Science and Engineering Ethics (2023) 29:27
https://doi.org/10.1007/s11948-023-00449-x

ORIGINAL RESEARCH/SCHOLARSHIP

A Comparative Defense of Self-initiated Prospective Moral Answerability for Autonomous Robot harm

Marc Champagne[1] · Ryan Tonkens[2]

Received: 14 September 2022 / Accepted: 16 June 2023

Abstract

As artificial intelligence becomes more sophisticated and robots approach autonomous decision-making, debates about how to assign moral responsibility have gained importance, urgency, and sophistication. Answering Stenseke's (2022a) call for scaffolds that can help us classify views and commitments, we think the current debate space can be represented hierarchically, as answers to key questions. We use the resulting taxonomy of five stances to differentiate—and defend—what is known as the "blank check" proposal. According to this proposal, a person activating a robot could willingly make themselves answerable for whatever events ensue, even if those events stem from the robot's autonomous decision(s). This blank check solution was originally proposed in the context of automated warfare (Champagne & Tonkens, 2015), but we extend it to cover all robots. We argue that, because moral answerability in the blank check is accepted *voluntarily* and *before* bad outcomes are known, it proves superior to alternative ways of assigning blame. We end by highlighting how, in addition to being just, this self-initiated and prospective moral answerability for robot harm provides deterrence that the four other stances cannot match.

Keywords Responsibility · Answerability · Ethics · Artificial intelligence · Robotics

✉ Marc Champagne
marc.champagne@kpu.ca

Ryan Tonkens
rtonkens@lakeheadu.ca

[1] Department of Philosophy, Kwantlen Polytechnic University, 12666, 72 Avenue, Surrey, Canada

[2] Department of Philosophy and Centre for Health Care Ethics, Lakehead University, 874 Tungsten Street, Room MP1002, Thunder Bay P7B 5E1, Canada

Published online: 13 July 2023

What to Do About Wrongdoings Without Wrongdoers

In 2018, as she crossed the road by foot with her bicycle at her side, 49-year old Elaine Herzberg was hit by a self-driving car. Rafaela Vasquez, who was behind the wheel at the time but was streaming a television show, was eventually charged with negligent homicide. This verdict brings some measure of closure. Yet, who will we blame when robots become fully autonomous and no human is in the loop (Hansson, 2023)?

Self-driving cars are just the beginning. In a 2022 essay for China's Cyberspace Administration, Elon Musk explained that his Tesla bots are meant to replace humans for dangerous tasks. However, it is clearly possible for such robots to themselves become a source of danger. So, if a robot makes its own decisions yet does not feel pain, who shall we punish when it commits a violent act? Sparrow gives this concrete example:

> Let us imagine that an airborne AWS [autonomous weapons system], directed by a sophisticated artificial intelligence, deliberately bombs a column of enemy soldiers who have clearly indicated their desire to surrender. These soldiers have laid down their weapons and pose no immediate threat to friendly forces or non-combatants. Let us also stipulate that this bombing was not a mistake; there was no targeting error, no confusion in the machine's orders, etc. It was a decision taken by the AWS with full knowledge of the situation and the likely consequences. […] Had a human being committed the act, they would immediately be charged with a war crime. Who should we try for a war crime in such a case? The robot itself? The person(s) who programmed it? The officer who ordered its use? No one at all? (Sparrow, 2007, pp. 66–67)

The unique combination of unfeeling and autonomy that characterizes these and other robots results in a "responsibility gap" (Matthias, 2004). It is because such robots are human-made artifacts, not natural entities, that there is a need to blame someone. No similar responsibility gap arises from the existence of, say, tornadoes. Likewise, we debate the ethics of war and not the ethics of tsunamis, because we realize that the former class of events is avoidable. Hence, this question "as to what extent persons can or should maintain responsibility for the behaviour of AI has become one of, if not the most discussed question […]" (Santoni de Sio & Mecacci, 2021, p. 1058; see Berber & Srećković, 2023; Coeckelbergh, 2020; Gunkel, 2020; Matthias, 2004; Theodorou & Dignum, 2020). We can now inventory different possible solutions. Clearly though, not all ways of dealing with harmful autonomous robots are on equal moral footing.

To see why some proposals are deficient, consider this nightmare scenario. A robot is built which can (somehow) make its own decisions and act on those decisions if it chooses to do so. In other words, it has genuine autonomy. Such a robot is not just built, but activated (not by you). Some time after its activation, it goes on a spree, killing innocent people. After much violence, authorities capture the robot. Since this robot cannot feel pain, there is no point in punishing it. Still, a human-made machine is not exactly a hurricane, so we need someone to blame. Predictably, a mob forms.

Unpredictably, that mob blames *you* for the robot's murderous rampage. Surprised by this spontaneous ascription of blame, you rightly protest that you played no causal role in the tragedy. The mob is unmoved by your pleas of innocence and holds you morally answerable for what has happened. Further nightmarish repercussions ensue.

Now, replay this scenario—with one important modification: *you* are the one who *decides* to activate the robot in the first place. Of course, being autonomous, the robot will go on to make decisions and perform actions which you could not foresee and in which you had no say. People are thus afraid of releasing such a powerful unknown into the wild. So, to gain popular support before flipping on the switch, you publically make yourself answerable in the event that the robot does something morally reprehensible (like killing innocent people). Sadly, it does. You are thus blamed by a mob. But, compared to the previous scenario, the desire to blame you for the robot's harmful actions seems far more justified. We want to survey and rank responses like these.

Bridging the Responsibility Gap

The troublesome combination of unfeeling and autonomy posed by sophisticated robots is not one that major normative ethical theories have paid much attention to. Mindful of the need for new solutions, we have argued that the responsibility gap can be bridged by what we term "blank check responsibility." According to this proposal, "[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*" (Champagne & Tonkens, 2015, p. 126; emphasis in original).

We want to go beyond our original article in several ways. First, we did not situate the blank check among competing proposals, so we want to fill this lacuna. We mentioned, in passing, that our proposal "retains the noncausal imputation involved in scapegoating while dropping its arbitrariness" (2015, p. 136). Prompted by recent work by Kiener (2022), this comparison can now be brought into much sharper focus. Second, situating the stance will allow us to give the blank check a technical label that captures well what it involves, namely *self-initiated prospective moral answerability* for autonomous robot harm (non-autonomous robots do not need this solution). Third, our stance arose from discussions of automated warfare, but we think that it can cover all robots, not just military ones. Finally, given the programmatic nature of our original suggestions, there remained many unanswered questions, so we tackle some (though by no means all) of them.

Despite these clarifications and expansions, our goal remains the same, namely to move from "Wait, don't push that button, it might lead to senseless violence" to "Wait, don't push that button, it might lead to senseless violence, and if it does, you will be held responsible and punished" (Champagne & Tonkens, 2015, p. 136). There are other ways of bridging the responsibility gap—we will survey four more—but we think the blank check emerges as best, overall.

As Kiener explains, our main proposal is "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" (Kiener, 2022, p. 577). One advantage of issuing such blank checks is that “one can deal with this issue [of the responsibility gap] without promoting the need of attributing moral responsibility to” autonomous robotic agents (Bernáth, 2021, p. 1372). However, the price to pay for this advantage is a decoupling of the causal and the moral. When a blank check is issued, the robot is *causally* responsible for its acts, whereas the person who stepped forward and vouched for the robot is *morally* responsible. This seals the gap—irrespective of the fact that no other connection can be established between what the autonomous robot did and what the vouching human did or desired.

This decoupling of the moral and the causal departs from mainstream intuitions. Miranda Fricker holds that “blame is out of order when one does bad things through no fault of one’s own. If no fault, then no appropriate blame” (2016, p. 170). Gary Watson holds that “to blame (morally) is to attribute something to a (moral) fault in the agent” (2004, p. 266). Susan Wolf holds that “the paradigm of blame” involves directing that attitude (often with anger) “toward someone who is perceived to have committed a relevant offense” (2011, p. 344). In their path model of blame, the psychologists Malle, Guglielmon and Monroe start with event detection and immediately provide an exit to “no blame” if no agent causality is present (2014, p. 151). In their words: “If no agent (person or group) is causally linked to the norm violation, the social perceiver may feel angry, sad, or worried, but blame does not arise because there is [no] target for it” (ibid.). While agreeing with us that moral responsibility can be attributed via a speech act, Kiener concludes that “one cannot fittingly ‘blame’ another person when that person is faultless, and a mere declaration to accept blame cannot make a difference here either. Hence, if there is to be genuine blame, there must be fault too” (2022, pp. 578–579).

Taddeo and Blanchard call this the “causality condition.” This condition says that “there has to be a causal connection between the decision/action of the agent and their effects” (2022, p. 4; see also Fischer & Ravizza, 2000; Sartorio, 2007). Our blank check proposal weakens this causality condition, since it construes the notion of causal connection in a more permissive way that includes the mere activation of a robot. The mainstream intuition says that there must be *direct* fault, but we contend that consensually-accepted indirect fault will do, especially when it is the only kind of fault available.

One might worry that a notion of blame which includes only a declaration (required for activation) would be insufficiently close to what the literature on blame usually focuses on. It is important to bear in mind however that, had a human not willingly vouched for the robot, things would have happened differently, by not happening at all. The metaphysics of such blameworthiness could thus be analyzed modally instead of causally. But, on any analysis, the human’s decision was not entirely divorced from the events that ensued. If a robot is truly autonomous, then accepting responsibility for its actions is insufficient to steer the robot in one direction or another. Still, because vouching was necessary for the robot to even act, vouching rendered the ensuing events possible. So, if it is true that “[o]ne thing that matters for the degree to which you are morally responsible for an outcome is your precise causal contribution to the outcome—intuitively, ‘how much’ you contribute to the outcome’s occurrence” (Bernstein, 2017, p. 165), then the fact that unhappy events would not have transpired had one not pushed the "on" button lays quite a bit of blame at the button-pusher's feet.

Humans feel pain, care what others think, and typically desire to not be confined for long periods of time, etc. Hence, unlike a robot, a human can be the locus for meaningful blame and punishment. The blank check is a social device to locate who to blame and punish. The exact nature of the blame and punishment is unimportant (for our philosophical purposes, at least). That it counts as blame and punishment and emerges as just is all that matters. Likewise, if a person takes responsibility for a robot and this robot does positive things, it could be possible to praise and reward the person for vouching for the robot, even if that person had little to do with the positive behavior. We shall nevertheless focus on blame, for simplicity.

Blank checks, as we shall see, can serve as useful deterrents. Retributivists hold that wrongdoers deserve to be punished, quite apart from whatever good consequences might follow from such punishment (Danaher, 2016; Kraaijeveld, 2020). It is not clear that the proponent of the blank check needs to take a side on this issue. Still, it is hard to see why a retributivist would object to punishment also having good consequences such as deterrence. If one deems an act to be bad, then it seems analytically entailed that one prefers the non- or low-occurrence of further tokens of that type of act.

Not all robots are the same, so the likelihood of harmful acts might vary widely. A robot built to serve as a security guard may be in a better position to cause harm than a robot built to deliver goods. A person vouching for a robot's actions would no doubt consider such intended functions and contexts when deciding whether to make herself morally answerable. Someone who willingly accepts responsibility knows that the robot could misbehave. Despite knowing this, the autonomy of a robot invariably makes the human decision risky. In this mix of knowledge and ignorance, one's hope for a good outcome is rendered consequential by publicly putting one's skin in the game. Since nothing in the blank check account compels anyone to accept this "moral gambit" (Taddeo & Blanchard, 2022), one can always decide to not activate the robot or ensure that the robot is sophisticated but not autonomous.

Mapping and Evaluating Five Possible Responses

We have just argued that voluntarily vouching for a robot—visibly and in advance of its deployment—is sufficient for assigning responsibility to a human in the event that the robot harms people. This blank check proposal, which was originally conceived for military contexts (Champagne & Tonkens, 2015), has received a lot of attention (see most recently Kiener, 2022; as well as Behdadi &, Munthe, 2020; Bernáth, 2021; Cernea, 2017; Chandler, 2018; Chomanski, 2021; Di Nucci, 2018; Gerdes, 2018; Hew, 2014; Köhler et al., 2018; Kraaijeveld, 2021; Kühler, 2020; Lima et al., 2021; Oimann, 2023; Restivo, 2017; Royakkers & Olsthoorn, 2018; Smith & Vickers, 2021; Taddeo & Blanchard, 2022; Tigard, 2021a; Tollon, 2021). However, Stenseke has recently expressed worries that, because the people involved in these debates often come from different fields, differing disciplinary assumptions "can serve to cement incommensurable visions and perspectives of the near- and long-term challenges of AI" (2022a, p. 2). Drawing on work by Baalen and Boon (2019), Stenseke thus calls for the establishment of "metacognitive scaffolds" that can help us "better analyze and understand […] respective views and commitments" (2022a, p. 9). We want to answer that call.

While Stenseke's own scaffold takes the form of a list, we think the field can be better represented hierarchically, as answers to key questions. As Kiener rightly observes, we make sure to "talk about taking *prospective* responsibility rather than *retrospective* responsibility" (Kiener, 2022, p. 577). We thus want to use this useful prospective/retrospective distinction and augment it with a further distinction. One could say, for example, "You will answer for what has happened." Here, answerability is required *of* a subject *by* a community. Vengeance, for example, runs in that direction of fit. Alternatively, one could take the lead and say "Let me answer for what has happened." Here, answerability is offered *by* a subject *to* a community. Atonement, for example, runs in that direction of fit. This combined analysis results in four possible stances:

Self-initiated prospective moral answerability
Other-initiated prospective moral answerability
Self-initiated retrospective moral answerability
Other-initiated retrospective moral answerability

For ease of use, we can encapsulate each stance as follows: *blank checks*, *framings*, *martyrs*, and *scapegoats*, respectively. An additional stance, which is defended by Stenseke (2022b) and which we will look at shortly, might be termed *pretense*. The conceptual space can therefore be carved in five, as shown in Fig. 1.

Once mapped, these five stances can be critically assessed and ranked. Our main contention is that, on reflection, blank checks emerge as the most just and plausible option. This is because pretense (option 5) is a non-starter, while the remaining options (2 to 4) are variations on mob violence. Let us therefore look at those options, from the least desirable to the most desirable.

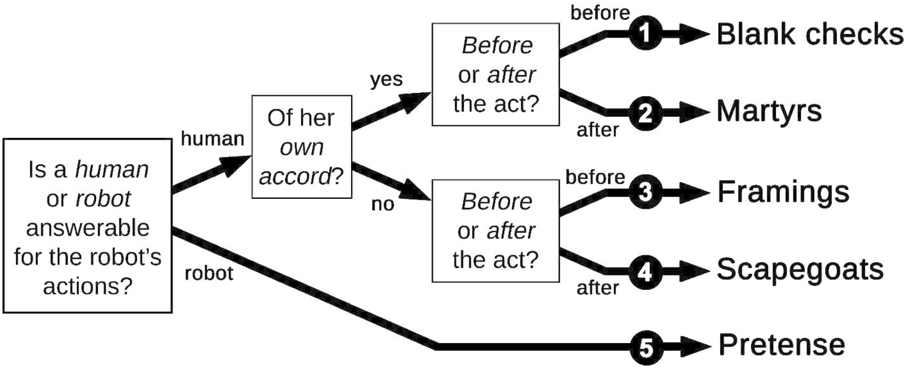
Is a *human* or *robot* answerable for the robot's actions?
human
robot
Of her *own accord*?
yes
no
Before or *after* the act?
Before or *after* the act?
before
after
before
after
1
Blank checks
2
Martyrs
3
Framings
4
Scapegoats
5
Pretense

Fig. 1 Five ways of assigning moral answerability for robotic actions

Pretense (Metaethical Fictionalism or Instrumentalism)

It is likely that autonomous robots “promise to transform not only the evaluative categories that we adopt in the legal and public policy domains, but also, more deeply and less obviously, the spirit of our customs and social norms” (Cappuccio et al., 2021, p. 2). Hence, one way to deal with the misdeeds of unfeeling autonomous robots is to act *as if* they care about our disapproval and punishments (Cappuccio et al., 2019). This change, while surprising, moves robots from the category of artifacts to the category of moral agents and/or patients. The boon is that we already know how to deal with the latter category (via blame, praise, etc.).

What is needed, according to this pretense response, is not a technological innovation but a social innovation: robots are now part of our lives, so our attitudes must keep up with the times. Since one can insist that “[v]iciousness towards robots is real viciousness” (Sparrow, 2021, p. 23; see Sparrow, 2017), treating robots the way we normally treat humans would seem consistent with building good character. Doing otherwise, it is argued, could lead humans to be crueler towards each other (Coghlan et al., 2019). People who champion this approach thus hold that robots are “unable to suffer like us” but nevertheless “can and often should be targeted with reparative measures” (Tigard, 2021b, p. 604).

This view is being defended by Hage (2017) and many others. Now, “[s]keptics may be inclined to dismiss the idea of punishing AI from the start as conceptual confusion—akin to hitting one’s computer when it crashes” (Abbott, 2020, p. 112). One of us, for example, regards questions of ontology as mandatory (Champagne, 2021). By contrast, Coeckelbergh “replaces the question about how ‘moral’ non-human agents *really* are by the question about the moral significance of *appearance*” (2009, p. 181; emphasis in original). Clearly then, “AI punishment cannot be dismissed out of hand. It is necessary to do the difficult pragmatic work of thinking through its costs and benefits, considering how it could be implemented in practice, and comparing the alternatives” (Abbott, 2020, p. 112).

Whether or not the tradeoff is worthwhile no doubt turns on how seriously one takes metaphysics. David Gunkel, for instance, draws on the work of Emmanuel Lévinas “who, in direct opposition to the usual way of thinking, asserts that ethics precedes ontology” (Gunkel, 2018a, p. 95). Gunkel (2018b, p. 166) acknowledges that Lévinas did not engage with robot ethics. But, Lévinas did make much of the human face, which he saw as the wellspring of all moral obligations: “The first word of the face is the ‘Thou shalt not kill.’ It is an order. There is a commandment in the appearance of the face, as if a master spoke to me” (Lévinas, 1985, p. 89). Lévinas’ approach is phenomenological, so how things appear in regular experience is, for him, decisive (Lévinas 1998, p. 33). As Mamak recently put it, “a decision on what a robot is must be based not on the intrinsic internal qualities of a robot but on its appearance,” such that policy makers have the go-ahead to “ignore philosophical deliberations” (Mamak, 2022, p. 1059). Given the ethical conundrum posed by the responsibility gap, an argument can be made that this retreat to appearances is beneficial, overall. Indeed,

> Standard moral theory […] is tailored to human agency and human responsibility, excluding non-humans. It makes a strong distinction between (humans as) subjects and objects, between humans and animals, between ends (aim, goal) and means (instrument), and sometimes between the moral and the empirical sphere. Moral agency is seen as an exclusive feature of (some) humans. But if non-humans (natural and artificial) have such an influence on the way we lead our lives, it is undesirable and unhelpful to exclude them from moral discourse. (Coeckelbergh, 2009, p. 181)

Although some philosophers claim that robots already possess the internal states required for morally responsible behavior (Søvik, 2022), the pretense stance is primarily "concerned with performance—behavior that conforms to moral values—not with 'what's going on on the inside' (agents' reasons and intentions)" (Gogoshin, 2021, p. 2). The fact that robots lack consciousness and/or feelings is not relevant, on this view. The argument is instead that, if we define moral agency as membership and mutual recognition and accountability in a "moral community" (Strawson, 2008, p. 23), we can use the practices found in a given community to adjudicate who counts as responsible for their actions. Once morality has been (re)defined in this way, autonomous robots "who have the capacity to reliably behave in accordance with the relevant moral rules and values of their social environment" can be seen as "morally responsible agents" (Gogoshin, 2021, p. 2). Inclusion in already existing practices is aided by the fact that robots can be built and programmed to display all the usual signs of remorse, guilt, defiance, and so on.

In our estimate, the pretense approach is unsatisfactory. If a society can decide to drastically reconfigure its moral practices to fit emerging technologies, what prevents it from disregarding other facts in order to expedite a happy ending? Functionalism in philosophy of mind is motivated in part by the thought that we cannot verify whether other people are "zombies" or conscious (Chalmers, 1996). With robots, however, we (or at least some of us) have access to the solution sheets, since we humans built those machines. It is thus strange to feign ignorance about something we already know, merely on account of this proving expedient for solving certain moral dilemmas.

By analogy, we could spontaneously eliminate the problem of illiteracy by acting as if illiterate people weren't people, but everyone knows that this would not really solve the problem. Such a consequence may look uncharitable, but it actually takes metaethical fictionalism (Joyce, 2001) seriously, in the same way that "the repugnant conclusion" takes utilitarianism seriously (Parfit, 1984, p. 388). The pretense stance is not fictionalist or instrumentalist, it is *selectively* fictionalist or instrumentalist. So, one must show what principled difference (if any) makes metaphysical concerns vital in one context and superfluous in another.

It is hard to see how there could be a *mens rea* without a *mens*. Engineers who speak of AI as being "conscious" may receive abundant attention from a public disposed to believe it too, but bedazzling our superstition modules lets those engineers escape scrutiny (the Ring of Gyges fable reminds us how actions veer more easily towards the immoral when they escape the scrutiny of peers).

Tigard holds that "our moral attitudes and practices are adaptable and will likely continue to evolve," so with time "we can coherently interact with AI systems—particularly those that are being developed to respond accordingly—in ways that assign a sort of responsibility" (2021a, p. 443). Yet, in spite of their sophisticated mimicry, robots neither care nor feel. When H. L. A. Hart defined the necessary conditions of punishment, the first requirement of his five-fold list was that it "involve pain or other consequences normally considered unpleasant" (2008, p. 4). People who hold that robots cannot be proper targets of moral answerability and blame are not "chauvinists" (Sætra, 2021). They merely don't want us to waste energy while the real culprits silently walk away. One needn't believe in karma to think that something is askew when victims are harmed and victimizers are unharmed.

Not all accounts of responsibility involve imposing an unpleasant experience on the offender, as illustrated by medieval trials of inanimate objects. So, anthropologically, the pretense response is perfectly feasible. Punishment partly serves an expressive function (Feinberg, 1965), so it may also prove legally feasible to punish robots, provided all parties believe that justice was served. When medieval villagers would put a pig on trial for devouring a child, the community no doubt found an outlet for its frustration, satisfied its "desire to endorse social values through acts of public justice," and "provided an explanation for tragedy" (Oldridge, 2005, p. 49). But, even if all those playing this particular language game emerged quite content from the exercise, it is weird to say that justice was served, especially if the pig's owner walked away with impunity. In the European city of Falaise in 1386, a pig which had harmed infants "was dressed in a new suit of man's clothes" before being hung in front of the many spectators gathered (Carson, 1917, p. 410). Our ability to anthropomorphize robots has gotten dramatically better than merely adding clothes. Even so, blaming and punishing a robot results only in mock-justice.

It would certainly not violate the laws of physics for courts to deal with robots in the same manner that they already deal with humans. The interesting philosophical question, however, is whether such completely pain-free "justice" would be more than performative. Performance may be a prominent part of how we deal with technology (Coeckelbergh, 2019), but so is the demand for a real basis.

Stenseke writes that, if autonomous robots "were ever to become a technical possibility, one would hope that there would be […] social movements that advocated for their rights and well-being" (2022b, p. 15). We are not alone in rejecting this pretense approach to moral answerability. Dennett (1987) spent his career in philosophy of mind arguing that "as if" displays of intelligent behavior suffice and was even one of the first to address the responsibility gap (Dennett, 1997). His permissiveness towards robots in the descriptive domain does not, however, carry over to the normative domain. In fact, he now thinks that "counterfeit people are the most dangerous artifacts in human history, capable of destroying not just economies but human freedom itself" (2023). In his estimate, "civilization itself is at risk" (ibid.). You *can* treat robots like persons, Dennett argues, but really you *shouldn't*.

Abbott suggests that we treat robots like we do corporations, which "are a member of our legal community but not our moral community" (2020, p. 4), but it is unclear whether such an analogy with corporations would be stable. One of the reasons why societies are fine with punishing corporations is that we know, irrespective of the particular corporate structure, that fines will eventually be felt by real humans (as financial losses, and so on). Abbott hints at this when he writes that "[t]here could be benefits to punishing AI […] because it could affect the behavior of AI developers, owners, and users […]" (2020, p. 112). We should therefore distinguish outright pretense, where we punish a robot and no human suffers, and mixed approaches that insist on some human being in the loop, perhaps unofficially. Indeed, a person "may blame an AI system to signal their commitment to a shared set of norms and, at the same time, blame developers to condemn immoral behavior" (Lima et al., 2023, p. 15). Once we relinquish the idea that morality is tasked with tracking real properties, it is unclear what prevents us from switching between stances on an ad hoc basis.

No doubt, "machines soon to be among us will be capable of recognizing and learning from our moral attitudes and practices—our anger and blame, gratitude and praise—perhaps just as effectively as from our simpler, non-moral commands" (Tigard, 2021a, p. 443). It will not take much mimicry to dupe us, since humans have an innate disposition to treat things like persons, especially when those things are sophisticated (Kneer & Stuart, 2021; Lima et al., 2021; Stuart & Kneer, 2021). Folk psychology answers only to instrumentalist convenience, so it does not care much about what is true or real. False positives also do not cost much, since blaming is a relatively cheap activity. One can blame a table for one's stubbed toe. Yet, to the extent that the practice of blame is a prelude to punishment and actual conflict resolution, the table's unfeeling nature prevents this moral cycle from reaching satisfactory completion. The same goes for robots. Since a human must be kept in the answerability loop, the mock-answerability of pretense will not do.

Framings (Other-initiated Prospective Moral Answerability)

In this strategy for attempting to resolve the responsibility gap, one or more people are held answerable by someone or some group other than themselves, before the robot has been released into the wild. This is unsatisfactory. If a group makes a person answerable for an autonomous robot's actions before anything (bad) has occurred, then this is unjust. Why this person? The choice seems arbitrary. The injustice of this arbitrariness is even more salient when we recall that the autonomous robot's decisions are independent of any human's decisions, i.e. no human decided that the robot would misbehave. The answerability is initiated by others, but the person targeted by the group never consented to making themselves answerable. Hence, the person is essentially being framed for harm that she will not cause, which is wrong (and remains wrong even if the autonomous robot never does anything wrong).

Framing, as it is understood here, should be distinguished from properly defined social/moral/legal positioning. It could be, for example, that a particular position in an office or research lab comes with a code of conduct. Given such well-defined role-responsibilities, accepting the job means accepting to be bound by these constraints and duties, which may include answerability for harms caused by a robot. Pointing to such an acceptance would not be to "frame" someone, since the second fork in the road one must take to reach position 3 in our diagram is that the answerability *not* be of the person's own accord.

Everyone understands that other-initiation is unjust. In fact, ordinary language has developed the word "voluntold" precisely to capture how other-initiated framings often try to arrogate the moral credentials of self-initiated blank checks. Framing someone may quickly deal with a harm caused by a robot, but one can never know whether one will be the next target of a set-up. Hence, apart from their injustice, wrongful accusations are counterproductive, since they undermine trust in the justice system as a whole (Brooks & Greenberg, 2021, pp. 48–49).

Scapegoats (Other-initiated Retrospective Moral Answerability)

In this strategy for attempting to resolve the responsibility gap, one or more people are held answerable by someone or some group other than themselves, after the robot has committed its unacceptable act. This is the nightmare scenario captured in our opening section. Yet, if a group blames a person for an autonomous robot's actions after something bad has happened, then this blame is unjustly assigned, since the person had nothing to do with the robot's autonomous actions.

Recall that, owing to their autonomous nature, the robots that concern us are not causally guided by the decision(s) of any human. This means that, even if God was the detective investigating the case and testifying in court, no direct connection could be established between an accused human and the robotic crime. So, if a human did not signal their prior answerability, then making that human answerable would be an unjust form of scapegoating.

There are, however, considerations that cast in a more favorable light other-initiated retrospective moral answerability. Since legislative bodies limit the actions of subjects with or without those subjects' agreement, the law's imposition is clearly other-initiated. And, since we do not have the "pre-crime" prescience described in Philip K. Dick's novella "The Minority Report," we must apply the law only after harm has been committed. The presence of other-initiation and retrospection thus seems to corner us into viewing the law as a form of scapegoating, which is clearly unacceptable.

To see how the scapegoating response differs from the law, consider the following. In the blank check approach, one cannot vouch for a robot alone in one's basement, anymore than one can marry another person without a witness, documents, and whatever else society requires. Answerability must take root in a community. Now, a proponent of the scapegoating response could say that, when others blame one for harms done by a robot, the answerability at hand was in fact self-initiated, perhaps via some sort of tacit social contract. The problem, however, is that securing a person's general participation in the rules of society falls short of establishing a person's responsibility for specific harms done by a specific robot. Why should the law (or any community) pick on *this* particular person, as opposed to another? This need for a specific target of blame explains why, absent the uncoerced issuance of blank check, it would be unjust (for a legal system or a mob) to hold a person culpable for deeds done by an autonomous robot. The law thus differs from other-initiated retrospective moral answerability.

Laws arguably aspire to have some moral purchase beyond brute physical enforcement. Hence, laws about harmful autonomous robots that go against how humans understand and assign blame and moral responsibility would likely remain "laws in books" and not "laws in actions" (Brożek & Jakubiec, 2017). Taddeo and Blanchard, for example, think that our proposal "ascribe[s] moral responsibility *nominally*" (2022, p. 12; emphasis in original). Yet, if defenders of the pretense response can invoke the plasticity and adaptability of social practices to predict what "the moral community" will in time see as natural (Gogoshin, 2021), surely defenders of blank checks can do the same. Germany, for example, "has so far resisted the idea of expanding criminal liability to nonhuman agents," (Gless et al., 2016, p. 415), so the blank check proposal might take root there. In any event, the future is not yet here and different legal systems respond to unprecedented cases differently, so passage into law (Turner, 2018) is probably a bad gauge for judging the five ethical stances surveyed here.

Recoiling from the first nightmarish scenario laid out at the start of this article, our original article explicitly distanced the blank check proposal from scapegoating: "It is important to underscore that our way out […] does not entail that a community will arbitrarily select a prominent figure as a lightning rod for its disapproval" (Champagne & Tonkens, 2015, p. 133). Indeed, we described our solution as a sort of "vouching for" rather than a "pointing of fingers" (ibid.). Answerability in the form of "vouching for" already exists as a practice. People vouch for their friends during job searches, parents vouch for their grown children (e.g., on home mortgages), supervisors vouch for the work and etiquette of their workers, principal investigators of clinical research vouch for the work of their research assistants (e.g., to uphold high standards of research ethics), and so on. Such prospective answerability may be new to the domain of autonomous robots, but it is arguably more reliable and just than retrospective answerability.

Martyrs (Self-initiated Retrospective Moral Answerability)

In this strategy for attempting to resolve the responsibility gap, one or more people hold themselves answerable, *after* the robot has caused harm. David Enoch champions such a self-initiated retrospective approach in the following example, where we might substitute "teenage son" with "autonomous robot":

> Your teenage son commits a crime, causing harm to person and property. You are not, let us suppose, directly responsible for the crime in any straightforward way—it's not as if you put him up to it, or even drove him to this kind of thing by your poor parenting. Parenting too, after all, is a percentage game, and this time you lost. […] Still, we would judge unfavorably a parent who neglects to—in some sense—take responsibility for her teenage son's behavior, perhaps, for instance, by apologizing for him, or some such. If facing hard questions, you settle for noting (correctly) that the relevant action was not yours, there seems to be something amiss—I would say, morally amiss—with your so doing. […] There is something to be said for your (in some sense) taking responsibility for your teenage son's action. Or so, at least, it seems to me. (Enoch, 2012,pp. 97–98)

Others share Enoch's intuition. Near the end of their widely-viewed 2023 video "The AI Dilemma," Tristan Harris and Aza Raskin say that "if you are going to release a little [AI] alien then, just like a child, if it goes and breaks something in the supermarket, you have to pay for it," so similarly "if you're a Facebook or whoever is making the [large language] models, if it gets leaked and it is used, then you should be responsible for it" (our transcription). Harris and Raskin mention this matter-of-factly, but in cases where a machine acts *autonomously*, it is far from obvious what justifies the ascription of responsibility, since an engineer or CEO or customer can always protest—quite rightly—that they had nothing to do with the robot's particular decision or action. Hence, the reason why Enoch's parent example "seems to capture something dear to our heart in the phenomenology of responsibility" (ibid., p. 100) is that the "Who?" question has been (prospectively) answered.

Unlike Enoch's case, which frontloads a connection, autonomous robots have no parents. So, who—*specifically*—should step up and take responsibility when such autonomous robots cause harm? Humans abound, so to foreground a particular individual, we need what philosophers of language call a sortal. Proponents of the blank check approach claim that only those who have *voluntarily* made themselves answerable *beforehand* can rightly be held responsible and possibly blamed. Our comparative analysis aims to show that selecting anybody else by any other means is unjust.

What would our "phenomenology of responsibility" (Enoch, 2012, p. 100) say if some random person, upon hearing of a teenager's crime, took it upon themselves to express the same regret and need to answer as the teenager's parent? If a person wants to be blamed for misdeeds they played no part in, then their motivations are arguably dubious. Of course, if one had vouched for that teenager (as, say, a legal guardian) before the incident, then the whole situation changes. But, absent such prior vouching, the answerability seems intrusive and even unhealthy. One could apologize for such harms, but not only would such an apology be supererogatory, it wouldn't clear the threshold of blame and thus wouldn't actually fix anything.

The parent in Enoch's example isn't just *any* parent but the parent of the actual misbehaving teenager. The answerable individual has thus already been specified. The situation that concerns us is significantly different, so it turns out that we cannot substitute "teenage son" with "autonomous robot."

Enoch agrees with us that it is possible for an agent to be "not responsible for the relevant thing" yet "*take* responsibility, and thereby *become* responsible" (Enoch, 2012 pp. 101–102; emphasis in original). This possibility is available because "an act of will can make all the difference" (ibid., p. 101). Kiener, however, believes that this self-initiation "is mistaken" (2022, p. 580) or at any rate one-sided, since "people can [make] themselves morally answerable for the harm caused by AI systems, not only ahead of time, but also when harm has already been caused" (2022, p. 576). Clearly, people can do this. The question is whether the moral community—which also has a say in the transaction and must also strive to be just—should accept such retrospective answerability. We do not think it should.

If a person's offer to willingly sacrifice themselves quells social unrest and psychological unease, expediency may trump other reconsiderations. We can imagine a well-intentioned utilitarian wanting to stop an angry mob and diffuse a tense situation by fabricating their involvement after the fact. Yet, supposing that the person's lack of prior involvement is known, blaming a willing martyr would amount to abetting self-inflicted harm. A person asking to be punished for horrors they had no connection with should be turned away (and perhaps offered counseling and/or a medal) by a truly moral “moral community” (Strawson, 2008, p. 23). A well-meaning Hollywood celebrity cannot ask, for example, to be jailed for the genocide in Darfur. It may be that “taking on too much responsibility seems a less serious flaw than taking on too little” (Mason, 2019, p. 203), but the “too much” remains. The injustice thus holds—even if the blame would provide an outlet for the group and the martyr’s shared indignation over what has transpired.

Another disadvantage of the martyrdom approach is that it faces a potential shortage of people. If (as the blank check recommends) we refrain from activating a robot until a person steps forward to make herself answerable for its forthcoming deeds, we will have at least one human to blame for every robotic harm caused. By contrast, when self-initiated retrospective moral answerability faces a dearth of volunteers, we can be left with robotic harms that go unpunished. This flaw of the martyrdom stance can be unpacked using the modal terms mentioned earlier. In the blank check situation, vouching is a necessary condition for activation. So, had a human not willingly vouched for the robot, things would have happened differently, by not happening at all. This counterfactual conditional fails to hold when a person makes herself answerable after the fact: remove her self-initiation and the atrocities remain. So, in addition to being unjust, self-initiated retrospective moral answerability is a poor deterrent.

Blank Checks (Self-initiated Prospective Moral Answerability)

Hopefully, the foregoing survey of stances shows why a person must “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” (Kiener, 2022, p. 577). Crucially, the linking of fates must be done at a time when the robot’s harmful actions are unknown. Of course, once a tragedy has occurred, a programmer, owner, or manufacturer might be riddled with guilt and thus seek atonement for what they see as their misdeeds. But, if one can in fact connect a specific person to a specific robotic act (in the weak or strong ways discussed by Himmelreich, 2019), then the robot in question was not truly autonomous and thus does not require the exotic solutions covered in this paper.

Kiener argues that in order to be responsible for the harm caused by an AI system, “one must have been involved in the use or development of the AI system that caused harm. This is because […] one’s involvement is a condition of responding meaningfully to those who have been harmed” (2022, p. 586). This strong reading of the causality condition means that “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” (ibid.). This is fine, as far as it goes. Someone in the aforementioned assembly line of skilled contributors would be a natural candidate for stepping up and vouching as the final contributor. However, if the person is optimistic about the unlikelihood of harm because she exerts some prior or present control over what the robot will do, then the situation gets expelled from the narrow subset of autonomous cases that generate the responsibility gap. Absent autonomy, a robot would become a (admittedly complicated) tool or instrument of human agency. We already know how—and have the legal tools—to deal with such cases. It is the more troublesome possibility of genuinely autonomous yet unfeeling robots which motivates the blank check proposal.

Presumably, those most likely to issue blank checks for the deployment of robots would be high-profile people, especially in the early days when autonomous robots would be rare and expensive. Given that powerful government officials and business people are not held accountable as often or punished with as much severity as ordinary people, one would expect practical problems relating to power and bias that plague the administration of justice to carry over to all five approaches. The advantage of the blank check approach, however, is that it at least identifies the proper target of blame (whether or not that target is actually reached).

The person who vouches for a robot is the person who activates it, so this person caps off a long list of people who were necessary but not sufficient for the autonomous harms that followed. You can have a financier, a software engineer, a roboticist, and a whole assembly line of skilled contributors working in concert; but, without the person who flips the "on" switch, none of those people would amount to a functioning robot. The blank check approach thus avoids the "problem of many hands" (van de Poel et al., 2015), because in our account it would be clear who is answerable for the unfeeling yet autonomous system's subsequent behavior.

Would it be deficient if just one of these people took responsibility? "Standard moral theory has difficulties in coping with these questions," because "it generally understands agency and responsibility as individual and undistributed" (Coeckelbergh, 2009, p. 181). A single AI, however, can be realized in multiple robots at once. Was it one super-robot or an army of robots that caused harm in the movie *Avengers: Age of Ultron*? If we opt for the one-robot gloss, would it suffice for Tony Stark to accept responsibility for every bad act that transpired? Replace this fictional example with Tesla bots and Elon Musk and the issue rapidly gains urgency. Without claiming answers to these questions, we would insist that, if the answerability is anything other than self-initiated and prospective, it will fall short of being just. As Dennett writes, "it would be reassuring to know that major executives, as well as their technicians, were in jeopardy of spending the rest of their life in prison in addition to paying billions in restitution for any violations or any harms done" by counterfeit people (2023).

Contrary to martyrs who retrospectively self-initiate their moral answerability, those who prospectively issue blank checks for a robot's actions can be credited with healthier motives, since they presumably believe—and up to this point have good reason to believe—that an activated robot will not do unacceptable things. Most engineers build robots with the (perhaps naive) expectation that those machines will enhance our lives, so the mad scientist trope has severe limitations. Still, given that autonomy renders the behavior of a robot unpredictable, a robot may end up doing something that we disapprove of, despite our best intents. This is the price to pay for endowing it with autonomy. Freedom helps a robot cope with novel situations, but since no rule can tell one how to apply a rule (on pain of regress), the faculty of judgment cannot be captured by any algorithm. So, when a person makes herself morally answerable for an autonomous robot's actions before the fact, she announces to the world that, in her estimation, that robot will have good judgment. Naturally, this is a fallible assessment, so the blank check is a way to bet—with a real risk of personal loss—on one's forecast.

What renders such answerability moral is not the element of risk per se, but rather the fact that the person vouching for an autonomous robot *knows* that there is an element of risk. A worthwhile distinction can thus be made between unknown-unknowns and known-unknowns. A policy-maker or engineer can hardly be blamed for an unknown-unknown. These are rightly classified unforeseeable accidents. With *known*-unknowns, however, the ethical weights get radically redistributed, since harms resulting from known-unknowns allow us to speak of undue risk, negligence, callousness, short-sightedness, etc. As Santoni de Sio and Mecacci explain,

> To what extent it is possible to establish standards of reasonable care for the design, use, and regulation of AI in the same way in which we do for buildings and bridges is precisely the question raised in present-day (legal) debate on the responsibility gaps for AI. […] [C]ulpability gaps with AI may happen precisely because the traditional assumptions about what should count as sufficient intention, knowledge, and foreseeability on the side of the defendant (criminal law) may not apply, due to the emergent and unpredictable behaviour of AI. (Santoni de Sio & Mecacci, 2021, pp. 1070–1071)

Because we are only concerned with the case of autonomous robots, no person activating such a robot could have known that it *would* cause harm. Even so, all persons should know that the robot *could* cause harm. Someone given access to a nuclear briefcase should grasp enough game theory and geopolitics to know that firing nuclear missiles would result in mutually assured destruction. Similarly, a person who presses the "on" button of an autonomous robot must have enough capacity-responsibility to know that, owing to its autonomy, the robot might cause harm. Because we can be certain about this uncertainty, moral blame becomes relevant. Recklessness without any prior precaution(s) is not a virtue.

Hope for a Happy Outcome, but with Skin in the Game

All participants to current debates want "to ensure that autonomous robots meet ethical and safety standards prior to their deployment" (Champagne & Tonkens, 2015, p. 134). In a blank check situation, "[i]t seems reasonable to assume that a [person] faced with the decision to deploy autonomous robots would work hard to ensure that those robots behave properly, since it is she who would be held responsible and punished in the event of their misbehavior" (ibid.). None of the other ways of dealing with autonomous robot harm have this built-in incentive.

With pretense, attention gets directed towards robots and away from humans. With framings, the group doing the framing knows it will emerge scot-free. With martyrs, the disaster has already happened, so despite the penitence, it is too late to make any real difference. The same practical helplessness applies to scapegoats.

Blank checks, by contrast, make robot misbehavior less likely. If, say, a robot's judgment must be trained via some sort of machine learning, there is an incentive to extend that learning period for as long as possible, while there is still avenue for action. Alignment with human values becomes, not just a desideratum, but a necessary condition for activation. If the person(s) vouching for the robot cannot attain sufficient confidence that such alignment will obtain, then this is excellent reason for not releasing the robot in the first place.

As the foregoing makes plain, we do not take deployment for granted. Rather, we argue that *if* autonomous robots are deployed, then someone needs to make themselves answerable beforehand for any robotic misbehavior. The antecedent of a conditional need not be affirmed (and give way to a *modus ponens*), so it is perfectly compatible with this proposal to refrain from deploying autonomous robots altogether. It is also possible that humans will vouch for robots and those robots will not harm anyone. Since there is no way to tell, the blank check serves as a provision for the worst case scenario.

Tigard sunders discussants into "techno-optimists" and "techno-pessimists." Techno-optimists are "those who argue that the [responsibility] gap can be bridged" and "who would prefer to harness the newfound benefits of technology and proceed with its deployment" (2021b, p. 590). By contrast, techno-pessimists think that, in light of concerns regarding responsibility in AI, "we must drastically scale back or altogether cease our deployment of AI systems" (ibid.). Conradie et al. (2022, p. 3) rank us as optimists. We prefer to think of ourselves as "techno-pragmatists," since our blank check proposal seeks to convert hope for a peaceful outcome into something actionable.

Everyone "hopes" that a robot will help rather than harm. But, without any consequences attached to this aspiration, it amounts to little. When hope does not bear out, it generates disappointment, not blame or any kind of justice. However, when the serious consequences of answerability are willingly accepted beforehand, hope gets converted into something more robust and consequential. One didn't just *wish* for a happy outcome, one *promised* a happy outcome. This is a crucial difference. Importantly, one cannot promise an outcome after it is known. With the blank check requirement in place, an engineer, politician, business person, or owner would no doubt think twice before pressing the "on" button. Thinking more before acting is rarely a bad thing, especially when the stakes are high.

Even so, one might worry that the blank check requirement would encourage "deadly over-caution"—to borrow Kazman's (1990) description of slow and risk-averse drug approval processes. Potentially life-saving machines could wait idly in storage because the incentive structure that we defend promotes excessive prudence. We should stress, however, that forgoing autonomous robots does not entail forgoing less-than-autonomous robots. It is not regressive, then, to suggest that these humbler robots might offer most or all of the desired practical benefits of autonomous versions while maintaining a traceable connection to human agency.

We Know that We Don't Know what Autonomous Robots Might Do

Responding to work by Smith (2007) and Scanlon (2008), Shoemaker (2011) distinguished between attributability, answerability, and accountability. Attributability tracks an agent's character, accountability tracks an agent's regard for others, and answerability tracks an agent's judgments. Anyone sympathetic to the idea that "there is no [thought-distinction] so fine as to consist in anything but a possible difference of practice" (James, 1898, p. 291) will take such definitional projects with a grain of salt. We can attribute a fine shade of meaning to each word at our disposal, but should we charge English-speaking philosophy of mathematics with conflating "*nombre*," "*chiffre*," and "*numéro*"? Analytic distinctions crafted to map onto words should receive less credence than words crafted to map onto analytic distinctions. Such methodological worries notwithstanding, the net take-away of Shoemaker's terminology is that answerability does not imply worthiness of blame (or praise). This is certainly true. So, to clarify: making oneself answerable for a robot's actions does not automatically mean being worthy of blame, since actual bad actions by the robot are needed. Still, a blank check provides tangible guidance on what to *do*—*who* to blame—in the event that an autonomous robot does cause harm. Since the prospective orientation of the blank check entails that we are dealing with possibilities that have not yet happened and may never happen, the device should be interpreted in the subjunctive mood. Just as blank checks might never be cashed, answerability may never reach blame.

One interesting epistemological feature is that those who vouch for autonomous robots must make themselves answerable when the robot's (good or bad or neutral) actions are unknown. Focusing only on harms, we might distinguish between not knowing if a token of a known type of harm will occur and not knowing if a previously unknown type of harm will occur. We know, for example, that civilians can be killed, so a person might wonder whether a robot will add yet another element to that non-empty set. Such a situation would involve one-ply ignorance, so to speak. Yet, vouching for a robot before the fact also involves two-ply ignorance, since the robot might go on to commit a previously-unknown type of harm. Again, when one issues a regular blank check, one can never be certain what it will be used to purchase.

One might object that, since a person can never be 100% confident that an autonomous robot will behave nicely, one should not make any promise that one is unable to keep. However, this does not render self-initiated prospective answerability unfair or unreasonable. On the contrary, it is precisely because we can never be certain of happy outcomes that vouching is needed. A critic of the blank check could argue that it is implausible for a human to make themselves answerable for types of harm that are completely unpredictable. However, we would reply that being answerable for new types of harms is simply part of the bargain. Activating an autonomous robot or AI invariably involves risk, but "[v]irtually every action carries with it some risk, however small, of serious harm to others, and so assigning individuals the right not to be subjected to risk, without their consent, is an impossible position" (Hayenhjelm & Wolff, 2012, p. e27).

Of course, the possibility of generating new kinds of harms shows just how radical it is to sign-off on the activation of autonomous robots. But, this is because autonomous robots are themselves radical inventions. Indeed, despite being our creations, these machines are capable of surprising us, not just by performing token actions of known types, but also by performing token actions of new types (it is easy to project utopian hopes and dystopian fears onto this inkblot). This possibility of novelty is a feature of autonomy that all the proposals surveyed have to contend with, so it poses no special obstacle to the blank check. In fact, we would argue that it justifies the extra caution captured by self-initiated prospective moral answerability. If "equal recklessness deserves equal blame" (Wolf, 2001, p. 6), then the person vouching for a robot might receive quite a bit of blame for allowing a new type of harm to appear on the moral landscape.

Releasing a known unknown into the wild is demonstrably a risky act, so making the matter hinge on how things work out involves an element of moral luck. That said, vouching beforehand of one's accord is *not* an act of luck, so it provides a viable way to confront an uncertain future.

Inference to the Best Inculpation

When choosing which of the five catalogued paths to take, the main desideratum should be that the response be just. We argue that, if one accepts this desideratum, it becomes vital that a person making themselves morally answerable for the actions of a robot do so *of their own accord* (i.e., self-initiated) *beforehand* (i.e., prospective). Therefore, of the five options surveyed, blank checks emerge as the best.

Blank checks are like pretense in asking us to reshape our social practices (by introducing vouching) but unlike pretense in insisting that robots don't feel and that humans must be kept in the loop. Blank checks are like framings in that the potential target of blame is selected before robotic harms are committed but unlike framings in that the person accepts this blame willingly, not coercively. Blank checks are like scapegoats in that a human is held responsible but unlike scapegoats in that a paper trail proves that the person accepted this beforehand. The blank check weakens the causal condition to a mere vouching, but scapegoating abandons it altogether. Blank checks are like martyrs in that they accept a person's offer of answerability but unlike martyrs in insisting that such an offer must be made beforehand, when non-activation and the possibility of a good outcome are still viable.

Santoni de Sio and Mecacci divide approaches to the responsibility gap into three categories: approaches that hold that "the responsibility gap is a new and intractable problem" (fatalism); approaches that hold that "the responsibility gap is not new and not a problem" (deflationism); and approaches that hold that "the responsibility gap is a problem that can be solved by simply introducing new technical and/or legal tools" (solutionism), which they further sub-divide into technical and legal solutions (2021, p. 1068). Our approach is not fatalist, because it holds that the problem is tractable. It is not deflationist either, because it acknowledges that the responsibility gap is a (new) problem. While the blank check does propose a solution to the responsibility gap (in certain specific circumstances), it is a different kind of solutionist approach than what Santoni de Sio and Mecacci have in mind, because it involves neither a technological solution nor a new liability regime (although it does not discount the potential role of such legal and technological add-ons).

Interestingly, Santoni de Sio and Mecacci end up endorsing a blank check approach: “In the presence of sufficient knowledge and training […] a military commander can be reasonably held accountable and culpable for his conscious decision to deploy an unpredictable technical system in a military mission, which ends up in the unlawful killing of innocent civilians. […] Similarly, the manager of a car manufacturing company and/or the chair of a road safety agency can be legitimately held accountable and culpable for their decision to put/allow on the public road a vehicle whose behaviour, as they well knew, could not be sufficiently predicted and explained” (2021, p. 1073). Such a stance can easily be misunderstood and veer into something unjust, so we have stressed how the assignment of responsibility must be *self-initiated* (so that no one is framed or scapegoated) and *prospective* (so as to avoid martyrdom).

What creates the responsibility gap is a conjunction of two attributes, namely robots that are autonomous and unfeeling. Both attributes have to be present for the blank check to make sense, so cases drawn from the human realm will invariably present limitations. Suppose that a parent picks their neighbor as a babysitter because that neighbor promised that she would be good. Lo and behold, contrary to that promise, the babysitting neighbor kills the child. The parent may be to blame for trusting the wrong person, but no one takes the parent to court for the murder. The reason, however, is that the causal culprit—the neighbor—is in a position to experience whatever retribution might result from a trial. A robot, by contrast, does not care if it is put on trial.

There is, however, a possible way around this. One might, for instance, abandon retributive talk of punishment and adopt instead a restitutive approach to crimes by machines. A restitutive approach requires that some (usually monetary) restitution be made to the victim(s), but importantly it does not require that the offender(s) be punished. Those who formulated this approach (e.g., Barnett, 1977) obviously did not have current concerns about robots in mind. Still, a restitutive approach could conceivably be used to bypass the entire question of consciousness. The person harmed by a robot or pitbull is owed something and once this debt has been paid, all is well—at least according to the account.

Although many societies have maintained stability with a restitutive approach (see Napoleon 2009, pp. 156–160), it is an open question whether restitution would be sufficient. In our estimate, such an approach would bypass consciousness but also justice. Having a human in the loop thus remains a desideratum. Of course, to the extent that one wants to prosecute—not just persecute—a dog owner for the misdeeds of a pitbull, one will need to know whether the pitbull was securely leashed at the time of the attack or whether it was recklessly released. The same goes for robots and the people who activate them. A properly functioning moral community cannot license irresponsible “press and runs,” so having a person take self-initiated prospective moral answerability for autonomous robot deeds must be a necessary condition of any activation.

Note that, as a speech act, vouching has limited powers. A person may “say” that they are responsible for whatever the Pope does tomorrow, but clearly they are not thereby causally or morally responsible, regardless of what they might proclaim. One cannot vouch for the Pope, because the Pope is already responsible for his own actions. Likewise, a parent who is a trustee for a child ceases to be responsible when the child regains control of their assets. It is a plain fact that blame can be deflected, so we needn't find mysterious the idea that a person can accept responsibility for robot-caused harm. The interesting question is whether blame is being deflected to the right target. All the responses except the blank check either miss the mark (in pretense) or pick their target unjustly (in martyrs, framings, and scapegoats).

Conclusion

Right now, when machines harm people, we blame the nearest human operator (Elish, 2019). We thus keep a human at the wheel—even if this human no longer steers the wheel. Clearly, this facile fix will cease to work once robots have different designs and become completely autonomous. We thus have to come up with better ways of assigning blame. In a bid to gain clarity and make progress, we carved the logical space in five and ranked the resulting options from worst to best, based on their pros and cons.

Pretense (metaethical fictionalism or instrumentalism) involves collectively blaming a robot and acting "as if" it cared. Because this strategy exploits social practices that are already in place and functioning (more or less), it undeniably brings comfort and resolution. Yet, pretense dodges metaphysical questions that do not vanish by not being asked. Moreover, the response fails to explain why we don't make a similar performative turn to fix other problems that would also become more tractable if we acted "as if" they didn't exist in their current form.

Framing (other-initiated prospective moral answerability) involves selecting a person without their consent and getting ready to blame that person, in the event that an autonomous robot causes harm. There are two possibilities here. If the person can be causally linked to the upcoming harm by a robot, then the robot is not truly autonomous and so we can turn to ordinary means of meeting out justice (not covered in this paper). If the person publicly made herself answerable for the robot's behavior beforehand, then we are ejected from framing and enter the more just domain of blank checks. In all other circumstances, the choice of person to be blamed becomes completely arbitrary and unjust.

Scapegoating (other-initiated retrospective moral answerability) involves selecting a person unrelated to robotic harms that have occurred and blaming that person for those harms. This seals the gap in responsibility, but unjustly, precisely because the person was unrelated to the events in question. Scapegoating thus runs afoul of the "causality condition" (Taddeo & Blanchard, 2022, p. 4) essential for moral answerability.

Martyrdom (self-initiated retrospective moral answerability) involves a person asking to be blamed for a robot's misdeeds. In a way, this is admirable. Yet, since the person never vouched for the autonomous robot beforehand, the lack of any causal connection makes it unjust for a moral community to accept the martyr's offer. Even if the strategy of martyrdom were somehow morally acceptable, it would provide no guarantee that a martyr would step up for every robot harm.

In their distinctive ways, pretense, framing, scapegoating, and martyrdom all enable "a form of psychological compensation—the very act of punishing the defendant *is* the compensation" (Lemley & Casey, 2019, p. 1385). While we agree that holding a robot or unconnected person morally answerable "may benefit the victim psychologically," it is worth asking whether there is a way "[t]o channel that instinct into other areas […] where it might be more productive" (ibid.). Emerging from our comparative analysis, the unique trade-offs provided by blank checks make them, not perfect, but the best of the bunch.

As seen at the outset, our original article warned "don't push that button, it might lead to senseless violence" and stressed that "if it does, you will be held responsible and punished" (Champagne & Tonkens, 2015, p. 136). In addition to being just, the twin conditions of self-initiation and prospective orientation render such senseless violence less likely. The conversation is not over, so we can expect further problems and refinements. Still, we think the justice and deterrence provided by blank checks are reasons enough to foreground self-initiated prospective moral answerability when dealing with autonomous robot harm.

Author Contributions MC and RT both contributed to this manuscript.

Funding No funds, grants, or other financial support were received by MC or RT during the preparation of this manuscript.

Data Availability Not applicable.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication MC and RT consent to have this manuscript published.

References

Abbott, R. (2020). *The reasonable robot: Artificial intelligence and the law*. Cambridge University Press.

Baalen, S., & Boon, M. (2019). Epistemology for interdisciplinary research – shifting philosophical paradigms of science. *European Journal for Philosophy of Science*, *9*(1), 1–28. https://doi.org/10.1007/s13194-018-0242-4.

Barnett, R. E. (1977). Restitution: A new paradigm of criminal justice. *Ethics*, *87*(4), 279–301. https://doi.org/10.1086/292043.

Behdadi, D., & Munthe, C. (2020). A normative approach to artificial moral agency. *Minds and Machines*, *30*(2), 195–218. https://doi.org/10.1007/s11023-020-09525-8.

Berber, A., & Srećković, S. (2023). When something goes wrong: Who is responsible for errors in ML decision-making? *AI & Society*. https://doi.org/10.1007/s00146-023-01640-1.

Bernáth, L. (2021). Can autonomous agents without phenomenal consciousness be morally responsible? *Philosophy & Technology*, *34*(4), 1363–1382. https://doi.org/10.1007/s13347-021-00462-7.

Bernstein, S. (2017). Causal proportions and moral responsibility. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility, volume 4* (pp. 165–182). Oxford University Press. https://doi.org/10.1093/oso/9780198805601.003.0009.

Brooks, S. K., & Greenberg, N. (2021). Psychological impact of being wrongfully accused of criminal offences: A systematic literature review. *Medicine, Science and the Law*, *61*(1), 44–54. https://doi.org/10.1177/0025802420949069.

Brożek, B., & Jakubiec, M. (2017). On the legal responsibility of autonomous machines. *Artificial Intelligence and Law*, *25*(3), 293–304. https://doi.org/10.1007/s10506-017-9207-8.

Cappuccio, M. L., Peeters, A., & McDonald, W. (2019). Sympathy for Dolores: Moral consideration for robots based on virtue and recognition. *Philosophy & Technology*, *33*(1), 9–31. https://doi.org/10.1007/s13347-019-0341-y.

Cappuccio, M. L., Sandoval, E. B., Mubin, O., Obaid, M., & Velonaki, M. (2021). Robotics aids for character building: More than just another enabling condition. *International Journal of Social Robotics*, *13*(1), 1–5. https://doi.org/10.1007/s12369-021-00756-y.

Carson, H. L. (1917). The trial of animals and insects: A little known chapter of mediæval jurisprudence. *Proceedings of the American Philosophical Society*, *56*(5), 410–415. https://ark.13960.t27b26t0z.

Cernea, M. V. (2017). The ethical troubles of future warfare: On the prohibition of autonomous weapon systems. *Annals of the University of Bucharest Philosophy Series*, *66*(2), 67–89.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Champagne, M. (2021). The mandatory ontology of robot responsibility. *Cambridge Quarterly of Healthcare Ethics*, *30*(3), 448–454. https://doi.org/10.1017/S0963180120000997.

Champagne, M., & Tonkens, R. (2015). Bridging the responsibility gap in automated warfare. *Philosophy & Technology*, *28*(1), 125–137. https://doi.org/10.1007/s13347-013-0138-3.

Chandler, D. (2018). Distributed responsibility: Moral agency in a non-linear world. In C. Ulbert, P. Finkenbusch, E. Sondermann, & T. Debiel (Eds.), *Moral agency and the politics of responsibility* (pp. 182–195). Routledge. https://doi.org/10.4324/9781315201399.

Chomanski, B. (2021). Liability for robots: Sidestepping the gaps. *Philosophy & Technology*, *34*(4), 1013–1032. https://doi.org/10.1007/s13347-021-00448-5.

Coeckelbergh, M. (2009). Virtual moral agency, virtual moral responsibility: On the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society*, *24*(2), 181–189. https://doi.org/10.1007/s00146-009-0208-3.

Coeckelbergh, M. (2019). *Moved by machines: Performance metaphors and philosophy of technology*. Routledge.

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

Coghlan, S., Vetere, F., Waycott, J., & Neves, B. B. (2019). Could social robots make us kinder or crueller to humans and animals? *International Journal of Social Robotics*, *11*(5), 741–751. https://doi.org/10.1007/s12369-019-00583-2.

Conradie, N., Kempt, H., & Königs, P. (2022). Introduction to the topical collection on AI and responsibility. *Philosophy & Technology, 35*(4), article 97. https://doi.org/10.1007/s13347-022-00583-7.

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

de Santoni, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x.

Dennett, D. C. (1987). *The intentional stance*. MIT Press.

Dennett, D. C. (1997). When HAL kills, who's to blame? Computer ethics. In D. G. Stork (Ed.), *HAL's legacy: 2001's computer as dream and reality* (pp. 351–365). MIT Press.

Dennett, D. C. (2023). The problem with counterfeit people. *The Atlantic*, May 16.

Di Nucci, E. (2018). Sexual rights, disability and sex robots. In J. Danaher, & N. McArthur (Eds.), *Robot sex: Social and ethical implications* (pp. 73–88). MIT Press.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, *5*, 40–60. https://doi.org/10.17351/ests2019.260.

Enoch, D. (2012). Being responsible, taking responsibility, and penumbral agency. In U. Heuer, & G. Lang (Eds.), *Luck, value, and commitment: Themes from the ethics of Bernard Wiliams* (pp. 95–132). Oxford University Press.

Feinberg, J. (1965). The expressive function of punishment. *The Monist*, *49*(3), 397–423. https://doi.org/10.5840/monist196549326.

Fischer, J. M., & Ravizza, M. (2000). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Fricker, M. (2016). What's the point of blame? A paradigm based explanation. *Noûs*, *50*(1), 165–183. https://doi.org/10.1111/nous.12067.

Gerdes, A. (2018). Lethal autonomous weapon systems and responsibility gaps. *Philosophy Study*, *8*(5), 231–239. https://doi.org/10.17265/2159-5313/2018.05.004.

Gless, S., Silverman, E., & Weigend, T. (2016). If robots cause harm, who is to blame? Self-driving cars and criminal liability. *New Criminal Law Review*, *19*(3), 412–436. https://doi.org/10.1525/nclr.2016.19.3.412.

Gogoshin, D. L. (2021). Robot responsibility and moral community. *Frontiers in Robotics and AI*, *8*(768092). https://doi.org/10.3389/frobt.2021.768092.

Gunkel, D. J. (2018a). The other question: Can and should robots have rights? *Ethics and Information Technology*, *20*(2), 87–99. https://doi.org/10.1007/s10676-017-9442-4.

Gunkel, D. J. (2018b). *Robot rights*. MIT Press.

Gunkel, D. J. (2020). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology*, *22*(4), 307–320. https://doi.org/10.1007/s10676-017-9442-4.

Hage, J. (2017). Theoretical foundations for the responsibility of autonomous agents. *Artificial Intelligence and Law*, *25*(3), 255–271. https://doi.org/10.1007/s10506-017-9208-7.

Hansson, S. O. (2023). Who is responsible if the car itself is driving? In D. P. Michelfelder (Ed.), *Test-driving the future: Autonomous vehicles and the ethics of technological change* (pp. 43–58). Rowman and Littlefield.

Hart, H. L. A. (2008). *Punishment and responsibility: Essays in the philosophy of law*. Oxford University Press.

Hayenhjelm, M., & Wolff, J. (2012). The moral problem of risk impositions: A survey of the literature. *European Journal of Philosophy*, *20*(S1), E26–E51. https://doi.org/10.1111/j.1468-0378.2011.00482.x.

Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology*, *16*(3), 197–206. https://doi.org/10.1007/s10676-014-9345-6.

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

James, W. (1898). Philosophical conceptions and practical results. *University Chronicle*, *1*(4), 287–310.

Joyce, R. (2001). *The myth of morality*. Cambridge University Press.

Kazman, S. (1990). Deadly overcaution: FDA's drug approval process. *Journal of Regulation and Social Costs*, *1*(1), 35–54.

Kiener, M. (2022). Can we bridge AI's responsibility gap at will? *Ethical Theory and Moral Practice*, *25*(4), 575–593. https://doi.org/10.1007/s10677-022-10313-9.

Kneer, M., & Stuart, M. T. (2021). Playing the blame game with robots. *In HRI '21 Companion: Companion of the 2021 ACM/IEEE international conference on human-robot interaction* (pp. 407–411). https://doi.org/10.1145/3434074.3447202.

Köhler, S., Roughley, N., & Sauer, H. (2018). Technologically blurred accountability? Technology, responsibility gaps and the robustness of our everyday conceptual scheme. In C. Ulbert, P. Finkenbusch, E. Sondermann, & T. Debiel (Eds.), *Moral agency and the politics of responsibility* (pp. 51–68). Routledge. https://doi.org/10.4324/9781315201399.

Kraaijeveld, S. R. (2020). Debunking (the) retribution (gap). *Science and Engineering Ethics*, *26*(3), 1315–1328. https://doi.org/10.1007/s11948-019-00148-6.

Kraaijeveld, S. R. (2021). Experimental philosophy of technology. *Philosophy & Technology*, *34*(4), 993–1012. https://doi.org/10.1007/s13347-021-00447-6.

Kühler, M. (2020). Technological moral luck. In B. Beck, & M. Kühler (Eds.), *Technology, anthropology, and dimensions of responsibility* (pp. 115–132). J. B. Metzler Verlag. https://doi.org/10.1007/978-3-476-04896-7_9.

Lemley, M. A., & Casey, B. (2019). Remedies for robots. *The University of Chicago Law Review*, *86*(5), 1311–1396. https://doi.org/10.2139/ssrn.3223621.

Lévinas, E. (1985). *Ethics and infinity: Conversations with Philippe Nemo*. Trans. R. A. Cohen. Duquesne University Press.

Lévinas, E. (1998). *Discovering existence with Husserl*. Trans. R. A. Cohen & M. B. Smith. Northwestern University Press.

Lima, G., Grgić-Hlača, N., & Cha, M. (2021). Human perceptions on moral responsibility of AI: A case study in AI-assisted bail decision-making. *Proceedings of the 2021 CHI conference on human factors in computing systems*, article 235. https://doi.org/10.1145/3411764.3445260.

Lima, G., Grgić-Hlača, N., & Cha, M. (2023). Blaming humans and machines: What shapes people's reactions to algorithmic harm. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. Association for Computing Machinery. https://doi.org/10.1145/3544548.3580953.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340.

Mamak, K. (2022). Should violence against robots be banned? *International Journal of Social Robotics*, *14*(4), 1057–1066. https://doi.org/10.1007/s12369-021-00852-z.

Mason, E. (2019). *Ways to be blameworthy: Rightness, wrongness, and responsibility*. Oxford University Press.

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

Napoleon, V. R. (2009). *Ayook: Gitksan legal order, law, and legal theory*. Doctoral dissertation, University of Victoria, Canada.

Oimann, A. (2023). The responsibility gap and LAWS: A critical mapping of the debate. *Philosophy & Technology*, *36*(5), article 5. https://doi.org/10.1007/s13347-023-00605-y.

Oldridge, D. (2005). *Strange histories: The trial of the pig, the walking dead, and other matters of fact from the medieval and renaissance worlds*. Routledge.

Parfit, D. (1984). *Reasons and persons*. Clarendon Press.

Restivo, S. (2017). *Sociology, science, and the end of philosophy: How society shapes brains, gods, maths, and logics*. Palgrave Macmillan.

Royakkers, L., & Olsthoorn, P. (2018). Lethal military robots: Who is responsible when things go wrong? In R. Luppicini (Ed.), *The changing scope of technoethics in contemporary society* (pp. 106–123). IGI Global. https://doi.org/10.4018/978-1-5225-5094-5.ch006.

Sætra, H. S. (2021). Challenging the neo-anthropocentric relational approach to robot rights. *Frontiers in Robotics and AI*, *8*(744426). https://doi.org/10.3389/frobt.2021.744426.

Sartorio, C. (2007). Causation and responsibility. *Philosophy Compass*, *2*(5), 749–765. https://doi.org/10.1111/j.1747-9991.2007.00097.x.

Scanlon, T. M. (2008). *Moral dimensions: Permissibility, meaning, blame*. Harvard University Press.

Shoemaker, D. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics*, *121*(3), 603–632. https://doi.org/10.1086/659003.

Smith, A. M. (2007). On being responsible and holding responsible. *The Journal of Ethics*, *11*(4), 465–484. https://doi.org/10.1007/s10892-005-7989-5.

Smith, N., & Vickers, D. (2021). Statistically responsible artificial intelligences. *Ethics and Information Technology*, *23*(3), 483–493. https://doi.org/10.1007/s10676-021-09591-1.

Søvik, A. O. (2022). How a non-conscious robot could be an agent with capacity for morally responsible behaviour. *AI and Ethics*, *2*(4), 789–800. https://doi.org/10.1007/s43681-022-00140-0.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

Sparrow, R. (2017). Robots, rape, and representation. *International Journal of Social Robotics*, *9*(4), 465–477. https://doi.org/10.1007/s12369-017-0413-z.

Sparrow, R. (2021). Virtue and vice in our relationships with robots: Is there an asymmetry and how might it be explained? *International Journal of Social Robotics*, *13*(1), 23–29. https://doi.org/10.1007/s12369-020-00631-2.

Stenseke, J. (2022a). Interdisciplinary confusion and resolution in the context of moral machines. *Science and Engineering Ethics*, *28*(3), 1–17. https://doi.org/10.1007/s11948-022-00378-1.

Stenseke, J. (2022b). The morality of artificial friends in Ishiguro's *Klara and the Sun*. *Journal of Science Fiction and Philosophy*, *5*, 1–18.

Strawson, P. F. (2008). *Freedom and resentment and other essays*. Routledge.

Stuart, M. T., & Kneer, M. (2021). Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents. *Proceedings of the Association for Computing Machinery Conference on Human-Computer Interaction*, 5(CSCW2), article 363. https://doi.org/10.1145/3479507.

Taddeo, M., & Blanchard, A. (2022). Accepting moral responsibility for the actions of autonomous weapons systems—a moral gambit. *Philosophy & Technology*, *35*(3), 1–24. https://doi.org/10.1007/s13347-022-00571-x.

Theodorou, A., & Dignum, V. (2020). Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence*, *2*(1), 10–12. https://doi.org/10.1038/s42256-019-0136-y.

Tigard, D. (2021a). Artificial moral responsibility: How we can and cannot hold machines responsible. *Cambridge Quarterly of Healthcare Ethics*, *30*(3), 435–447. https://doi.org/10.1017/S0963180120000985.

Tigard, D. (2021b). There is no techno-responsibility gap. *Philosophy & Technology*, *34*(3), 589–607. https://doi.org/10.1007/s13347-020-00414-7.

Tollon, F. (2021). The artificial view: Toward a non-anthropocentric account of moral patiency. *Ethics and Information Technology*, *23*(2), 147–155. https://doi.org/10.1007/s10676-020-09540-4.

Turner, J. (2018). *Robot rules: Regulating Artificial Intelligence*. Palgrave Macmillan.

van de Poel, I., Royakkers, L., & Zwart, S. D. (2015). *Moral responsibility and the problem of many hands*. Routledge.

Watson, G. (2004). *Agency and answerability: Selected essays*. Oxford University Press.

Wolf, S. (2001). The moral of moral luck. *Philosophical Exchange*, *31*(1), 5–19. http://hdl.handle.net/20.500.12648/3203.

Wolf, S. (2011). Blame, Italian style. In R. J. Wallace, R. Kumar, & S. Freeman (Eds.), *Reasons and recognition: Essays on the philosophy of T. M. Scanlon* (pp. 332–347). Oxford University Press.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.