Ethical Theory and Moral Practice (2022) 25:575–593
https://doi.org/10.1007/s10677-022-10313-9

Can we Bridge AI's responsibility gap at Will?

Maximilian Kiener[1]

Accepted: 4 July 2022 / Published online: 29 July 2022

Abstract
Artificial intelligence (AI) increasingly executes tasks that previously only humans could do, such as drive a car, fight in war, or perform a medical operation. However, as the very best AI systems tend to be the least controllable and the least transparent, some scholars argued that humans can no longer be morally responsible for some of the AI-caused outcomes, which would then result in a responsibility gap. In this paper, I assume, for the sake of argument, that at least some of the most sophisticated AI systems do indeed create responsibility gaps, and I ask whether we can bridge these gaps *at will*, viz. whether certain people could *take* responsibility for AI-caused harm simply by performing a certain speech act, just as people can give permission for something simply by performing the act of consent. So understood, taking responsibility would be a genuine normative power. I first discuss and reject the view of Champagne and Tonkens, who advocate a view of taking *liability*. According to this view, a military commander can and must, ahead of time, accept liability to blame and punishment for any harm caused by autonomous weapon systems under her command. I then defend my own proposal of taking *answerability*, viz. the view that people can makes themselves morally answerable for the harm caused by AI systems, not only ahead of time but also when harm has already been caused.

Keywords Artificial Intelligence · Responsibility gap · Normative Powers · Answerability · Liability

1 Introduction

The best artificial intelligence (AI) is often the least controllable and the least transparent. It is the least controllable because, among other reasons, some of its processes are too fast to be monitored in real time (Coeckelbergh 2020; Sparrow 2007), and it is the least transparent because some of its technological bases, e.g. deep neural networks, create 'black

✉ Maximilian Kiener
maximilian.kiener@philosophy.ox.ac.uk

[1] Faculty of Philosophy, University of Oxford, University College, High Street, OX1 4BH Oxford, UK

boxes' with impenetrably complex algorithms (Bathaee 2018; Wang et al. 2019). However, insofar as the ability to control and anticipate outcomes is key to human moral responsibility, a pressing question emerges: how could anyone ever be morally responsible when such sophisticated yet uncontrollable and opaque AI causes harm, e.g. when an autonomous car runs over a pedestrian, when a surgical robot paralyses a patient, or when an autonomous weapon system kills a non-combatant?

This question triggered a controversial debate on whether AI creates 'responsibility gaps', i.e. situations in which no one is morally responsible for the harm caused by AI.[1] Some scholars have argued that the use of AI creates responsibility gaps (Danaher 2016; De Jong 2020; Matthias 2004; Sparrow 2007) precisely because it removes control and understanding, whereas others have denied the existence of such gaps (Himmelreich 2019; Johnson 2015; Tigard 2020). Still others have endorsed an intermediate position, arguing that gaps exist but that they are narrower than usually thought (Nyholm 2020, Chap. 3; Schulzke 2013; Simpson and Müller 2016).

In this paper, I will assume – for the sake of argument – that at least some of the most sophisticated AI systems do create responsibility gaps, and I will ask whether we can bridge these gaps *at will*, viz. whether we can *take responsibility* for AI-caused harm simply by performing a certain speech act, viz. communicating the intention of hereby making oneself responsible, which I understand to be analogous to how people give permission for something (via consent) simply by communicating the intention of hereby giving a permission. So understood, taking responsibility would be a genuine 'normative power'. Thus, if this normative power exists, people could simply *make* themselves morally responsible for something they would otherwise not be responsible for. I will argue that a certain form of such a normative power does indeed exist, and I thereby aim to address not only a pressing challenge in current AI ethics but also to shift some of the central paradigms in the debate on moral responsibility more generally.

I will proceed as follows: I will first discuss and reject the view of Champagne and Tonkens (henceforward 'CT'), who advocate a view of *taking liability*. According to this view, a military commander can and must, ahead of time, accept liability to blame and punishment for any harm caused by autonomous weapon systems under her command. I will then defend my own proposal of *taking answerability*, viz. the view that people can makes themselves morally answerable for the harm caused by AI systems, not only ahead of time, but also when harm has already been caused. Finally, I will outline the extent to which the normative power of taking answerability can close AI's responsibility gaps and also explain how my account can remain attractive even to those who deny the existence of genuine responsibility gaps.

But before I proceed, let me clarify two aspects of this paper. First, the only responsibility gap that can be addressed by means of a normative power will be the gap concerned with responsibility as answerability, rather than responsibility as liability to blame or punishment. Second, even in addressing the responsibility-as-answerability gap, the effect of the proposed normative power may be limited in the end.

[1] For other formulations of what a 'responsibility gap' is, see Himmelreich (2019) and List (2021).

2 CT's proposal: the normative power of taking prospective liability

2.1 Outline of CT's view

CT focus on AI in military contexts, especially on autonomous weapon systems, and they accept that those systems create responsibility gaps (or 'vacuums' as they also call them). Gaps arise, they argue, because '(1) a programmer [or any other human person in the development and use of the AI] cannot be justifiably held responsible for the actions of a truly autonomous robot and (…) (2) holding a nonsentient robot responsible is meaningless and unsatisfactory' (Champagne and Tonkens 2015, p. 134). However, CT still claim that 'there is a way to seal the responsibility vacuum' (p. 126). More precisely, they argue for.

> 'the possibility of "blank check" responsibility: A person (or persons) of sufficiently high military or political standing could accept responsibility for the actions (normal or abnormal) of all autonomous robotic devices—even if that person could not be causally linked to those actions besides this prior agreement.' (p. 126)

And they add:

> 'we regard it as fair and reasonable for the commanding officer to willingly and freely assign responsibility to herself, ahead of time.' (p. 133)

Thus, CT argue that a person can *accept* or *take* responsibility for something they would otherwise not be responsible for. It is a person's act of will or communication that creates this person's responsibility for AI-caused harm in the first place. CT call such a taking of responsibility 'noncausal imputation' (pp. 127, 136) and describe it as entering into an 'explicit social contract' (p. 136). So understood, taking responsibility is a *normative power*: a person takes responsibility for something simply by performing a certain speech act, just as a person gives permission for something simply by performing the act of consent.

More specifically, CT talk about taking *prospective* responsibility rather than *retrospective* responsibility. In other words, people take responsibility only 'ahead of time', viz. for harm that certain AI systems may cause in the future, but not for harm that has already been caused.

In addition, CT talk about responsibility in terms of *liability to blame and punishment.* To take responsibility, they explain, is 'to accept blame (or praise) for whatever robotic acts transpire in war' (p. 132), to become 'liable to receive blame for any self-initiated robot acts deemed immoral' (p. 132), and - as a result - to be 'subject to punishment' (p. 134) for the robots' harmful actions.

Finally, CT argue that certain people, i.e. the 'suitably ranked and sufficiently informed person or (persons)' (Champagne and Tonkens 2015, p. 133), are not only *able* to, but also *have* to, take responsibility. In other words, it is 'a nonnegotiable condition for accepting the position of authority' (p. 132) of, say, the most highly ranked military commander or the politician in charge of a military operation, that these people also take responsibility, ahead of time, for 'whatever robotic acts transpire in war' (p. 132). CT do not specify who else might be *able* to take responsibility for robotic acts, e.g. whether the remote computer programmer might be able to do so too, but they are clear that certain people *must* take responsibility.

Summarising these points, we can see that CT present an account of taking *prospective blame*-responsibility as a *non-negotiable condition* of certain positions of authority, like the position of military commander. To support their view, CT present two main arguments. First, they claim that we already accept something like taking responsibility by an act of will in the realm of medical consent:

> 'It is standard to use informed consent as part of our gauge of the ethics of intersubjective practices (e.g. patient recruitment for experimental research), and with such consent in place, (at least some) responsibility is thereby transferred to the agent doing the consenting.' (p. 133)

CT here argue that medical consent not only creates permission but also shapes responsibility. When a patient consents to a medical procedure, he not only permits his physician to proceed but also takes responsibility for the risks inherent in the medical procedure, viz. the patient foregoes the right to complain if certain risks materialise, at least as long as the medical procedure is performed consistently with professional standards.[2] In a similar way, they argue that people can take responsibility, or can consent to taking responsibility, in the context of AI too.

Second, CT claim that we also already take responsibility when accepting new roles. For instance, when accepting a job at a university, I impose on myself certain obligations concerning teaching, administration, and research that I would otherwise lack. CT argue that the idea that a military commander could take responsibility when accepting the role of commander is very similar. Thus, CT present their proposal as a combination and further development of two already accepted ideas, i.e. the idea that consenting can imply taking responsibility and the idea that we take responsibility when assuming new roles.

2.2 Criticism of CT's view

CT present an intriguing proposal. Yet, their proposal faces two dilemmas and, therefore, we ought to reject it.

2.2.1 Dilemma 1: Either Blame is Inappropriate or CT's Normative Power is Redundant.

In CT's view, when a military commander takes responsibility, the commander will have to accept 'blame' (Champagne and Tonkens 2015, p. 132) and be 'subject to punishment' (p. 134) for the robots' actions, e.g. the unlawful killing of non-combatants. Yet, the military commander may still be without fault. What is more, it is precisely in such cases of faultless causation of harm where CT's 'blank check'-responsibility is supposed to close the arising responsibility gaps.

But how could others genuinely *blame* the commander without implying fault? As Fricker explained: 'blame is out of order when one does bad things through no fault of one's own. If no fault, then no appropriate blame' (Fricker 2016, p. 170). In Gary Watson's view, 'to blame (morally) is to attribute something to a (moral) fault in the agent' (Watson 2004, p. 266). And Wolf says that blame concerns 'judgments and attitudes we may form in response to faulty behavior' (Wolf 2011, p. 332). Thus, one cannot fittingly 'blame' another person

[2] For a discussion of this point in the context of medical AI, see also Kiener (2021, p. 710).

when that person is faultless, and a mere declaration to accept blame cannot make a difference either. Hence, if there is to be genuine blame, there must be fault too.

The rationale behind this position is that, conceptually speaking, we cannot distinguish moral blame from, say, sadness or disappointment without indicating that blame registers (at least perceived) moral fault in its target. If, as *cognitive* theories of blame hold (cf. Hieronymi 2004; Watson 1996), blame is essentially a negative moral judgment about a person or their conduct, then such a judgment requires referring to some moral fault, or otherwise it could not hold up its distinctive moral negativity. If, as *emotional* theories of blame hold (cf. Menges 2017; Shoemaker 2017), blame essentially involves specific forms of emotional engagement, most notably resentment or anger, then some kind of moral fault in the addressee of blame is needed too, or otherwise anger and resentment become unfitting. Or if, as *conative* theories of blame hold, and in particular Scanlon's version (Scanlon 2008), blame registers some moral impairment to an interpersonal relationship, then we also need to refer to a moral fault in the blamed person, or otherwise it would become implausible why there this moral impairment Scanlon has in mind, viz. why someone would have reason to modify their relation to the blamed person by, for instance, not considering them a friend anymore (Scanlon 2008, p. 129). Hence, across these major approaches to blame, some reference to fault is required to make an account of blame not only *intensionally* adequate, i.e. provide a sound explanation of what blame *is* (as opposed to sadness or disappointment), but also *extensionally* adequate, i.e. provide an account that correctly separates blame from other similar, yet distinct reactions or judgments (such as sadness or disappointment).

Following these approaches to blame, I will assume that a person is 'at fault' when they committed a harmful or wrongful act, or omitted to perform some required act, without justification or excuse.[3] In other words, fault is the moral stain that results from conduct which is neither justified nor excused. So understood, fault may arise in a wide range of cases, potentially including cases where people act from a flawed character, even when they could not control their character. Yet, importantly, fault requires more than just doing something that causes harm, e.g. taking a risk that later materialises. This is because when, say, a military commander takes a risk in employing an autonomous weapon system, the commander could, in principle, be justified in doing so (if, after all, it was the best thing to do, all things considered) or be excused (e.g. because the commander was non-culpably ignorant of a crucial aspect of the situation). In either case, the commander would not be at fault, morally speaking, and not blameworthy, even though the risk materialised and harm was done. Cases where the commander is excused and faultlessly causes harm resemble Bernard Williams's famous example of the 'lorry driver who, *through no fault of his*, runs over a child' (Williams 1981, p. 28, emphasis added). Just like Williams's unlucky lorry driver, CT's unlucky commander may cause severe harm while intentionally engaging in a certain activity or taking a certain risk. Yet, both the driver and the commander may have displayed the utmost diligence and caused the harm through no fault of their own. But CT's account applies to such cases of faultless harming and demands that the commander be subjected to blame. What is more, as already mentioned, it is *precisely* in these cases of faultless harming where the responsibility gap arises and CT's account on 'blank check' responsibility purports to close it.

[3] Due to the clause on omissions, note that fault does not require an actual causal link between an outcome and a specific act.

But once we clarified that blame presupposes fault, we can see that CT face a dilemma: either (i) blame is inappropriate in the absence of fault and, thus, the act of taking responsibility cannot extend liability to blame; or (ii) blame is appropriate in the presence of fault, but then the act of taking responsibility is redundant since appropriate blame tracks fault and not prior acts of taking responsibility at will. Either way, CT's view that one can make oneself liable to blame by an act of will is mistaken. One cannot fittingly 'blame' a faultless person, and a declaration to accept blame cannot make a difference either, just as one cannot fittingly hate someone merely because they declared themselves to be worthy of hate.

However, this objection to CT's proposal not only applies to the liability to *blame*, but also to the liability to *punishment*, since punishment is also connected to fault. As Duff explained:

> 'A criminal trial aims to determine whether someone accused of committing such a wrong is provably guilty; a conviction does not just record in neutral tones that the defendant committed the wrong without justification or excuse, but condemns or blames him for committing it. The punishment to which he is then normally sentenced communicates that blame to the offender and to others; this is part of what distinguishes criminal punishment from other kinds of burden that states can impose, and criminal law from other kinds of legal regulation.' (Duff, 2021, p. 1)

Thus, even when we think about (legal) punishment, rather than (moral) blame, a similar dilemma arises: either (i) punishment is inappropriate in the absence of fault and, thus, the prior act of taking responsibility cannot extend penal liability; or (ii) punishment is appropriate in the presence of fault, but then the act of taking responsibility is redundant, since appropriate punishment tracks fault and not the prior decision to extend one's penal liability. Thus, it seems that CT's focus on liability to blame and punishment creates a fundamental problem in their view, with regards to both moral blame and legal punishment.[4]

2.2.2 Dilemma 2: Either the Arguments from Strict Liability are Decisive against CT's Proposal or CT's Proposal cannot close Responsibility Gaps.

However, CT might have a partial reply to my objections so far. Even if, in morality, there is no fault-free blame, in the criminal law there is the possibility of *strict* (i.e. blame-free) penal *liability*. Thus, could CT defend their view at least from this legal perspective?

By way of further elaboration, '[o]ffences of strict liability require proof that the defendant performed the prohibited conduct, but do not require proof that the defendant was blameworthy' (Herring 2020, p. 210). Applied to the commander: it would only need to be

[4] I here agree with Schulzke, who also argues that commanders can be responsible for the harm caused by autonomous weapon systems. In Schulzke's view, and as in my view, the commanders' responsibility, understood as liability to blame (Schulzke 2013, pp. 215, 216), does not depend – as CT claim – on an act of taking responsibility alone, but rather on the commanders' specific involvement in the use of AWS. As Schulzke elaborates: 'Commanders should be held responsible for sending the AWS into combat with unjust or inadequately formulated ROE [Rules of Engagement], for failing to ensure that the weapons can be used safely, or for using AWS to fight in unjust conflicts, as all of these conditions that enable or constrain an AWS are controlled by the commanders.' (Schulzke 2013, p. 215). In all these cases, Schulzke says, commanders are responsible *when and because* they did not take 'adequate precautions' (p. 214) or showed 'negligence of not guarding against foreseeable dangers' (p. 214). Thus, the grounds of blame-related responsibility have to involve at least 'negligence' (pp. 213, 214), if not recklessness, which means, they have to involve fault. This is where Schulzke and I diverge from CT, who claim that an act of taking responsibility alone could ground liability to blame or punishment.

shown that the autonomous weapon system caused harm and thereby violated a legal norm when the commander was on duty, not that the commander was in any way blameworthy or at fault. However, I need to be clear at this point: the question concerning CT's proposal is not whether strict liability itself would be justified in cases of the harm caused by autonomous weapon systems. The question is, more specifically, whether military commanders could simply decide whether to impose such strict liability on themselves, ahead of time. But to address the latter question, we can still consider the former question first, i.e. the question about strict liability *per se*, and then evaluate how the related arguments change once we focus on the normative power of imposing strict liability, instead of traditional strict liability.

Support for strict liability *per se* is weak in the context of autonomous weapon systems because some of the key arguments against strict liability apply their full force. Firstly, strict liability offences normally concern '"less serious" offences [that] play the role of regulating people's behaviour so that society can work effectively, rather than indicating that the defendant has behaved in a morally reprehensible way. These regulatory offences often do not require proof of mens rea because they do not carry the weight of moral censure that more serious crimes carry' (Herring 2020, p. 211). Yet, as CT claim, military commanders could be responsible for clear war crimes, such as the 'killing spree' (Champagne and Tonkens 2015, p. 125) of a robot which shoots 'clearly surrendering' (p. 128) people. Thus, to this extent, a strict liability offence is inappropriate in this context.

Secondly, in strict liability offences, the costs of dangerous activities are normally placed on those who stand to benefit most from them (Turner 2019, p. 91). Yet, it is not the commander who would benefit most from the use of autonomous weapon systems. The commander uses those systems to protect human soldiers and engages in military action for her country. CT argue that military commanders receive '[s]ocial prestige' (Champagne and Tonkens 2015, p. 125) in their occupation. But the intricacies and numerous interests at stake in war are far too complex to assume that the use of autonomous weapon systems primarily benefits those deciding their use. Thus, to this extent, strict liability is inappropriate too.

Thirdly, most strict liability offences do not carry a serious stigma and can therefore be justified (Carpenter 2003, pp. 221–223; Herring 2020) (See also R v Hughes [2013] 1 WLR 2461). Yet, killing innocent people would come with a serious stigma and, to this extent too, strict liability for the harms of autonomous weapon systems is inappropriate.

In light of these points, the case against strict liability *per se* seems decisive. But how do these arguments change if it is no longer a question of strict liability *per se* but rather one of a *normative power* to impose such strict liability on oneself, as CT suggest? It seems that the core of these arguments against strict liability is that strict liability imposes a serious burden on the person being held liable, e.g. an accusation of serious wrongdoing and stigma, without giving that person any meaningful control over whether they will be subject to such a burden. Introducing a normative power and prior agreement can mitigate this aspect: it is up to a person to decide whether or not to accept such strict liability in the future, e.g. it is the commander's decision whether to accept a certain role which involves the use of autonomous weapon systems, or whether to decline that role. Is this enough to save CT's proposal?

Unfortunately, introducing such a normative power does not provide sufficient control to mitigate the arguments against strict liability. According to CT's proposal, the commander would need to decide, ahead of time, whether to accept strict liability for *any* harm caused by autonomous weapon systems when she is on duty. But in this prior decision, the commander could not reasonably foresee what harm she is likely to become liable for. Moreover, the control the commander has is further limited by the fact that she can only exercise the normative power of taking responsibility *once*, namely when deciding whether to accept her role as a commander. As soon as she *is* the commander, there is no further possibility of taking or rejecting responsibility at will so that, *within the role of being a commander*, a person is strictly liable in the traditional sense. For this reason, CT's normative power does not give people significant control. It only provides the amount and the type of control that people already have in other areas where they can decide, ahead of time, whether to engage in certain activities to which the law attached strict liability.

But what if we drop the claim that military commanders must also, as a condition of accepting their role, accept strict liability ahead of time and we grant the commanders more discretion over whether to take responsibility *within* their role? This would create more control and depart from traditional models of strict liability. Unfortunately, however, this path would lead CT to a second dilemma: to mitigate the force of the standard arguments against strict liability (which seem decisive against CT's proposal), CT need to grant the military commander considerably more discretion over whether she will face strict liability; the greater the discretion, particularly within the role of being a commander, the less decisive these standard arguments against strict liability will be. But to close the responsibility gap, CT cannot grant any significant discretion at all, otherwise it would be completely contingent whether liability is taken. Thus, CT's proposal faces an insurmountable problem: they need to do both at once, increase and decrease the discretion for the commander.

I therefore conclude that even the strict penal liability model is not a defensible version of CT's proposal. The major obstacle in CT's proposal is the focus on responsibility in terms of *liability*, either to moral blame or legal punishment. Thus, if we are to develop a better version of a normative power of taking responsibility, this is where change is needed.

3 My proposal: the normative power of taking Retrospective Answerability

To make a fresh start, I will now move the focus away from responsibility, understood as liability to blame and punishment, to responsibility, understood as moral answerability. I define moral answerability (henceforward 'answerability') as an obligation to answer and explain, normally one's own harmful conduct, and thereby respond to those who have the standing to demand such an explanation. So understood, answerability may also lead to further obligations, such as the obligation to apologise, to follow up on the well-being of those who have been harmed, to take precautionary measures so that similar harm will not recur, and so on. This list is not exhaustive but shows that answerability, as I understand it, can be defined in terms of certain obligations associated with moral responsibility.

This understanding of answerability slightly departs from the philosophical orthodoxy. For instance, Shoemaker argued that being answerable for something implies that *what one is answerable for* reflects in some way one's 'quality of judgment'. Accordingly, answerability presupposes a certain connection to our judgment and thereby makes certain so-called reactive attitudes appropriate (Shoemaker 2015, Chap. 2). By contrast, in my view, answerability need not imply that what I am answerable for reflects my quality of judgment.

To be morally answerable only requires that I am under a moral obligation to explain and answer to others. So understood, my take on answerability is closest to Antony Duff's view, which also describes answerability in terms of owing certain others an explanation and refrains from specifying the answerable person's quality of judgment or eligibility as a target of reactive attitudes (Duff 2009).

In what follows, I will rely on this notion of answerability and argue that, under certain circumstances, there exists a normative power of making oneself morally answerable in this way, both prospectively and retrospectively, for the harm caused by sophisticated AI systems.[5] I will present two main arguments for this view.

3.1 Argument from analogy to promises

David Owens argues that the normative power of 'promising exists because it serves our *authority interest'* (Owens 2012, p. 146). In this section, I extend Owens's view on promising to support the normative power of taking responsibility too. More specifically, I claim that, if we accept at least the outline of Owens's influential view on promising, we should also accept my specific view on taking responsibility as answerability.

Authority interests, Owens explains, are *normative* interests, viz. interests in normative aspects, such as the existence of certain moral obligations. More specifically, authority interests are 'interests in *controlling* what others are obliged to do' (Owens 2012, p. 142; emphasis added). Promises serve those authority interests because when one receives and accepts a promise, one gains control over a particular moral obligation that another person owes to one. As the promise-receiver, I can now decide whether to release you from your promise and thereby control whether you are morally obliged to act as promised. Such control is valuable because it can balance power-asymmetries in personal relationships (since the promise-giver transfers some of their control over the situation to the promise-receiver), it can create morally valuable relationships (since promising may express a commitment to another person's needs), it can assist social co-ordination and, as Shiffrin argued, because promises are integral to intimate relationships such as friendships or the parent-child relationship (Shiffrin 2008). Therefore, Owens concludes: 'it is *good* for people to have authority, that the simple possession of authority over certain matters makes their lives go better (*ceteris paribus*). The normative force of a promise derives from this fact about human well-being' (Owens 2012, p. 150).

However, Owens's account does not stop at the interests of the *promise-receiver* (the promisee). The rationale behind the normative power of promising also includes the interests of the *promise-giver* (the promisor). As Owens explains:

> 'we all have an interest in being able to satisfy the needs of others should we so wish. Where the promisee has an interest in acquiring authority over the promisor, the promisor often has an interest in being able to grant this authority, an interest in having the power to make promises.' (Owens, 2012, p. 146)

[5] So far, I have assumed that taking responsibility requires the performance of an explicit speech act. In formal contexts such as the military, this assumption seems justified. However, my proposal can remain open to the possibility that people may take responsibility in ways other than performing an explicit speech act. For instance, in informal contexts, a person might take responsibility simply by apologising. See for instance Enoch (2012).

In other words, if – as receivers of promises – we have important authority interests, then – as givers of promises – we also have an interest in being able to satisfy others' authority interests, since the power to satisfy the important interests of others can be a vital asset in our lives, especially when it comes to forming and maintaining valuable personal relationships. Thus, Owens explains that the normative power exists because it serves the important authority interests of those receiving promises, as well as the correlated interests of those being able to satisfy these authority interests.

I will now devise an argument similar to Owens's to support the normative power of taking responsibility. More precisely, I will claim that there is a so-called *answerability interest*, and that the *answerability interest* is to *taking responsibility* what the *authority interest* is to *promising*, so that, if we accept that authority interests ground the normative power of promising, we should also accept that answerability interests ground the normative power of taking responsibility.

Answerability interests, like authority interests, are *normative* interests, viz. interests in normative aspects, such as the existence of certain moral obligations. More specifically, answerability interests are, like authority interests, interests in *controlling* those moral obligations. Yet, answerability interests do not concern the moral obligations we receive through promises, but rather the moral obligations that arise with moral answerability. Earlier, I defined moral answerability as an obligation to answer and explain, and I claimed that such answerability also leads to further obligations, such as the obligations to follow up on those who have been harmed, to apologise and thereby acknowledge the moral significance of the harm, or to take precautionary measures to prevent similar future events, etc. Thus, I understand answerability interests as interests in the *existence of* and in *one's control over* the moral obligations connected to the moral answerability of others.

So described, we can see why answerability interests are important. When we suffer harm through other people's action, we have an interest that *some* people not only plan to or *de facto* explain to us what happened, follow up on us, etc., but also that they *morally owe* this to us. The existence of such moral obligations is important because, when other people have these obligations, this signifies that our having been harmed is not just some natural misfortune or a case of a damaged object, i.e. something to be filed in an insurance company's drawer, but rather something that requires interpersonal moral acknowledgment. Moreover, it is important to us that we can also control those moral obligations of others by deciding whether or not to release them from these obligations, just as it is important for us to control others' obligations in this way in the context of promising. This is because such control lifts us from the status of mere *patients*, viz. those who suffered harm, to that of genuine (moral) *agents*, viz. those who have an active say in the moral aftermath of the harm. Being denied this active role would degrade us, objectify us, and insult our status as one being worthy of respect.

Normally, our answerability interests are satisfied because human-induced harm normally comes with some people's moral answerability anyway. Yet, sometimes our answerability interests may remain unsatisfied. If AI really creates responsibility gaps, then there is a human-induced harm without moral answerability, which leaves our answerability interests frustrated.[6] It is only the normative power of taking responsibility at will, viz. the power to impose moral answerability on oneself, that could create moral answerability and thereby

[6] Examples include autonomous weapon systems killing civilians, surgical robots paralysing patients, or autonomous cars injuring pedestrians.

satisfy the answerability interests of those who have been harmed. Thus, just as only the normative power of *promising* can serve some of our *authority interests*, only the normative power of *taking responsibility* can serve some of our *answerability interests*.

But it is not just about the interests of the people being harmed - it is also about the interests of everyone else, and in particular those who have been involved in the causation of harm. Some of the latter, even if they are not morally responsible (judged by the facts of the situation), may be in a unique position to *take* responsibility. If they do, they can provide comfort to those who have been harmed and help them overcome their distress and suffering. What is more, if a person freely imposes on himself the moral obligations and burdens associated with answerability in order to help others (when that person would otherwise not be morally answerable), he displays virtue. In taking responsibility, a person acts like the generous person described by Wolf and MacKenzie, viz. a person who 'voluntarily benefits or tries to benefit others at cost to herself' (Wolf 2001, p. 14), refusing to be a person who is 'more concerned with saving face than helping others' (MacKenzie 2017, p. 110). Thus, just as Owens explained that authority interests are not only significant for those who possess them but also for those who can satisfy them, I claim in a similar way that answerability interests are not only significant for those harmed parties who possess them but also for those who can satisfy them. Satisfying answerability interests provides a chance to benefit others, to form valuable moral relationships, and to develop one's moral excellence.

Hence, to come full circle: the normative power of taking responsibility satisfies answerability interests when they would otherwise be left unsatisfied, just as the normative power of promising serves authority interests when these would otherwise be left unsatisfied. Therefore, we can support the normative power of taking responsibility in the context of AI in a similar way to Owens's original account of promising: just as authority interests ground the normative power of promising, answerability interests ground the normative power of taking responsibility.

This Owens-style argument also sharpens the contour of this normative power of taking responsibility. To begin with, this argument specifically supports the focus on responsibility as *answerability*, rather than liability. Since the interests of the person *taking* responsibility are equally important in this argument, and not just the interests of the person being harmed, a normative power of taking liability could not be sufficiently supported. After all, as argued in my critique of CT's proposal, there are strong objections against strict liability in cases of responsibility gaps, and such objections receive their force from not taking sufficient account of the interests of the person subject to liability.[7]

In addition, my Owens-style argument would include the power of taking *retrospective* responsibility, viz. taking responsibility for a harmful event, even if it has already occurred. I hereby differ from CT, who only focused on taking *prospective* responsibility, viz. taking responsibility for things in the future. Retrospective responsibility is included because answerability interests apply equally in cases in which harm has already occurred: even if it has already occurred, those harmed still have equal, or even stronger, answerability interests in the existence of moral answerability. Yet, taking *prospective* answerability will remain significant too.

[7] This is not to say that the people harmed might not have some 'retributivist interests' as (Danaher 2016) put it, viz. an interest in seeing someone punished for the harm. But since these interests would mainly reside with the people harmed, and not extend to those being punished, such interests are unlikely to contribute to an Owens-style argument that includes the interests of all parties involved.

Furthermore, my argument suggests that not everyone *can* take responsibility. I claimed that taking responsibility is grounded in answerability interests. But if so, the power of taking responsibility must also be limited to those who can, at least in some basic way, satisfy others' answerability interests. To do so, I claim, one must have been involved in the use or development of the AI system that caused harm. This is because being involved in the harm not only puts one in a special position to explain what happened, but also because one's involvement is a condition of responding meaningfully to those who have been harmed. After all, if I were to claim responsibility for some AI-related harm to which I have no connection at all, the victims of the harm may understandably perceive such conduct as completely unrelated to them, and perhaps even as insulting or as an attempt to divert attention away from those who should take responsibility instead. Thus, the power of making oneself morally answerable for the harm caused by an AI-system is restricted to those who have been involved in the development and use of that AI.

But before I proceed, a few clarifications are in order. First, the type of involvement in the development and use of AI that I am concerned with is still insufficient to ground responsibility on its own, since I continue to assume that there is a responsibility gap. Second, such involvement can take place at different stages, as my phrase 'involvement in the *development and use* of AI' indicates. For instance, someone could be involved as a computer scientist in the development of an AI system or as an end-user in the final application. Third, involvement can be usefully analysed in terms of two important metrics. The first is *proximity*, i.e. the temporal or spatial distance between an agent's conduct and the AI-related outcome. End-users normally show greater proximity to an outcome than the computer scientists who developed the system. The second metric is *relevance*, which concerns the importance of a person's involvement. Here, the computer scientists could sometimes have made a more important contribution to the AI's outcome than the end-user, at least when the latter uses the system only for its previously determined purpose. These two metrics are particularly useful in settings with multiple agents, as in the context of AI, and can help to identify where taking responsibility has its greatest value. The greater the proximity (to the outcome) and relevance (of the contribution), the more likely it is that a person can satisfy an answerability interest and, thus, the higher the value of this person's act of taking responsibility.[8]

3.2 Contractualist Argument

In the previous section, I used the description of answerability interests as part of an Owens-style argument in favour of a normative power of taking responsibility. In this section, I will show that my substantial points about answerability interests can also be part of a contractualist argument.

[8] At this point, a comparison with Enoch (2012) might be helpful. Enoch discussed the idea of taking responsibility in the context of Bernard Williams's famous lorry driver case and argued that one could only take responsibility for what is in the *penumbra* of one's agency, but not for something that is completely unrelated to oneself, such as the movements of the planets. I agree with the basic idea in Enoch's account that taking responsibility requires some form of agential involvement. However, there are also important differences between Enoch's view and mine. In addition to the fact that Enoch does not consider AI-related contexts at all, my view on taking responsibility differs from Enoch's by, firstly, substituting the vague notion of 'penumbra' of agency with a more definite description of 'involvement in the development and use of AI' and, secondly, by using the notion of such involvement to explain who could best satisfy answerability interests, rather than considering it an independent condition.

Contractualists think that correct moral principles are identified by some hypothetical agreement or decision-making. For instance, Scanlon argues that we can identify the correct moral principles by focusing on what 'no one could reasonably reject as a basis for informed, unforced, general agreement' (Scanlon 1998, p. 153).[9] Normally, such principles concern general requirements and prohibitions. But, as Watson pointed out, they need not be so restricted and can also include principles that confer normative powers on us. In fact, as Watson argued, it seems one could not reasonably reject a principle that endows people with the power of promising:

> 'The idea is that, in virtue of various interests and values, no one could reasonably reject a principle that conferred the power to bind one's will in this way. (…) The significance of expectations, coordination, the value of assurance, as well as the interest in forging mutual commitments, would be registered in the standpoint of reasonable rejection. In this way, promise-making and promise-keeping turn out to answer to deep human needs, just as we understand them to do.' (Watson, 2009, p. 164)

But Watson's contractualist rationale not only applies to the normative power of promising. It can also be extended to the power of taking responsibility, as one could not reasonably reject a principle that allows people, under certain circumstances, to take answerability by an act of will. If what I have said in the last section is roughly correct (viz. if a normative power of taking responsibility, as described, is needed to satisfy important interests of those who are harmed, as well as the interests of others), then this speaks in favour of a principle allowing such a power to be the basis of our living together, and it speaks against a reasonable rejection of such a principle. Thus, even if one rejects Owens's method of grounding normative powers directly in certain normative interests, one could still accept my substantial points about answerability interests as part of a contractualist defence of the normative power of taking responsibility.

Moreover, it seems that the grounds for a reasonable rejection of the normative power of taking responsibility are already eliminated by the specifications of taking responsibility as presented in the last section. To begin with, one might reasonably reject a focus on liability to blame and punishment, due to the argument given in Sect. 1. Yet, since I focus on answerability, rather than liability, this objection is ruled out. Moreover, one might reasonably reject a principle that allows *anyone* to take responsibility, no matter how remote they are to the harm caused, because such a taking of responsibility would be arbitrary and could be insulting. Yet, with a suitable specification, and in particular a restriction on who is able to satisfy answerability interests, this objection can be ruled out too. Hence, it seems that the specification of taking responsibility, as I understand it, rules out grounds for reasonable rejection and thereby creates contractualist support for the normative power as I described it.[10]

[9] In this section, I will mainly focus on Scanlon's formulation of contractualism, but my arguments may equally apply to other versions, such as Rawls's contractualism. Rawls does not talk about what no one could reasonably reject, but rather – more positively – what certain parties would *accept* in a situation behind a veil of ignorance. See (Rawls 2005).

[10] These arguments concern taking responsibility for harmful, as opposed to beneficial, conduct. Therefore, my arguments in favour of taking responsibility will remain restricted to harm. Whether taking responsibility for beneficial action is possible will depend on whether a related normative power could also serve some of our *normative* interests or whether contractualism would adopt a principle that allows taking responsibility

4 Two questions (and an objection)

Before I conclude, let me address two pressing questions: why would anyone ever take responsibility for something bad, given that doing imposes a burden on oneself? And relatedly, if it remains optional to take responsibility, do we not risk leaving the responsibility gap wide open?

As I outlined earlier in my discussion of Owens, there are virtue-related or relationship-related reasons to take responsibility. By taking responsibility, we can satisfy answerability interests of others and benefit them. Doing so, while accepting a burden for oneself, can be virtuous as well as create and maintain morally valuable relationships. Thus, insofar as we care about virtue, moral excellence, and our relationships with others, we can have strong reasons to take responsibility in certain circumstances.

This claim should not come as a surprise: the normative power of taking responsibility is similar to the normative power of promising: both concern the imposition of obligations on oneself and the creation of benefits for others; and just as promising is valuable because it assists forming and maintaining relationships with others as well as due to the way it can give others normative control over a situation, taking responsibility can be valuable for similar reasons too.

In addition, however, taking responsibility could also be morally *obligatory*, at least in cases in which the harm to another party is severe and there is no one else who could take responsibility in a meaningful way. Adam Smith argued that simply walking away from such cases, leaving the harmed party on their own, could humiliate and offend them. Smith said: ‘To make no apology, to offer no atonement, is regarded as the highest brutality’ (Smith 2009, p. 125 (Part II. Section III. Chapter II.)). In a similar way, I suggest, not taking responsibility in certain circumstances would humiliate and offend those who have been harmed, potentially glossing over the moral significance of the harm done to them. But if so, it will be morally obligatory to take responsibility in these cases because failing to do so can offend others. Thus, there can also be *deontic* (i.e. obligation-related) and not only aretaic (i.e. virtue-related) reasons to take responsibility. [11]

Finally, we can have reason to take responsibility because doing so affirms our practical identity as real-world agents. Wolf explained that there is value in ‘expressing our recognition that we are beings who are thoroughly in-the-world, in interaction with others whose movements and thoughts we cannot fully control, and whom we affect and are affected by accidentally as well as intentionally, involuntarily, unwittingly, inescapably, as well as voluntarily and deliberately’ (Wolf 2001, p. 14). (Wolf 2001, p. 13). But if so, we should not shy away from the harm we faultlessly caused but rather express, by taking responsibility, that our agency had a morally significant impact, even if such impact and our action alone would not lead to moral responsibility.

But even if these reasons for taking responsibility exist, one might wonder how likely it is that people will *actually* take responsibility and close the responsibility gap. If individuals

for beneficial action. I am doubtful about the prospects of such a view but I will remain neutral on this point for the purposes of this paper.

[11] My argument relies on the more general view that, at least sometimes, it can be obligatory to exercise a normative power (e.g. promise, take responsibility, etc.) while the normal effect of exercising the normative power (e.g. promise: obligation, taking responsibility: responsibility) is still absent. For a discussion of this view, see Enoch (2012, pp. 105–106).

are free to decide whether to take responsibility, there is indeed no guarantee that someone will actually take responsibility. Thus, whether and how much the responsibility gap will be closed is a contingent fact. It may just be that the responsibility gap becomes a little narrower.

To close responsibility gaps in important areas may therefore require returning to an idea from CT, according to which taking responsibility *ahead of time* is a condition for taking on certain roles or engaging in certain activities. Earlier, I argued that CT's own account could not legitimately demand such prospective responsibility because it would thereby introduce unjustified strict penal liability. However, my own account does not face the same objection, since I focus on responsibility as *answerability* and not on responsibility as *liability.* Thus, in my view, we can make taking responsibility (as answerability) a condition of certain positions, thereby introducing a form of *strict answerability*, without thereby falling prey to the objections against *strict liability.*[12] Hence, there is greater room for regulating taking prospective responsibility in my view and thereby ensuring that responsibility gaps will be closed more effectively. Yet, even here, we cannot completely guarantee that the responsibility gap will be fully closed and we should therefore acknowledge the limited, yet still very important role, that taking responsibility can play.

But there may be a final objection to my view. Earlier, I claimed that the interests of the victims of harm play an important role in supporting the normative power of taking responsibility. However, this argument might go too far for my purposes: the victims' interests alone may be strong enough to create certain obligations, so that a separate act of taking responsibility would no longer be needed for those obligations to arise. If so, the act of taking responsibility would be redundant.

To respond to this objection, let me clarify the interests under consideration. In my earlier argument, I focused on what I called the victims' *answerability* interests. Answerability interests are *normative* interests, i.e. interests that other people *have an obligation* to do certain things, rather than *factual* interests, i.e. interests that other people simply *do* these things. However, no matter how strong the *normative* interests of person A are that person B has a certain obligation towards them, such interests alone cannot create that obligation. For instance, no matter how strong my neighbour's normative interest is that I am obliged to paint his house, this interest alone does not create such an obligation for me. Similarly, no matter how strong the victims' normative interest is that someone be obliged to explain, apologise, and follow-up, such interest alone cannot create these obligations. Therefore, taking responsibility is not redundant.

However, there may be a stronger version of this objection once we consider the victims' *factual* interests, i.e. the victims' interests in *actually receiving* an explanation, apology, and follow-up, as opposed to the *normative* interests in *being owed* these things. If victims' factual interests are strong enough, they alone could be reason enough to require satisfaction by whomever is in a position to provide it. Moreover, such reasons could even create an obligation, at least if one accepts the principle that it is obligatory to help others when doing so would greatly benefit them and not impose on them any significant burden (cf. Singer

[12] To clarify: strict *answerability* and strict *liability* both apply to someone solely by virtue of his conduct and independent from mental states such as intent, negligence, or recklessness. Yet, strict answerability means that, as a result of such conduct, a person is obliged to answer to others whereas strict liability means that a person can be blamed (in the moral context), or convicted, punished, respectively obliged to pay compensation (in the legal context).

1972). Hence, in such cases, an act of taking responsibility could be unnecessary for certain obligations to arise.

For the sake of argument, I will grant that positive obligations can arise from factual interests and I thereby disregard libertarian (or liberal) objections to such a position. But even then, it is doubtful whether factual interests can generate the *right kinds* of obligation. To begin with, positive obligations arising from the interests or needs of others are normally *imperfect* obligations in the sense that they do not correspond to specific claim-rights. But having such claim-rights, and being able to control the corresponding obligations, is exactly what the victims have a normative interest in, or so I have argued.

Moreover, factual interests cannot generate the obligation to apologise, because giving a genuine and fitting apology for something requires that one is also responsible for it, and not just that apologising would have positive consequences. As Govier and Verwoerd explain, 'to apologize for an action is to admit that (…) one was responsible for it' (Govier and Verwoerd 2002, p. 69) and Shuman adds: 'to be meaningful, an apology must (…) acknowledge responsibility' (Shuman 2000, p. 185). (See also (Cunningham 2014; Gill 2000; Harris et al. 2006; Smith 2008). Thus, without responsibility, what is meant to be an apology just ends up as an expression of regret or disappointment. The prospect of satisfying certain interests alone cannot make an apology any more appropriate, let alone obligatory, than the prospect of satisfying certain interests can make punishing someone appropriate.

However, even if we set aside these points too and assume that factual interests can generate the right obligations (i.e. *perfect* obligations and an obligation to apologise), the act of taking responsibility would still not be redundant. A comparison to promising, which also concerns imposing obligations on oneself, can show why.

Suppose my friend really needs my help to move house and he also helped me in the past. My friend's strong factual interest in my help, together with the norms of friendship, make it obligatory for me to help him. Thus, no further promise is needed to create this obligation for me. However, even if I subsequently *also* promise to help him, this promise is still significant. To begin with, my promise can still specify an otherwise vaguer obligation, potentially increase its stringency, and make it an explicit part of our interaction. But most importantly, the fact that there are *some* cases where interests alone create an obligation does not imply that this is always so. In fact, in most cases, my friend's interests do not create an obligation for me so that my power of promising and creating an obligation at will still retains its full value in these cases.

What is true of promises in these regards is also true of taking responsibility. Even when the victims' factual interests create obligations on their own, taking responsibility can still add value. In taking responsibility, a person can specify an otherwise vague obligation, increase the stringency of an obligation, and make it an explicit part of a future interaction. What is more, in the context of AI, there are often numerous people who could satisfy the victims' interests so that the act of taking responsibility could also provide victims with a specific addressee whom they would otherwise lack. But most importantly, and just as in the case of promising, the fact that there are *some* cases where interests alone create certain obligations does not imply that this is always so. In fact, in most cases the victims' interests will not create obligations, which means that the power of taking responsibility will retain its full value in these cases.

Thus, even if we concede that factual interests alone can generate the right kinds of obligation (which I argued is not the case), the normative power of taking responsibility is still not redundant and retains its value.

5 Conclusions

In this paper, I accepted that certain AI systems cause responsibility gaps, and I asked whether we could close those gaps *at will*, viz. by exercising a normative power of taking responsibility. I first rejected Champagne and Tonkens's view of *taking prospective liability*, a view according to which a military commander can, ahead of time, accept liability to blame and punishment for any harm caused by autonomous weapon systems under his command. Instead, I defended a view of *taking retrospective answerability*, viz. the view that people can make themselves morally answerable for the harm caused by AI systems, not only ahead of time, but also when harm had already been caused.

My view shows that a normative power of taking responsibility can close AI's responsibility gap, but only within certain limitations. Since I specifically focus on answerability, my view can close an *answerability* gap, but not necessarily a *liability* gap. Yet, ensuring answerability is an important step in addressing the responsibility gap and, what is more, the normative power I defended can close this answerability gap more effectively than CT's proposal could ever close a liability gap. Since I focus on answerability (as opposed to the much more burdensome liability), taking such answerability can be made more easily a condition of accepting certain roles. After all, when being held answerable, a person would be addressed as a rational moral agent, deserving of a genuine moral conversation, and not – as in cases of being held strictly liable – be a mere passive object of blame or punishment. In addition, I also introduced the possibility of taking *retrospective* responsibility, in addition to *prospective* responsibility, thereby providing another route to closing the responsibility gap. Furthermore, given the reasons people have to exercise the normative power of taking responsibility as answerability, it is more likely that people will actually make themselves answerable than it is that they would make themselves liable to blame and punishment.

Finally, I have based my arguments on the assumption that responsibility gaps exist, so that 'taking' responsibility would amount to 'creating' responsibility when there would otherwise be none. However, my approach could also remain attractive to those who deny responsibility gaps. These scholars could think of taking responsibility, not as a way of 'creating' responsibility, but rather as a way of 'transferring' responsibility from, say, a group of people where it is highly unclear whether anyone has more than a marginal share of responsibility to someone who will then take on that responsibility as the primary holder. Taking responsibility in this sense could also satisfy important answerability interests of others and lead to a contractualist principle that one could not reasonably reject. Hence, taking responsibility as a normative power, understood in the sense of 'transferring' responsibility, could be supported even in the absence of responsibility gaps. Exploring such arguments in detail, however, will be a task for another occasion.

Acknowledgements An earlier version of this paper received the CEPE IACAP Best Paper Award 2021 and benefitted from the reviewers' comments. I also wish to thank audiences at the universities of Edinburgh, Hamburg, Oxford, and Salzburg, including Hannah Altehenger, Holger Baumann, Gunnar Björnsson, Susanne Burri, Vincent Conitzer, Thomas Douglas, Johannes Himmelreich, Anneli Jefferson, Sebastian

Köhler, Peter Königs, Hélène Landemore, Gabriel de Marco, Leonhard Menges, Vincent C. Müller, Sven Nyholm, Carina Prunkl, Peter Schaber, Maureen Sie, Judith Simon, John Tasioulas, Shannon Vallor, Tillmann Vierkant, John Zerilli, and others. I am also very grateful for the astute and constructive reports from the reviewers for this journal.

Conflicts of Interest No conflicts of interests.

References

Bathaee Y (2018) The artificial intelligence black box and the failure of intent and causation. Harv J Law Technol 31(2):889–938

Carpenter CL (2003) On statutory rape, strict liability, and the public welfare offense model. Am UL Rev 53:313

Champagne M, Tonkens R (2015) Bridging the Responsibility Gap in Automated Warfare. Philos Technol 28(1):125–137. doi:https://doi.org/10.1007/s13347-013-0138-3

Coeckelbergh M (2020) Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics 26(4):2051–2068. doi:https://doi.org/10.1007/s11948-019-00146-8

Cunningham MJ (2014) States of apology. Manchester University Press

Danaher J (2016) Robots, law and the retribution gap. Ethics Inf Technol 18(4):299–309. doi:https://doi.org/10.1007/s10676-016-9403-3

De Jong R (2020) The retribution-gap and responsibility-loci related to robots and automated technologies: A reply to Nyholm. Sci Eng Ethics 26(2):727–735

Duff A (2009) Legal and Moral Responsibility. Philos Compass 4(6):978–986. doi:https://doi.org/10.1111/j.1747-9991.2009.00257.x

Duff A (2021) Criminal Responsibility without Blame? *Manuscript*

Enoch D (2012) Being Responsible, Taking Responsibility, and Penumbral Agency. In: Heuer U, Lang G (eds) Luck, Value, and Commitment: Themes from the Ethics of Bernarnd Williams. Oxford University Press, pp 95–131

Fricker M (2016) What's the point of blame? A paradigm based explanation. Noûs 50(1):165–183

Gill K (2000) The Moral Functions of an Apology. Philosophical Forum 31(1):11–27. doi:https://doi.org/10.1111/0031-806X.00025

Govier T, Verwoerd W (2002) The Promise and Pitfalls of Apology. J Soc Philos 33(1):67–82. doi:https://doi.org/10.1111/1467-9833.00124

Harris S, Grainger K, Mullany L (2006) The pragmatics of political apologies. Discourse & society 17(6):715–737. doi:https://doi.org/10.1177/0957926506068429

Herring J (2020) Criminal law: text, cases, and materials, Ninth edn. Oxford University Press, Oxford

Hieronymi P (2004) The force and fairness of blame. Philosophical Perspect 18:115–148

Himmelreich J (2019) Responsibility for Killer Robots. Ethical Theory and Moral Practice 22(3):731–747. doi:https://doi.org/10.1007/s10677-019-10007-9

Johnson DG (2015) Technology with no human responsibility? J Bus Ethics 127(4):707–715

Kiener M (2021) Artificial intelligence in medicine and the disclosure of risks. AI Soc 36(3):705–713

List C (2021) Group agency and artificial intelligence. Philos Technol 34(4):1213–1242

MacKenzie J (2017) Agent-Regret and the Social Practice of Moral Luck. Res Philosophica 94(1):95–117

Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6(3):175–183. doi:https://doi.org/10.1007/s10676-004-3422-1

Menges L (2017) The emotion account of blame. Philos Stud 174(1):257–273

Nyholm S (2020) Humans and robots: Ethics, agency, and anthropomorphism. Rowman & Littlefield Publishers

Owens D (2012) Shaping the normative landscape. Oxford University Press
Rawls J (2005) Political Liberalism. Columbia University Press
Scanlon T (1998) What we owe to each other. Cambridge, Mass.; London: Belknap Press of Harvard University Press
Scanlon T (2008) Moral dimensions: permissibility, meaning, blame. Belknap Press of Harvard University Press, Cambridge, Massachusetts
Schulzke M (2013) Autonomous weapons and distributed responsibility. Philos Technol 26(2):203–219
Shiffrin SV (2008) Promising, intimate relationships, and conventionalism. Philosophical Rev 117(4):481–524. doi:https://doi.org/10.1215/00318108-2008-014
Shoemaker D (2015) Responsibility from the margins, First edn. Oxford University Press, Oxford
Shoemaker D (2017) Response-dependent responsibility; or, a funny thing happened on the way to blame. Philosophical Rev 126(4):481–527
Shuman DW (2000) The role of apology in tort law. Judicature 83(4):180–189
Simpson TW, Müller VC (2016) Just War and Robots' Killings. Philosophical Q 66(263):302–322. doi:https://doi.org/10.1093/pq/pqv075
Singer P (1972) Famine, affluence, and morality.Philosophy & Public Affairs,229–243
Smith A (2009) The theory of moral sentiments. Penguin, London
Smith N (2008) I was wrong: the meanings of apologies. Cambridge University Press
Sparrow R (2007) Killer Robots. J Appl Philos 24(1):62–77. doi:https://doi.org/10.1111/j.1468-5930.2007.00346.x
Tigard DW (2020) There is no techno-responsibility gap. Philos Technol 1–19. doi:https://doi.org/10.1007/s13347-020-00414-7
Turner J (2019) Robot rules: regulating artificial intelligence. Palgrave Macmillan
Wang F, Kaushal R, Khullar D (2019) Should health care demand interpretable artificial intelligence or accept "black box" medicine? Ann Intern Med 59–61. doi:https://doi.org/10.7326/M19-2548
Watson G (1996) Two faces of responsibility. Philosophical Top 24(2):227–248
Watson G (2004) Agency and Answerability: Selected Essays. Oxford University Press
Watson G (2009) Promises, reasons, and normative powers. In: Sobel D, Wall S (eds) Reasons for Action. Cambridge University Press, pp 155–178
Williams B (1981) Moral luck: philosophical papers, 1973–1980. Cambridge University Press
Wolf S (2001) The moral of moral luck. Philosophic Exch 31(1):2–16
Wolf S (2011) Blame, Italian Style. In: Jay RKR, Wallace, Freeman S (eds) Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon. Oxford University Press