The risks of autonomous machines: from responsibility gaps to control gaps

Hindriks, Frank; Veluwenkamp, H.M.

DOI
10.1007/s11229-022-04001-5

Publication date
2023

Document Version
Final published version

Published in
Synthese: an international journal for epistemology, methodology and philosophy of science

Citation (APA)
Hindriks, F., & Veluwenkamp, H. M. (2023). The risks of autonomous machines: from responsibility gaps to control gaps. *Synthese: an international journal for epistemology, methodology and philosophy of science*, *201*(1), Article 21. https://doi.org/10.1007/s11229-022-04001-5

Synthese (2023) 201:21
https://doi.org/10.1007/s11229-022-04001-5

ORIGINAL RESEARCH

The risks of autonomous machines: from responsibility gaps to control gaps

Frank Hindriks[1] · **Herman Veluwenkamp**[1,2]

Received: 9 June 2022 / Accepted: 4 December 2022

Abstract
Responsibility gaps concern the attribution of blame for harms caused by autonomous machines. The worry has been that, because they are artificial agents, it is impossible to attribute blame, even though doing so would be appropriate given the harms they cause. We argue that there are no responsibility gaps. The harms can be blameless. And if they are not, the blame that is appropriate is indirect and can be attributed to designers, engineers, software developers, manufacturers or regulators. The real problem lies elsewhere: autonomous machines should be built so as to exhibit a level of risk that is morally acceptable. If they fall short of this standard, they exhibit what we call 'a control gap.' The causal control that autonomous machines have will then fall short of the guidance control they should emulate.

Keywords Autonomous machine · Control gap · Direct responsibility · Indirect responsibility · Meaningful human control · Responsibility gap

Who is to blame when a self-driving car harms another agent? Not the car itself, because it is not a moral agent.[1] But the passengers are not suitable targets of blame either. After all, they had no control over what happened. According to some, no one is to blame for harms caused by self-driving cars and autonomous machines more generally. At the same time, it is appropriate to attribute blame for this kind of harm. Hence, there is a responsibility gap: although there is reason to attribute blame, it is

[1] Anderson and Anderson (2007) and Floridi and Sanders (2004) defend a minimalist notion of moral agency that might be applicable to (future) self-driving cars. However, these minimalist notions do not entail moral blameworthiness. In contrast, Tigard (2020) argues that we sometimes do regard machines as responsible or, more precisely, answerable. However, he also maintains that our responses to them resemble those we display towards young children and psychopaths, which means they are not answerable in the same way as adult human beings.

✉ Herman Veluwenkamp
h.m.veluwenkamp@rug.nl

1 University of Groningen, Groningen, The Netherlands

2 Technical University Delft, Delft, The Netherlands

Published online: 07 January 2023

impossible to do so (Matthias, 2004; Sparrow, 2007). Those who believe that there can be responsibility gaps recognize how puzzling they are. In fact, Andreas Matthias (2004), who is typically credited with introducing the term, takes them to pose a threat to the consistency of our practices of attributing moral responsibility.[2] Furthermore, he regards autonomous machines as inherently unpredictable, as no one has control over them. This might mean, he proposes, that we should refrain from using them altogether (see also Sparrow, 2007).

A careful consideration of autonomous machines and the harms they might cause supports a different diagnosis, or so we argue. First, some harms are blameless. To support this, we compare self-driving cars to ordinary cars that get into an accident. And we propose that both kinds of accidents can be without fault. Second, in other cases, people are to blame indirectly. Think, for instance, of those who designed and produced the machine. We go on to argue that the real problem lies elsewhere: sometimes the level of risk to which an autonomous machine exposes others is too high. If this is the case, it exhibits a control gap. The machine will then be inadmissible. This means that, unless the risk level is brought down to a morally acceptable level, its use should be prohibited.

We start in Sect. 1 by discussing the notion of a responsibility gap. In Sect. 2, we argue that there are no responsibility gaps: either no one is to blame, or the blame that should be ascribed can be attributed to people who bear indirect responsibility. And in Sect. 3, we propose the notion of a control gap, which exists if the risk of harm to which an autonomous machine exposes others is too high.

1 Responsibility gaps are incoherent

Autonomous machines are a blessing and a curse. On the one hand, they make it easier for human beings to achieve their goals, as we can delegate tasks to them. They also bring new goals within reach. Think, for instance, of firefighter robots who can go where firefighters cannot or of helicopters that can fly on Mars without there being any humans there. On the other hand, they come with new risks. Matthias (2004) argued that the risks of autonomous machines are particularly severe when they rely on machine learning, which is inherently unpredictable. Because of this, it is impossible to foresee the consequences of their actions. And nobody is able to assume responsibility for harms caused by autonomous machines (ibid., p. 177). Thus, human beings lack control over autonomous machines. This is particularly worrisome because they might get out of control. Even if the dystopian overtones of Matthias' line of reasoning are out of proportion, he clearly identifies the core of the argument for responsibility gaps.

According to this argument, responsibility gaps are due to two facts about autonomous machines. First, they are agents that make decisions and perform actions independently of other agents. This means that other agents lack direct control over them. Second, autonomous machines are (currently) not able to process moral considerations. More technically, they are not responsive to moral reasons (Fischer & Ravizza, 1999). This in turn means that they are not moral agents. It follows that, if

[2] Villiers (2002) used the term 'responsibility gap' before Matthias (2004).

an autonomous machine causes harm, it cannot be held responsible.[3] However, it is problematic to hold other agents responsible, so it is argued, because they were not involved in the decision-making process. And they did not have direct control over what it did. Thus, it appears that nobody is responsible. But this conclusion is difficult to accept, because the harm should not have occurred. Because of this, it does seem appropriate to blame someone for it.

Against the background of this argument, a responsibility gap can be defined as follows[4]:

> [*RG*] A responsibility gap exists exactly if an autonomous machine causes harm and no one is to blame for it, even though blame is appropriate.

The underlying idea is that responsibility gaps occur when human beings have, so to say, given up their seat. As they lack control, they cannot be blamed. However, as someone has been harmed, blame is appropriate. Thus, this blame remains unattributed.[5] [RG] is the conception of responsibility gaps that Matthias seems to have in mind (2004). It is also alluded to by David Gunkel when he talks about the way autonomous machines "open up fissures in the way responsibility comes to be decided, assigned, and formulated" (2020, p. 313). Moreover, Filippo Santoni de Sio and Jeroen van den Hoven seem to suggest this idea by remarking that autonomous machines cause responsibility to "evaporate" (2018, p. 2). In light of this, we refer to responsibility gaps conceived in this way as 'responsibility remainders.'[6]

The claim that autonomous machines give rise to responsibility gaps has received a lot of attention, to such an extent that 'the question so as to what extent persons can or should maintain responsibility for the behaviour of AI has become one of, if not the most discussed question in the growing field of so-called ethics of AI' (Santoni de Sio and Meccaci 2021, p. 1058). This claim has triggered both inflationary and deflationary responses. Practically, Matthias (2004, p. 175) argued that we might have to refrain from using autonomous machines altogether, although he acknowledged that the costs of doing so are high. Theoretically, he maintained that responsibility remainders 'cannot be bridged by traditional concepts of responsibility ascription' (ibid.). Furthermore, they pose 'a threat to … the consistency of the moral framework

[3] See footnote 1 for a discussion of accounts of moral agency that reject the inference to moral responsibility.

[4] For recent discussions of responsibility gaps, see Coeckelbergh (2016), Danaher (2016), Gunkel (2017), Hellström (2013), Hevelke and Nida-Rümelin (2015), Himmelreich (2019), Köhler (2020), Köhler et al. (2017), Nyholm (2018a), Purves et al. (2015), Robillard (2018), Roff (2013), Santoni de Sio and Mecacci (2021), Sparrow (2007) and Tigard (2020).

[5] *RG* is concerned with the possibility that *some* blame is appropriate while *none* can be attributed. In Sect. 3, we consider semi-autonomous machines, which have partial control over what they do. It has been argued that responsibility gaps can also arise with respect to them. If so, the problem is that *some* blame can be attributed to the operators, but *not enough* (Köhler et al., 2017).

[6] As we focus on the question who, if anyone, is to blame for harm caused by an artificial agent, we abstract, for the most part, from the epistemic problem whether it is possible to determine who is responsible (Mecacci & Santoni de Sio, 2020; Nyholm, 2020). We also set aside the many gaps that have been postulated since Matthias (2004) that do not pose a threat to the coherence of our concepts or practices, such as the public accountability gap, the retribution gap and the achievement gap (see Santoni de Sio & Mecacci, 2021; Danaher, 2016; Danaher & Nyholm, 2020).

of society' (ibid., p. 176). Thus, responsibility gaps present an insoluble problem that comes with great costs.

The deflationary response denies that there is a genuine problem (Simpson & Müller, 2016, Köhler et al., 2017, and Tigard, 2020). Sebastian Köhler et al. (2017) argue that responsibility gaps do not exist because the agent who deploys the autonomous machine is to blame for any harm caused. But their argument concerns agents who do so 'in full knowledge' of the 'harmful consequences' they thereby risk. This reveals that it is incomplete: it says nothing about cases in which information is partial (or about cases of blameless harm, for that matter). Furthermore, it is too deflationary. As we argue in Sect. 3, there is a deep and pressing problem. But it has been misidentified: instead of responsibility gaps, there can be control gaps.

Our argument against responsibility remainders is that the very notion is incoherent. The thing to note is that, if blame is appropriate, then there is reason to attribute it. And if there is reason to ascribe blame, it must be possible to do so. To be sure, it might not always be fully transparent who is to blame or how blame is to be distributed to particular agents. But if all relevant information were available, it could be done. The crucial step is that from reason to possibility. It is supported by a consideration that is intimately related to the widely accepted claim that ought implies can. As Bart Streumer (2007) puts it, one cannot have reason to do the impossible. It follows that, if blame is appropriate, it must be possible to attribute it. Conversely, if it is impossible to attribute blame, then it is inappropriate to do so. Hence, the kind of discrepancy involved in responsibility remainders cannot arise.

But if the very concept is incoherent, how can it be that responsibility gaps are taken so seriously? Part of the answer is provided by the inflationary response. The notion has always been taken to push the boundaries of our imagination. For instance, Filippo Santoni de Sio and Jeroen Van den Hoven worry that 'human responsibility will [..] evaporate' (2018, p. 2). We believe that this strains credulity. To demystify the notion, we need to dig deeper. Recall that, according to Matthias (2004), responsibility gaps reveal an inconsistency in our moral framework. They challenge the fundamental assumption that the amount of blame that can in fact be attributed (B) must be identical to the amount that is appropriate (B^+): necessarily, $B = B^+$. The reason for this is that, if there is a responsibility gap, then $B^+ > B$. And these two claims are inconsistent. In response, Matthias concludes in effect that we have to reject the fundamental assumption. But this assumption is more plausible than the challenge that is supposed to undermine it. We propose instead that the inconsistency is located in the very concept of a responsibility gap itself (at least if it is understood as [RG]). It cannot be that $B^+ > B$, as it is impossible for there to be any unattributable blame.

By hypothesis, $B = 0$. There will be unattributable blame if $B^+ > 0$. Thus, there are two reasons for which someone might think there is a responsibility gap ($B^+ > B$). First, she mistakenly believes that no blame can be attributed (it is not the case that $B = 0$). Second, she falsely believes that blame is appropriate (it is not the case that $B^+ > 0$). We discuss both mistakes in Sect. 2. There, we argue that the first mistake is made by those who fail to appreciate that agents might be responsible for the harm indirectly. The second mistake is due to a failure to see that some harms are blameless. To avoid this second mistake, it is important to see why blame is believed to be appropriate. Some invoke intuitions about situations in which autonomous machines cause harm.

For instance, Köhler, Roughley and Sauer assume that, when harm is caused, some amount of blame is 'prima facie fitting.' (2017, p. 54) Note, however, that what seems prima facie fitting can turn out to be inappropriate after closer inspection.

An alternative approach to responsibility gaps compares autonomous agents who cause harm to human beings who do so in similar circumstances. There is a responsibility gap in this sense, if the human being were to blame. This idea supports a second conception of responsibility gaps:

> [*RG**] A responsibility gap exists exactly if an autonomous machine causes harm, no one is to blame for it, but blame would be appropriate had it been caused by a human being.[7]

We can find this conception explicitly in Himmelreich (2019, p. 734), but is also alluded to by Sparrow when he notes, in relation to war crimes due to autonomous weapon systems, that "[h]ad a human being committed the act, they would immediately be charged with a war crime" (2007, p. 66). In contrast to *RG*, this definition is as such perfectly coherent. After all, there is nothing mysterious about the actual situation being different from how things could have been. Yet, it is assumed that, if there is such a difference, there is a problem, similar to when a debt cannot be collected. In this case, there is no unattributable blame. But there is a deficit that needs to be solved. In light of this, we refer to this second conception of responsibility gaps as 'the deficit conception.'[8]

According to this deficit conception, the situation in which the harm occurs is to be compared to that in which a human agent causes the same harm under circumstances that are as similar as possible. Thus, the actual world in which an autonomous machine causes harm is to be compared to the closest world in which that harm is caused by a human being. The extent to which the human being is to blame serves as a reference point (B^*). A responsibility deficit exists exactly if: $B^* > B$. At this point, the claim is merely that there is a difference between the actual world and a nearby possible world. As mentioned, this is not problematic as such. To illustrate this, compare a very young child who breaks a vase to an adult who does so under basically the same circumstances. Both are accidents. But it may be that the adult should have been more careful. If so, the child is not to blame while the adult is. This difference is perfectly intelligible. Furthermore, such differences across cases can be insightful, for instance when considering their legal ramifications. Similarly, deficits can also inform liability for harm caused by autonomous machines.[9]

Even so, the deficit conception plays a problematic role in the debate. Its proponents combine the observation that the two are different with the following assumption: the

[7] This definition is very similar to Himmelreich's (2019) description of a responsibility gap. He introduces the notion of a minimal agent, which is an agent that has intentional but lacks moral agency. Given this notion, he holds that "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" (734).

[8] Philip Pettit (2007, p. 194) refers to a responsibility gap as 'a responsibility deficit.' However, instead of with autonomous machines, he is concerned with the attribution of responsibility to organizations or collective agents.

[9] We thank an anonymous referee for pointing this out to us.

amount of blame that is appropriate in this nearby possible world is a suitable point of reference for the actual world. This 'reference-point assumption' comes down to the claim that the amount of blame that is appropriate in the nearby possible world is equal to the amount of blame that is appropriate in the actual world ($B^* = B^+$). However, there is no good reason to think that this assumption is correct, as Himmelreich seems to acknowledge (2019, fn. 14). What is even worse, the reference point assumption renders *RG** inconsistent for the same reason as *RG*. For there to be a deficit, the amount of blame appropriate in the nearby possible world has to be more than the blame that is in fact appropriate ($B^* > B$). We already established that the amount of blame that can in fact be attributed must be identical to the amount that is appropriate ($\mathrm{B} = B^+$). And, these to claims together are inconsistent with the reference-point assumption ($B^* = B^+$). Hence, the deficit conception—in combination with the reference-point assumption—suffers from the same problem that proved to be fatal for the remainder conception.

There is just one difference, which is not without significance. The deficit conception makes explicit the source of intuitions about blameworthiness. Below we take issue with the claim that autonomous machines should be compared to human beings. We will argue that it is more pertinent, in this context, to compare them to non-autonomous machines, such as household appliances. Both are artificial. And both are created, maintained and regulated by human beings. The thing to appreciate is that non-autonomous machines raise the same two issues we highlighted before: blameless harm and indirect blame. We will argue that, once they are properly taken into account, it becomes apparent that all blame that is appropriate can be attributed. This means that our responsibility practices are perfectly coherent and that there is no need to consider the possibility of responsibility gaps.

2 Responsibility gaps are unmotivated

2.1 Blameless harm

The notion of a responsibility gap is not only *incoherent*, but also *redundant*. We support this second claim by arguing that all blame that is appropriate can be attributed. More specifically, either the harm caused by an autonomous machine is blameless, or someone else is to blame for it indirectly. This reveals that the notion of a responsibility gap is *unmotivated*. Our argument provides indirect support for the main argument presented in Sect. 1, if only because it makes it easier to see that it is indeed correct. More importantly, it prepares for the claim that we defend in Sect. 3: instead of responsibility gaps, there can be control gaps. For this purpose, we need the notion of a morally acceptable level of risk. We defend and develop this notion by comparing cases of blameless harm and indirect blame involving human beings to those involving autonomous machines. And we argue that harm caused by an autonomous machine is blameless if the level of risk to which it exposed others was morally acceptable and that people are indirectly responsible for such harm if that risk level was higher than it should have been.

To argue that some harms caused by autonomous machines are blameless, we compare an autonomous or self-driving car with a non-autonomous car with a human driver (a level 5 car to a level 0).[10] Suppose the self-driving car is produced by car manufacturer CM and that its operating system is called 'Ariadne' (we use that name also for the car as such). The ordinary car to which we compare Ariadne is driven by Bertram. Now, suppose that Bertram is driving on a road when a pedestrian, Chloe, suddenly appears from behind a bunch of trees and runs across the street. It is too late for Bertram to stop. His car hits Chloe and she is hurt. Is Bertram to blame for the accident? It seems not. Chloe stepped on the street so suddenly that he could not have avoided her. What about Chloe? It so happens that she checked the road before disappearing behind the trees. She failed to spot Bertram's car due to a non-culpable oversight. Hence, the harm is blameless.

Even though she is an autonomous machine, it seems to us that Ariadne can also be implicated in harms that are blameless. Suppose that the accident was caused by Ariadne instead of Bertram. As before, Chloe is not to blame. Ariadne will not be blameworthy either because she is not a moral agent. But why think that blame is appropriate at all? The circumstances under which the harm occurs is identical to those under which Bertram caused an accident. And just as Bertram, Ariadne was not prepared for them. They were so challenging that the harm was unavoidable. Bertram could not reasonably have been expected to avoid the harm. Similarly, given the state of the technology, it might be unreasonable to expect Ariadne to do so. If so, blame is not appropriate in either case.

Bertram was not to blame for the harm he caused because he was excused. Although he caused harm, he was not at fault. Someone who has an excuse has done something that is wrong, but some defeating condition undermines the ground for attributing blame to him (Wallace, 1994, 2019). If no one else is to blame, the harm is blameless. But how can harm be blameless when it is caused by an autonomous machine, which cannot be excused?

Now, we expect autonomous machines to pose fewer risks to others than human beings. For instance, self-driving cars will be allowed on the road only if they are less prone to accidents than human drivers. But the risk of harm will never be zero, as there are no risk-free technologies. Thus, there will be some level of risk that is morally acceptable within a society.[11] In light of this, we propose that, if Ariadne meets this standard and functions properly, any harm she causes is blameless. Such situations form a strict subset of those in which Bertram causes harm without being blameworthy for doing so. Thus, the comparison between Ariadne and Bertram reveals that some

[10] The Society for Automotive Engineering (SAE) distinguishes five levels of automation. Level 0 is an ordinary non-autonomous car that is controlled directly by a human agent. On levels 1 to 4, some control can be delegated to the vehicle. Human agents can monitor what the car does and intervene when this is needed or desired. Except when they intervene, their control is indirect. At higher levels, the car will do certain things by itself even when a human agent takes control. This means that the direct or operational control she has when she intervenes is partial. On level 5, a human driver is not necessary. In fact, a steering wheel is optional and all human agents who occupy the car are considered to be passengers. At this level, cars are fully automated. They are autonomous machines.

[11] Sparrow and Howard (2017) and Nyholm (2018b) argue that, once self-driving or level 5 cars are substantially safer than level 0 cars, it is morally problematic to keep on using the latter.

harms for which blame cannot be attributed are blameless ($B^+ = 0$). Hence, there is no need to consider the possibility of responsibility gaps.

If Ariadne were a human agent, she would be excused because of the fact that Chloe suddenly appeared from behind a bunch of trees and ran across the street. However, as Ariadne is not a moral agent, she does not need an excuse. As an autonomous machine, she is not susceptible to blame. In practice, we actually have higher expectations of autonomous machines. The relevant standard is the level of risk that is morally acceptable within the society at issue. The idea is that, when an autonomous machine functions normally, it exposes others to a certain risk of harm. Suppose that it causes harm, but there is nothing out of the ordinary. Although the harm will be disconcerting, there will be no reason for concern beyond this if the actual risk of harm to which it exposed others did not exceed the level that is morally acceptable. In such a situation, I propose, it is unproblematic that no blame can be attributed. Importantly, this account does not require recourse to the notion of an excuse.

The underlying intuition is simple. Sometimes an accident is just that: an accident. In such cases, the best explanation of the harm invokes the circumstances in which it occurred rather than the agent who caused it. However, unfortunate it is, there is no need to attribute blame to anyone.

2.2 Indirect blame

Autonomous machines are technical artifacts. As such, they are designed, engineered, programmed, manufactured and regulated. These processes involve what we call 'enablers', which include designers, engineers, software developers, manufacturers and regulators. They create and maintain the conditions under which autonomous machines can act. Regulators make the laws and policies about the admissibility of the relevant technology. Thus, enablers make it possible for autonomous machines to function in society in an acceptable manner. Because of this, they are the salient candidates for indirect responsibility. One of the concerns that the argument for responsibility gaps raises is that autonomous machines make decisions and perform actions independently of other agents. This means that other agents lack direct control over them. And it rules out that enablers bear direct responsibility. But not indirect responsibility.

The famous example of indirect responsibility is that of a drunk driver who kills a pedestrian. As he was incapacitated at the time, he has an excuse. However, he is responsible for getting into the car while drunk. And because of this, he is to blame for her death (Robichaud & Wieland, 2019; Rosen, 2004; Smith, 1983; Zimmerman, 1997 rejects the idea). In such cases, the excuse clears the agent from direct responsibility but not from indirect responsibility. Again the question arises: How can this notion apply to autonomous machines given that they cannot have excuses?

The drunk driver was blameworthy because he was responsible for drunk driving. This suggests that, for an agent to be indirectly responsible, it should have acted differently at a prior point, to avoid that the situation in which the harm occurred ever arose (Wallace, 1994, 2019).

Now, the drunk driving example is such that the agent who has an excuse is the same as the agent who bears indirect responsibility. But this need not be the case. Suppose that someone else got him drunk or cut the brake line of his car. It seems that this agent will then be to blame for the death of the pedestrian, even though she was not directly implicated in the accident. When an agent is excused for causing harm, the blame for the harm transfers to whomever facilitated or enabled the harm and is thereby responsible for the circumstances that excuse the agent who caused it.

To see how the notion of indirect responsibility might apply to autonomous machines, we compare them to non-autonomous machines, such as kitchen appliances. Both are designed, engineered, manufactured and regulated. And in both cases, the enablers are not directly involved in situations where harm occurs. But their actions can be conducive to the circumstances under which they come about. Suppose you use and clean your kitchen appliance properly following the maintenance instructions. In spite of this, it causes a short-circuit and a fire in your kitchen. Who is to blame? The kitchen appliance is not supposed to malfunction in this way. Suppose that the manufacturer could and should have prevented it from doing so. Then it is to blame for the harm, albeit indirectly. The same can be seen when we consider ordinary non-autonomous cars. Manufacturers actually take responsibility when harms occur because of failures by compensating the victims, recalling the cars and fixing the defect free of charge.[12] More generally, enablers bear all the blame that is to be attributed for harm caused by non-autonomous machines. We propose that this holds for autonomous machines as well: if blame is appropriate, it falls on the enablers.

Importantly, blame will be appropriate only if the level of risk to which they expose others is morally unacceptable. As just discussed, for an agent to be indirectly responsible for a harm in these cases, he must be directly responsible for the circumstances under which it could occur. Furthermore, those circumstances must have been subpar: because of them, the autonomous machine exposed others to an unacceptably high level of risk. Thus, suppose that Ariadne's software is updated. In many respects this update is an improvement. However, it makes her worse at recognizing pedestrians. Even when there is enough time for her to stop, pedestrians get hurt. After a brief investigation, CM recommends owners to revert to the previous software version, while it continues to investigate the problem more thoroughly. Unless there are mitigating circumstances, CM will be to blame, including in particular its software developers. The problem is that Ariadne poses a morally unacceptable level of risk to others. This is why the software developers are to blame.

Now, suppose that it is not feasible to bring the level of risk down to an acceptable level, or that there was not enough evidence that this standard was met. In these cases, the autonomous machines should have been deemed inadmissible: its use should have been prohibited in contexts where others might get hurt. Suppose, for instance, that the algorithms used are self-learning and opaque, and therefore unpredictable (Matthias, 2004; Santoni de Sio & Mecacci, 2021). If no reliable estimate can be given of the risks to which they expose others, the machine should not be deployed. And if it is, the regulators, perhaps together with other enablers, will be to blame.

[12] For an example involving a number of car manufacturers see:https://www.nhtsa.gov/equipment/takata-recall-spotlight.

The upshot is that, for an autonomous machine to be admissible, the level of risk to which it poses others has to be morally acceptable. If this is the case, any harm caused will be blameless. And if it is not the case, enablers will be to blame indirectly. Hence, the notion of a responsibility gap is unmotivated, as all blame that is appropriate can be attributed.

3 Control gaps

3.1 Insufficient control

Autonomous machines do present a deep and urgent problem for society, which concerns the level of risk to which they expose others. The key issue is what risk level is morally acceptable. In practice, this raises difficult questions. Is the level of risk to which autonomous machines actually expose others at or below this level? And when do enablers have enough evidence that this is the case? However, we are concerned with the nature and significance of the problem. We consider cases in which, we assume, the risk level is unacceptably low. And we argue that, if this is the case, autonomous machines exhibit a control gap.

Now, a normally functioning moral agent does not exhibit a control gap, or so we argue. To support this claim, we first discuss what moral agency is. Just as intentional agents, moral agents can engage in purposeful action. What distinguishes them from mere intentional agents is that they possess moral competence. This means that normative considerations, such as harm and fairness, play an important role in their decision making and that they guide their actions. More precisely, an agent has normative competence if it has the ability to decide and act in a way that is appropriately sensitive to normative considerations (Wallace, 1994). A rather influential way of understanding what this means is in terms of guidance control: an agent possesses guidance control exactly if it is (moderately) responsive to moral reasons, which means that it is receptive and (at least weakly reactive) to such considerations (Fischer & Ravizza, 1999).[13]

Guidance control is a threshold notion. This means that an agent needs to be sufficiently, i.e. moderately, reason-responsive in order to be a moral agent. Only if this is the case does it make sense to blame it. Furthermore, it characterizes the ways in which an agent responds to normative considerations and thereby the role they play in perception, deliberation and action. As such, the notion is best understood in functionalist terms, which means that it applies not only to human beings, but at least potentially also to collective agents and robots.[14]

Autonomous machines are intentional agents. However, they can (currently) not be moral agents, as they do not possess guidance control. It is possible to program a machine such that it avoids obstacles across a range of circumstances. This enables it to

[13] Shoemaker (2015) and Sripada (2015) characterize reasons-responsiveness as one of two influential accounts of moral agency, the other being the deep-self approach.

[14] Functionalism is sometimes contrasted with standardism, the view that moral agency should be present in an agent in the same way as it is found in humans (Behdadi & Munthe, 2020). For a discussion of moral agency as guidance control in the context of collective agents, see (Hindriks, 2018).

regularly avoid causing harm. However, to be genuinely responsive to harm, it should also be able to understand what it means to harm another agent. Such understanding requires the ability Furthermore, it needs to be able to take the perspective of another agent along with the concept of well-being and see how things are from the other's point of view (Darwall, 1998). Autonomous machines are not able to engage in perspective-taking and lack the concept of well-being. Because of this, they are not as such suitably sensitive to harm as such. Notwithstanding the hype which surrounds AI research, current machine learning algorithms are closer to basic statistical inference than to a proper understanding of the world (Marcus & Davis, 2019; Fjelland, 2020).

A control gap exists whenever the risk level of an autonomous machine is too high. There will then be a discrepancy between the control the autonomous machine actually has and the control it should have. But what kinds of control are at issue here? In enacting a decision, an autonomous agent exerts causal control. Furthermore, its behavior should emulate that of a suitably motivated moral agent. This means that it should approximate the behavior of a well-intentioned agent with guidance control[15]:

> [*CG*] An autonomous machine exhibits a control gap exactly if there is a discrepancy between the causal control it has and the guidance control it should have or emulate.

Thus, the normative standard that an autonomous machine has to meet is set by the risk level that is (deemed to be) morally acceptable. A risk level typically consists of two elements: the probability of the hazard-related incident or exposure occurring and the severity of the harm caused by it. The control that an autonomous machine has is deficient if it poses too high a level of risk (for example, in situations involving children).

There are two ways in which a control gap can be decreased, if not avoided altogether. First, by adding proactive and reactive barriers (Van Nunen et al., 2018). Proactive safety barriers prevent the occurrence of an undesirable event, while reactive safety barriers are in place to control or to mitigate the consequences of the event. This is the classical approach of safety engineering (Sklet, 2006). Second, by improving the extent to which the machine can emulate guidance control. This can be done, for instance, by programming the agent such that the proxies it uses for harm correlate better with actual harm. Consider Ariadne once again. Software developers and engineers should, for instance, enable her to recognize pedestrians and avoid hitting them when they step on the road. The better Ariadne is calibrated, the smaller the control gap that Ariadne exhibits. Once her level of risk is morally acceptable, she will not

[15] In the context of AI, minimalist conceptions of moral agency have been proposed. Floridi (2013, see Gunkel 2012 for a criticism of Floridi's ideas that were already present in Floridi and Sanders 2004), for example, holds that if an agent "can cause moral good or evil" (Floridi 2013, p. 147), then it can for this reason be identified as a moral agent. However, such a minimalist conception is less useful for the purpose of this paper for at least two reasons. First, identifying an agent as a moral agent in this way does nothing to alleviate the worry that there might be a control gap. On the contrary, if an agent can cause moral good or evil, then this is precisely a reason to make sure that there are no control gaps. Secondly, from the fact that something is a minimal moral agent in Floridi's sense, we cannot infer that it is a suitable target for responsibility attributions (Floridi 2013, p. 151). This makes the minimalist account also less suitable in the context of responsibility gaps. See also footnote 1 and 13.

exhibit such a gap at all. She will then be able to avert a wide range of harms, even though she does not recognize them as such.

As we see it, control gaps have been mistaken for responsibility gaps. Suppose an autonomous machine exhibits a control gap and causes harm. Although no one was directly implicated in the situation where the harm was brought about, the enablers will be responsible for this. The fact that the machine exhibits a control gap implies that they have not suitably equipped the autonomous machine with the causal control it should have. But they might not have been in a position to do so. For instance, they might have employed machine learning, which is notoriously opaque. Furthermore, it might have been unclear what level of risk was (deemed to be) morally acceptable and what counts as sufficient knowledge in this respect. Policies and regulations might have lagged behind. Because of this, enablers might not have known enough about what they were supposed to do (Santoni de Sio & Mecacci, 2021).

Under such circumstances, enablers will have excuses, which limit the extent to which they can be blamed. It will certainly be less than the extent to which a human being would be to blame had he caused the harm at issue. And it will also be less as compared to the responsibility they would incur for non-autonomous machines assuming that in their case the questions mentioned have relatively clear answers. This is obviously unsatisfactory, which is why it has been taken to imply the existence of responsibility gaps. However, the problem is emphatically not that the blame that is appropriate cannot be attributed. There is a problem with how things are, not with the amount of attributable blame. In particular, the autonomous machine does not have sufficient control over its actions and their consequences. It exhibits a control gap. Because of this, it exposes others to a level of risk that is too high. That gap ought to be closed such that this goes down to a level that is morally acceptable.[16]

3.2 Meaningful human control

Control gaps are not limited to autonomous machines as we conceive of them. Machines can also exhibit them when they have moral agency and when they are semi-autonomous, in which case they are partially under the control of operators. Moral agents possess guidance control. And guidance control encompasses the abilities that constitute causal control (Fischer & Ravizza, 1999). If machines that are moral agents exhibit a control gap, they suffer from a deficiency in their guidance control. As a consequence, they will be unable to avoid harm in some situations where they should do so. It may be, for instance, that they appreciate the moral significance of harm and realize that they should prevent it. At the same time, their ability to avert it is subpar.

By way of analogy, consider someone who suffers from Tourette's syndrome along with coprolalia. He utters obscene words or makes socially inappropriate and derogatory remarks due to an unwanted urge to do so. But he acknowledges that he should not do so. He recognizes that it would be better if he could suppress the urge or not have it at all. Even so, he is unable to comply with his own moral standards. Such an agent

[16] The problem of how to close the control gap resembles what is known as 'the AI control problem', which concerns super-intelligent AI (Bostrom, 2014; Russell, 2019).

exhibits a control gap. However, he is not to blame for his inappropriate behavior. It is not an eligible option for him to avoid all the situations in which he might end up acting in an inappropriate manner. In contrast, machines that are moral agents can and should refrain from operating under conditions where they pose an unacceptable level of risk.

Semi-autonomous machines have partial control over their actions. Think, for instance, of a car with lane assist. The operator determines the lane in which the car drives. The car helps him to stay in that lane. Semi-autonomous machines can have full control over what they do some of the time, as when a plane is on autopilot during part of a flight. A key consideration to make a machine semi-autonomous is if it is not safe enough to be fully autonomous. This is why it came as a shock that a Tesla vehicle could easily be tricked into operating without a driver.[17] If it is too risky for a machine to function entirely on its own, it exhibits a control gap. In order for it to be admissible, we propose, it should be designed such that the operator can fill this gap.

According to an increasingly popular view, semi-autonomous machines should be under meaningful human control (see Horowitz & Scharre, 2015, Roff & Moyes, 2016 and Ekelhof, 2019).[18] We argue that this notion can be used so as to make our proposal more concrete. The key idea is that the operator of a semi-autonomous machine must be in a position to suitably interact with it. Although the notion is still under construction, a few things are widely agreed upon. To begin with, the operator must be aware of the possibilities and constraints of the machine. Conversely, the machine should be attuned to the physical and mental limitations of the human operator. Most importantly, the operator should be in a position to intervene, to adjust the actions of the machine, for instance so as to avert harm. To this end, the operator should be able to disengage certain automatic systems when appropriate and to overrule the machine in certain situations. But not, for instance, if the intervention would require super human reaction times. At the same time, there should be certain actions that the machine should not be able to initiate, as the Tesla example illustrates.

Meaningful human control is a notion of control that goes beyond mechanisms that have been labeled as "human-in-the-loop." Someone may, for example, be present "in-the-loop" and therefore able to intervene causally in the operations of the system, while not having sufficient knowledge to actually influence the process in a meaningful way. For example, if a human is trained to press a button when the AI system tells it to, then the human is "in-the-loop," but there is no meaningful human control over the system (Horowitz & Scharre, 2015). Meaningful human control should also be distinguished from "human-on-the-loop" governance mechanisms, at least if it is understood as the capability for human intervention in all decisions that the AI system makes. Meaningful human control is compatible with an inability to influence the operation of a system in

[17] After a fatal crash involving a Tesla vehicle with no one in the driver seat, it was discovered that it could easily be tricked into operating without a driver. The Consumer Reports team said: 'It was a bit frightening when we realized how easy it was to defeat the safeguards, which we proved were clearly insufficient.' See https://www.bbc.co.uk/news/technology-56854417.

[18] Nyholm (2018a) discusses similar issues in terms of 'collaborative' and 'supervised agency' (see Köhler, 2020 for a critical discussion).

cases where the system is responsive to the relevant moral reasons (Cavalcante Siebert et al., 2022; Veluwenkamp, 2022).

To give content to the moral dimension of the notion, Santoni de Sio and Van den Hoven (2018) propose to explicate meaningful human control in terms of guidance control. The idea is that semi-autonomous machines form socio-technical systems together with the operators and that such systems act as if they possess guidance control. There may be situations in which machines emulate such control on their own. In such cases, they track moral reasons, which means that they are 'demonstrably and verifiably responsive to the *human* moral reasons relevant in the circumstances' (ibid., p. 7). But in other situations, the operator should be able to intervene such that reasons only he recognizes are decisive for what it does.

It is widely believed that meaningful human control prevents responsibility gaps (Santoni de Sio and Van den Hoven, 2018, Cavalcante Siebert et al., 2022). However, if our earlier arguments are correct, this cannot be right: all blame that should be attributed can in principle be attributed. We propose, instead, that it serves to close control gaps. More precisely, the point of designing a semi-autonomous machine so as to be under meaningful human control is to enable operators to suitably intervene and thereby be in a position to fill the control gap. When this has been achieved, the socio-technical system as a whole can act as if it possesses guidance control. This in turn means that the operator is able to appropriately guide the machine so as to decrease the level of risk in situations where the performance of the semi-autonomous machine is wanting. If he fails to do so, he may well be to blame for the harm that ensues.

In sum, both semi-autonomous and autonomous machines can have flaws due to which their ability to prevent harm is subpar. When this is the case, there will be a control gap. And the risk of harm to which it exposes others will be too high. The machine will then be inadmissible. In principle, the enablers are to blame for any harm caused, as they have failed to ensure that its risk level is morally acceptable. Now, suppose that an autonomous machine does not exhibit a control gap. In that case, any harm it causes will be blameless. Imagine next that a semi-autonomous machine that is under meaningful human control causes harm. If it is not blameless, then its operator will be blameworthy for failing to prevent it. Thus, the ultimate challenge concerning such machines is not the attribution of responsibility. Instead, it is the construction of control.

4 Conclusion

In this paper we have argued that the current focus on responsibility gaps should be shifted towards control gaps. The main reason for this is that responsibility gaps do not exist. To show this we have given two arguments: an a priori argument which shows that the notion is incoherent and a general argument which shows that in alleged cases of responsibility gaps either no one is to blame, or the blame that should be ascribed can be attributed to people who bear indirect responsibility. As we see it, in many of the morally problematic cases, there is a control gap: i.e., a discrepancy between the causal control the autonomous machine has and the guidance control it should have

or emulate. This discrepancy causes others to be exposed to a level of risk that is too high. The control gap ought to be closed such that the level of risk goes down to a level that is morally acceptable. To achieve this, we propose to design autonomous machines so as to be under meaningful human control.

Acknowledgements We gratefully acknowledge helpful comments from Hein Duijf, Niels de Haan, Sven Nyholm and David Schweikard.

Funding The research by Herman Veluwenkamp is supported by the Delft Digital Ethics Centre.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine, 28*(4), 15.

Behdadi, D., & Munthe, C. (2020). A normative approach to artificial moral agency. *Minds and Machines, 30*(2), 195–218.

Bostrom, N. (2014). *Superintelligence: Paths, dangers*. Oxford University Press.

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., et al. (2022). Meaningful human control: Actionable properties for AI system development. *AI Ethics*. https://doi.org/10.1007/s43681-022-00167-3

Coeckelbergh, M. (2016). Responsibility and the moral phenomenology of using self-driving cars. *Applied Artificial Intelligence, 30*, 748–757.

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

Danaher, J., & Nyholm, S. (2020). Automation, work and the achievement gap. *AI and Ethics, 1*(3), 227–237.

Darwall, S. (1998). Empathy, sympathy, care. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 89*(2/3), 261–282.

Ekelhof, M. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy, 10*(3), 343–348.

Fischer, J. M., & Ravizza, M. (1999). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications, 7*(1), 10. https://doi.org/10.1057/s41599-020-0494-4.

Floridi, L. (2013). *The ethics of information*. Oxford University Press UK.

Floridi, L., & Sanders, J. W. (2004). On the morality of artifcial agents. *Minds and Machines, 14*(3), 349–379.

Gunkel, D. J. (2012). *The machine question: Critical perspectives on Ai, robots, and ethics*. MIT Press.

Gunkel, D. J. (2017). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology, 22*(4), 307–320.

Gunkel, D. J. (2020). A vindication of the rights of machines. In *Machine Ethics and Robot Ethics* (pp. 511–530). Routledge.
Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and Information Technology, 15*, 99–107.
Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for crashes of autonomous vehicles: An ethical analysis. *Science and Engineering Ethics, 21*, 619–630.
Himmelreich, J. (2019). Responsibility for killer robots. *Ethical Theory and Moral Practice, 22*, 731–747.
Hindriks, F. (2018). Collective agency: Moral and amoral: Collective agency: Moral and amoral. *Dialectica, 72*(1), 3–23. https://doi.org/10.1111/1746-8361.12215
Horowitz, M., & Scharre, P. (2015). *Meaningful human control in weapon systems: A primer*. Center for a New American Security.
Köhler, S. (2020). Instrumental robots. *Science and Engineering Ethics, 26*(6), 3121–3141.
Köhler, S., Roughley, N., & Sauer, H. (2017). Technologically blurred accountability? Technology, responsibility gaps and the robustness of our everyday conceptual scheme. In *Moral agency and the politics of responsibility* (pp. 51–68). Routledge.
Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.
Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Vintage.
Mecacci, G., & Santoni de Sio, F. (2020). Meaningful human control as reason- responsiveness: The case of dual-mode vehicles. *Ethics and Information Technology, 22*(2), 103–115.
Nyholm, S. (2018a). Attributing agency to automated systems: On human-robot collaborations and responsibility-loci. *Science and Engineering Ethics, 24*, 1201–1219.
Nyholm, S. (2018b). The ethics of crashes with self-driving cars: A roadmap, II. *Philosophy Compass, 13*(7), e12507.
Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield Publishers.
Pettit, P. (2007). Responsibility incorporated. *Ethics, 117*(2), 171–201.
Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice, 18*, 851–872.
Robichaud, P., & Wieland, J. W. (2019). A puzzle concerning blame transfer. *Philosophy and Phenomenological Research, 99*(1), 3–26.
Robillard, M. (2018). No such thing as killer robots. *Journal of Applied Philosophy, 35*, 705–717.
Roff, H. (2013). Killing in war: Responsibility, liability, and lethal autonomous robots. In F. Allhoff, N. Evans, & A. Henschke (Eds.), *Routledge handbook of ethics and war: Just war theory in the 21st century*. London: Routledge.
Roff, H. M., & 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*.
Rosen, G. (2004). Skepticism about moral responsibility. *Philosophical Perspectives, 18*, 295–313.
Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.
Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*. https://doi.org/10.1007/s13347-021-00450-x
Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*, 15.
Shoemaker, D. (2015). *Responsibility from the margins*. Oxford University Press.
Sklet, S. (2006). Safety barriers: Definition, classification, and performance. *Journal of Loss Prevention in the Process Industries, 19*(5), 494–506.
Smith, H. M. (1983). Culpable ignorance. *Philosophical Review, 92*, 543–571.
Simpson, T. W., & Müller, V. C. (2016). Just war and robots' killings. *The Philosophical Quarterly, 66*(263), 302–322.
Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.
Sparrow, R., & Howard, M. (2017). When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport. *Transportation Research Part C: Emerging Technologies, 80*, 206–215.
Streumer, B. (2007). Reasons and impossibility. *Philosophical Studies, 136*(3), 351–384.
Sripada, C. (2015). Moral responsibility, reasons and the self. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility* (pp. 242–264). Oxford University Press.
Tigard, D. W. (2020). There is no techno-responsibility gap. *Philosophy & Technology, 34*(3), 589–607.

Van Nunen, K., Swuste, P., Reniers, G., Paltrinieri, N., Aneziris, O., & Ponnet, K. (2018). Improving pallet mover safety in the manufacturing industry: A bow-tie analysis of accident scenarios. *Materials, 11*(10), 1955.

Veluwenkamp, H. (2022). Reasons for meaningful human control. *Ethics and Information Technology, 24*(4), 51. https://doi.org/10.1007/s10676-022-09673-8

Villiers de, E. (2002). Who will bear moral responsibility? *Communicatio, 28*(1), 16–21.

Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Harvard University Press.

Wallace, R. J. (2019). *The moral nexus*. Princeton University Press.

Zimmerman, M. J. (1997). Moral responsibility and ignorance. *Ethics, 107*, 410–426.