Ethical Theory and Moral Practice (2019) 22:731–747
https://doi.org/10.1007/s10677-019-10007-9

Responsibility for Killer Robots

Johannes Himmelreich[1]

Accepted: 21 May 2019 / Published online: 11 June 2019

Abstract
Future weapons will make life-or-death decisions without a human in the loop. When such weapons inflict unwarranted harm, no one appears to be responsible. There seems to be a responsibility gap. I first reconstruct the argument for such responsibility gaps to then argue that this argument is not sound. The argument assumes that commanders have no control over whether autonomous weapons inflict harm. I argue against this assumption. Although this investigation concerns a specific case of autonomous weapons systems, I take steps towards vindicating the more general idea that superiors can be morally responsible in virtue of being in command.

Keywords Moral philosophy · Causation · Moral responsibility · Responsibility gap · Hierarchical groups · Artificial intelligence

1 Introduction

Future weapons systems will be autonomous. They might decide on their own, without delegating these decisions to a human supervisor, whether or not to engage potential targets and whether or not to inflict lethal harm.[1] The prospect of such autonomous weapons systems (AWS) has prompted various objections. Some of these objections contend that deploying AWS is morally problematic as such (cf. Burri 2017, p. 164). This paper addresses an objection of this sort, called the *responsibility gap argument* (Matthias 2004; Sparrow 2007). A responsibility gap exists when an AWS harms someone, but no one is responsible. Such a gap is problematic, or so argue proponents of the argument, because it violates "a fundamental condition of fighting a just war" (Sparrow 2007, p. 67), echoing the concern of Michael Walzer that "there can be no justice in war if there are not, ultimately, responsible men and women" (1977, p. 287).

[1]To be clear, I expect that only some, not all, future weapons systems will be autonomous. I assume that AWS decide at least in a thin sense of "decide," in which also a driverless car decides to stop when a light is about to turn red.

✉ Johannes Himmelreich
jhim@stanford.edu

[1] McCoy Family Center for Ethics in Society, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA

In this paper, I respond to this argument by taking a closer look at the notion of control.[2] The responsibility gap argument rests on the proposition that commanders are not responsible because they have insufficient or no control over whether an AWS inflicts harm. I argue against this proposition. I first assume, for the sake of the argument, that such a responsibility gap would be problematic and that responsibility requires control, and then argue that a commander exercises control in a way that suffices to partly ground their responsibility. A commander has control over a general outcome or state of affairs (what I will call a "probabilistic outcome"), even when they lack control over the AWS' particular targeting decisions (what I will call a "particular outcome"), and even when the AWS fails to carry out a command it was given. Or in other words: I argue that even if a commander has no control over a certain harm, a commander has control over a risk of harm[3]; and that a lack of control is not the reason for why responsibility gaps might arise.

The plan of this paper is as follows. I first reconstruct the responsibility gap argument with a focus on a commander of an AWS, setting aside other possible subjects of responsibility. I then argue, contrary to what the responsibility gap argument assumes, that a commander does in fact control a mission's outcome in an abstract but relevant sense, even if the commander has not authorized engagements specifically. My argument consists, in equal parts, of a definition of "control" and distinctions of things that are under a commander's control.

2 Background

To motivate the topic, consider the following hypothetical case as an example of how AWS can give rise to responsibility gaps.[4]

> A future AWS, an autonomous drone, bombs a column of enemy soldiers who have indicated their desire to surrender. The drone's commander gave orders to patrol the region and engage legitimate targets. But the drone wrongly identified the surrendering soldiers as legitimate targets.

Who is responsible in this case? Had the decision to bomb the targets been made by a human – a role that for today's remotely controlled drones is played by what some military forces call the "Joint Terminal Attack Controller" – then this person would likely be responsible. But absent a human decision maker, contends the responsibility gap argument, no one, in particular no individual, is responsible. When an AWS makes targeting decisions, then responsibility gaps will arise and the use of AWS is hence morally problematic as such (Matthias 2004; Sparrow 2007, 2016; Roff 2013). Amplified by the work of organizations such as Human Rights Watch (2012) and the Campaign to Stop Killer Robots (2017), the responsibility gap argument has enjoyed significant uptake in public deliberation.[5]

Responsibility gaps have already been addressed from different angles.[6] Some suggest that the commander may be responsible (Lin et al. 2008; Hellström 2012; Roff 2013; Nyholm

[2] In other words, I concentrate on the control condition for moral responsibility and set aside the epistemic condition (cf. Fischer and Ravizza 1998, p. 12).

[3] This claim pertains only to cases in which a commander has an actual choice, at least, between either deploying an AWS or not deploying it, such that the former but not the latter option carries risks of harm.

[4] This case should not be confused with a case due to Sparrow (2007), which I discuss towards the end of the paper.

[5] Some advocacy groups call it an "accountability gap."

[6] Responsibility may lie with developers (Lokhorst and van den Hoven 2011), politicians (Steinhoff 2013), or the AWS itself (Hellström 2012; Burri 2017, p. 73). Responsibility might be shared (Schulzke 2013; Robillard 2018), or "a new kind of ... responsibility" might be required (Pagallo 2011, p. 353).

2017), but this view has rarely been argued for in full.[7] This paper explains a commander's responsibility in terms of control, defining "control" explicitly, and thereby employing a concept that is widely believed to be closely related to moral responsibility.[8]

Before getting started, let me clarify what I mean by "responsibility." By "responsibility" and its cognates I understand what an agent acquires because of what she has done or brought about that grounds permissions of other agents to react to this agent in certain ways (Scanlon 2008, pp. 128–31; Shoemaker 2011, 2015). In this sense, responsibility is backward-looking. This contrasts with the forward-looking use of "responsibility," which refers to obligations to, for example, manage risks, perform certain actions, or produce certain outcomes. When an agent *is* responsible in this backward-looking sense of the term, then (some) others are justified in *holding* her responsible. The practice of holding an agent responsible has several aspects (Shoemaker 2011, 2015). Its foremost aspect is evaluative. The attitudes and responses involved in holding a person responsible often express or constitute a judgment not only of the person's action but of the person herself. This is responsibility as attributability, in reaction to an agent's character or her practical commitments. A person's responsibility moreover justifies taking a certain stance towards this person and forming evaluative or emotive attitudes, such as blame, praise, or resentment, as a part of this stance. This is sometimes called responsibility as accountability, in reaction to the respect and regard with which an agent acted in her relationships to others. In addition to justifying attitudinal stances, moral responsibility involves assessing and questioning the reasons the agent took to justify her actions. This aspect can be called responsibility answerability.[9]

Although these different aspects of responsibility can be distinguished, I remain largely neutral with respect to different analyses of what moral responsibility is, such as whether to be responsible just is to be blameworthy. Instead of the analytical question of what responsibility is, this paper concerns the grounding question of why someone is responsible. I assume that responsibility need not always involve accountability and blameworthiness. Instead, responsibility might involve that an agent is answerable, which is not a light matter insofar as the agent is under obligations to explain her actions and, perhaps, to even apologize for harms that ensued (cf. Burri 2017, p. 177).

[7] Santoni de Sio and van den Hoven (2018) offer an account of meaningful human control, to which my account is an alternative, as I explain below. Lin et al. (2008) as well as Roff (2013, p. 357) focus on legal instead of moral responsibility and consider the possibility that a commander is responsible only as one among many options (next to, for example, the responsibility of developers). They do not aim to offer an argument for or against a commander's responsibility neither do they develop an account for why a commander would (not) be responsible. Nyholm (2017), similar to my approach, suggests to investigate responsibility by drawing on "hierarchical models of collaborative agency, where some agents within the collaborations are under other agents' supervision and authority." But Nyholm (2017, p. 1203) admits that "a fully worked-out theory is not offered" in his paper.

[8] By contrast, Hellström (2012) rests his explanation of a commander's responsibility on the concept of autonomous power, which "denotes the amount and level of actions, interactions and decisions the considered artifact is capable of performing on its own." Unlike control, autonomous power plays no role in existing discussions of moral or legal responsibility. Yet, the account that I propose here is compatible with that of Hellström (2012) and can be seen as spelling out an alternative way of understanding the idea of autonomous power.

[9] Shoemaker (2011, 2015), as others, distinguishes these (attributability, answerability, accountability) as different forms of responsibility. I do take an official view as to whether there are different kinds or forms of responsibility or if, instead, there is only one kind of responsibility that comes in different degrees. In order to remain neutral about this issue while nevertheless incorporating Shoemaker's distinction in some form, I opt for the language of "aspects" of responsibility.

3 Responsibility Gaps

Responsibility gaps are a problem not only for AWS but, for example, also for group agents (Pettit 2007; Braham and van Hees 2011; Duijf 2018). Generally speaking, responsibility gaps seem to occur in the vicinity of *merely minimal agents* that have intentional agency but that lack moral agency. Merely minimal agents are intentional agents in the sense that they can form beliefs, decide, and act but they cannot be responsible for their actions.[10] Plausible examples of merely minimal agents are artificial agents such as AWS or computer programs, but also group agents (cf. Sparrow 2016, p. 108; Danaher 2016, p. 301; Albertzart 2017; Thompson 2018).[11]

I take it as a basic assumption that AWS are merely minimal agents (Sparrow 2007, p. 74; Hellström 2012, p. 101; Nyholm 2017, p. 1209).[12] AWS are intentional agents making decisions based on fairly complex reasoning (Sparrow 2016). To speak of computer systems as believing, intending, or deciding something will become increasingly plausible with advances in artificial intelligence. This might be either because one subscribes to a certain kind of functionalist analysis of what intentions and beliefs are, or because one takes the efficiency of attributing such mental states to these systems as sufficient for them having these states. But beyond this intentional agency, AWS are not agents, or autonomous, in any thick or moral sense, such as the sense featured in Kantian moral theories. This view about AWS as merely minimal agents seems broadly shared in the literature.[13]

To my knowledge, the literature contains no definition of "responsibility gap" in terms of necessary and sufficient conditions. I suggest that one important kind of responsibility gap occurs when no one is responsible for an action of merely minimal agents.[14] More precisely, a situation gives rise to a responsibility gap if and only if (1) a merely minimal agent does x, such that (2) no one is responsible for x; but (3) had x been the action of a human person, then this person would be responsible for x.

Let us call the first condition the *Minimal Agency* condition. This necessary condition distinguishes responsibility gaps from occurrences such as floods or landslides that do not appear morally problematic in the same way. The second condition is the *Responsibility Void* condition, which just states that, as things stand, no one can be responsible for x. The third is the *Lack of Moral Agency* condition, stating that the responsibility void would not have arisen if someone with moral agency, such as a human person, had performed the action. The last two

[10] We can understand "agency" in one of two ways. First, we can understand "agency" as a relation between an agent and an action representing *who* did *what*. This is intentional agency. Second, we can understand "agency" as a predicate representing the property of *being an agent*. Many usages of "agency" in this predicative sense often require more than standing in the agency relation.

[11] Although some argue that some group agents might be responsible and they might thereby avoid responsibility gaps (Pettit 2007; List and Pettit 2011, chap. 7; Duijf 2018).

[12] Robillard (2018, p. 707) observes that this assumption is widely shared, if only tacitly. In fact, a popular textbook on artificial intelligence (AI) defines AI as "as the study of agents" (Russell and Norvig 2010, p. viii).

[13] For example Sparrow (2016, p. 108) writes that "even if the machine is not a full moral agent, it is tempting to think that it might be an 'artificial agent' with sufficient agency, or a simulacrum of such, to problematize the 'transmission' of [the human operator's] intention."

[14] However, this understanding of "responsibility gap" seems to over-generate because it picks out actions by animals, which are another kind of merely minimal agents, as leading to responsibility gaps. This raises the question of why, if at all, responsibility gaps are morally problematic. I assume, for the sake of the argument, that responsibility gaps are morally problematic at least in the case of AWS.

conditions are often taken to suggest that there is a "deficit in the accounting books" in that no one is in fact responsible for something for which we otherwise have grounds to hold someone responsible (Pettit 2007, p. 194). This, in turn, leads to a practical problem: a "retribution gap" such that retributive urges cannot be assuaged because no culpable wrongdoers can be found (Danaher 2016).[15]

To establish that no one can be responsible, the responsibility gap argument often proceeds by elimination, ruling out each plausible candidate who could be responsible until none is left. Each candidate is shown to fail to meet a necessary condition for moral responsibility (Sparrow 2007; Roff 2013). When it comes to the commander, the arguments invoke that *responsibility requires control*. That is, if *a* is responsible for *x* then *a* must have had control over *x*. The responsibility gap argument says that the commander cannot be responsible for the bombing because they exerted insufficient or no control over it. After all, or so goes the argument, it was the AWS that decided to bomb the soldiers. Matthias (2004, p. 177) contends that "nobody has enough control over the machine's actions to be able to assume the responsibility for them." And Sparrow (2007, p. 71) writes that the commander is not responsible, since otherwise "[m]ilitary personnel will be held responsible for the actions of machines whose decisions they did not control."

In sum, the commander is not responsible because they have insufficient or no control over the outcome and the AWS is not responsible because, as a merely minimal agent, it fails to meet some other necessary condition for moral responsibility (Sparrow 2007, p. 72; Roff 2013, p. 354; Johnson and Axinn 2013, pp. 135–37; Purves et al. 2015). If we restrict plausible candidates for responsibility to only the commander and the AWS, then by elimination we get a responsibility void where no one is responsible for the outcome that the targets are bombed. But someone ought to be responsible. Had a human bombed the targets, they would normally be responsible. We face a responsibility gap.

This argument should give us some pause. Assuming that responsibility gaps are morally problematic, it is understandable that many see this argument as a strong case against developing "killer robots" or even against conducting research into artificial intelligence generally (Human Rights Watch 2012). Those who make these demands think that the responsibility argument is not only valid but that it is sound. But I think it is not sound. The commander *has* some sort of control sufficient to make them responsible. Although a commander has no control over the particular outcome and certain bombings by the AWS, a commander has control over a probabilistic outcome and hence the risk of imposing harm. I argue that this is sufficient to partly ground their responsibility and thereby defeat the responsibility gap argument.

To motivate this view, consider an analogy. As far as intentional agency is concerned, an AWS is akin to a subordinate human soldier (cf. Nyholm 2017). There is good reason to think that a commander is responsible for what their soldiers do (Walzer 1977, pp. 316–23). The United States Army even goes so far as to declare that "commanders are responsible for everything their command does or fails to do" (2014). How plausible is the idea that commanders can be morally responsible for what their soldiers do? To answer this question, we need to think about the notion of "control" in order to understand how a commander may have a mission's outcome under their control although they are not directly involved in battle.

[15] I want to register my hesitation in thinking that responsibility gaps are problematic as such. See note 14.

4 Control

My view is that there is a plausible sense of "control" on which it is true that the commander has control over something related to the bombing. Whereas some authors have suggested a view along these lines suggesting that a commander has control (Lin et al. 2008, p. 59; Hellström 2012, p. 104; Roff 2013, p. 358; Nyholm 2017; Santoni de Sio and van den Hoven 2018), others deny that the commander has control (Matthias 2004, pp. 175–77; Sparrow 2007, p. 72; Roff 2013, p. 357).[16] This divergence calls for a clearer understanding of the term. One way of understanding "control" is as follows.[17]

Robust Tracking Control. An *a* has control over whether an outcome *x* occurs if

1. there is an order *a* can give, such that
2. if *a* were to give this order, then *x* would occur (in all relevantly similar situations), and
3. if *a* were not to give this order, then *x* would not occur (in all relevantly similar situations).

Intuitively, the idea behind this definition of "control" is one of *tracking* (cf. Nozick 1981, pp. 172–85). The conditionals 2 and 3 capture this idea: some outcome *x* would occur if *a* were to give the order and it would not occur if *a* were not to give the order. The occurrence of the outcome hence tracks whether *a* gives an order. Instead of this counterfactual formulation, a closely equivalent formulation can be given in terms of probabilities. But it should be noted that to control an outcome is not just to raise the probability of its occurrence. To control an outcome, it is also required that when no order is given, the outcome does not occur.

The definition above builds in a certain degree of robustness. This is achieved by assessing the truth of the conditionals in a way that differs from the standard semantics of conditionals (Lewis 1973, p. 24; List and Menzies 2009).[18] Specifically, the first conditional requires not only that the outcome *x* occurs in the actual situation in which *a* gives the order; it is moreover required that *x* would occur in *all relevantly similar situations* in which *a* gives the order.[19] Similarly, the second conditional requires that without the order, the outcome would not have occurred in all relevantly similar situations.

Despite this robustness, robust tracking control accommodates that things can go wrong, and that control need not be "perfect." In other words, it allows for risky actions and mistakes. It allows for risky actions in that the outcomes, in contrast to events, can represent disjunctive descriptions. You can have control over an outcome *x* even if *x* is naturally described by "either E_1 or E_2 happens," where "E_1" and "E_2" are descriptions of particular events in natural language. Moreover, robust tracking control allows for mistakes in that outcomes can include consequences that are unintended.[20] Both of these points will be made clearer with examples in sections 5 and 6.

[16] For how my approach differs from these, see notes 7 and 8.
[17] I state only a sufficient condition for control because the necessary part is not needed for my argument.
[18] In the standard way, the first conditional is true already if *a* in fact gives an order and *x* occurs.
[19] As is standard with applications of such semantics for counterfactuals, the question of how "all relevantly similar situations" is defined must be set aside.
[20] This is because robust tracking control does not include a condition referring to the content of the order or to the descriptions of the outcomes, let alone the relation between the two.

Finally, a clarification on the notion of "giving an order." This notion is a convenience and not essential to the definition. Although, control is usually exercised though orders, especially in a military context, not all kinds of control involve orders. The notion of "giving an order" should be understood liberally so that AWS will be given orders in some sense, in a form that will most likely involve a description of the mission and its objectives in a way the AWS can parse.

The notion of control that robust tracking control defines differs significantly from an alternative account of meaningful human control that has recently been put forth by Santoni de Sio and van den Hoven (2018).[21] The account of Santoni de Sio and van den Hoven takes a "variety of moral input" and, more specifically, requires the AWS to be responsive to moral reasons. By contrast, robust tracking control requires that an outcome tracks whether an order has been given. Whereas the account of Santoni de Sio and van den Hoven is morally demanding especially for the AWS (requiring responsiveness to moral reasons), robust tracking control is modally demanding especially for the commander (requiring occurrence of an outcome to track an order across a range of possible worlds).[22] These two rival conceptions of control complement each other as alternative understandings of "control" and alternative responses to the responsibility gap argument. To the extent that it is harder for the commander to be in "control" in the sense of robust tracking control, understanding "control" as robust tracking control concedes more ground to proponents of the responsibility gap argument.[23] I argue that commanders can be said to be in control of something that happens as a result of their command, even under this relatively demanding notion of control.

Let us now consider a case that is simpler than the case above of the AWS that wrongly identifies surrendering soldiers as legitimate targets. I wish to use the following simpler case to illustrate the point that a commander can be in control of what an AWS does on their command.

> **Command 1.** A commander orders an AWS to bomb a certain enemy compound at a specific time. The AWS bombs the compound.

The AWS executes the order that the commander gave and the compound is bombed. Assuming that the AWS functions reliably, the compound would be bombed in all relevantly similar situations in which the commander gives this order to bomb the compound. Furthermore, the compound would not be bombed unless the commander gives the order to do so. Hence, the commander has control over this outcome. There is no second commander who is in charge of the region in which the compound is located. More generally, as an idealized model of a military organization, I assume that the individuals on each level of the military hierarchy have domains of control that do not overlap.[24]

[21] Nevertheless, there are broad similarities between the account of Santoni de Sio and van den Hoven and my account. First, both accounts are concerned with the same issue: the relation that partly grounds agents' moral responsibility. Second, both accounts formulate control as tracking following Nozick (1981, pp. 172–85).

[22] Relatedly, the account of Santoni de Sio and van den Hoven is modelled after what Fischer and Ravizza (1998) call "guidance control," whereas robust tracking control is modelled after what Fischer and Ravizza call "regulative control."

[23] Fischer and Ravizza (1998) argue that instead of the relatively demanding notion of regulative control, on which robust tracking control is modelled, only the weaker notion of guidance control is necessary for responsibility.

[24] This sets aside the so-called overdetermination problem to which definitions in terms of counterfactual conditionals are notoriously susceptible.

With this case we can already clear up one misunderstanding surrounding the responsibility gap argument. Command 1, paired with the above definition of "control," demonstrates that it is not true as a general rule that a commander has no or insufficient control over what an AWS does. In Command 1 the commander has control over whether the target is bombed. It might be said that the commander is in some sense the agent of the bombing even though they do not perform the bombing.

5 Authorization

Admittedly, the case of Command 1 differs in many ways from the original case above, so let us add a complication. Future AWS will make decisions about targeting and engagement without a commander giving their explicit authorization for particular bombings. The commander will give orders beforehand but will otherwise be out of the loop. Although holding permanent communication with the AWS might be possible, doing so has operational disadvantages (cf. US Department of Defense 2012; Sparrow 2016).

5.1 Responsibility for a Bombing

Some argue that control requires that an AWS does not move "from one 'attack' to another without so being ordered by a human to do so" (Roff and Moyes 2016, p. 5). Before this backdrop, that a commander does not explicitly authorize a bombing, appears to be a problem. The next case describes such a bombing without particular authorization.

> **Command 2.** A commander orders an AWS to patrol a large region and engage legitimate targets. During the mission, communication is not maintained. The AWS identifies a potential target, which can only be engaged immediately. The AWS takes the target to be legitimate and engages it. The target turns out to be a legitimate target.

In this case, according to the definition of "control" above, the commander does not have control over whether this particular target is bombed. This is because it is not the case that this particular target would be bombed if the commander were to give the order. Instead, there plausibly are similar situations in which the commander gives the order, but the AWS decides against bombing this particular target in favor of bombing some other target, or no target at all. Hence, this particular bombing does not track the order.

Yet the commander seems to have control over something – but over what? At the least, even if the commander does not have control over which particular targets are bombed, they have control over whether or not they give an order. By way of this action, they control whether some targets *might* be bombed. Although the commander has no control over any *particular* bombing, the commander has control over a more abstract outcome of whether some targets might be bombed. Along these lines, two outcomes can be distinguished. The AWS is deployed in both outcomes but beyond this deployment, the two outcomes differ in what results.

> **Outcome *A*:** this particular target is bombed.
> **Outcome *B*:** some target is bombed or no target is bombed.

A is a special case of *B*. Let each outcome be represented by a set of possible worlds. The set that represents outcome *A,* that the AWS is deployed and this *particular* target is bombed, is a strict subset of the set that represents outcome *B,* that the AWS is deployed and *some* target is bombed. I will call outcome *A* a *particular outcome*, in contrast to outcome *B*, which I call a *probabilistic outcome*. This idea of probabilistic outcomes is by no means new. Probabilistic outcomes as I defined them here are front and center in probability and decision theory.

When we grant the intelligibility of probabilistic outcomes, then, despite a lack of explicit authorization, there is an outcome over which the commander has control. If the commander were to give the order, then some targets might be bombed; otherwise, no targets would be bombed. As far as the condition that *responsibility requires control* is concerned, there is some outcome for which the commander may be responsible: outcome *B*.

5.2 A Trilemma

But this conclusion that the commander is responsible for outcome *B* might seem wrong. You might object that the commander is in fact responsible for outcome *A* that this *particular* target is bombed. Yet, insofar as the commander has control only over outcome *B*, and insofar as responsibility requires control, the commander can be responsible only for outcome *B*. Before addressing this objection head-on, I find it helpful to see that this objection proceeds on the following trilemma. One of the following three claims must be false because together they are inconsistent: (1) The commander does not have control over outcome *A*. (2) Responsibility requires control. (3) The commander is responsible for outcome *A*.

This trilemma brings out a central part of the dialectic of the responsibility gap argument. It is a trilemma because, in the face of inconsistency, something has to give, yet you can respond to the responsibility gap argument in any of at least three different ways. By rejecting any of the three claims you open a door towards an escape route from the responsibility gap argument.

Denying any of the first two claims accommodates the objection and concedes that the commander is actually responsible for outcome *A* (the third claim of the trilemma). But, more broadly, each of the three escape routes may seem plausible on reflection. In fact, each of the three escape routes has been argued for in that each of the three claims of the trilemma has been denied in existing literature.

The first claim of the trilemma is denied by any view that counts the commander as having control over outcome *A*. A view to this effect can be found in Fischer and Ravizza (1998), who adopt a definition of "control" that is in some ways weaker than the one I have offered above.[25] On their view, the commander does have control over outcome *A* and can hence be responsible for outcome *A*.

The second claim of the trilemma is denied from at least three sides. First, proponents of resultant moral luck can be read as denying that responsibility requires control. They contend that an agent can be responsible for how things turn out even if how things turn out is beyond the agent's control. Proponents of resultant moral luck hence reject the basic premise of the responsibility gap argument that responsibility requires control. This makes room for the claim that the commander is responsible for outcome *A*.

[25] Fischer and Ravizza (1998) distinguish between guidance control and regulative control and argue that only guidance control is necessary for moral responsibility. When "control" is understood as guidance control the commander seems to have control over outcome *A*. See also Santoni de Sio and van den Hoven (2018).

Second, proponents of a rational-relations view contend that the things for which an agent is responsible reflect something about the agent and that the relation between an agent and the thing for which the agent is responsible is one of rational reflection, not one of control (e.g. Wolf 1993; Smith 2005). Hence, proponents of a rational-relations view of moral responsibility will also reject the basic premise of the responsibility gap argument that responsibility requires control.[26] Moreover, they will likely argue that the commander is responsible for outcome *A* insofar as this outcome reflects the commander's judgements and attitudes.

Third, proponents of what is called tracing theories of moral responsibility (or derivative moral responsibility) contend that an agent can be responsible for an outcome over which they did not have control if this outcome can be traced back to a "benighting action" over which the agent had control (e.g. Smith 1983; Ginet 2000; Montminy 2018). Because proponents of tracing theories hence contend that agents can be responsible for consequences beyond their control, they deny that responsibility requires control. They would argue that the commander can be responsible for outcome *A* insofar as outcome *A* can be traced back to (or derives from) something that was under the commander's control.[27]

Of course, each of these views raises deep questions. How should "control" best be understood? Should we accept resultant moral luck? Is moral responsibility grounded in tracing or a rational relation and not in voluntary control? How should each of these relations be understood? But the point of this short overview is only to illustrate that the existing literature on moral responsibility already provides several ways of avoiding the responsibility gap argument.

5.3 Response to the Objection

Let me now return to the objection that the commander should be responsible for outcome *A* and not only for outcome *B*. The view that I defend here is that there are good reasons to deny the intuition that the commander is responsible for outcome *A*. In terms of the trilemma, I reject the third claim. Related positions in the literature are put forward by proponents of internalist theories of moral responsibility according to which agents are responsible only for things such as their willings, attitudes, or their quality of will (e.g. Scanlon 2015, p. 96; Khoury 2018).[28] This line of response is attractive because it grants to the proponent of the responsibility gap argument the basic assumption that responsibility requires control as well as the relatively strong definition of "control" used above.[29] At the same time, this third line of response needs to explain why we should discount the intuition that the commander is responsible for outcome *A*.

[26] They might argue that responsibility requires rational control. But they reject that responsibility requires volitional control, which is the notion used in the responsibility gap argument.

[27] Insofar as a proponent of a tracing theory distinguishes between direct responsibility (for things directly under an agent's control) and derivative responsibility (for things traceable to things under an agent's control), a version of the responsibility gap argument returns: Commanders are only derivatively but not directly responsible for what an AWS does. But if this is a problem at all, it has little to do with AWS. On a tracing theory, all responsibility is derivative responsibility. I am grateful to an anonymous referee for pressing me to clarify this point.

[28] For the purposes of this paper, I do not side with the proponents of this view. Instead, I develop an independent response that is compatible with much of what internalists contend (e.g. that investigations looking for the specific objects of responsibility are somewhat irrelevant) although my response also denies a central internalist claim (that agents are only responsible for things such as their willings, attitudes, or their quality of will).

[29] Internalists do not always accept that responsibility requires control.

Before responding to the objection, let us look first at the case in favor of the objection, that is, at the case in favor of the view that the commander is responsible for the particular outcome *A*. This view finds support in two considerations. First, there might be a basic intuition that the commander is responsible for the bombing. Second, there is evidence from language use and assertions such as "the commander is responsible for the bombing," which seem to suggest that the commander is responsible for outcome *A*. I briefly address each of these considerations in turn.

First, we should distinguish between the normative and the metaphysical content of the intuition that the commander is responsible for the bombing. Whether the commander is blameworthy and hence accountable is a normative issue. By contrast, which action or outcome the commander is accountable for, or what explains this accountability or blameworthiness is an issue in the metaphysics of responsibility. Similarly, whether a commander is answerable is a normative issue. What action, event or outcome the commander is answerable for is a metaphysical issue. Although I grant that our intuition is clear with respect to the normative issues, I doubt that the intuition is clear about the metaphysics of responsibility. Internalists would argue that the intuition might just be that the commander is responsible *simpliciter* or that the commander is responsible for *something* (cf. Zimmerman 2002, pp. 567–70; Khoury 2018). This is their way of emphasizing the normative content and discounting the metaphysical content of the intuition. I prefer to discount the metaphysical content of the intuition by offering a reinterpretation which seems particularly plausible when we focus on answerability. It seems plausible to demand of the commander to justify why they took a risk by giving an order. This, after all, seems to be what they did. The view I have given so far can well accommodate this idea. As far as the requirement of control is concerned, the commander can be responsible because of the probabilistic outcome that was under their control. In this way, the view can account for the normative issues that the intuition conveys: the commander is responsible, or answerable, for taking a risk.

Second, against the evidence from language use, it should be observed that we often talk of someone being responsible for *taking a risk*. The idea that agents can be responsible for taking risks is naturally captured by responsibility for probabilistic outcomes. Thinking of agents as having control over a probabilistic outcome can make sense of their responsibility. For an illustration, consider the following case.

> **Random Killing** A killer has two victims at his mercy. The killer can operate a machine that randomly kills exactly one of the two victims. The killer operates the machine and victim 1 is killed.

The killer has no control over the particular killing of victim 1. They only have control over the probabilistic outcome that victim 1 *or* victim 2 is killed. Analogously to Command 2, we can distinguish between two outcomes, a particular outcome and a probabilistic outcome.

> **Outcome *A*:** victim 1 is killed.
> **Outcome *B*:** victim 1 is killed or victim 2 is killed.

It seems undeniable that the killer is responsible. But are they responsible for outcome *A* or are they responsible for outcome *B*? I think we should say three things in response. First, the killer is responsible, and their responsibility is partly grounded in their control over outcome *B*. With this response, we uphold the principle that responsibility requires control. Second, the idea that

the killer's responsibility is grounded in their control over outcome *B* is consistent with intuitions about what assessments and treatments of the killer are appropriate. With this response, we salvage the normative content of our intuitions, and hence much of what we mean, by saying "the killer is responsible for victim 1's death." Finally, although statements such as "the killer is responsible for victim 1's death" refer to outcome *A* in their surface structure, their truth might still be consistent with the idea that the killer's responsibility is grounded in their control over outcome *B*.[30] These considerations should allow us to significantly discount the force of the intuition that the commander is responsible for outcome *A*.

Of course, there is an important disanalogy between the killer and the commander. Whereas we can suppose that the commander has good intentions, the killer has clearly bad intentions – as much is suggested by calling them a "killer." The reason why the killer is responsible, you might argue, is because they had bad intentions. This raises the question of whether the commander can be responsible even though they, unlike the killer, did not have bad intentions. This is the question I turn to next.

6 Success

In the case that putatively gives rise to a responsibility gap, what happened differed from what the commander hoped to happen. The objective of the mission was to kill legitimate targets. The soldiers had surrendered and were thus illegitimate targets, but the AWS killed the soldiers nevertheless. The fact that the outcome of the mission contradicted the mission's objective – or, in other words, that the outcome contradicted what the commander ordered, hoped, or intended to happen – is another salient feature of the original case to which I now return.[31]

Some might argue that this feature of the case speaks against the commander's responsibility. If commander is not responsible because they ordered the opposite of what happened, the responsibility gap argument bites again. The idea behind this response is that responsibility requires success, that is, that what happens does not contradict the agent's plans or intentions. But in cases of individual action, this idea seems plainly implausible.[32] Although the situation is more complex in the case of acting on orders, it seems that commanders can be responsible if a mission turns out differently from what they hoped for, ordered or intended. Consider the following case.

> **Failed Rescue** A climber is injured. If they are not rescued, they will die. However, the rescue is risky, and everyone knows this. A mountain rescue commander orders their rangers to attempt the rescue. The rescue fails. The climber and one ranger die.

This case is similar to the case of AWS in that it describes a failed mission in the sense that what happened contradicts the objective of the mission and what the commander hoped for. But it seems that the commander is responsible even if they are not blameworthy. The commander can be called on to answer for their decision to take the risk in attempting the rescue. This captures the idea that "we must answer for the harms that we cause even if we

[30] It depends on the semantics of such responsibility statements.
[31] A mission can be successful (its objective is achieved), unsuccessful (something results that contradicts the mission's objective), or neither successful nor unsuccessful (in all other cases, such as the mission being aborted).
[32] Suppose the killer in Random Killing hopes to kill victim 2 but victim 1 is killed instead. The fact that the outcome contradicts the killer's intention is not a reason against their responsibility.

cause them […] through non-culpable accident, inadvertence, or mistake" (Duff 2009, p. 305). Answerability might also involve an obligation of the commander to apologize to those who were harmed as a result of the commander's decision (cf. Burri 2017, p. 177). Of course, a commander might not only be answerable but also blameworthy. But this depends on more details of the Failed Rescue case than I can go into here. Among other things, whether the commander is blameworthy will depend on what evidence was available to them, whether they decided to take risks in a way consonant with their role obligations, and several other factors. If the commander of the mountain rescue mission was negligent or reckless in making their decision, they are blameworthy.

7 Responsibility

So far, I have argued for three ideas. First, a commander may have control over an outcome that they do not bring about themself. Second, even if the *particular* outcome is not under their control, a commander may have control over a probabilistic outcome, which partly grounds their moral responsibility. Third, a commander may be responsible for this outcome even if things did not turn out as they hoped. Each of these features can be found in the original case, in which the AWS bombs soldiers who indicated their desire to surrender.

According to the responsibility gap argument, the commander is not responsible for the bombing because the commander does not have control over the bombing. We can now see that this is a misconception insofar as there is an outcome over which the commander has control. Again, we can distinguish two outcomes.[33]

Outcome *A*: these particular targets are bombed.
Outcome *B*: some targets are bombed or no targets are bombed.

As in the case of Command 2, the commander of the AWS has control over outcome *B* but not over outcome *A*. The analogy to cases involving risk, such as Random Killing or Failed Rescue, suggests that this limited control suffices. Even in cases of risk, responsibility is plausibly partly grounded in an agent's control.[34] The necessary condition that responsibility requires control can hence be met. Finally, the original case has the additional complication that the actual outcome contradicts what the commander hoped for. But here the case of Failed Rescue suggests that this does not stand in the way of the commander's responsibility.

So far, I concentrated on necessary conditions for responsibility. To argue that the commander is in fact responsible, a sufficient condition for responsibility is needed. Here is not the place to expound a theory about when someone is responsible. But are there other necessary conditions for responsibility that the commander fails to meet?

Moral responsibility classically requires not only control but also certain epistemic states (cf. Fischer and Ravizza 1998, pp. 12–13). I have argued that the commander meets a necessary condition of control, but the commander might not meet the epistemic condition. Some proponents of the responsibility gap argument suggest that the commander is not

[33] Although omitted in their description, the AWS is deployed in each of these.
[34] The claim is *not* that how things turn out makes a difference to an agent's responsibility. In this respect my claim differs importantly from claims defended by proponents of resultant moral luck.

responsible because they lack foresight (Roff 2013, p. 357; Danaher 2016, p. 305). Replacing control with foreseeability, awareness, and the like is one means of reestablishing the conclusion of the responsibility gap argument. A commander might not be responsible because they fail to meet an epistemic condition that is necessary for moral responsibility.

I am not convinced that such a shift to epistemic conditions of moral responsibility can give us a sound argument (Sparrow 2007, p. 70). Could the commander have foreseen that the AWS might bomb illegitimate targets? It seems plausible that the epistemic standards to which an agent is held rise with the stakes. Moreover, the epistemic standards also plausibly depend on role obligations especially when your role involves making decisions on behalf of or affecting others. Although determining these standards merits a paper of its own, it seems plausible that the possibility that illegitimate targets are bombed is foreseeable. Even if AWS were less prone to mistakes than human soldiers or long-range weapons, no one – neither human nor AWS – is perfect when it comes to distinguishing legitimate from illegitimate targets. Mistakes might happen. This insight is trivial enough for anyone to foresee. Thus, a commander is likely to meet the epistemic condition of foreseeability.[35]

That mistakes will happen is foreseeable not only in principle but also in practice. As a psychological fact, any commander is likely aware of the possibility that things can go wrong. Accordingly, just as they would with any new technology, military forces will subject AWS to tests to determine their reliability before deploying them (cf. US Department of Defense 2012). Following common military practice, commanders will be made aware of the probability of malfunctions. Hence, a commander is likely to meet the epistemic conditions for responsibility. To be clear, this is not a conclusive argument for the commander's responsibility but it suggests that an alternative responsibility gap arising from epistemic conditions of responsibility is less likely than one might think (cf. Burri 2017, p. 177).

The AWS in the original case at the beginning of this paper commits an *error of belief*. That is, the AWS intends to kill only legitimate targets but has a false belief about which targets are legitimate. By contrast, we can also imagine that an AWS goes rogue and thereby shows an *error of intent* by intending to kill illegitimate targets. To the extent that it is possible for AWS to form malicious intentions, some might worry that errors of intent lead to a distinct form of responsibility gap. In fact, the literature on the responsibility gap sometimes discusses a case that can be read in this way (cf. Sparrow 2007, p. 66). What can be said about responsibility gaps arising from errors of intent?

The arguments I have given above apply broadly also to errors of intent. As far as the commander's control is concerned, it does not seem to matter whether the AWS believed the targets to be legitimate or not. Just as there is a risk that the AWS commits an error of belief, the risk that the AWS commits an error of intent is a risk that the commander decides to take. However, as before, if the control condition does not stand in the way of the commander being responsible, an interlocutor may turn to epistemic conditions. The question is then whether it was foreseeable that the AWS would commit an error of intent. The considerations given above that a malfunction is foreseeable were neutral about the reasons and the way in which an AWS deployment might go wrong. Hence, the considerations should apply mutatis mutandis to errors of intent.

[35] Likewise, Sparrow (2007, p. 70) argues that mere unpredictability of AWS is no sufficient reason that the commander is not responsible. He writes: "If the autonomy of the weapon *merely* consists in the fact that its actions cannot always be reliably predicted … then [e]mploying AWS …is like using long-range artillery. … [R]esponsibility for the decision to fire remains with the commanding officer."

8 Conclusion

Although AWS give rise to several serious moral quandaries, the responsibility gap is not one of them. When an AWS harms someone, the commander may be responsible – at least as far as this concerns the condition that the commander must be in control. I began by reconstructing the argument for the responsibility gap, detailing how neither the commander nor the AWS appears responsible for an illegitimate bombing. The AWS is not responsible because it is not a moral agent. The commander is not responsible because the bombing was not under their control, or so goes the argument. But despite its initial plausibility, this argument is not sound.

When we define "control" carefully and distinguish between particular and probabilistic outcomes, we can see that a probabilistic outcome that involves the bombing is in fact under the commander's control. In this way, the situation of a commander resembles that of any agent who is deciding under risk. With the help of a trilemma, I have illustrated that, beyond the solution I suggest here, there are in fact several ways to avoid the responsibility gap. I concentrated on defending the view that the commander is responsible in a serious way, even if they are not responsible for the bombing as such.

However, control is at most a partial ground of moral responsibility. Moreover, the discussion of control was restricted to the case of commanders giving orders; and I understood "responsibility" as answerability and not only as blameworthiness or accountability. Naturally, this paper hence had to leave several questions unanswered. Specifically, under what conditions can a commander also be blameworthy? How improbable can a particular outcome be while still part of a probabilistic outcome? In contrast to an individual commander giving orders, what difference does it make that a military organization is in fact a complex structure of collaboration and collective decision making? These are important questions that need to be addressed. In this way, my argument that responsibility gaps are not what makes AWS morally problematic stands at the beginning and not at the end of investigations into the morality of AWS.

Acknowledgements I have benefitted from presentations and discussions of this paper at the London School of Economics, the Australian National University, the Graduate Reading Retreat of the Stockholm Centre for the Ethics of War and Peace, the Future of Just War conference in Monterey, the Humboldt University Berlin, the University of Sheffield, and the Frankfurt School of Finance & Management. I am also grateful for conversations with and/or comments by Gabriel Wollner, Christian List, Susanne Burri, Helen Frowe, Ying Shi, Seth Lazar, Matthew Adams, Sebastian Köhler, and Christine Tiefensee, as well as two anonymous referees for this journal.

References

Albertzart M (2017) Monsters and their makers: group agency without moral agency. In: Reflections on ethics and responsibility. Springer, Cham, pp 21–35

Braham M, van Hees M (2011) Responsibility voids. Philos Q 61:6–15. https://doi.org/10.1111/j.1467-9213.2010.677.x

Burri S (2017) What is the moral problem with killer robots? In: Strawser BJ, Jenkins R, Robillard M (eds) Who should die. Oxford University Press, Oxford

Campaign to Stop Killer Robots (2017) The problem. http://www.stopkillerrobots.org/the-problem/. Accessed 21 Feb 2017

Danaher J (2016) Robots, law and the retribution gap. Ethics Inf Technol 18:299–309. https://doi.org/10.1007/s10676-016-9403-3

Duff RA (2009) Strict responsibility, moral and criminal. J Value Inq 43:295–313. https://doi.org/10.1007/s10790-009-9183-7

Duijf H (2018) Responsibility voids and cooperation. Philos Soc Sci forthcoming 48:434–460. https://doi.org/10.1177/0048393118767084
Fischer JM, Ravizza M (1998) Responsibility and control: a theory of moral responsibility. Cambridge University Press, Cambridge
Ginet C (2000) The epistemic requirements for moral responsibility. Noûs 34:267–277. https://doi.org/10.1111/0029-4624.34.s14.14
Hellström T (2012) On the moral responsibility of military robots. Ethics Inf Technol 15:99–107. https://doi.org/10.1007/s10676-012-9301-2
Human Rights Watch (2012) Ban "killer robots" before It's too late. In: Human Rights Watch. https://www.hrw.org/news/2012/11/19/ban-killer-robots-its-too-late. Accessed 28 Oct 2015
Johnson AM, Axinn S (2013) The morality of autonomous robots. J Mil Ethics 12:129–141. https://doi.org/10.1080/15027570.2013.818399
Khoury AC (2018) The objects of moral responsibility. Philos Stud 175:1357–1381. https://doi.org/10.1007/s11098-017-0914-5
Lewis D (1973) Counterfactuals. Wiley-Blackwell, Oxford
Lin P, Bekey G, Abney K (2008) Autonomous military robotics: risk, ethics, and design. California Polytechnic State University
List C, Menzies P (2009)Non-reductive physicalism and the limits of the exclusion principle. J Philos 106:475–502
List C, Pettit P (2011) Group agency: the possibility, design, and status of corporate agents. Oxford University Press, Oxford
Lokhorst G-J, van den Hoven J (2011) Responsibility for military robots. In: Patrick L, Keith A, Bekey GA (eds) Robot ethics. The MIT Press, Cambridge
Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6:175–183. https://doi.org/10.1007/s10676-004-3422-1
Montminy M (2018) Derivative culpability. Can J Philos:1–21. https://doi.org/10.1080/00455091.2018.1441361
Nozick R (1981) Philosophical explanations. Harvard University Press, Cambridge
Nyholm S (2017) Attributing agency to automated systems: reflections on human–robot collaborations and responsibility-loci. Sci Eng Ethics 24:1–19. https://doi.org/10.1007/s11948-017-9943-x
Pagallo U (2011) Killers, fridges, and slaves: a legal journey in robotics. AI & Soc 26:347–354. https://doi.org/10.1007/s00146-010-0316-0
Pettit P (2007) Responsibility incorporated. Ethics 117:171–201
Purves D, Jenkins R, Strawser BJ (2015) Autonomous machines, moral judgment, and acting for the right reasons. Ethical Theory Moral Pract 18:851–872. https://doi.org/10.1007/s10677-015-9563-y
Robillard M (2018) No such thing as killer robots. J Appl Philos 35:705–717. https://doi.org/10.1111/japp.12274
Roff HM (2013) Responsibility, Liability, and Lethal Autonomous Robots. In: Allhoff F, Evans N, Henschke A (eds) Routledge handbook of ethics and war: just war theory in the 21st century, Routledge, London, p 352
Roff HM, Moyes R (2016) Meaningful human control, artificial intelligence and autonomous weapons. In: Briefing paper prepared for the informal meeting of experts on lethal autonomous weapons systems, UN convention on certain conventional weapons. p 2
Russell SJ, Norvig P (2010) Artificial intelligence: a modern approach. Prentice Hall
Santoni de Sio F, van den Hoven J (2018) Meaningful human control over autonomous systems: a philosophical account. Front Robot AI 5. https://doi.org/10.3389/frobt.2018.00015
Scanlon T (2008) Moral dimensions: permissibility, meaning, blame. Harvard University Press, Cambridge
Scanlon T (2015) Forms and conditions of responsibility. In: Clarke R, McKenna M, Smith AM (eds) The nature of moral responsibility: new essays. Oxford University Press, Oxford
Schulzke M (2013) Autonomous weapons and distributed responsibility. Philos Technol 26:203–219. https://doi.org/10.1007/s13347-012-0089-0
Shoemaker D (2011) Attributability, answerability, and accountability: toward a wider theory of moral responsibility. Ethics 121:602–632
Shoemaker D (2015) Responsibility from the margins. Oxford University Press, Oxford
Smith H (1983) Culpable ignorance. Philos Rev 92:543–571. https://doi.org/10.2307/2184880
Smith AM (2005) Responsibility for attitudes: activity and passivity in mental life. Ethics 115:236–271. https://doi.org/10.1086/426957
Sparrow R (2007) Killer Robots. J Appl Philos 24:62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x
Sparrow R (2016) Robots and respect: assessing the case against autonomous weapon systems. Ethics Int Aff 30:93–116. https://doi.org/10.1017/S0892679415000647
Steinhoff U (2013) Killing them safely: extreme asymmetry and its discontents. In: Strawser BJ (ed) Killing by remote control: the ethics of an unmanned military. Oxford University Press, Oxford
Thompson C (2018) The moral Agency of Group Agents. Erkenn 83:517–538. https://doi.org/10.1007/s10670-017-9901-7

US Department of Defense (2012) Autonomy in weapon systems
US Department of the Army (2014) Army regulation 600–20: Army command policy
Walzer M (1977) Just and unjust wars: a moral argument with historical illustrations. Basic Books, New York
Wolf S (1993) Freedom within reason. Oxford University Press, Oxford
Zimmerman MJ (2002) Taking luck seriously. J Philos 99:553. https://doi.org/10.2307/3655750