Ethics and Information Technology (2022) 24: 36
https://doi.org/10.1007/s10676-022-09643-0

ORIGINAL PAPER

Artificial intelligence and responsibility gaps: what is the problem?

Peter Königs[1]

Accepted: 9 February 2022 / Published online: 24 August 2022

Abstract
Recent decades have witnessed tremendous progress in artificial intelligence and in the development of autonomous systems that rely on artificial intelligence. Critics, however, have pointed to the difficulty of allocating responsibility for the actions of an autonomous system, especially when the autonomous system causes harm or damage. The highly autonomous behavior of such systems, for which neither the programmer, the manufacturer, nor the operator seems to be responsible, has been suspected to generate responsibility gaps. This has been the cause of much concern. In this article, I propose a more optimistic view on artificial intelligence, raising two challenges for responsibility gap pessimists. First, proponents of responsibility gaps must say more about when responsibility gaps occur. Once we accept a difficult-to-reject plausibility constraint on the emergence of such gaps, it becomes apparent that the situations in which responsibility gaps occur are unclear. Second, assuming that responsibility gaps occur, more must be said about why we should be concerned about such gaps in the first place. I proceed by defusing what I take to be the two most important concerns about responsibility gaps, one relating to the consequences of responsibility gaps and the other relating to violations of jus in bello.

Keywords Artificial intelligence · Responsibility · Jus in bello

Introduction

Recent years have seen rapid progress in artificial intelligence and in the development of autonomous systems that rely on artificial intelligence. Some of the autonomous systems already in use or about to be developed include self-driving cars and other autonomous vehicles, autonomous weapons systems, work robots, as well as medical and legal AI systems. Autonomous systems hold the promise of performing many tasks with greater speed, accuracy and reliability than humans or conventional machines, thereby reducing costs, boosting productivity, increasing safety and liberating us from routine work.[1] On the downside, critics have noted that it is difficult to assign responsibility for what an autonomous system does, especially in case of an accident or other adverse events. The behavior of an autonomous system that uses complex algorithms is too autonomous and too unpredictable for any of the human agents who have contributed to it to be responsible for it, or so it has been alleged. This has led scholars to speak of a 'responsibility gap' (Danaher, 2016; Matthias, 2004) and to discourage the use of autonomous systems on this ground, especially of autonomous weapons systems (Sparrow, 2007, 2016).[2] A lively and ongoing debate is revolving around how to address this problem.[3]

The project of this paper is to present a more optimistic view on artificial intelligence by dispelling such concerns about responsibility gaps. I proceed by formulating two challenges that those sounding the alarm about alleged gaps in responsibility must address. Responsibility gap pessimists, as I will call them here, must establish two things: First,

✉ Peter Königs
p.koenigs@fs.de

[1] Frankfurt School of Finance & Management, Frankfurt, Germany

[1] See Danaher, 2019.

[2] To be precise, Danaher identifies a *retribution* gap, which is a special case of a responsibility gap.

[3] For general discussions of the responsibility gap, refer to Baum et al., 2022; Chomanski, 2021; Danaher, 2022; Himmelreich & Köhler, 2022; Gunkel, 2020; Johnson, 2015; Köhler et al., 2018; Köhler, 2020; Kraaijeveld, 2020; Nyholm, 2018; Santoni de Sio & Mecacci, 2021; Tigard, 2021. Notable contributions on the responsibility gap specifically in the military context include, but are not limited to, Burri, 2018; Hellström, 2013; Himmelreich, 2019; Leveringhaus, 2018; Noorman & Johnson, 2014; Purves, Jenkins, & Strawser, 2015; Robillard, 2018; Simpson & Müller, 2016; Taylor, 2021. Contributions in the Topical Collection of *Philosophy & Technology* on 'AI and Responsibility' are referenced in this paper, but they appeared too late to be adequately considered here.

they must give us reason to believe that responsibility gaps may plausibly emerge when autonomous systems are used. Second, they must explain why the existence of such gaps is problematic. The two challenges presented in this paper concern these two components of responsibility gap pessimism. The challenge I formulate in the first part of the paper (Sect. 2) concerns the emergence of responsibility gaps. I will not categorically deny the existence of responsibility gaps, but I will suggest that the nature and emergence of responsibility gaps has remained somewhat opaque.[4] After proposing an account of what responsibility gaps are, I will suggest that there are situations in which we can be confident that responsibility gaps do *not* emerge. This raises the question in which situations they *do* emerge. Until believers in responsibility gaps specify the circumstances in which they occur, responsibility gaps remain somewhat of a philosophical mirage. This is the first challenge.

In the second part of the paper (Sects. 3 and 4), I will argue that responsibility gaps, assuming they exist, are just not that problematic.[5] I will proceed by engaging with what I take to be the two most serious worries about responsibility gaps. The first worry is that responsibility gaps would have disastrous consequences, as they would incentivize harmful behavior. The second worry is that the use of intelligent military systems will violate *jus in bello*. By defusing these two concerns, my discussion will put pressure on responsibility gap pessimists to say more about why we should think that the responsibility gaps that might plausibly emerge are something to be worried about to begin with. This is the second challenge. In a nutshell, then, the claim of this paper will be the following: It is unclear whether and when responsibility gaps occur, but if they do occur, we need not be too concerned about them.

I am, in this paper, focusing on responsibility issues that arise as a result of the autonomy of autonomous systems. I am not concerned with the so-called problem of many hands (Poel et al., 2015; Thompson, 1980). Some scholars have suggested that similar problems with the allocation of responsibility arise whenever some outcome is the result of the contributions of a large number of agents ('hands'). The problem of many hands is not, however, peculiar to outcomes produced by autonomous systems, arising also in more conventional contexts (see e.g. Poel et al., 2012). Although this article focuses on responsibility gaps that are due to a system's acting autonomously, some of the findings will likely carry over to the problem of many hands.

Responsibility gaps and their elusiveness

Before I can embark on the project of defusing concerns about responsibility gaps, we need a better grasp of what we are talking about when we are talking about such gaps. In this section, I will propose a characterization of what responsibility gaps are, and, based on this characterization, formulate the first challenge.[6]

I take responsibility gaps to be in the first instance about moral responsibility, as opposed to legal responsibility. Moreover, I agree with Sebastian Köhler, Neil Roughley and Hanno Sauer that those who have posited responsibility gaps are best understood as concerned primarily with the accountability aspect of responsibility (Köhler et al., 2018, p. 52; see also Baum et al. 2022, pp. 5-6). That is, someone is responsible for something when she can justly be blamed or praised for it.[7]

A distinct yet related concept is what I call moral liability for damages. A person is morally liable for damages she has caused if she has a moral duty to compensate the victim for these damages. The victim then has a moral right to compensation. Moral liability is distinct from legal liability, that is, from having a legal obligation to compensate for damages caused.

Moreover, I take the term 'responsibility gap' to refer primarily to situations in which the following two conditions are met:

> No Responsibility: An autonomous system carries out some action without anyone—neither the programmer, the manufacturer, the operator, nor the autonomous system itself—being responsible for this action and its consequences.
> Autonomy: This absence of responsibility is due to the system's autonomy.

[4] Authors who have denied that responsibility gaps emerge include Burri, 2018, pp. 175–177; Himmelreich, 2019; Köhler et al., 2018; Lauwaert, forthcoming; Robillard, 2018; Simpson & Müller, 2016, pp. 305–307; Tigard 2021.

[5] The possibility that responsibility gaps might not be problematic has been noted by Himmelreich (2019, n14 and n15) and Robillard (2018, p. 708). That there might even be something positive about responsibility gaps has been suggested by Danaher (2022).

[6] Matthias' original characterization of responsibility gaps is, in my view, rather unclear (Matthias, 2004).

[7] See e.g. Sparrow, 2007, p. 71. There is disagreement among responsibility scholars as to whether there are one or several concepts of responsibility. Champions of the latter view believe that accountability is a concept of responsibility in its own right, alongside other concepts of responsibility, all of which are needed to do justice to our discourse about responsibility (Shoemaker, 2011; Watson, 1996). Champions of the former view believe that one concept of responsibility can capture our entire discourse about responsibility, including the accountability aspect (Smith, 2012, 2015). That responsibility is (at least in part) about blameworthiness and praiseworthiness can be agreed on by members of both camps.

No Responsibility should be understood to involve a normative statement. It states that nobody is responsible in the sense that it would be *unjust* or *unfair* to blame anyone. Some version of this condition is accepted by nearly everyone who has written on responsibility gaps (see e.g. Danaher, 2016; Himmelreich, 2019, p. 734; Köhler et al., 2018, p. 54; Sparrow, 2007, 2016).[8] Autonomy specifies in addition that the absence of responsibility must be due to the system's autonomous behavior. We do not speak of a responsibility gap if the absence of responsibility is unrelated to the system's autonomy, say, because it is due to people's lack of a free will, to them being controlled by external forces, etc. Rather, the idea is that it is the system's autonomy—its ability to operate independently from human intervention in complex environments and to make decisions that are not based on concrete human instructions—that cancels people's responsibility.[9] At the same time, the machines are not deemed sophisticated or 'agent-like' enough to be themselves possible bearers of responsibility.[10]

Some might want to add a descriptive condition stating that there is some kind of popular demand for responsibility:

> Demand: There is a demand for responsibility in that people have a desire to hold somebody to account.

The motivation for adding this condition is that a responsibility gap may be thought to be not just the absence of someone who is responsible but, precisely, a *gap* in responsibility. That is, a responsibility gap may be thought to involve some kind of *mismatch* or *discrepancy*. One way of understanding this mismatch or discrepancy is that there is a desire or demand for responsibility that is not met (see e.g. Danaher, 2016).[11] In what follows, I will assume that the conjunction of No Responsibility and Autonomy captures the essence of what a responsibility gap is, their fulfillment being necessary and sufficient for the occurrence of such a gap, while the Demand condition is optional. It makes sense to speak of a responsibility gap and to consider whether such a gap is problematic even when, for whatever reason, there is no desire on anyone's part to hold somebody to account.[12] I will, however, briefly return to Demand in the concluding section of this paper.[13,14]

As mentioned in the introduction, the first task for responsibility gap pessimists is to establish that responsibility gaps may indeed occur. Assuming the above account of responsibility gaps, they must show that there are situations in which an autonomous system carries out some action without anyone being responsible for this action and its consequences because of the system's autonomy.

The challenge, now, consists in specifying the circumstances in which this is the case. To see the problem, consider that although it is conceivable that responsibility gaps arise in *some* situations, the idea that no one is *ever* responsible for what an autonomous system does clearly must be rejected. It is fair to assume that even those who postulate the existence of responsibility gaps must accept that there are certain situations in which a human agent is responsible

[8] Köhler et al. could be read as defending a descriptive version of No Responsibility, as they characterize responsibility gaps as situations in which nobody can justly be held responsible *according to "our traditional everyday practice of responsibility ascription"* (2018, p. 54 and passim; emphasis added). Thus conceived, the problem of responsibility gaps could be dealt with simply be reforming our practices of responsibility ascription. But surely, many who have been concerned about responsibility gaps have been concerned not about shortcomings of our actual practices of ascribing responsibility but about the possible absence of someone who bears responsibility. The normative version of No Responsibility is therefore preferable, and Sebastian Köhler has confirmed that this is how their account is meant to be interpreted (personal communication).

[9] How to characterize the autonomy of an autonomous system is itself a matter of some debate. For two helpful discussions, refer to Hellström, 2013; Noorman & Johnson, 2014.

[10] In line with much of the literature, I am assuming that responsibility is entirely canceled rather than merely reduced. This fits better with the talk of a responsibility gap. Also, if there is a problem with responsibility gaps, this problem can be assumed to be most serious when responsibility is absent rather than merely reduced.

[11] The term 'demand' sometimes carries a normative connotation. I use it in a purely descriptive sense.

[12] I am here deviating from Danaher's suggested definition, according to whom responsibility gaps (or *retribution* gaps) are best understood as a "mismatch between the human desire for retribution and the absence of appropriate subjects of retributive blame." (2016, p. 299).

[13] Unlike Danaher and myself, Köhler et al. have defended a normative version of Demand, according to which it is "surely appropriate" to hold somebody responsible (Köhler et al., 2018, p. 54 and passim). In my view, this is another legitimate way of characterizing responsibility gaps. Notice that the problematic nature of responsibility gaps would be built into the definition. For there to be a responsibility gap would mean for there to be a *problematic* absence of responsibility. If this definition were used, the claim of this paper would quite simply be that responsibility gaps do not exist at all.

[14] A brief note on Himmelreich's definition, which is similar to mine but still relevantly different: According to him, "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." (2019, p. 734) Condition (3) is related to my Autonomy condition. It is also supposed to capture the mismatch element of responsibility gaps. A problem with (3), however, seems to be that whether a human person would be responsible for x had x been the action of this person is unclear and depends on the specifics of the counterfactual scenario, which are left undefined. For any outcome produced by an intelligent system, there is a counterfactual world in which the same action could have been performed by a human person who would have been responsible for it, and one in which, for some reason or other, she would not have been responsible for it. This is why I think it preferable to simply state that the absence of responsibility must be due to the systems autonomy and to capture the, in my view optional, mismatch element in terms of the Demand condition.

for harm caused by such machines.[15] In particular, it is plausible to assume that the No Responsibility condition is not met when harm or damage is caused by an autonomous system as a result of human carelessness (negligence or recklessness) or because the autonomous system was intended to cause harm (malice).

Consider first negligence, which is one kind of careless behavior.

> Negligent Engineer
> An engineer has developed a new highly autonomous vehicle (think a self-driving car or an autonomous drone). Once she sends the vehicle on its mission, she is unable to control its actions or to predict with any precision how it will behave. But she makes no effort to anticipate and reduce possible risks. When the engineer finally dispatches the vehicle, it makes a mistake, injuring innocent bystanders.[16]

The fact that the vehicle acts autonomously and unpredictably does not mean that the engineer is not responsible. On the contrary, I submit that we would judge her responsible because she failed to use reasonable care. It was negligent of her to send the vehicle on its mission without putting a significant effort into identifying and reducing risks to innocent bystanders. She should have done more to foresee and limit the risk of harm. This renders her responsible. To claim that there was a responsibility gap in Negligent Engineer is not plausible, because No Responsibility is not met.

A fortiori, the engineer is blameworthy if she acts recklessly, that is, if she is aware of a high risk of the vehicle injuring innocent bystanders and dispatches it nonetheless. And she is clearly blameworthy if she causes harm on purpose (malice), that is, if she is aware of a high risk and dispatches the vehicle because she *seeks* to cause harm.

This highlights the, perhaps obvious, fact that at least *sometimes* people can justly be held responsible for the harmful actions of an autonomous system, namely when these people fail to exercise due care (negligence or recklessness) or intend to cause harm (malice). This much should be agreed upon even by those who have posited the existence of responsibility gaps. Surely, for instance, Sparrow would not want to say that the commander of an autonomous weapons system that kills civilians because the commander completely fails to consider possible risks or intends to kill them is blameless due to the weapon's autonomous behavior.[17] If responsibility gaps arise, they arise only in *some* situations, namely in situations in which the outcome is not the result of carelessness or malice. This is what I will refer to as the plausibility constraint on the emergence of responsibility gaps. Responsibility gaps can plausibly arise only in situations in which the negative outcome is not due to carelessness or malice. For otherwise, the No Responsibility condition would not be met.[18]

Now, the introduction of this plausibility constraint does not as such rule out the existence of responsibility gaps. It is merely a *constraint* on when responsibility gaps may plausibly be expected to arise. But it does raise the question what kind of human behavior the system's autonomy exculpates, given that is does not seem to exculpate negligent, reckless or malicious behavior. Put differently, what is the sort of human behavior that is blameless if it leads to an autonomous system causing damage but which, ceteris paribus, would be blameworthy if it led to a conventional machine causing damage?

Proponents of responsibility gaps thus face a dilemma: Either they reject the plausibility constraint and maintain that the autonomy of an autonomous system creates responsibility gaps by rendering all kinds of behavior blameless, including negligent, reckless or malicious behavior. We would then have some idea of the situations in which responsibility gaps arise and which behavior they exculpate. But this view is extremely implausible, and to my knowledge no promising defense of it has been offered. Or they accept the plausibility constraint by acknowledging that there are still situations in which those handling autonomous systems do bear responsibility, especially situations in which they fail to act with due care or act with malice. But this renders the situations in which responsibility gaps arise elusive. These would have to be situations in which people who handle autonomous systems and who handle them with due care

[15] The literature on responsibility gaps is somewhat unclear on the question whether such gaps arise always or only in some circumstances.

[16] I am stipulating that the vehicle is devised, engineered, programmed and operated entirely by this one individual in order to sidestep issues related to the problem of many hands. A similar scenario is used by Robillard (2018, p. 709).

[17] Indeed, at one point, he seems to admit that negligence is blameworthy (Sparrow, 2007, p. 69). Only Danaher insists that mere negligence is not sufficient for retributive blame (Danaher, 2016, p. 302). Among responsibility scholars, the view that negligence does not exculpate is shared by many (see e.g. Raz, 2010; Sher, 2009; Shiffrin, 2017) but not everyone. One reason why negligent wrongdoing might be blameless is that negligence, unlike recklessness, is characterized by the *absence* of certain mental states (King, 2009). But if this is true, that is, if negligent wrongdoing is blameless per se, the absence of responsibility would not be due to the autonomous system's autonomy.

[18] The observation that the negligent use of autonomous systems is blameworthy has also been eloquently made by Burri, 2018, pp. 175–177; Himmelreich, 2019; Köhler et al., 2018; Köhler, 2020, pp. 3133–3139; Robillard, 2018, p. 709; Simpson & Müller, 2016, pp. 306–307. Unlike some of them, however, I do not want to rule out the possibility of responsibility gaps. I am merely observing that they do not arise when there is negligence or other carelessness involved.

bear no responsibility *because of the system's autonomy*. These situations need to be specified.

This is the first – still unmet – challenge that responsibility gap pessimists must address. Many have been concerned about the possibility of responsibility gaps, but the precise circumstances in which they may emerge have remained conspicuously underspecified. Unless this challenge is met, there is little reason to believe that responsibility gaps are 'a thing'.

In the remainder of this paper, however, I will assume that such gaps may indeed arise and formulate the second challenge: Even if such gaps arise—subject to the above plausibility constraint—it is unclear why we should be concerned about them.

Responsibility gaps and their consequences

The intuitively most compelling concern about responsibility gaps revolves around their possible harmful consequences. Sparrow, though primarily focusing on deontological principles of just war theory, points out that there are "weighty consequentialist considerations" speaking against allowing autonomous weapons systems that generate responsibility gaps: "An inability to identify those responsible for war crimes would render their prosecution moot, […] with disastrous consequences for the ways in which wars are likely to be fought." (Sparrow, 2007, p. 67) He objects specifically to the use of autonomous systems in the military domain, but his 'consequentialist' worry is not confined to war crimes and military autonomous systems. The same worry arises for non-military autonomous systems and is a recurring theme in the literature on artificial intelligence and responsibility. A similar logic informs, for instance, Deborah Johnson's observation that if people can be held responsible, this will "keep the pressure on developers to ensure the safety and reliability of such devices." (2015, p. 714) And Santoni de Sio and Mecacci have recently observed that responsibility gaps "are concerning insofar as the more persons designing, regulating, and operating the system can legitimately (and possibly systematically) avoid blame for their wrong behaviour, the less these agents will be incentivised to prevent these wrong behaviours." (2021, p. 1063).[19]

As I understand it, then, the worry is that if nobody is responsible, people are under no pressure to minimize harm and damage. Lacking incentives to minimize harm and damage, programmers, manufacturers, and operators of autonomous systems will fail to exercise due care or might even cause harm on purpose, because they can do so with impunity. Their careless or intentionally harmful behavior will have disastrous consequences for the rest of society. The worry is thus one about incentives to act with due care and to avoid harm, or rather about the lack of disincentives to act carelessly and to intend harm. Responsibility gaps erode a socially beneficial incentive structure.

Two further clarifications may be helpful to better understand this concern. First, while it is a plausible supposition that an absence of responsibility eliminates disincentives to act without due care, the nature of these disincentives needs to be specified. I find it useful to distinguish three types of disincentives that may be thought to be lacking if responsibility gaps exist: (1) blame (2) moral liability for damages (3) punishment. Second, it is important to note that we are not talking about the disincentives people are *actually* facing but those that they may *justly* be subjected too. The reasoning is that, when there is a responsibility gap, it would be *unjust* to blame people, to require them to pay for damages or to punish them. They would lack these disincentives because we ought not to do what is unjust.

We thus get three versions of this concern:

(1) If responsibility gaps exist, it would be unjust to blame people, because it is unjust to blame people who do not bear any responsibility. People would thus lack this disincentive to act with due care.
(2) If responsibility gaps exist, it would be unjust to require people to pay for damages, because it is unjust to require people who do not bear any responsibility to pay for damages. People would thus lack this disincentive to act with due care.
(3) If responsibility gaps exist, it would be unjust to punish people, because it is unjust to punish people who do not bear any responsibility. People would thus lack this disincentive to act with due care.

As far as I can see, each version of this worry is unfounded. The notion that people should be allowed to act carelessly, let alone to intentionally cause harm, with impunity because of potential responsibility gaps is untenable. The reason why no such pessimistic conclusion follows has, of course, to do with the finding that the emergence of responsibility gaps is subject to the above introduced plausibility constraint. If responsibility gaps arise at all, they do not arise in situations in which people act without due care.

Consider first blame. Blaming people can discourage them from engaging in socially harmful behavior.[20] It is a

[19] On responsibility gaps and incentives, see also Baum et al., 2022, p. 6; Chomanski, 2021.

[20] A useful distinction here is that between two senses of holding people responsible (see Smith, 2007). It can refer to the judgment that someone is blameworthy. But it can also refer to the act of blaming this person. It is the unpleasantness of the latter that provides the incentive to avoid blameworthy behavior.

natural thought, then, that responsibility gaps will remove this incentive to refrain from engaging in such behavior. But it follows directly from the above introduced plausibility constraint on the emergence of responsibility gaps—namely that they do not plausibly arise in situations of carelessness or malice—that this concern is unfounded. To be sure, if it *were* the case that responsibility gaps rendered negligent, reckless or malicious behavior blameless, people might be more inclined to engage in such behavior, knowing that they could not be justly blamed for it. But that an autonomous system's high degree of autonomy exculpates such behavior is precisely what is so implausible to assume. If negligent, reckless or malicious behavior led to an autonomous system causing harm, whoever engaged in this behavior—the manufacturer, the programmer, the operator, etc.—clearly is blameworthy. It is therefore also just to subject this person to blame and to discourage such behavior in this way. I am not here assuming that responsibility gaps do not arise at all. I am merely assuming that their emergence is subject to the above introduced plausibility constraint. They do not exculpate careless or malicious behavior.

Something similar applies to (2) and (3). The idea behind (2) is that it would be unjust to require manufacturers, programmers or operators to pay for the damages they cause through their carelessness, a requirement that might otherwise deter them from acting in this way.[21] Version (3) captures Sparrow's worry that responsibility gaps would render the prosecution of war criminals moot, with disastrous consequences for how wars will be fought. The reasoning seems to be that if nobody is responsible, commanders of autonomous weapons will not be sufficiently disincentivized from committing war crimes, because it would be unjust to punish ("prosecute") them for doing so. As a consequence, more war crimes will happen.[22] Again, however, both (2) and (3) are based on the implausible assumption that the autonomous nature of autonomous systems would exculpate careless or malicious behavior, an assumption that we should reject as per the plausibility constraint.

Note also that (2) and (3) are more contentious than (1). The claim that is unjust to blame people who do not bear any responsibility is close to tautological, given that responsibility is understood in terms of blameworthiness. By contrast, (2) and (3) involve the more contentious claims that it is unjust to require people who are not responsible to pay for damages or to punish them, respectively. This exposes (2) and (3) to an additional objection. Not only is it implausible to assume that careless or malicious behavior might not be blameworthy when autonomous systems are involved. In addition, one might question the assumption that it is wrong to require people who do not bear responsibility for some adverse outcome they have caused to pay for damages or to punish them. Even if it *were* the case that the autonomy of autonomous systems rendered careless or malicious people blameless, this need not necessarily mean that it would be unjust to require these people to pay for damages they have caused.[23] By the same token, it might be justifiable to punish careless or malicious behavior in an effort to deter such behavior, even if this behavior should not be blameworthy.

Be that as it may, the principal reason why the argument from consequences fails is that it implausibly assumes that responsibility gaps would exculpate and, as a consequence, incentivize negligent, reckless or malicious behavior. Responsibility gaps may exist, but they surely do not occur when people act carelessly or with malice. Programmers, manufacturers and operators of autonomous systems may justly be subjected to blame and censure, liability charges, and punishment when they act in such a way. This will disincentivize socially harmful behavior. Again, I am not denying that responsibility gaps might exist. I am open to their existence, with the caveat that the conditions of their emergence have so far remained somewhat underspecified. I am merely disputing that they arise in the kind of situations in which they would have to arise in order for the consequentialist concern to be valid.

Responsibility gaps and war

Another concern about responsibility gaps, outlined by Sparrow (2007), revolves around a very special type of autonomous system, namely autonomous weapons systems. Autonomous weapons systems are an interesting case study in their own right. But the discussion of Sparrow's reasoning will have implications beyond the military domain.

[21] As Santoni de Sio and Mecacci observe, "victims of unjust harm will be less likely to receive compensation." (2021, p. 1063) Note that I focus on moral liability as opposed to legal liability. Some scholars have expressed concerns about legal liability gaps (Matthias, 2004; Roff, 2013). In my view, however, the question is not whether people are *legally* required to pay for damages, which depends on the contingent and changeable legal framework in a country, but whether it would be *just* to require them to pay for damages. The legal framework should piggyback on what is just.

[22] I am here talking about punishment as a deterrent, rather than retributive punishment, as I am skeptical about the moral permissibility of retribution (see Königs, 2013). Still, I hope that my discussion goes some way towards addressing Danaher's worry that responsibility gaps undermine people's trust in the legal system (2016, pp. 307–308). Danaher is concerned that people will be upset about the lack of retributive punishment meted out by the criminal justice system as a result of responsibility gaps. If my reasoning is correct, there is at least no reason not to punish the careless or intentionally harmful use of autonomous systems. This would not be retributive punishment, but it might satisfy people's desire for justice. Apart from this, I do not think we should cater to people's retributive desires.

[23] Similarly, Burri, 2018, p. 177.

We have seen that Sparrow, like many others, harbors consequentialist concerns about responsibility gaps, which were dealt with above. But his principal objection is that there is a conflict between the principles of *jus in bello* (justice in war) and the absence of a person who is responsible for damage caused by these weapons: "[I]t is a fundamental condition of fighting a just war that someone may be held responsible for the deaths of enemies killed in the course of it. In particular, someone must be able to be held responsible for civilian deaths." (Sparrow, 2007, p. 67; similarly Roff, 2013)[24] The responsibility gap generated by autonomous weapons systems renders their use unethical, as it entails a violation of *jus in bello*.[25]

Sparrow's case against autonomous weapons systems is perhaps the most influential articulation of moral concerns about artificial intelligence and responsibility. But it, too, fails to withstand scrutiny. As in the previous section, rather than to deny that responsibility gaps might emerge, I will suggest that responsibility gaps are not as problematic as they have been made out to be. Although the emergence of responsibility gaps should be taken to be subject to a plausibility constraint, I am happy to accept that there may still be situations in which, say, a military general or a commander of an autonomous weapons system is not responsible for what this weapons system does because of its autonomy (although the circumstances in which this happens await clarification). From a just war theory perspective, however, such gaps in responsibility are just not that worrisome.

Consider the three most widely accepted principles of *jus in bello*, of which Sparrow cites the first and the second:

1. Discrimination: Belligerents must always distinguish between military objectives and civilians, and intentionally attack only military objectives.
2. Proportionality: foreseen but unintended harms must be proportionate to the military advantage achieved.
3. Necessity: The least harmful means feasible must be used. (Lazar, 2017, capitalization added)

Sparrow, as I read him, thinks that, by generating responsibility gaps, the use of autonomous weapons systems is at variance with *jus in bello* in two different ways.

First, he suggests that responsibility gaps entail a violation of the above principles or make it impossible to comply with them:

> The assumption and/or allocation of responsibility is [...] vital in order for the principles of *jus in bello* to take hold at all. The principle of discrimination, for instance, which requires that combatants distinguish between legitimate and illegitimate targets, assumes that we can specify who is responsible for attacks that may violate it. (Sparrow, 2007, p. 67)

But it is not clear why responsibility gaps should clash with the above principles of *jus in bello*. Either autonomous weapons systems are programmed and deployed in such a way that the conditions specified by these principles are met, or in a way that leads to violations of these conditions. In the former case – if civilians are not targeted, foreseen but unintended harms are proportionate, and the least harmful means feasible are used – it is trivially true that the three requirements of *jus in bello* have been met and that the war is, in this respect, just.[26] It is thus clearly possible to comply with these principles of *jus in bello* even when nobody is responsible.[27] In the latter case – if civilians are targeted, the foreseen but unintended consequences are disproportionate, or excessively harmful means are used – it is again trivially true that the principles of *jus in bello* have been violated and that the war is, in this respect, unjust. But the absence of someone who can justly be held to account plays no part in the violation of these principles. If it *were* the case that someone could justly be held responsible for the weapons' actions, the principles would *still* be violated. The principles count as violated in virtue of the fact that civilians were killed, the foreseen but unintended harms were not proportionate, or the means used were not the least harmful feasible, not in virtue of the fact that no one can be held responsible. Responsibility is not relevant to the compliance with the three most commonly invoked principles of *jus in bello*.

Second, however, Sparrow seems to hold that the requirement that someone be responsible for harm caused by autonomous weapons systems is a principle of *jus in bello* in its own right. That is, he suggests adding another principle to the set of classic principles of *jus in bello* listed above, roughly along the following lines:

4. Responsibility: Someone must be responsible for the deaths (or, perhaps more broadly, the harm) caused.

This interpretation is supported by his assertion that the "condition [that someone be responsible] may be thought of

[24] Sparrow is echoing Michael Walzer, who, at one point, asserts that "[t]here can be no justice in war if there are not, ultimately, responsible men and women." (Walzer 1977, p. 288) He offers little support for this claim.

[25] Autonomous weapons systems are distinct from remotely controlled (but non-autonomous) weapons systems, and they raise different ethical questions (see e.g. Strawser, 2010).

[26] I am adding ‚in this respect ' to acknowledge that there may still be violations of the principles of *jus ad bellum* or *jus post bellum*. For a skeptical view on the ability of autonomous weapons systems to comply with the three classic principles of *jus in bello*, refer to Miller, 2016, ch. 10.

[27] Similarly, Taylor, 2021. See also Purves et al., 2015, pp. 854–855.

as one [of] the requirements of *jus in bello*" (2007, p. 67).[28] If this condition is accepted as one of the principles of *jus in bello*, the use of autonomous weapons systems, assuming that it generates responsibility gaps, would indeed constitute a violation of *jus in bello*. The crucial question then becomes why we should accept this additional principle of *jus in bello*.

Sparrow's reasoning seems to be that this principle is supported by the same considerations that form the moral foundation of the three above-mentioned classic principles of *jus in bello*, namely respect for the humanity of our enemies. He contends that it is this respect for our enemies that "underlies" (2016, p. 110) or "underpins" (p. 112) the principles of justice in war. To kill our enemies, especially civilians, without anyone being responsible for it would be, just like violating any of the three classic principles of *jus in bello*, to disrespect our enemy and thus to negate the basic value underlying *jus in bello*. It is easiest to quote him at length on the issue of respect:

> It is a minimal expression of respect due to our enemy – if war is going to be governed by morality at all— that someone should accept responsibility, or be capable of being held responsible, for the decision to take their life. If we fail in this, we treat our enemy like vermin, as though they may be exterminated without moral regard at all. The least we owe our enemies is allowing that their lives are of sufficient worth that someone should accept responsibility for their deaths. (2007, p. 67)

If this is why autonomous weapons systems violate *jus in bello*, they would do so even when the three classic principles of *jus in bello* are complied with. But the case for accepting Responsibility is weak.

First, considerations about what we would signal or express by using autonomous weapons seem relatively unimportant, providing at best a weak pro tanto reason against their use. Weighing goods is notoriously difficult, but if we had to choose between sending a message of respect by renouncing autonomous weapons and, say, saving the life of a single human being, soldier or civilian, the choice should be obvious. The importance of signaling respect pales in comparison to that of the other goods at stake in war.

Moreover, respect for one's enemies can be expressed in many different ways. One way may indeed be to refrain from risking to kill someone unless somebody could justly be held responsible for it. But there are other ways to express respect. For instance, the government could erect monuments to honor those killed by autonomous weapons, or the commanders of autonomous weapons could, every once in a while, observe a minute of silence to express their respect for the dead. There are countless other ways of expressing respect. Arguably, the most powerful way of expressing respect for the lives of civilians is to do everything in one's power to avoid killing them, and this might even require the use of autonomous weapons (Jenkins & Purves, 2016, p. 396). That is, if we wish to express respect for our enemies, especially civilians, there are ways of doing this that are compatible with deploying autonomous weapons, even if they generate responsibility gaps.

Finally, the fact that the absence of someone who can be held responsible expresses a lack of respect is, as Sparrow himself acknowledges, a conventional fact (Sparrow, 2016, p. 109). It expresses a lack of respect because of certain contingent, socially constructed conceptions about the meaning of actions, which happen to be shared in our society. The moral significance of such conventions is limited. If there should be weighty reasons in favor of using of such weapons, perhaps their more ethical behavior on the battlefield (Arkin, 2010), we should seek to overcome the convention that prevents us from going ahead. This is a general problem with arguments that appeal to the symbolic meanings of actions. If some otherwise beneficial action or policy is objected to on the grounds of what it conventionally symbolizes or expresses, we should alter this convention or, indeed, ignore it altogether (see Brennan & Jaworski, 2015).

The second line of reasoning, then, which posits an independent principle of *jus in bello* to the effect that someone must be responsible for the deaths caused in war, is unconvincing, too. I wish to emphasize that I have not positively argued for the use of autonomous weapons. There may well be other reasons to be wary of 'killer robots'. But the problem is not that they violate *jus in bello* by generating responsibility gaps.

Although the above discussion has revolved around military robots, its significance is not confined to the military domain. Some might think that the idea that it is disrespectful if someone is harmed without anyone being responsible applies similarly to harms caused by non-military autonomous systems, such as self-driving cars or work robots. This idea would thus provide a much more principled reason to object to the development and use of autonomous systems, irrespective of the domain in which they are used. The reasons, however, why Sparrow's argument from disrespect does not withstand scrutiny carry over to the non-military domain, too, defusing also this more general, domain-independent objection to autonomous systems.

Concluding these two sections, even if responsibility gaps exist, it is unclear why we should be concerned about them. Assuming that their emergence is subject to the suggested plausibility constraint, they seem quite innocuous. This, then, is the second challenge. Responsibility gap pessimists must give us reason to believe that responsibility

[28] This is also how Steinhoff interprets him (2013, pp. 179–180).

gaps, should they exist, are really something to be concerned about.

Conclusion

If my analysis is correct, it warrants at least cautious optimism about artificial intelligence and responsibility gaps. To be sure, nothing I have said conclusively rules out the existence of problematic responsibility gaps. But if sound, my analysis suggests that there is little positive reason to believe that problematic responsibility gaps exist. I am under no illusion, however, that this article settles the issue. Having formulated my critique of responsibility gap pessimism in terms of two challenges, it is open to those who have voiced concerns about responsibility gaps to demonstrate that these challenges can be met. Responsibility gap pessimists would have to, first, specify the circumstances in which responsibility gaps occur given the suggested plausibility constraint on their occurrence, and, second, explain how *these kinds* of responsibility gaps are cause for concern.

One option would be to exploit the fact that I have only considered two concerns about responsibility gaps. Even if these concerns are unwarranted, there might well exist other problems with responsibility gaps that were overlooked. For instance, one might pursue the idea that responsibility gaps are problematic simply in virtue of the fact people's desire to blame somebody will have to go frustrated. Even if the Demand condition should not be built into the definition of what responsibility gaps are, the contingent fact that people may have the desire to blame someone in a responsibility gap situation, which cannot be satisfied, could explain what is problematic about responsibility gaps. I do not find this plausible. For one thing, the mere frustration of such a desire hardly warrants the high level of concern that has pervaded the debate about artificial intelligence and responsibility. For another thing, it is certainly odd, from a moral point of view, that we should wish for there to be more blameworthiness in the world just for people's desire to blame not to go frustrated. Still, it is in principle open to proponents of responsibility gaps to attempt to develop an argument along these lines or to identify other problems with responsibility gaps that were not accounted for above.[29] Note, though, that even if some such problem could be identified, its identification would only address the second of the two challenges. Responsibility gap pessimists would still have to address the prior challenge of showing that responsibility gaps are even a thing. They would first have to specify the circumstances that give rise to such gaps given the suggested plausibility constraint on their emergence.

By way of conclusion, I wish to emphasize that even if, as I have suggested, artificial intelligence does not produce problematic responsibility gaps, this should not make us overlook the basic fact that allocating responsibility is not easy. When people interact with intelligent systems – producing them, programming them, selling them, using them, etc.—it may be difficult to determine to what extent these people are to be held responsible for an outcome caused by an intelligent system.[30] Relatedly, while it may be agreed that programmers, manufacturers and operators of autonomous systems must exercise due care in order to escape the charge of negligence or recklessness, it may not be clear what exactly due care requires. There is thus bound to be some responsibility injustice, as one might call it, in that there will be both false positives and false negatives. People will on occasion be held responsible to a higher or lower degree than appropriate, as already happens today when no intelligent systems are involved. This problem is epistemic rather than metaphysical. It does not consist in the actual (metaphysical) absence of a responsible agent but in the (epistemic) difficulty of correctly determining how responsible people are. This difficulty does therefore not amount to a genuine responsibility gap. Such gaps were defined as involving an actual absence of a responsible agent, in accordance with how such gaps are usually understood in the literature. Still, the epistemic problem *is* a problem, which is worth bearing in mind.[31] Like the problem of many hands, the epistemic problem is not specific to intelligent systems.[32] Even when no artificial intelligence is involved, it is difficult to determine how blameworthy people are or where negligence begins in decision-making under risk. Exactly how much care would a group of engineers have to exercise before engaging in geoengineering in order to be blameless should something go wrong? It is hard to tell. But the problem may be particularly acute when autonomous systems are involved, given the novelty and unpredictability of AI technology. The above considered engineer, who makes no effort to anticipate and reduce possible risks, is surely culpably negligent. But given the novelty and unpredictability of AI

[29] Two discussions of why responsibility gaps might be problematic by Baum et al. (2022, pp. 6–8) and Himmelreich & Köhler (2022, pp. 6–7) appeared too late to be adequately considered here.

[30] I am grateful to the reviewers of this manuscript for pressing me to discuss this problem.

[31] One might be tempted to avoid false positives by refraining altogether from holding people responsible. Although this would lead to an absence of responsibility *allocation*, it would not entail an actual absence of a responsible agent as necessary for there to be a genuine responsibility gap. Also, refraining altogether from holding people responsible seems like an unjustified overreaction. What I think we should do is cautiously approximate the just allocation of responsibility.

[32] It is discussed by Köhler et al. (2018), who do not exclusively focus on AI technology.

technology, it is difficult to say how much effort she ought to put into reducing risks and which risks she can reasonably be expected to anticipate. We can, I think, confidently make the formal claim that responsibility gaps do not arise in situations of negligence (let alone recklessness or malice). But putting flesh on this claim by spelling out precisely what negligence amounts to is, admittedly, difficult. Thus, while the above discussion warrants cautious optimism about AI-induced responsibility gaps as I suggested to define them, it does not alter the basic fact that achieving responsibility justice is difficult. When autonomous systems are involved, achieving responsibility justice may be particularly difficult.

Acknowledgements For helpful feedback and discussions, the author wishes to thank the reviewers of this paper, Susanne Burri, Niël Conradie, Johannes Himmelreich, Gregor Hochstetter, Max Kiener, Sebastian Köhler, Leo Menges, Saskia Nagel, Sven Nyholm, Daniel Tigard as well as the audience at the ECAP10 in Utrecht.

Funding information Work on this paper was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2023 Internet of Production—390621612.

References

Arkin, R. C. (2010). The case for ethical autonomy in unmanned systems. *Journal of Military Ethics, 9*(4), 332–341.

Baum, K., Mantel, S., Schmidt, E., & Speith, T. (2022). From responsibility to reason-Giving explainable artificial intelligence. *Philosophy & Technology, 35*(1), 12.

Brennan, J., & Jaworski, P. M. (2015). Markets without symbolic limits. *Ethics, 125*(4), 1053–1077.

Burri, S. (2018). What Is the Moral Problem with Killer Robots. In B. J. Strawser, R. Jenkins, & M. Robillard (Eds.), *Who Should Die? The Ethics of Killing in War* (pp. 163–185). Oxford University Press.

Chomanski, B. (2021). Liability for robots: Sidestepping the gaps. *Philosophy & Technology, 34*(4), 1013–1032.

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

Danaher, J. (2019). *Automation and Utopia: Human Flourishing in a World without Work*. Harvard University Press.

Danaher, J. (2022). Tragic choices and the virtue of techno-Responsibility gaps. *Philosophy & Technology, 35*(2), 26.

Gunkel, D. J. (2020). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology, 22*(4), 307–320.

Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and Information Technology, 15*(2), 99–107.

Himmelreich, J. (2019). Responsibility for killer robots. *Ethical Theory and Moral Practice, 22*(3), 731–747.

Himmelreich, J. & Köhler, S. (2022). Responsible aI through conceptual engineering. *Philosophy & Technology, 35*(3), 60.

Jenkins, R., & Purves, D. (2016). Robots and respect: a response to robert sparrow. *Ethics & International Affairs, 30*(3), 391–400.

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

King, M. (2009). The problem with negligence. *Social Theory and Practice, 35*(4), 577–595.

Köhler, S. (2020). Instrumental robots. *Science and Engineering Ethics, 26*(6), 3121–3141.

Köhler, S., Roughley, N., & Sauer, H. (2018). Technologically blurred accountability. In C. Ulbert, P. Finkenbusch, E. Sondermann, & T. Diebel (Eds.), *Moral Agency and the Politics of Responsibility* (pp. 51–68). Routledge.

Königs, P. (2013). The expressivist account of punishment, retribution, and the emotions. *Ethical Theory & Moral Practice, 16*(5), 1029–1047.

Kraaijeveld, S. R. (2020). Debunking (the) Retribution (Gap). *Science and Engineering Ethics, 26*(3), 1315–1328.

Lauwaert, L. (2021) Artificial intelligence and responsibility. *AI & Society, 36*(3), 1001–1009.

Lazar, S. (2017) War. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition)

Leveringhaus, A. (2018). What's so bad about killer robots? *Journal of Applied Philosophy, 35*(2), 341–358.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

Miller, S. (2016). *Shooting to Kill: The Ethics of Police and Military Use of Lethal Force*. Oxford University Press.

Noorman, M., & Johnson, D. (2014). Negotiating autonomy and responsibility in military robots. *Ethics and Information Technology, 16*(1), 51–62.

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

Poel, I., & v. d., Royakkers, L., & Zwart, S. (Eds.). (2015). *Moral Responsibility and the Problem of Many Hands*. Routledge.

Poel, I., Fahlquist, J. N., Doorn, N., Zwart, S., & Royakkers, L. (2012). The problem of many hands: climate change as an example. *Science and Engineering Ethics, 18*(1), 49–67.

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice, 18*(4), 851–872.

Raz, J. (2010). Responsibility and the negligence standard. *Oxford Journal of Legal Studies, 30*(1), 1–18.

Robillard, M. (2018). No such thing as killer robots. *Journal of Applied Philosophy, 35*(4), 705–717.

Roff, H. M. (2013). Killing in war: Responsibility, liability, and lethal autonomous robots. In F. Allhoff, N. G. Evans, & A. Henschke (Eds.), *Routledge Handbook of Ethics and War: Just War Theory in the Twenty-First Century* (pp. 352–364). Routledge.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology, 34*(4), 1057–1084.

Sher, G. (2009). *Who knew? Responsibility without awareness*. Oxford University Press.

Shiffrin, S. (2017) The Moral Neglect of Negligence. In D. Sobel, P. Vallentyne, & S. Wall (Eds.), *Oxford Studies in Political*

Philosophy (Vol. 3) (pp. 197–228). Oxford: Oxford University Press

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

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

Smith, A. M. (2007). On being responsible and holding responsible. *The Journal of Ethics, 11*(4), 465–484.

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

Smith, A. M. (2015). Responsibility as answerability. *Inquiry, 58*(2), 99.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.

Sparrow, R. (2016). Robots and respect. *Ethics and International Affairs, 30*(1), 93–116.

Steinhoff, U. (2013). Killing Them Safely: Extreme Asymmetry and Its Discontents. In B. J. Strawser (Ed.), *Killing By Remote Control: The Ethics of an Unmanned Military* (pp. 179–207). Oxford University Press.

Strawser, B. J. (2010). Moral predators: the duty to employ uninhabited aerial vehicles. *Journal of Military Ethics, 9*(4), 342–368.

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. M. (1980). Moral responsibility of public officials: The problem of many hands. *American Political Science Review, 74*(4), 905–916.

Tigard, D. (2021). There is no techno-responsibility gap. *Philosophy & Technology, 34*(3), 589–607.

Walzer, M. (1977). *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. Basic Books.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics, 24*(2), 227–248.