers, and differentialists. First, we acknowledge the moral problem raised by the pessimists (de Jong, 2020; Matthias, 2004; Roff, 2013) and claim, contrary to trivializers (Königs, 2022; Tigard, 2021b), that there is a real problem to be
addressed. We also agree with pessimists that previous ways of dealing with RGs are inadequate, but contrary to them, we believe that the issue is solvable. Furthermore, while our account partly converges with certain solutionist camps—technical, human, collective, and machine—we differ in some important respects. For example, our theory differs for human solutionists—e.g., moral gambit or "blank check" liability—in our claim that knowing in advance that something may go wrong as a result of our actions is not enough for responsibility. Essentially, we believe "gambit" (Taddeo & Blanchard, 2022) is the wrong way of framing the gap as it plays on the uncertainty, which is, irrelevant as far as moral responsibility is concerned. Starting from the "uncertainty" side of things paints a false picture that is exploited for moral purposes: it frames the RG as an epistemic challenge,

and not a moral one. Regardless of the odds, if you use an AM to promote your ends, you are responsible. Moreover, our theory avoids the scapegoating of "blank check" liability (Champagne & Tonkens, 2015), and also addresses the retributionist gap by identifying the person who is responsible. With that said, this does not mean that "control gaps" and other epistemic conditions are irrelevant for addressing RGs, but rather that: control gaps and epistemic conditions only become relevant as a consequence of user's moral responsibility. In a similar vein, while we do not believe that any technical solutions (e.g., Saxon, 2016; Hindriks & Veluwenkamp, 2023) can solve the moral issue with RGs, they can still serve to inform users of the risks involved, and help them to make an informed decision of whether to use an AM.

But we also agree with some human solutionists (Taddeo & Blanchard, 2022): that collective solutionists' way of addressing RG fails to respect the conditions under which individual agents are fairly and deservedly held responsible. Like gambit or blank checks play on uncertainty, collective solutionists play on the ambivalence of the "problem of many hands"; when many actors have contributed to an unfortunate outcome, and it becomes difficult to identify who is responsible (Oimann, 2023). Of course, the problem of many hands—like moral uncertainty— is a genuine concern; and while it may be present in RGs, it should not be conflated with the normative problem.

In contrast to machine solutionists, our solution does not depend on future technology that makes it possible for machines to take responsibility. That being said, our answer is consistent with machine solutionists: it is possible that long-distant future AMs could be capable of being held responsible for their actions, following the conditions described in §3. If so, there will be no responsibility gap. But until then, RG is a real issue that needs to be addressed. On that note, we also have to entertain the possibility that AMs may never reach the state that bears full responsibility for their actions.

Furthermore, we agree with differentialists (e.g., Santoni de Sio & Mecacci, 2021) that there are many dimensions and nuances to respect in the RG debate. However, contrary to them, we do not believe that it is necessary to import all distinctions in the theoretical debate about moral responsibility into debates about RG.[16] The main problem with the introduction of AMs in different domains—the problem that makes RGs of critical and urgent importance—is "who should be held responsible if things go wrong?". Thus, the most crucial aspect of RG is related to responsibility in the accountability sense, and by focusing on this sense of responsibility we can give a simple answer that sidesteps unnecessary complications. Ultimately, we believe that it is only against a background of a solution to *this* part of the RG that other aspects of responsibility—e.g., public and forward-looking responsibility—becomes relevant to address.

There is one account that in some respects comes close to the theory presented here. Sebastian Köhler (2020) calls AMs "minimal agents" and suggests a solution to the RG that centres on an analogy with trained dogs. According to him, trained dogs are always under the supervision of the person who uses them; therefore, if something bad happens, the user should be held responsible. If on the other hand, it turns out that there was something wrong in the training of the dog, the trainer should be held responsible. The same can be said in the case of AMs: they always act under the supervision of the user, and if some avoidable harm is caused by the machine, the user should be held responsible—supposing that there was nothing wrong in the design or manufacturing process. However, the problem with Köhler's solution is that it, unlike our proposal, only applies to supervised machines and not to machines that are completely unsupervised. While the former—supervised

[16] For a philosophical discussion of different senses of responsibility see Shoemaker (2011).

AMs—can create some weaker RGs, the stronger and most challenging cases arise in the latter, when little to no supervision occurs. In other words, Köhler only closes a small part of the RG but fails to address the stronger RG that arises from the use of unsupervised AMs. Nevertheless, he also considers cases where an AM's actions are unpredictable. But, in this case, his answer is a variant of the *human solutionists* appeal to foreknowledge (e.g., moral gambit). According to him when the action of the machine is unpredictable it:

> is users and designers who choose to use an AI to solve a problem or perform a task while *knowing* that it is impossible to predict whether and how the problem will be solved or task will be performed. […] But, all of these issues put the responsibility of designers and users square on the table, rather than interfering with their capacity for being responsible. (Köhler, 2020, p. 3137) (our emphasis)

As we argued earlier, this way of dealing with RG is problematic: just *knowing* that something may go wrong if I do an action, does not automatically make me responsible for the consequences.

5. Conclusion

To conclude, when AMs cause unnecessary harm, and this is not due to recklessness or negligence in the design and manufacturing process, we should hold responsible those who use these machines for their aims. While this might leave questions about the epistemic conditions for responsibility designations in specific cases open, it provides a basis for closing the responsibility gap. We also considered a relevant objection: holding users responsible for AMs' actions will disincentivize manufacturers to produce safer machines. We showed that this objection is wrongheaded because it treats the responsibility gap as an excuse. Finally, we also highlighted some of the strengths of our theory. First, it gives systematic answers to the different domains where AMs are developed and deployed (e.g., AMs and LAWs). Second, it encourages the responsible use of AMS, which also motivates manufacturing companies to develop safer machines. Ultimately, we believe that "responsibility before freedom" subsumes previous solutionist endeavours to "close" the responsibility gap, all while answering to the original concern raised by pessimists.

References

Amoroso, D., & Giordano, B. (2019). Who is to blame for autonomous weapons systems' misdoings? *Use and Misuse of New Technologies: Contemporary Challenges in International and European Law*, 211-232.

Behdadi, D., & Munthe, C. (2020). A normative approach to artificial moral agency. *Minds and Machines*, *30*, 195-218.

Bovens, M. A. P. (1998). *The quest for responsibility: Accountability and citizenship in complex organisations*. Cambridge university press.

Champagne, M., & Tonkens, R. (2015). Bridging the Responsibility Gap in Automated Warfare. *Philosophy & Technology*, *28*(1), 125-137. https://doi.org/10.1007/s13347-013-0138-3 Chengeta, T. (2016). Accountability gap: Autonomous weapon systems and modes of responsibility in international law. *Denv. J. Int'l L. & Pol'y*, *45*, 1.

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, *18*(4), 299- 309.

de Jong, R. (2020). The retribution-gap and responsibility-loci related to robots and automated technologies: A reply to Nyholm. *Science and engineering ethics*, *26*(2), 727-735.

Galliott, J. (2020). No Hands or Many Hands? Deproblematizing the Case for Lethal Autonomous Weapons Systems. In S. C. Roach & A. E. Eckerts (Eds.), *Moral Responsibility in Twenty-First Century Warfare: Just War Theory and the Ethical Challenges of Autonomous Weapons Systems* (pp. 155-179). State University of New York Press.

Goldman, A. H. (1979). The paradox of punishment. *Philosophy & Public Affairs*, 42-58. Goldman, A. H. (1982). Toward a new theory of punishment. *Law and Philosophy*, *1*(1), 57-76.

Hindriks, F., & Veluwenkamp, H. (2023). The risks of autonomous machines: from responsibility gaps to control gaps. *Synthese*, *201*(1), 21.

Howard, M. (2017). When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport. *Transportation Research Part C: Emerging Technologies*, *80*, 206-215. https://doi.org/https://doi.org/10.1016/j.trc.2017.04.014

Johnson, D. G. (2015). Technology with no human responsibility? *Journal of Business Ethics*, *127*(4), 707-715.

Köhler, S. (2020). Instrumental robots. *Science and engineering ethics*, *26*(6), 3121-3141. Königs, P. (2022). Artificial intelligence and responsibility gaps: what is the problem? *Ethics and Information Technology*, *24*(3), 36.

Lagioia, F., & Sartor, G. (2020). AI systems under criminal law: a legal analysis and a regulatory perspective. *Philosophy & Technology*, *33*(3), 433-465.

List, C. (2021). Group agency and artificial intelligence. *Philosophy & Technology*, *34*(4), 1213-1242.

Locke, J. (1689). *The second treatise of civil government (2015)*. Broadview Press. Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*, 175-183.

McMahan, J. (2004). The ethics of killing in war. *Ethics*, *114*(4), 693-733.

Morris, C. W. (1991). Punishment and Loss of Moral Standing. *Canadian Journal of Philosophy*, *21*(1), 53-79.

Nozick, R. (1974). *Anarchy, state, and utopia*. New York: Basic Books.

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and engineering ethics*, *24*(4), 1201-1219.

Oimann, A.-K. (2023). The Responsibility Gap and LAWS: a Critical Mapping of the Debate. *Philosophy & Technology*, *36*(1), 1-22.

Otsuka, M. (1994). Killing the innocent in self‐defense. *Philosophy & Public Affairs*, *23*(1), 74-94.

Roff, H. M. (2013). Killing in war: Responsibility, liability, and lethal autonomous robots. In *Routledge Handbook of Ethics and War* (pp. 352-364). Routledge.

Ross, W. D. (1930). *The right and the good (2002)*. Oxford University Press.
Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, *34*, 1057-1084.

Saxon, D. (2016). Autonomous drones and individual criminal responsibility. In *Drones and Responsibility* (pp. 17-46). Routledge.

Shoemaker, D. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics*, *121*(3), 602-632.

Simmler, M., & Markwalder, N. (2019). Guilty robots?–rethinking the nature of culpability and legal personhood in an age of artificial intelligence. Criminal Law Forum,

Simmons, A. J. (1991). Locke and the Right to Punish. *Philosophy & Public Affairs*, 311-349.

Simpson, T. W., & Müller, V. C. (2016). Just war and robots' killings. *The Philosophical Quarterly*, *66*(263), 302-322.

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

Sparrow, R. (2007). Killer robots. *Journal of applied philosophy*, *24*(1), 62-77. Sparrow, R., &

Taddeo, M., & Blanchard, A. (2022). Accepting moral responsibility for the actions of autonomous weapons systems—a moral gambit. *Philosophy & Technology*, *35*(3), 78.

Taylor, I. (2021). Who Is Responsible for Killer Robots? Autonomous Weapons, Group Agency, and the Military‐Industrial Complex. *Journal of applied philosophy*, *38*(2), 320-334. Thompson, D. F. (1980). Moral responsibility of public officials: The problem of many hands. *American Political Science Review*, *74*(4), 905-916.

Thomson, J. J. (1991). Self-defense. *Philosophy & Public Affairs*, 283-310.

Tigard, D. W. (2021a). Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible. *Camb Q Healthc Ethics*, *30*(3), 435-447. https://doi.org/10.1017/s0963180120000985

Tigard, D. W. (2021b). There Is No Techno-Responsibility Gap. *Philosophy & Technology*, *34*(3), 589- 607. https://doi.org/10.1007/s13347-020-00414-7

Watson, G. (1996). Two Faces of Responsibility. *Philosophical Topics*, *24*(2), 227-248.

Wellman, C. H. (2012). The Rights Forfeiture Theory of Punishment. *Ethics*, *122*(2), 371-393.

Wood, A. W. (1990). *Hegel's Ethical Thought*. Cambridge University Press.