Journal of
APPLIED PHILOSOPHY

*Journal of Applied Philosophy*
*doi: 10.1111/japp.12765*

# AI and Responsibility: No Gap, but Abundance

**MAXIMILIAN KIENER**

ABSTRACT *The best-performing AI systems, such as deep neural networks, tend to be the ones that are most difficult to control and understand. For this reason, scholars worry that the use of AI would lead to so-called* responsibility gaps*, that is, situations in which no one is morally responsible for the harm caused by AI, because no one satisfies the so-called control condition and epistemic condition of moral responsibility. In this article, I acknowledge that there is a significant challenge around responsibility and AI. Yet I don't think that this challenge is best captured in terms of a responsibility* gap*. Instead, I argue for the opposite view, namely that there is responsibility* abundance*, that is, a situation in which* numerous *agents are responsible for the harm caused by AI, and that the challenge comes from the difficulties of dealing with such abundance in practice. I conclude by arguing that reframing the challenge in this way offers distinct dialectic and theoretical advantages, promising to help overcome some obstacles in the current debate surrounding 'responsibility gaps'.*

## 1. Introduction

Amazon's Alexa tells a ten-year-old girl to touch a live plug with a penny, encouraging her to do something that could lead to severe burns or even the loss of an entire limb.[1] A self-driving Tesla tests its fully autonomous mode but it fails to recognise an upcoming bend, killing two people.[2] Several companies are aiming to develop robots that could remove the human surgeon from certain medical operations altogether, creating opportunities as well as risks for future healthcare.[3]

These examples illustrate the conflicts and opportunities posed by the now rapid expansion of AI. AI has advanced to the point where it can now do things that were once exclusive to humans, including communicating freely with others (e.g. Alexa), driving cars, and performing complicated medical tasks. Moreover, AI can *outperform* humans in some of these areas: AI may soon be the safer driver and is already outperforming humans in fields like medical diagnosis.[4]

However, even the best AI is not perfect. Sometimes, things will go wrong; for example, sometimes conversational AI manipulates children, an autonomous car kills people, and a surgical robot paralyses a patient. And in such cases, many scholars worry that the use of AI causes so-called *responsibility gaps*, that is, situations in which no one is morally responsible for the harm caused by AI, because no one seems to have the required control or foresight to be morally responsible for the harmful outcomes of AI systems.[5]

In this article, I acknowledge that there is a significant challenge around responsibility and AI. Yet I don't think that this challenge is best captured in terms of a responsibility *gap*. Instead, I argue for the opposite view, namely that there is responsibility *abundance*, that is, a situation in which *numerous* agents are responsible for the harm caused by AI,

and that the challenge comes from the difficulties of dealing with such abundance in practice. I shall proceed in four steps. I first outline a novel form of genuine moral responsibility, which I call 'strict moral answerability', and then, second, I present an argument for its existence. Third, I show how strict moral answerability leads to an abundance of (rather than a gap in) responsibility in the context of AI, and fourth, I use this result to reframe the challenge around responsibility and AI in terms of a novel dilemma. I conclude by arguing that reframing the challenge in this way offers distinct advantages, promising to help overcome some obstacles in the current debate surrounding 'responsibility gaps'.

## 2. Strict Moral Answerability

I understand responsibility in terms of moral answerability, rather than in terms of blameworthiness.[6] I *define* a morally answerable agent as someone 'who can intelligibly be asked to "answer for" her attitudes and conduct', where 'to answer for' means 'to give her (justificatory) reasons for thinking, feeling, or acting in the way she has'.[7] Thus, morally answering is not about giving purely causal explanations of one's conduct (e.g. by pointing to the psychological or neurological causes of one's action). Rather, it is about explaining one's motivation and the good one saw in one's conduct. For this reason, a morally answerable person plays a very different role compared to an eyewitness, who merely provides a detached report on what happened. Furthermore, there are times when it is not only reasonable to expect an answer from the responsible party, but they are also obliged to provide one when asked, particularly by those they have harmed and who have the right to seek such an answer.[8]

Alongside this *definition* of moral answerability, I propose the following sufficient *condition*: a person is morally answerable for some harm if that person caused the harm and had a prospective obligation to guard against such harm. Thus, there are two constituents: a causation clause and an obligation clause. This means that answerability arises when a person causes harm during an activity, where the nature of the activity obliges them to guard against said harm. But it does *not* mean that answerability is restricted to those situations where guarding against some harm alone would already preclude the harm, that is, guarantee that the harm will not occur. Conversely, answerability is not restricted to situations where the existence of harm would imply that there must have been some form of infringement, violation, or other insufficient regard for the prospective obligation. It is a key feature of my proposed view that answerability can arise and persist even if a person fully satisfies their obligations, but still causes harm due to bad luck.

My proposed *definition* of moral answerability is widely accepted, whereas the sufficient *condition* is novel and sets my view apart from others. It is also this condition that turns my proposed account into a view about *strict* moral answerability, where 'strict' means, following R.A. Duff's account,[9] that there is no requirement of fault, moral deficiency, negligence, or ill will on the part of the answerable person.[10] The condition is met whenever a person causes a particular type of harm, which they had an obligation to guard against.

Let me now illustrate my proposal. Bernard Williams's scenario of the lorry driver serves as a prime example of strict moral answerability as I see it. In this case, a lorry driver, through no fault of his own, but owing to sheer bad luck, tragically runs over a child.[11] Williams originally presented this case to illustrate 'agent-regret', a poignant feeling of sorrow combined with the desire that one's actions had not contributed to such a tragic

outcome. But I shall now repurpose this example and transfer it from an investigation into the rationality of emotions to an investigation of responsibility. To do so, let us assume there was no negligence involved, and the specific outcome was neither intended nor foreseeable. Despite this, the driver still caused the accident and, like everyone who operates vehicles, had a prospective obligation to prevent accidents. This obligation is inherent in a driver's role. Consequently, the lorry driver satisfies both the causation and obligation clauses of my view, and he is therefore strictly morally answerable for the accident.[12]

As a result, it is not only reasonable and appropriate to ask him to answer for his role in the accident, but certain others also have the standing or right to demand an answer. The driver is then obliged to explain why he acted as he did, assure others he acted with due care, and acknowledge his responsibility to certain others – in his case the victim and their family. When things are clear – for example, when it is known that the driver was not at fault – the 'answer' or 'explanation' expected can be short but sufficient. Something like 'I did my best, but I failed, and that failure was beyond my control' not only provides the necessary facts, but also initiates the interpersonal assurance that the driver meant no harm and that he adheres to shared ethical values. Sometimes 'answering' may merely require a person to acknowledge his responsibility or to apologise.[13] Thus, answerability, properly understood, is not identical to explainability and does not require the answerable person to be able to give a particularly sophisticated or detailed report of what happened.[14]

So understood, this view on strict answerability differs from the standard model in moral philosophy. Answerability theorists in current moral philosophy hold that being morally answerable for something requires that this something *reflect one's judgment*. Shoemaker says that '[o]ne is an answerable agent … in virtue of one's quality of judgment'.[15] For Shoemaker, some *harmful* action or outcome reflects an agent's evaluative judgment if, afterwards, the agent either endorses their reason for the action (despite causing harm) or if it makes sense for that person to look back at their decision-making process and make changes to prevent similar incidents in the future.[16] However, the lorry driver neither endorses any reason to hit pedestrians nor has any grounds to change his conduct, insofar as – I have assumed – he caused the harm through no fault of his own. Thus, Shoemaker explicitly states that the driver is not morally answerable.[17]

Therefore, my view contrasts with those other views in two main ways. First, it differs in what it concludes about situations like Williams's: while those other views do not consider the lorry driver morally responsible for the accident, my view does. Second, it differs in how it explains moral answerability. In my analysis, the condition for moral answerability is not a reflection of judgment, but some form of agential causation. Specifically, moral answerability hinges on whether an individual's actions caused a type of harm they had an obligation to guard against.

But my proposed view on strict *answerability* must not be confused with strict *liability*. Strict liability is a legal term and means a person can be convicted, punished, or obliged to pay compensation solely by virtue of having performed a certain act (i.e. the *actus reus*), regardless of their state of mind at the time (*mens rea*), such as intention, recklessness, or negligence. Beyond my concern with morality, not law, there are further differences. Most notably, strict liability primarily pertains to conviction, punishment, or its moral equivalent, blame, rather than a person's obligation to provide an explanation or justification, which I focus on. However, what my view shares with strict liability is the strictness condition: in both cases, liability and answerability arise solely from the performance of a certain act, regardless of *mens rea*. It is just that in cases of strict answerability, a person can

still avert liability (to legal redress or moral blame) by showing that they exercised due diligence, whereas in strict liability evidence of due diligence makes no difference, and so obligations to guard against risks are not considered.[18]

Before I apply this notion of responsibility as strict moral answerability to AI, however, it is necessary to provide an argument for its existence, given that most philosophers do not recognise it as a form of genuine moral responsibility.

## 3. A Novel Argument for Strict Moral Answerability

I argue that strict moral answerability can be supported by expanding on insights from John Gardner's work, in particular Gardner's view that when we fail to conform to reasons that apply to us, those reasons persist in exerting a certain normative force:

> If one does not fully conform to a reason – if one does not do exactly what it is a reason to do – the reason does not evaporate. It does not evaporate even though one was justified in not conforming to it. It does not evaporate even though it is now too late fully to conform to it. Instead it now counts as a reason for doing the next-best-thing.[19]

In the first instance, reasons favour certain actions: a reason to take one's kids to the beach favours, trivially, taking one's kids to the beach. Whether we are also *required* to do so also depends, of course, on the presence or absence of countervailing reasons. However, as Gardner points out, if we fail to conform to our reasons, then – in the second instance – our reasons do not 'evaporate'. Rather, their demands are transformed into something else, namely into a demand for doing the 'next best thing': having failed to take the kids to the beach today, we now ought to take them to the beach tomorrow or, if this is not possible, compensate them in some other way or at least apologise.

Here is why I think Gardner is right. When we do the next best thing, we strive to minimise our departure from the un-conformed-to reasons and express our recognition of the reasons we had in the first place. But most importantly, we thereby reaffirm our status as moral and rational agents, that is, as beings to whom these reasons apply and who ought to conform to them as much as possible. Thus, if we think of ourselves as moral and rational agents, we cannot but accept the normative persistence of un-conformed-to reasons. *Qua* rational and moral agents, it is inconceivable that we could simply brush off non-conformity to reasons in the first instance and not aim to minimise our departure from their demands in the second.

I aim to build upon Gardner's insights and transfer them to our example of the lorry driver. In so doing, I shift the focus slightly. My primary concern goes beyond the general discussion of *reasons*, which was central to Gardner's account. Instead, I consider, more specifically, *prospective obligations* – namely those that pertain to guarding against types of harm. Moreover, I will also suggest a different reading of 'non-conformity'. I assume that the lorry driver met all reasonable expectations in attempting to prevent collisions with pedestrians. In this respect, he completely 'conformed to' his prospective obligations. However, there is another sense of non-conformity, which concerns the discrepancy between what the driver aimed to do with his precautionary measures on the one hand and the unfortunate outcome that occurred due to his bad luck on the other. Focusing on such

cases, I ask what the 'next best thing' would be for the driver, or what it is that he morally ought to do after the accident.

To approach this question, it can help to identify first what the driver *must not* do, namely dismiss the incident as mere bad luck and state that it has nothing to do with him. Cornford aptly describes such behaviour as 'callous',[20] Wolf calls it 'appalling',[21] and Piovarchy also critiques it sharply: 'If the driver … treats the child's death as simply unlucky and rejects any new duties, he displays a level of disregard for which we can blame him'.[22] We can blame the driver because such disregard constitutes disrespect and leads to what is known as 'expressive harm', that is, harm that occurs when someone fails to respond appropriately after causing harm.[23] Thus, in exploring the normative aftermath of the lorry driver case, we can lead our general investigation into what the 'next best thing' for the driver is towards the more specific investigation into the actions required of the driver to avert disrespect and expressive harm.

This is where strict moral answerability becomes important. According to my view of strict answerability, the victim's parents are entitled to demand that the lorry driver answer for his role in the accident. They can demand that he acknowledge his involvement, explain the circumstances from his perspective, describe how he met or failed to meet his responsibilities (such as taking precautionary measures), and that he provide assurances, such as restating his respect for the moral rights of others, reaffirming his commitment to shared ethical values, and helping the victims view the incident as a regrettable event rather than an act of negligence or disrespect.

In acknowledging and discharging his strict moral answerability, the driver finds a means of avoiding disrespect and averting expressive harm. To begin with, strict moral answerability allows the driver to recognise a shift in the relationship to the victim, respectively their family: while they might have been strangers prior to the accident, they are now connected by what happened in a very specific moral sense, one that requires second-personal recognition and thus opposes the sort of neglect inherent in disrespect. After all, answerability is itself a moral relationship between those who ought to provide an answer and those who are entitled to receive it. Moreover, strict moral answerability counteracts the harm asymmetry in the accident, that is, the disproportionate allocation of harm to the victim,[24] by acknowledging an opposite asymmetry, that is, one that gives the victim (or their family) the upper hand in terms of a certain right or entitlement to demand an answer from the driver. Thus, strict moral answerability can restore a potentially disrupted moral balance that could result in expressive harm. Finally, in fulfilling his strict moral answerability, the driver acknowledges that he owes something to the victim and their family, and that he needs to comply with their requests for an explanation or justification. In so doing, he respects their dignity as persons and validates them – to use a phrase from Rawls – as 'self-authenticating sources of valid claims',[25] that is, beings who are not only entitled to assert claims on their behalf but also generate these claims themselves and are their primary originators, including claims to explanations and justifications. For these and other reasons, strict moral answerability is a key tool for the driver to avoid disrespect and expressive harm, and it thus outlines a core part of the 'next best thing' for the driver to do.

On this basis, I conclude that strict moral answerability is a key component of the moral aftermath of situations where we cause harm that we have an obligation to guard against. It is needed to avoid disrespect and avert expressive harm, and thus is core to, or at least part of, the 'next best thing' that the driver ought to do. For the same reason, it constitutes an

important moral mechanism through which we navigate the complexities in the aftermath of harm. Thus, when we extend Gardner's initial ideas to Williams's lorry driver, we can conclude that the driver is morally answerable and, more precisely, that he is *strictly* morally answerable as no specific *mens rea* or fault was required. To be clear: this is not a knockdown argument. Rather, it is best understood as an inference to the best explanation. I propose strict moral answerability as the best account of the moral aftermath in cases like the lorry driver. Thus, we have reason to assert that strict moral answerability should form part of our responsibility framework.

This outline of, and support for, strict moral answerability completes the first half of my article. In the next section, I shall proceed to apply strict moral answerability to contexts of AI.

## 4. Applying Strict Moral Answerability to AI

Suppose it is no longer a lorry operated by a human driver, but rather an autonomous, AI-operated lorry that hits a pedestrian. Who would be responsible in this scenario? Is there anyone who resembles our lorry driver and satisfies the two clauses in my sufficient condition of strict moral answerability, that is, the condition that one is strictly morally answerable for some harm if one caused the harm (causation clause) and had an obligation to guard against such harm (obligation clause)?

By way of illustration, suppose there is a fatal accident with an autonomous vehicle, and many people across different stages of the vehicle's development and use played a role. Software engineers developed the algorithms that controlled the vehicle's motions, provided the integration of subsystems, and handled the sensor data fusion under rare environmental conditions. A quality assurance team conducted routine tests to simulate outlier scenarios that could expose AI system vulnerabilities in real-world settings. Regulatory compliance officers aimed to ensure that the AI-based software met existing standards and that these standards were updated to cover adequately the emerging technological complexities in autonomous vehicles. Users operated the vehicle, relied to varying degrees on the autonomous system, and maintained a certain, potentially limited, level of alertness. User-experience designers created the interface that communicates the system status to users, thereby determining the number of potential alerts and influencing the driver's level of attention. Finally, safety consultants advised on risk assessment models that estimate the probability and impact of simultaneous system failures. Although this description remains an oversimplification, it already shows how much more complex the situation can be in the context of AI as compared to Williams's lorry driver case. Against this background, let us revisit the two conditions of strict moral answerability, building on the more general debates around causation and obligations, and introducing them to the debate on AI.

### *4.1. Causation*

#### *4.1.1 Overdetermination*

A first possibility of divergence between Williams's lorry driver and AI contexts is that, in the latter, we may encounter causal overdetermination. I assume that an outcome, such as some harm, is overdetermined if and only if two or more distinct and actual conditions are

individually sufficient, but not (individually) necessary, for that outcome to occur.[26] By way of using a widespread example, suppose two assassins each fire a bullet which hits the same victim simultaneously, leading to the victim's death. The victim's death would not have occurred in the absence of *both* bullets but would still have occurred if *either* bullet was fired without the other. This example is a special case of overdetermination, often called *symmetric* overdetermination, because (at least) two causes operate at the same time, creating a situation of, as Bernstein puts it, 'double the causation'.[27]

A similar scenario can occur in the context of AI. Consider Carol, who is the lead AI engineer in developing the neural networks and decision-making algorithms of a self-driving vehicle. Even though Carol exercised all the diligence that one could expect and tested the systems rigorously, an unexpected anomaly in the data-processing pipeline causes the vehicle to accelerate unexpectedly. Meanwhile, Dan, who operates the vehicle and is supposed to monitor its behaviour, as well as intervene if necessary, has set the system's driving parameters. However, Dan does not know that his specific configurations contribute to the acceleration of the vehicle in certain situations. Neither Carol's neural network nor Dan's settings are inherently flawed, nor do they make them culpable in any way, akin to the lorry driver's blamelessness in Williams's original example. Yet each of their actions independently suffices to cause the crash, illustrating a case of symmetric overdetermination too.

If we follow the previous description of such cases of overdetermination, understood as cases of 'double the causation', and apply my causation clause, then *both* assassins as well as *both* Carol and Dan 'caused' the death. Therefore, (other things being equal) they are all also *fully* responsible for the outcome. But note the following important qualification, which Zimmerman reminds us of:

> To say that someone is *fully* responsible is not to say that he is *solely* responsible; responsibility is not to be cut up, like a pie, so that the more people that join in a wrongdoing [or harming], the less responsibility to be allocated to each. On the contrary, responsibility may be multiplied.[28]

Zimmerman's distinction between full and sole responsibility provides an important clarification of my view too.[29] Whoever meets the causation clause of my view is on track to being *fully* responsible, yet there is no implication that they will be *solely* responsible as well. I endorse Zimmerman's rejection of the pie model of responsibility and argue that, in cases of symmetric overdetermination, many people can be responsible for the same event.

Moreover, Zimmerman later also provides a convincing rationale as to why this is a plausible view: the mere presence of other wrongdoers or causers of harm does not provide any excuse, exemption, justification, or other responsibility-undermining explanation. Thus, the mere presence of other causing agents cannot affect one's responsibility either. Therefore, Carol and Dan are fully, yet not solely, responsible. They *share* responsibility in the sense where *shared* responsibility does not entail *distributed*, *or divided*, responsibility.

This conclusion is also further supported by some of my earlier considerations. After all, the situation Carol and Dan find themselves in significantly resembles that of the lorry driver. Like Williams's lorry driver, who faultlessly caused harm, Carol and Dan are similarly faultless. They, like the lorry driver, experience and may justifiably express agent-regret, are obliged to provide explanations and justifications for their actions to the victims

of the accident upon request, and would cause disrespect and expressive harm if they denied the accident had anything to do with them. Consequently, Carol and Dan are strictly morally answerable for the accident, even though they find themselves in a situation of symmetric overdetermination, which was not the case in Williams's original lorry driver scenario. This conclusion would also hold if we added further participants. The number of symmetric overdeterminers does not matter (for the reasons outlined by Zimmerman) and therefore does not undermine the individual answerability of each of these causal overdeterminers.

Yet, there is also another type of overdetermination, which may arise in contexts of AI too and therefore requires our attention, namely cases of *pre-emptive* overdetermination. Such cases are well illustrated by the canonical examples from Harry Frankfurt. In these examples, there is an evil neuroscientist who is prepared to interfere with an agent's brain and make them perform a certain act if the agent shows the slightest sign of not performing that very act on their own. Yet, as it happens, the agent performs the act independently and no interference is needed.[30] This is a case of overdetermination too because there are two distinct conditions, that is, the agent's motivation and the neuroscientist's interference, which would be individually sufficient but not necessary for the outcome to occur. Yet the neuroscientist does not actually *bring about* the outcome, but merely *guarantees* the outcome with their presence in the background.[31] The neuroscientist, unlike either of the two assassins or either Dan or Carol in my previous examples of symmetric overdetermination, only plays the role of a pre-empted backup rather than the role of a second operating cause.

Applying the causation clause of my view then leads to the conclusion that the pre-empted backup does not meet the causation clause and thus will not be responsible for what happens. After all, the pre-empted agent can resort to a so-called *denial*: they can simply state that they did not do the thing in question and thus avert responsibility. This is also the correct result: in Frankfurt cases, it is the self-determining agent (rather than the non-interfering neuroscientist in the background) who is responsible for their own action.[32] Thus, my view satisfactorily deals with cases of pre-emptive overdetermination. In such cases, only those agents who actually cause an outcome can be considered potential bearers of responsibility.

In the context of AI, pre-emptive overdetermination involving an autonomous vehicle could look as follows: Dan, the vehicle's operator or user, adjusts the control panel settings so that the AI will boost efficiency in traffic-navigation within a certain margin. Simultaneously, the software developer Carol designs an autonomous backup safety system, which is supposed to intervene if the other systems fail. Later, when the vehicle is in operation, a glitch in Dan's settings makes the vehicle speed up. However, had Dan's settings not caused the acceleration, a software malfunction in Carol's backup system would have done so. Thus, Carol's system is a pre-empted backup. Here, only Dan, not Carol, is comparable to Williams's initial lorry driver. Only Dan, not Carol, fittingly experiences and expresses agent-regret. Only Dan, not Carol, is obliged to answer to those who were harmed. And only Dan, not Carol, would be disrespectful and cause expressive harm if he did not answer to the demands for an explanation and justification. Thus, strict moral answerability would only be with Dan, not Carol.

Hence, dealing with cases of overdetermination in the context of AI requires that we distinguish between *symmetric* overdetermination, where all causing agents are

responsible, and *pre-emptive* overdetermination, where only those who *bring about* an outcome, but not those who merely *guarantee* an outcome, are responsible.

#### *4.1.2 Joint Causation*

Now, consider joint causation: an outcome, such as some harm from AI, is jointly caused if and only if two or more distinct and actual conditions are jointly sufficient and individually necessary for that outcome to occur (as contrasted with overdetermination where the conditions are *individually* sufficient but not necessary). For example, suppose there are two assassins again, each firing a bullet at the same victim, but this time *both* bullets are required to kill the victim; either bullet on its own would be insufficient. Bernstein provides a good slogan for such cases when she says: 'Joint causation is often thought of as a kind of "causal teamwork", whereas overdetermination is often thought of as "double the causation"'.[33] Thus, the relation between each assassin and the outcome is now different: whereas they individually *caused* the outcome in cases of symmetric overdetermination, they now only make a *causal contribution* to the outcome. This change requires that we reconsider each assassin's responsibility, and by extension the responsibility of people in AI insofar as their situation is one of joint causation too.

In fact, cases of joint causation seem most prevalent in the context of AI, where the partial contributions of developers and users may accumulate and jointly cause harmful outcomes. Here is a simplified possible scenario in the context of autonomous vehicles: two factors cause a crash, and each of them was necessary but only jointly sufficient for the outcome. The first factor comes from Carol, who designed the vehicle's software to increase speed whenever its sensors detect clear road conditions. However, the software occasionally misinterprets brief gaps in fog as safe and, in such cases, this can lead to a risky increase in speed. The second factor is related to the vehicle's maintenance and concerns Dan as the owner. Dan recently purchased aftermarket tyres because these tyres are more fuel-efficient and Dan wanted to improve his sustainability footprint. The fact that these tyres have reduced traction on wet or slick surfaces did not concern him as Dan trusts the automated operation of the vehicle even under rare and difficult weather conditions. Unfortunately, one day, while driving in a foggy environment, the sensors briefly detect a clearing, and therefore Carol's software triggers an increase in speed. At the same time, the reduced traction of Dan's new tyres makes it difficult for the vehicle to handle this increase in speed. The vehicle skids and eventually crashes. Neither Carol's software nor Dan's choice of tyres alone would have led to the crash; only their coincidence under rare conditions did.

To evaluate responsibility in joint causation, Bernstein begins with what she calls an 'Intuitive Doctrine: you are only morally responsible for what you cause'.[34] Bernstein then adds the claim that 'if one is morally responsible only for the outcome that one causes, one should also only be morally responsible for the *part* or *proportion* of an outcome that one causes'.[35] Bernstein says the latter claim is just a 'precisification'[36] of the Intuitive Doctrine and eventually leads us to the following principle:

> *Proportionality*: An agent's moral responsibility for an outcome is proportionate to her actual causal contribution to the outcome.[37]

Unfortunately, Bernstein's view conflates two importantly different aspects, namely 'causing part of an outcome' and 'partly causing an outcome'. Consider the second version of the two assassins scenario again. Each assassin partly causes one single, indivisible

outcome, that is, death, as opposed to each of them causing separate parts of a compound outcome. Yet Bernstein's view (that is, that 'one should … only be morally responsible for the *part* or *proportion* of an outcome that one causes')[38] assumes the latter and is thus unable to explain how both assassins can share the responsibility for the *death*, rather than just bear individual responsibility for the specific wounds that their separate bullets cause.[39] On Bernstein's precisification view, then, the object of responsibility could only ever be part of the victim's death. Yet we are unable to identify a distinct 'part' of death and, what is more, this view fails to capture the fact that each individual assassin is responsible for the *entire death*, rather than just parts of it.

Although Bernstein's argument is flawed because of the conflation of 'causing part of an outcome' and 'partly causing an outcome', her conclusion, *Proportionality*, can be rescued. To see how, we need to use an insight from Kaiserman's work on responsibility and causation. Kaiserman stated a widely accepted principle, which he called *Individualism*:

> *Individualism*: An agent A's degree of responsibility for some outcome o is fully grounded in facts about A, o and the relations between them.[40]

*Individualism* has a restrictive dimension because it holds that a person's responsibility can only be affected by those three aspects that it is 'fully grounded in', namely 'A' (facts about that person), 'o' (the outcome this person is alleged to be responsible for), and 'r' (the relation between 'A' and 'o'). Accordingly, just adding further assassins or operating causes, as in my previous examples of symmetric overdetermination, would not affect people's responsibility. Thus, Kaiserman confirms my previous conclusion.

But *Individualism* also highlights an important difference between cases of symmetric overdetermination and cases of joint causation. The relation 'r' is different in these types of cases: people *fully cause* ($r_1$), an outcome in symmetric overdetermination, but only *causally contribute* ($r_2$) to an outcome in joint causation. Thus, if responsibility is grounded in r, and thus affected by r, someone's responsibility in a case of joint causation will differ from someone's responsibility in a case of symmetric overdetermination, and Bernstein's *Proportionality* offers a plausible explanation of how: 'An agent's moral responsibility for an outcome is proportionate to her actual causal contribution to the outcome',[41] where the causal contribution describes the relation 'r'.

But spelling out what being *proportionate* means, that is, determining whether people are more or less responsible in cases of joint causation than in cases of symmetric overdetermination, proves to be a thorny issue. As Bernstein notes, the assessment of responsibility in cases of joint causation will depend on one's theory of causation.[42]

Suppose one subscribes to a so-called 'productive' theory of causation, according to which causing is transferring energy. According to this theory, the assassins, as well as Carol and Dan in our AI example, would transfer greater energy in cases of overdetermination than in cases of joint causation. The energy that each agent transfers in overdetermination cases is sufficient for the outcome, whereas the energy that each of them transfers in cases of joint causation is insufficient and only necessary. Hence, this suggests that each assassin's, as well as Carol's and Dan's, responsibility is reduced in cases of joint causation.

However, if one chooses a Lewisian 'counterfactual dependence' theory of causation, according to which causing requires difference-making in the counterfactual sense, then the agents in joint causation cases bear greater responsibility. There is no counterfactual

dependence between the death and either of the two individual shootings in overdetermination cases, or between the accident and either Carol's or Dan's action, whereas there is such counterfactual dependence in cases of joint causation.

Thus, Bernstein concludes: 'Whether the assassins in [cases of overdetermination] … bear more moral responsibility than those in [cases of joint causation] … depends on which causal concept one employs'.[43] Therefore, the claim that one's responsibility is proportionate to one's causal contribution does not yet tell us the exact degree of responsibility, but merely that there is a connection between responsibility and causal contribution, and that there is at least *some* responsibility in cases of joint causation.

Degrees of responsibility could mean different things, including degrees of blameworthiness (if there was a harmful or wrongful action) or the range of liabilities one acquires through one's responsibility. In this article, however, I am concerned with answerability. Answerability is a binary rather than scalar notion: one either is or is not answerable. Thus, even if a person's desert for blame or their range of liabilities varies according to their causal contribution, this will not change *the fact that* they are answerable. Yet degrees of responsibility could be understood as making a person's obligation or reasons to answer for their conduct more or less stringent. Accordingly, a person with a higher degree of responsibility would have a stronger obligation or stronger reasons to comply with demands for an explanation or justification than a person with a lower degree of responsibility. And higher degrees of responsibility may also make it appropriate to approach the person with such higher responsibility first and demand an explanation or justification before approaching those with lower degrees of responsibility. This leads to a two-fold conclusion: degrees of responsibility do not affect *that* a person is answerable but may change how *strong* a person's obligation or reasons to answer are.

This is also the correct ethical assessment. Recall the joint causation of Dan and Carol, and add as many other contributing causes as you wish, to come closer to the complex reality of AI development and use. Perhaps the imperfectly integrated sensors misinterpreted the environmental data, the warning systems inadequately alerted the user, the safety protocols failed to compensate, and so on. Each party involved in the chain that led to the accident – engineers, testers, regulators, users, designers, and consultants – made an incremental contribution to the harmful outcome. Still, all people involved can fittingly experience and express agent-regret. They are also all obliged, upon request, to explain and justify their conduct to those who were harmed and who are entitled to an explanation. If the engineers, testers, regulators, users, designers, and consultants simply walked away and denied involvement, they would engender disrespect and expressive harm, just as the lorry driver in Williams's original case would. Their indifference and refusal to answer would be a morally inappropriate response to what happened. Thus, the minuteness of a contribution does not preclude one's strict moral answerability in cases of accidents. Each participant, in reflecting on their role, must consider and answer to the question of how even minor oversights or decisions contribute to outcomes of vast ethical significance. Therefore, even though the role of the numerous people involved in the chain that led to the accident are very different from that of the initial lorry driver, this does not preclude the significant aspects relevant to strict moral answerability and the need to avoid disrespect and expressive harm persisting in both scenarios. Therefore, even if there is just a causal contribution, rather than full causation, this is enough to satisfy the causation clause of strict moral answerability.

On this basis, I conclude that everyone involved in joint causation of harm is answerable for what they did. However, there may be some variation in how strongly they are obliged to answer. Moreover, *pace* Bernstein, I argue that people can be answerable for a single, indivisible outcome to which they causally contributed, and not only for parts of some compound outcome.

This conclusion now allows me to specify my causation clause: it requires that a person cause *or* causally contribute to some outcome. This occurs in cases of (symmetric) overdetermination as well as in cases of joint causation. Therefore, insofar as the roles of the people involved in the use and development of AI resemble the roles of my protagonists in the examples of symmetric overdetermination or joint causation, the people in AI (potentially very many) will bear responsibility as answerability for a harmful outcome.

Now that I have discussed the causation condition, I shall proceed to the obligation clause of my view.

### *4.2. Obligations*

All people involved in the development and use of AI have the obligation to guard against certain types of harm, for example, that the autonomous car hits a pedestrian, that the hiring algorithm discriminates against certain applicants, or that the medical AI misdiagnoses patients. Thus, I shall assume that everyone involved in the development and use of AI shares an obligation of this general character, namely the obligation to guard against the risk normally associated with an activity.

But to discharge this general obligation, different people may also be required to do different things, given their different involvement in the development and use of AI. For instance, consider once again the context of autonomous driving. Programmers and software engineers are obliged to implement enough redundancies and fail-safes in the AI systems and thereby enable them to anticipate and mitigate potential failures under unexpected conditions. Quality assurance engineers are obliged to conduct extensive and rigorous testing that includes atypical scenarios to ensure system reliability and safety in real-world applications. Regulatory compliance officers are obliged to keep abreast of technological advancements and update safety regulations accordingly so that they reflect the capabilities and risks associated with the state of the art. Users of autonomous vehicles have an obligation to operate these vehicles only within prescribed limits and to maintain alertness to take control whenever necessary. User-experience designers are obliged to create clear, intuitive interfaces that effectively communicate the system's status and potential risks, ensuring that users can make informed decisions as quickly and as easily as possible. Safety consultants have an obligation to develop detailed risk assessments that consider not only typical but also rare scenarios. Corporate executives are also obliged to attach adequate importance to ethical considerations and safety in their business practices so that commercial goals will not compromise important ethical values and principles in the context of autonomous driving. This shows that all these obligations depend on the specific role one plays in the development and use of some AI system, as well as on the specific skillset one either has or can be expected to have in relation to guarding against certain types of harm.

Taking these points together, I assume that everyone in the development and use of AI shares the general obligation to guard against certain risks of harm and that they possess some more specific obligations, tailored to their position and expertise. What is important,

however, is that all these obligations concern specific types of harm, rather than an obligation not to cause harm, without further qualification. Thus, whenever people cause or causally contribute to harm that they have an obligation to guard against, either an obligation shared with many others or an obligation specific to their role, they are strictly morally answerable.

This position leads to a two-fold result. On the one hand, the obligation clause specifies who is responsible for the harm caused by AI. My view identifies those people with specific (often role-based) obligations to guard against certain types of harm as meeting the sufficient condition of strict moral answerability and thus being eligible bearers of responsibility. On the other hand, the shared general obligation to guard against certain types of harm, such as hitting people with autonomous cars, as well as the existence of further versatile role obligations, make it likely that many people in the development and use of AI will meet the obligation clause of my proposed view.

In conclusion, then, many people will meet the condition for strict moral answerability when AI causes harm: they meet the causation clause (either by causing or causally contributing) and they meet the obligation clause (by possessing various general and specific obligations not to cause certain types of harm). Thus, whenever a self-driving car causes an accident, or some other AI system causes harm, many of those involved in the development and use of this AI system are likely to become strictly morally answerable for the harmful outcome.

Thus, I arrive at the main idea of this article: in the context of AI, we do not face a responsibility *gap*; we face responsibility *abundance*.[44]

## 5. Re-Interpreting the Challenge from Responsibility and AI

This change of perspectives – that is, from a responsibility gap to responsibility abundance – does not mean that the challenge concerning responsibility and AI disappears. In fact, the abundance of responsibility is a problem on its own. Once numerous agents are responsible, it becomes impractical to hold all of them responsible. In practice, we simply lack the means to hold everyone answerable. But at the same time, we cannot just single out some of them because doing so would be arbitrary and unjust. Therefore, with responsibility abundance there is a dilemma: (i) the great number of agents makes it unfeasible to hold everyone responsible; and (ii) just singling out some of them would be arbitrary and unjust. In other words, it seems that we *cannot* do (i) and we *should not* do (ii). Thus, whereas the problem with a responsibility *gap* is that we lack what we want, the problem with responsibility *abundance* is that we cannot use what we have.

Reframing the challenge about responsibility and AI in terms of an abundance of responsibility offers several advantages, two of which are especially noteworthy. First, there is a *dialectic* advantage: the dilemmatic structure focuses our attention and highlights specific means to tackle the problem, that is, address one of the two horns of the dilemma and establish answerability as a feasible and just social practice. This more specific description of the challenge contrasts with a rather hazy understanding of why a responsibility gap is a problem in the first place. While many address responsibility gaps, only very few attempt to explain why it is a serious problem, and some even argue that it is not problematic at all,[45] despite some strong intuitions to the contrary. Thus, my proposal allows

us to circumvent the ambiguity of the challenge's nature and also highlights specific ways to address it.

Second, there is a *theoretical* advantage: my proposal provides more robust grounds for a new way of thinking about responsibility, including ideas about 'taking responsibility'. Enoch and myself argued that there is a normative power of taking responsibility, that is, the ability to make oneself responsible for something by mere declaration.[46] One of Enoch's central examples is Williams's lorry driver, who is not responsible immediately after the accident but can, and ought, to take responsibility by an act of will or declaration. I developed this idea in the context of autonomous weapons systems and claimed that suitably positioned people can make themselves responsible for the harm caused by AI simply by declaration. However, the strongest objection against such views is that responsibility cannot be created out of the blue. Just saying 'I hereby make myself responsible' does not automatically confer responsibility where it previously did not exist. It is at this point that my proposal in this article can make Enoch's view and my view on 'taking responsibility' more palatable to those who endorse the aforementioned objection. This is because my proposal in this article can show that there has always been some responsibility, that is, strict moral answerability. Accordingly, taking responsibility need not necessarily be understood as creating responsibility *ex nihilo* but rather as making it possible for others to hold one responsible without having to contend with horn (ii) of the aforementioned dilemma and thereby acting unjustly. Thus, my proposal allows us to understand the normative change brought about by the act of taking responsibility differently, namely by giving others permission to hold one answerable without running into arbitrariness or injustice. So understood, my approach facilitates new theoretical grounds for more constructivist thinking about responsibility, where some forms of moral responsibility can be taken and negotiated, not simply discovered like a brute moral fact.

But before I conclude, I also need to address an objection. Some may still be unconvinced by the responsibility abundance. They may think that strict moral answerability is not genuine moral responsibility,[47] or that even if strict moral answerability is genuine moral responsibility, there is still a lack of other types of responsibility, most notably those types associated with blame; and either way, responsibility gaps persist.

In response, I want to emphasise that a responsibility gap implies two aspects: first, the *absence* of responsibility, and second, the *need* for responsibility. Mere absence is not enough as, for example, we would not talk about responsibility gaps regarding the movements of the planets; we simply do not expect there to be any human responsibility involved here. But with the development of new technologies, such as AI, things are different: we imply that there must be some kind of responsibility, otherwise something is missing.

On the basis of this bi-partite structure of the claim about the responsibility gap (i.e. the absence of responsibility and the need for responsibility), there are also two corresponding ways of denying a responsibility gap: denying the *absence* of responsibility or denying the *need* for responsibility. My proposal could target both aspects. So far, I have presented my view as the claim that strict moral answerability is genuine moral responsibility, and I thereby denied the *absence* of moral responsibility. But in addition, and independently, my proposal could also be interpreted as showing that strict moral answerability (whether *genuine* moral responsibility or not) precludes the *need* for any other types of moral responsibility. This is because, once we manage to establish a feasible and just scheme of holding each other (strictly) answerable, there will be no need for further practices of blame. Such

a scheme of strict answerability can be premised on mutual recognition, equality, and respect. It is capable of creating and maintaining social trust in the development and use of technologies, and it encourages due diligence and inclusivity. If this is so, it is not clear what roles or functions other types of responsibility, especially those associated with blaming or punishing people, could have. It is unclear why we would still *need* blame and other forms of responsibility whenever there is strict answerability.

However, it's important to clarify that my proposal does not entirely exclude blame: when answerability practices or investigations reveal fault on the part of certain agents, blame may still be warranted and can play a valuable role. The key point is that blame remains an open question and should not be sought or considered valuable unless it arises naturally from the findings of answerability practices or investigations.

Thus, what started as an objection to my proposal – that is, that the responsibility gap persists – is now reframed as a serious challenge addressed to the responsibility gap theorists. This challenge is to explain not only how strict moral answerability cannot be classed as genuine moral responsibility or why it excludes other types of moral responsibility, but also why there is a real need for other types of moral responsibility, *after* we established a scheme of strict moral answerability. Unless this challenge is addressed, the defence of responsibility gaps is incomplete.

## 6. Conclusion

This article proposed a change of perspective, from responsibility gaps to responsibility abundance. I explained and defended the idea of strict moral answerability in the first half of this article and applied it to the context of AI in the second half. My proposed change of perspective offers dialectic and theoretical advantages: it allows us to address the challenge posed by responsibility and AI with greater precision and it breaks new ground for thinking about responsibility. So understood, framing the challenge in terms of a responsibility abundance promises to overcome some of the main obstacles in the current debate and helps us to gain a fresh perspective on responsible AI.

*Maximilian Kiener, Hamburg University of Technology, Hamburg, Germany.* maximilian.kiener@tuhh.de

## Acknowledgements

I am particularly grateful to the two anonymous reviewers and the editor of this journal for their exceptionally insightful and constructive feedback. I also extend my thanks to the audiences at the University of Hamburg, the University of Oxford, the University of Neuchâtel, TU Eindhoven, LMU Munich, the University of Southampton, RU Bochum, and the interdisciplinary European seminar series LiaNs, where I could present and discuss earlier versions of this article. A special thanks also goes to Matthew Braham, Jonas Bozenhard, Mark Coeckelbergh, Rafaela Hillerbrand, Thomas Krödel, Peter Niesen, Sven Nyholm, Judith Simon, Shannon Vallor, and Valentin Weber for their feedback. Finally, I thank the Leverhulme Trust for invaluable support through an Early Career

Fellowship (ECF-2021-176). Open Access funding enabled and organized by Projekt DEAL.

## NOTES

1 BBC News, "Alexa."
2 Reuters, "Two Die."
3 Thomas, "FDA-Approved."
4 Kiener, "Artificial Intelligence"; Bathaee, "Artificial Intelligence."
5 Sparrow, "Killer Robots"; Coeckelbergh, "Artificial Intelligence"; Nyholm, *Humans*; Kiener, "Can We Bridge?"
6 This section draws on my exposition of strict moral answerability in Kiener, "Strict Moral Answerability."
7 Smith, "Responsibility," 103; Scanlon, *What We Owe*, chap. 6.
8 Duff, *Answering*, chap. 1. So understood, moral answerability often serves an investigative role: we demand answers to find out whether people lived up to their obligations. However, moral answerability can also persist when there is no uncertainty about a person's due diligence. Consider the lorry driver case, outlined below. Even if the child's family already knew exactly what happened and (let us assume) knew that the driver took due care, they still retain an entitlement, and possibly an emotional need, to hear that very fact *from the driver*, and the driver still owes them an answer in this regard. The family's knowledge does not undermine the driver's obligation to comply with their requests for justification and explanation, and thus does not undermine the driver's answerability.
9 The term 'strict' is borrowed from the law where it denotes that there is no requirement of *mens rea*, such as intention, recklessness, negligence, or types of fault or ill will.
10 I am greatly indebted to Antony Duff, who first introduced the idea that moral responsibility as answerability can be strict; Duff, *Answering*. For the purposes of this article, however, I shall rely on my own exposition of this idea and not concern myself with the differences between my account and Duff's view. For a more in-depth comparison, see Kiener, "Strict Moral Answerability."
11 Williams, *Moral Luck*.
12 However, stating that the lorry driver is answerable does not exclude the possibility that others, including the parents of the child, are also answerable for their conduct.
13 Kiener, "Varieties of Answerability," and Hubbs, "Answerability," even argued that there can be answerability without an answer.
14 Here, the driver is answerable despite his lack of fault. But this is not to deny that answerability is often an open-ended inquiry where it is unclear whether someone was at fault and where the certainty about the absence of fault, as simply assumed in the lorry driver case, is not a given.
15 Shoemaker, *Responsibility*, 82. See also Smith, "Responsibility," 103; Scanlon, *What We Owe*, 276.
16 Shoemaker, *Responsibility*, 67. Smith and Scanlon hold similar views; Smith, "Responsibility," chap. 6; Scanlon, *What We Owe*, chap. 6.
17 Shoemaker, *Responsibility*, 67–68.
18 Cf. Kiener, "Strict Moral Answerability."
19 Gardner, "Wrongs and Faults," 57.
20 Cornford, *Structures*, 45.
21 Wolf, "Moral," 12.
22 Piovarchy, "Blame," 148. There is some divergence between Piovarchy and Gardner's account here. Whereas Piovarchy suggests that there may be *new* duties, Gardner may argue that the *old* duties are still there and now simply ask for the 'next best thing'. The debate as to whether there are new or old duties, or reasons, does not matter for my purposes. I shall only focus on the question of what people ought to do 'next' in cases like the lorry driver.
23 Cf. Piovarchy, ibid., 155–6.
24 This is, of course, not to deny that the driver may suffer psychological trauma too.
25 Rawls, *Political Liberalism*, 32, 96.
26 Cf. Funkhouser, "Frankfurt Cases"; Sartorio, "Two Wrongs."
27 Bernstein, "Causal Proportions," 170.
28 Zimmerman, "Sharing Responsibility," 355, emphasis added.

29 This is true even though Zimmerman focuses on blameworthiness-responsibility rather than answerability-responsibility as I do.
30 Frankfurt, "Alternate Possibilities."
31 Funkhouser, "Frankfurt Cases," 350.
32 This is, of course, compatible with the view that we may hold the evil neuroscientist responsible for other things such as his conduct more generally and taking the role in the background.
33 Bernstein, "Causal Proportions," 170.
34 Ibid., 165. For the purposes of this article, I sidestep challenges to this view in accounting for responsibility for omissions.
35 Ibid., 167.
36 Ibid., 167.
37 Ibid., 167.
38 Ibid., 167.
39 See also Kaiserman, "Causal Contribution," on the notion of causal contributions to a single outcome.
40 Kaiserman, "Responsibility," 3598.
41 Bernstein, "Causal Proportions," 167.
42 Ibid., 170 ff.
43 Ibid., 172.
44 Note that those who reject causation as a condition of responsibility may face an even greater abundance of responsibility in the context of AI, as they could include people with relevant obligations, even if they are outside the causal chain. These cases are beyond the scope of this article, but they show how my view about a responsibility abundance could apply in other types of case too, which could prove even more significant for the context of AI.
45 Königs, "Artificial Intelligence"; Danaher, "Tragic Choices."
46 Enoch, "Being Responsible"; Kiener, "Can We Bridge?"
47 Shoemaker, "On Criminal and Moral Responsibility."

## References

Bathaee, Yavar. "The Artificial Intelligence Black Box and the Failure of Intent and Causation." *Harvard Journal of Law & Technology* 31, no. 2 (2018): 889–938.

BBC News. "Alexa Tells 10-Year-Old Girl to Touch Live Plug with Penny." December 28, 2021. https://www.bbc.com/news/technology-59810383.

Bernstein, Sara. "Causal Proportions and Moral Responsibility." *Oxford Studies in Agency and Responsibility* 4 (2017): 165–182.

Coeckelbergh, Mark. "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability." *Science and Engineering Ethics* 26, no. 4 (2020): 2051–68. https://doi.org/10.1007/s11948-019-00146-8.

Cornford, Andrew. *The Structures of the Criminal Law*. New York: Oxford University Press, 2011.

Danaher, John. "Tragic Choices and the Virtue of Techno-Responsibility Gaps." *Philosophy & Technology* 35, no. 2 (2022): 1–26.

Duff, Antony. *Answering for Crime: Responsibility and Liability in the Criminal Law. Legal Theory Today*. Oxford: Hart, 2009.

Enoch, David. "Being Responsible, Taking Responsibility, and Penumbral Agency." In *Luck, Value, and Commitment: Themes from the Ethics of Bernard Williams*, edited by Ulrike Heuer and Gerald Lang, 95–131. Oxford: Oxford University Press, 2012.

Frankfurt, Harry G. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66, no. 23 (1969): 829–839. https://doi.org/10.2307/2023833.

Funkhouser, Eric. "Frankfurt Cases and Overdetermination." *Canadian Journal of Philosophy* 39, no. 3 (2009): 341–369.

Gardner, John. "Wrongs and Faults." In *Appraising Strict Liability*, edited by A. Simester, 51–80. Oxford: Oxford University Press, 2005.

Hubbs, Graham. "Answerability without Answers." *Journal of Ethics and Social Philosophy* 7, no. 3 (2013): 1–15.
Kaiserman, Alex. "Causal Contribution." Paper presented at the Proceedings of the Aristotelian Society, 2016. https://philpapers.org/rec/KAICC
Kaiserman, Alex. "Responsibility and the 'Pie Fallacy'." *Philosophical Studies* 178, no. 11 (2021): 3597–3616.
Kiener, Maximilian. "Artificial Intelligence in Medicine and the Disclosure of Risks." *AI & Society* 36, no. 3 (2021): 705–713.
Kiener, Maximilian. "Can We Bridge AI's Responsibility Gap at Will?" *Ethical Theory and Moral Practice* 25, no. 4 (2022): 575–593. https://doi.org/10.1007/s10677-022-10313-9.
Kiener, Maximilian. "Strict Moral Answerability." *Ethics* 134, no. 3 (2024): 360–386.
Kiener, Maximilian. "Varieties of Answerability." In *The Routledge Handbook of Philosophy of Responsibility*, edited by Maximilian Kiener, 204–216. London and New York: Routledge, 2023.
Königs, Peter. "Artificial Intelligence and Responsibility Gaps: What Is the Problem?" *Ethics and Information Technology* 24, no. 3 (2022): 1–11.
Nyholm, Sven. *Humans and Robots: Ethics, Agency, and Anthropomorphism*. Lanham: Rowman & Littlefield, 2020.
Piovarchy, Adam. "Blame in the Aftermath of Excused Wrongdoing." *Public Affairs Quarterly* 34, no. 2 (2020): 142–168.
Rawls, J. *Political Liberalism*. New York: Columbia University Press, 2005.
Reuters. "Two Die in Tesla Car Crash in Texas with 'No One' in Driver's Seat, Police Say." *Guardian*, April 19, 2021. https://www.theguardian.com/technology/2021/apr/19/two-die-in-tesla-crash-no-one-in-drivers-seat-police.
Sartorio, Carolina. "Two Wrongs Do Not Make a Right: Responsibility and Overdetermination." *Legal Theory* 18, no. 4 (2012): 473–490.
Scanlon, T. M. *What We Owe to Each Other*. Cambridge: Belknap Press of Harvard University Press, 2000.
Shoemaker, David. "On Criminal and Moral Responsibility." In *Oxford Studies in Normative Ethics*, Vol 3 edited by Mark Timmons, 154–178. Oxford: Oxford University Press, 2013.
Shoemaker, David. *Responsibility from the Margins*, 1st ed. Oxford: Oxford University Press, 2015.
Smith, Angela. "Responsibility as Answerability." *Inquiry* 58, no. 2 (2015): 99–126.
Sparrow, Robert. "Killer Robots." *Journal of Applied Philosophy* 24, no. 1 (2007): 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.
Thomas, Liji. "FDA-Approved Surgical Robots Trend toward Autonomy, Study Finds." *News Medical*, April 29, 2024. https://www.news-medical.net/news/20240429/FDA-approved-surgical-robots-trend-toward-autonomy-study-finds.aspx
Williams, Bernard. *Moral Luck: Philosophical Papers, 1973–1980*. Cambridge: Cambridge University Press, 1981.
Wolf, Susan. "The Moral of Moral Luck." *Philosophic Exchange* 31, no. 1 (2001): 2–16.
Zimmerman, Michael J. "Sharing Responsibility." *American Philosophical Quarterly* 22, no. 2 (1985): 115–122.