The point of blaming AI systems

Forthcoming in *Journal of Ethics and Social Philosophy*

Hannah Altehenger (Konstanz), Leonhard Menges (Salzburg)

Abstract: As Christian List (2021) has recently argued, the increasing arrival of powerful AI systems that operate autonomously in high-stakes contexts creates a need for "future-proofing" our regulatory frameworks, i.e., for reassessing them in the face of these developments. One core part of our regulatory frameworks that dominates our everyday moral interactions is *blame*. Therefore, "future-proofing" our extant regulatory frameworks in the face of the increasing arrival of powerful AI systems requires, among others things, that we ask whether it makes sense to *extend* our blaming practices to these systems. In the paper, we argue for the admittedly surprising thesis that this question should be answered in the affirmative: contrary to what one might initially think, it can make a lot of sense to blame AI systems, since, as we furthermore argue, many of the important functions that are fulfilled by blaming humans can also be served by blaming AI systems. The paper concludes that this result gives us a good *pro tanto* reason to extend our blame practices to AI systems.

Keywords: AI, blame, moral responsibility, responsibility gap

1. Introduction

One key feature of both our present age and the decades to come is that we face the increasing arrival of powerful AI in many important domains of our lives. Many authors have argued that this raises new and deep ethical challenges (for overviews see Noorman 2020; Müller 2020). One of the philosophically most interesting is, as Christian List has recently put it, that "we may have to adjust some of our conventional anthropocentric approaches to morality" (List 2021,

1215). Or, in other words, the arrival of powerful AI suggests "that our moral theories and regulatory frameworks should be 'future-proofed'" (List 2021, 1240), i.e., reassessed in the face of these developments.

One core part of our regulatory frameworks that can be found almost universally across human societies are our practices of praise and blame (see Sommers 2012). Praising others for (what we perceive to be) commendable behavior and blaming them for (what we perceive as) transgressions is one of the key forms that our "regulatory interactions" can take.

Hence, one important part of "future-proofing" our extant regulatory frameworks in the face of the increasing arrival of powerful AI is to ask whether it makes sense to *extend* these practices and, in particular, our practice of *blame* to these systems.[1] This is the question that we shall focus on in this paper.

Our main claim is that, contrary to what one might initially think, this question should be answered in the affirmative, i.e., we shall argue that it can make sense to blame AI systems. More specifically, we shall defend the claim that we have a *pro tanto* reason to extend our blaming practices to these systems.

To support this claim, we shall proceed as follows: in the next section (sec. 2), we will present in more detail the claim that the increasing presence of AI systems creates a need for future-proofing our regulatory practices. We contend that future-proofing blame is one key element in such an endeavor that List himself has overlooked, and we also clarify how our paper relates to the so-called responsibility gap debate, which has recently received much attention in AI ethics (e.g., Matthias 2004; Sparrow 2007; Himmelreich 2019; Köhler 2020; Nyholm 2020, chap. 3; Danaher 2022). In the main part of the paper (sec. 3), we first discuss how to proceed to answer the question of whether it makes sense to extend our blaming practices to AI systems. We propose that this issue shall be settled by focusing on the *functions* that these practices fulfill. We then argue that our blaming practices can fulfill several valuable

[1] Like many other philosophical works on "regulatory practices", we shall focus on blame rather than praise.

functions when targeting AI systems, which suggests that we have at least a *pro tanto* reason to extend those practices to these systems. Before concluding, we will discuss how the issue of whether it makes sense to blame AI systems relates to the issue of whether AI systems can be blameworthy (sec. 4).

2. Preliminaries

The claim that the increasing arrival of AI gives rise to deep ethical challenges is a commonplace. Things become more interesting, though, once we ask why *exactly* the ethical challenges raised by AI systems seem to be of a more fundamental nature than, say, the challenges raised by the increasing reliance on "traditional" machines since the Industrial Revolution. Here is what we take to be the most convincing answer to these questions: unlike the machines that arrived on the scene during the Industrial Revolution, we now face the increasing arrival of systems that have the ability to (i) operate relatively autonomously in largely uncontrolled environments and (ii) make "high-stakes" decisions (List 2021, 1218; see also, e.g., Müller 2020, sec. 1.2; Nyholm 2020, chap. 2). List illustrates this point in the following passage:

> If a system has only limited capacities, such as a robotic floor cleaner or a pre-programmed factory robot, or if its use has no serious spill-over effects beyond a restricted environment, as in the case of an automated train in a tunnel, then it does not give rise to qualitatively novel risks, compared to earlier technologies. (…) By contrast, if an AI system operates relatively freely in a largely uncontrolled environment, as in the case of a driverless car or a fully autonomous drone, or if it can make high-stakes decisions on its own, as in the case of some medical, financial, and military systems, then the societal implications are qualitatively novel. *We are then dealing with artefacts as genuine decision-makers, perhaps for the first time in human history*. (List 2021, 1218 our emphasis)

If the development of novel AI systems were restricted to sophisticated vending machines or systems that can autonomously assemble IKEA furniture, then few of us would feel that the ethical challenges these systems raise were qualitatively novel. But the development of AI systems also includes entities like driverless cars, autonomous air vehicles, medical helper AI systems, diagnostic devices, and financial trading systems. Unlike the

former, these systems all operate in “high-stakes contexts”, where the occurrence of some amount of serious harm seems inevitable. However, due to their increasingly autonomous mode of functioning, it will be increasingly difficult, if not impossible, to hold some human being responsible for that harm.

According to List, all of this “suggests that our moral theories and regulatory frameworks should be ‘future-proofed’” (2021, 1240), i.e., reassessed in the face of these developments. List also provides a sketch of how AI systems could be held responsible for the harm they cause:

> The proposed form of AI responsibility may, in turn, have to be underwritten by certain assets, financial guarantees, and/or insurance, so that, in the event of a harm, the system or its legal representatives can be made to pay appropriate fines and compensation. (List 2021, 1230)

The passage just quoted arguably captures *some* of our practices of holding each other responsible. However, imposing fines and demanding compensation for perceived transgressions clearly does not exhaust these practices. Another crucial practice that seems to dominate our everyday moral interactions and that List’s account of holding AI responsible omits is *blame*. Hence, future-proofing our responsibility practices in a *comprehensive* way would also require reassessing our blaming practices, and, more specifically, asking whether it makes sense to extend these practices to AI systems. It is this task that our paper focuses on.

However, before moving on to this task, two clarifications are in order. First, we need to clarify what kind of AI systems we are interested in. Secondly, we need to explain how our main concern relates to the so-called responsibility gap debate.

Regarding the first issue, we are merely interested in those AI systems that qualify as intentional agents in a minimal sense of the term. Following List, we shall assume that minimal intentional agency requires

(i) “representational states (which encode an entity’s ‘beliefs’ about how things are)”
(ii) “motivational states (which encode its ‘desires’ or ‘goals’ as to how it would like things to be)” and, finally,

(iii) "a capacity to interact with its environment on the basis of these states so as to 'act' in pursuit of its desires or goals in line with its beliefs." (List 2021, 1219)

We shall furthermore assume that many already-existing and even more near-future AI systems meet the conditions for minimal intentional agency.[2]

Some may object that no further discussion is needed, once this assumption is in place: (minimal) intentional agency, the objection goes, is sufficient, both for blameworthiness and for its making sense to be the target of blame.[3]

We have two replies to this objection. One says that there are many entities which fulfill the above conditions for minimal intentional agency but which are such that, intuitively, it seems to be an open question whether they fulfill the conditions for blameworthiness or whether blaming them makes sense. Toddlers, people with severe cognitive disabilities, psychopaths, as well as wild non-human animals, qualify as minimal intentional agents (given the above understanding of minimal intentional agency). Intuitively, however, it seems at least to be an open question whether they satisfy the conditions for blameworthiness and whether blaming them makes sense.

Our second reply is that the distinction between minimal intentional agency on the one hand and the kind of agency that is necessary for blameworthiness or for its making sense to be the target of blame on the other is not only intuitive; it is also one that is commonly made in different philosophical debates. Authors who are skeptical about blameworthiness or the justifiability of blame, for example, are, typically, not skeptical about (minimal) intentional agency. Consider Derk Pereboom's (2014) skepticism about a specific kind of blameworthiness—what he calls blameworthiness in the "basic desert sense". Pereboom

---

[2] Let us forestall a possible misunderstanding: in presupposing that many already-existing and even more near-future AI systems have representational states and motivational states, it may seem that we have made a highly contested assumption, namely, that many current and even more near-future AI systems "have minds". But this way of putting the matter is misleading. To be sure, the claim that many existing and near-future AI systems have belief- and desire-like states seems to entail that they have minds *in a minimal sense*. However, this should not be confused with the claim that such systems can have full-fledged, human-level minds, complete with phenomenally conscious states, the capacity for self-consciousness, verbal abilities, emotions, and a rich network of diverse propositional attitudes. That many or, indeed, any existing and near-future AI systems have minds of this kind is *not* what we are presupposing. Many thanks to Peter Schulte for helpful advice on this point.

[3] Many thanks to an anonymous reviewer for urging us to address this objection.

argues that luck or determinism undermine the sort of agency that is necessary for this kind of blameworthiness. But he does not argue that these factors undermine (minimal) intentional agency. Similarly, many authors in AI ethics in general and the responsibility gap debate in particular share our assumption that the relevant AI systems, i.e., those systems that are claimed to generate responsibility gaps, are intentional or, as it is also sometimes put, 'autonomous' agents in a minimal sense of these terms (see, e.g., Sparrow 2007, 65, 74; Danaher 2016, 301; Nyholm 2018, 1207–9; Burri 2018, 165–66; Himmelreich 2019, 734; Köhler 2020, 3124; Königs 2022, 36). Those authors assume or argue that AI systems are agents in some minimal sense and contend that it is, nonetheless, inappropriate or even impossible to blame them when they cause unjustified harm.

The considerations offered in the preceding should be enough to show that the above objection fails: even if one assumes that an entity satisfies the conditions for minimal intentional agency, it is still an interesting, open question whether it satisfies the conditions for blameworthiness or whether blaming it makes sense.

Let us turn next to our second clarification, namely, how our paper relates to the responsibility gap debate. We shall be primarily concerned with the issue of whether it makes sense *to blame* AI systems rather than with the issue of whether AI systems can be *blameworthy*. We would like to emphasize that these are distinct questions. For it could turn out that AI systems can be blameworthy, but it does not make sense to blame them, and it could also turn out that it makes sense to blame AI systems even if they cannot be blameworthy.[4] Many authors in the responsibility gap debate ask who, if anyone, can be *blameworthy (responsible)* if an AI system causes some unjustified harm (e.g., Matthias 2004; Sparrow 2007; Himmelreich 2019; Köhler 2020; Nyholm 2020, chap. 3; Kiener 2022). The focus of our paper will thus be different from theirs. However, some authors within this debate are (also) concerned with the question of whether we can *blame* AI systems.[5] In particular,

[4] We shall expand on the sense of "making sense" that is at issue here in the next section. Moreover, we will take up the issue of AI *blameworthiness* again in section 4.
[5] Many thanks to an anonymous reviewer for bringing this point to our attention.

these theorists have argued that blaming AI systems is *not possible*. An argument to this conclusion says, roughly, that blaming is a form of harming and that it is impossible to harm AI systems (see Solum 1992, 1245–46; Sparrow 2007; Danaher 2016). We will discuss this particular line of thinking in section 3.1. In general, though, the remainder of this paper should make clear that we disagree with the claim that it is impossible to blame AI systems and the considerations that we shall offer in the next section can be read as an argument against this view.

3. How Blaming AI Systems Makes Sense

To a first approximation, *blame* can be characterized as "a reaction to something of negative normative significance about someone or their behavior" (Tognazzini and Coates 2018, sec. Introduction). There are many controversies surrounding the exact nature of blame (for overviews, see Coates and Tognazzini 2012; 2013; Tognazzini and Coates 2018; Smith 2022; Menges forthcoming). However, for the purposes of this paper, it will be best to stay neutral on this issue. Together with many theorists working on blame, we shall assume that manifestations of blame can be quite diverse. Among other things, they can take the form of openly expressed anger, unexpressed feelings of resentment or even seemingly dispassionate acts of relationship-modification (e.g., calmly de-friending someone on one's social media account) (see, e.g., Smith 2022, sec. 2).

With this minimal understanding of blame in place, let us ask next how we should proceed in order to settle the issue of whether it makes sense to extend our blaming practices to AI systems. We propose that the best answer to this question is to focus on blame's *functions*. Or, somewhat more precisely, proceeding from the assumption (to be substantiated in a moment) that our blaming practices have several valuable functions, we put forward the following suggestion: to decide whether it makes sense to extend our blaming practices to AI systems, we should ask whether these practices can still fulfill enough of their valuable functions when targeting AI systems.

Our suggestion relies on two background assumptions which, however, seem very plausible (as we shall argue next). The first is as follows:

(1) Our blaming practices fulfill several valuable functions.[6]

As mentioned previously, there is much controversy about the exact nature of blame. However, most theorists seem to agree that blame has certain valuable functions or, as it is more commonly expressed, "has a point" (see Watson 1987, 230; see also Macnamara 2015a, 219; Fricker 2016; Wang 2021). We shall elaborate on what these functions are in the remainder of this section. For now, we merely want to stress that the assumption that our blaming practices fulfill certain valuable functions seems to be widely shared among theorists working on blame.[7] (Note that if blame possessed no valuable functions, it would be hard to understand why so many philosophers try to show that blaming people can be appropriate even if determinism is true—if it "had no point", then everybody should be happy to get rid of it.)

Our second background assumption can be put as follows:

> (2) If our blaming practices would still fulfill their valuable functions in targeting entities of type x (or, at least, enough of these functions for them to still "have a point"), then we have a *pro tanto* reason to extend these practices to entities of type x.

Claim (2) seems very intuitive, at least assuming that one does not read into it something stronger than it says. Claim (2) does *not* say that we ought, all things considered, to extend our blaming practices to entities of type x, if, in targeting entities of type x, our blaming practices would fulfill (enough of their) valuable functions.[8] Nor does it say that we would have sufficient reason to do so. Instead, claim (2) makes a much more modest claim, namely, that, in this

---

[6] To clarify, we use the term "function" in a minimal sense of "what a thing does" and, consequently, the term "valuable functions" in the sense of "the positive effects a thing has." Or, to put the same point in a slightly different and somewhat colloquial manner: what we are interested in when we talk about the "valuable functions" of our blaming practices are the "cool things that blame does for us." We are grateful to Sebastian Köhler for urging us to be clearer on this point and for suggesting that we express this point in this manner.

[7] Note that the assumption that blame fulfills certain (valuable) functions is independent from the claim that blame can ultimately only be defined in terms of its functions (this is, roughly, the view of McKenna 2013; Fricker 2016; Shoemaker and Vargas 2021). One can accept the former assumption, while rejecting the latter.

[8] Here and in the following we use the expression "enough of their valuable functions" as a shorthand for "enough valuable functions for our blaming practices to still 'have a point'".

case, we would have a *pro tanto* reason to extend these practices to entities of type x (which then may or may not be outweighed by other reasons against such an extension).

In the following, we shall argue that our blaming practices would fulfill several valuable functions when targeting AI systems (and clearly enough of their valuable functions to still "have a point") and that we, therefore, have at least a *pro tanto* reason to extend them to these systems.

3.1 Retribution

It may seem natural to claim that one valuable function of our blaming practices is *retribution*, i.e., that one valuable feature of these practices is that they help ensure that the guilty "get what they deserve".

Could appealing to this function support the claim that our blaming practices would fulfill valuable functions in targeting AI systems? We are skeptical about this, for two reasons.[9]

First, we are skeptical about the idea that the retribution function is a *valuable* function. Our skepticism is motivated by a general anti-retributivist stance, i.e., we would reject the idea *that there is something (non-instrumentally) good in a guilty party's being harmed*, which is at the very core of retributivist thinking (for an overview, see Walen 2021).

Secondly, there is reason to doubt that the retributive function could still *be fulfilled* if the blamee was an AI system (see also Sparrow 2007, 71–73; Danaher 2016). After all, in order for this function to be fulfilled, it is necessary that a blaming response can in some way be *harmful* for the target, since, as was just mentioned, the idea that there is something good about a guilty party's *being harmed* is at the very core of retributivist thinking. Now, there is no difficulty seeing how a blaming response can be harmful if the target is a human being: few of us like to be blamed by others. Indeed, it often *feels quite uncomfortable*, if not *somewhat painful* to be the recipient of blame. But it is much more difficult to see how blame could harm

[9] A view that may be somewhat similar to ours is expressed by Gogoshin (2021, 9).

AI systems. There is a complicated debate about the nature of harm, but it seems plausible that for something to be harmful, it must at least do one of the following: cause bad (painful) experiences, frustrate desire, set back some interest, or diminish an agent's quality of life. First, however, it is difficult to see how blaming responses should lead AI systems to have painful experiences, since these systems plausibly lack phenomenal consciousness (at least those that are currently around and that will be around in the near future).[10] Second, while we are very sympathetic to the assumption that AI systems can have desires,[11] it is difficult to see how blame, as a general matter of fact, should frustrate these desires: while it does seem plausible that the vast majority of human beings has some desire(s) which are frustrated by instances of blame, making the same assumption about AI systems would seem to require a fair amount of undue anthropomorphizing. Third, it is far from clear what it means to say that AI systems have interests or a quality of life. In view of all this, it is considerably difficult to see how our blaming responses would still retain their harmful character in targeting AI systems and, consequently, how they could still fulfill their retributive function.[12]

It would be too hasty to conclude from this, though, that we have no reason to extend our blaming practices to AI systems. This is because, as the remainder of the paper will show, prospects look much brighter once we turn to further (valuable) functions of these practices.

3.2 Modification of behavior

While the retributive function is essentially backward-looking, there is a further important function of blame that is essentially forward-looking, namely, modifying the future behavior of the blamee (see, e.g., McGeer 2013, sec. 2.3).

---

[10] For an argument in support of this claim, see, e.g., the reasoning put forward by List (2021, 1237–38).

[11] In fact, we defend this view in our unpublished manuscript "How Robots Can Be Blameworthy" (co-authored with Peter Schulte).

[12] The reasoning that we have just offered is admittedly sketchy. Hence, we do not claim to have shown that it is impossible that blame's retributive function can be fulfilled when the blamee is an AI system. The point we wish to make is a weaker one: at least for those AI systems that are currently around and that will be around in the foreseeable future, it seems much more plausible to assume that this function *cannot* be fulfilled than to assume that it can.

In order for blame to fulfill its behavior-modification function when targeting an AI system, the latter would obviously have to possess some kind of feedback mechanism. More specifically, the system would have to be able to recognize instances of blame as such and to process them in a way that would eventually lead to behavior modification. In principle, this may happen in two ways: the first way is "classic reprogramming". Imagine that, once an AI system has "registered" a number of blaming responses directed at it, it sends a corresponding signal, which then leads to re-programming, i.e., a human supervisor assesses these responses and, if judged appropriate, makes some fitting alterations to the system's priorities. The second way is autonomous machine learning. Imagine that after a training phase with a sufficiently large "blame database", an AI system uses further instances of blame directed at it to itself update its database with desirable responses. We are not the first to maintain that autonomous machine learning may one day lead to "blame-sensitive" AI. In particular, Dane Gogoshin (2020; 2021) and Daniel Tigard (2021a) have recently contended that relevant reinforcement learning mechanisms may allow for the construction of AI systems which can modify their behavior in reaction to our blaming responses.[13]

There are obviously some pros and cons to both approaches and some significant technical challenges to overcome in order to implement them. However, we would like to stress, in line with the aforementioned treatments of the matter, that there do not seem to be any in-principle obstacles here. Registering instances of blame and treating them as a source of feedback ultimately just amounts to a form of learning. Hence, on the plausible assumption that learning in AI systems is possible, and that further substantial progress will be made in that domain in the coming decades, it seems plausible that, at some future point at least, AI systems can be construed that can use our blame responses as a source for learning. And once this point will be reached, there do not seem to be any obstacles to the fulfillment of blame's behavior-modification function.

[13] Both Gogoshin and Tigard in turn draw on Wallach and Allan's (2009) work on artificial moral cognition.

Interestingly, there are even respects in which the fulfillment of this function may be *easier* if the blamee is an AI system rather than a human being: first, unlike in the case of human beings, the fulfillment of blame's behavior-modification function can't be thwarted by episodes of *akrasia*. Once a relevant episode of learning has been completed, the system will adapt its overt behavior accordingly. Second, humans sometimes respond to being blamed in destructive ways such as counter-blaming or playing the "blame game", seeking fault elsewhere, and so on (see, e.g., Pettigrove 2012; Pereboom 2021, chap. 1). A well-programmed AI can avoid these responses.

Suppose, though, that our assessment in this section was overly optimistic and that, contrary to what we've just claimed, it is unlikely that blame can fulfill its behavior-modification function when targeting AI systems (because no or only very few AI systems will ever possess the relevant learning mechanisms). Would this mean that extending our blaming practices to AI systems would be pointless? In the remaining sections, we will argue that this would not follow. As we will show, our blaming practices have several additional valuable functions, some of which can be fulfilled surprisingly well when the blamee is an AI system.

3.3 Conversation

As several theorists have stressed, blame seems to possess another important function which may be somewhat less obvious than the retribution- and behavior-modification function. This is the function of initiating or sustaining conversations about the negative normative or evaluative status of what happened—henceforth referred to as "normative conversations" (see, e.g., Watson 1987; McKenna 2013; McGeer 2013; Macnamara 2015b; Mason 2019, chap. 5; Wang 2021; for a similar point, see also Tigard 2021b). This function can be fulfilled by open statements, but also by less explicit forms of communication (e.g., a raised eyebrow can also start a normative conversation).

A normative conversation initiated or sustained by an instance of blame can be valuable in many respects. It can give the targets of blame reasons to act differently in the future and help them to further develop their ability to respond to relevant reasons (e.g., Vargas 2013;

McGeer 2019). It provides an opportunity for targets of blame to explain or even justify what they did, to learn about how we perceive their conduct, and to ask for forgiveness (e.g., McKenna 2013; Fricker 2016). These are important processes because we need a peaceful way to deal with the "normative ruptures" in our social webs. For instance, when we directly blame a friend for telling a mean joke about us, we start a conversation with her about what she did. We communicate that we found her behavior unacceptable and, thereby, start an exchange of our views about the reasons and values that are at issue. Ideally, she will ask for forgiveness and, thereby, try to restore our friendship.

Can blame fulfill the function of initiating or sustaining a conversation about the negative normative or evaluative status of what happened when the blamee is an AI system? Regarding current AI systems, this seems implausible. A key worry regarding these systems is that not even their designers are able to understand why they come to a certain conclusion and not to a different one (see, e.g., Müller 2020, sec. 2.3). In that case, having a normative conversation is impossible. We cannot converse with someone about the normative status of what they did who is unable to explain, much less justify, what they did.[14]

This situation may change in the future. A lot of energy is currently being put into theorizing about and engineering so-called transparent or explainable AI (XAI) (Floridi et al. 2018; Bathaee 2018; Langer et al. 2021; Baum et al. 2022; for a critical discussion of the need for XAI in the medical sector see London 2019).[15] In a nutshell, the idea is to build AI systems that allow the users, engineers, regulators, and so on to understand how and why the system comes to a certain decision or proposal.[16] Now, an XAI system in this sense is not yet a system

[14] Some may object that recent successes of large language models like ChatGPT show that normative conversations between humans and AI systems are already happening. First, however, these systems, too, cannot explain or justify how they came to their decisions. Second, it seems unclear whether they can ask for forgiveness and be forgiven. However, insofar as these are key aspects of normative conversations, there is reason to doubt that such conversations between humans and current AI systems are already possible.

[15] In this context, see also Daniel Tigard (2021b)'s recent suggestion that we should design (what he calls) technologically-answerable systems, i.e., systems which have the ability to provide their users with answers as to why a certain behavioral output occurred.

[16] One way to achieve this is to equip the AI system with an "ethical black box" analogous to a flight data recorder that records its decision-making process (see, e.g., Winfield and Jirotka 2017).

with which one can have the same kind of normative conversation that we know from our direct interactions with human wrongdoers. That the system can make us understand why and how it comes to a decision does not yet guarantee that it understands us when we challenge its decisions, that it learns from our blame, that it asks for forgiveness, and so on. Perhaps such a fully "conversable" AI system can be engineered (List 2021, 1228–32 is optimistic about this). But independently of this, we would like to offer the following novel line of reasoning: even if the prospect of conversable AI does not turn out to be realistic and even if there will never be a fully transparent AI system, there would still be an important sense in which blame can fulfill its function of initiating and sustaining a normative conversation when the blamee is an AI system.

Our starting point is the observation that, in everyday life, we often initiate or sustain a conversation about the negative normative or evaluative status of what people did who can neither explain nor justify their conduct, for example when we discuss our histories. In our communities, it is important for us to converse with each other about the wrongdoings of, for example, American slaveholders or German Nazis, despite the fact that the transgressors, given that they are no longer living, are unable to explain or justify their behavior or to ask for forgiveness.[17] The value of these conversations cannot be that it helps the transgressors develop their rational abilities, change their behavior, or understand what we think about what they did. Rather, the value of these conversations lies in *helping us today*. That is, these conversations help us to develop our capacity to respond to relevant moral reasons, to not do what these transgressors did, and to understand how the world perceives their conduct. These are important issues. To converse about the normative or evaluative status of what certain transgressors did thus plays important roles even if these transgressors cannot be part of the conversation.

[17] A parallel argument could be run for human agents whose psychological make-up is such that playing a constructive part in a normative conversation is very difficult (if not impossible) for them, e.g., agents with narcissistic personality disorder.

The same can be true when the blamee is an AI system. Even if we cannot converse with a self-driving car that prioritizes driving its customers home quickly over the safety of pedestrians, we can converse with each other about the normative or evaluative status of what the car does. This can play important roles in developing our normative reasoning abilities, changing future conduct, and sharing how we perceive the normative and evaluative world.

Thus, regardless of whether XAI will ever be fully realized and even if AI systems never achieve the status of conversable entities, there still is a sense in which blame can fulfill its valuable function of initiating and sustaining normative conversations when the blamee is an AI system.

3.4 Protest

As several theorists have argued, another important function of blame is to enable a specific form of *moral protest* (e.g., Talbert 2012; Smith 2013; Pereboom 2021, chap. 2). The core idea here is that, by blaming another party, we can "stand up for [ourselves]" (or others) and "put something important on record" (Talbert 2012, 106), namely, roughly speaking, that the way the other party has treated us (or the third person we are standing up for) was not okay. Or, as Smith has put it, one key aim (or function) of our blaming responses is to "*register* the fact that the person wronged did not deserve such treatment" (Smith 2013, 43) and "to prompt moral recognition and acknowledgment of this fact on the part of the wrongdoer and/or others in the moral community" (Smith 2013, 43).

The latter qualification is important since it highlights the fact that the protest function of blame can be fulfilled even if it is unlikely, or even impossible, to gain moral recognition from the transgressor herself and, we may add, even if the transgressor is unlikely, or even unable, to modify her behavior in response to our blame. Indeed, according to Matt Talbert, "such protest is *meant largely for the protester and for his fellow sufferers*" (p. 107, our emphasis) and its "intelligibility depends [not] on whether anyone will be converted to a better moral point of view" (Talbert 2012, 107). By protesting, we make it clear to us and those around us that we are standing up for something. It is not necessary that the party whose conduct we protest

does or can understand our protest, or respond to it, or reform their behavior in reaction to it. We can protest the behavior of a cruel dictator who will never learn about our protest just as we can protest against what the American slaveholders or German Nazis did even if they are long dead. Or, as we may also put it, the protest function of blame is more about the protesters and those who learn about the protest than about the party whose conduct we protest.

In view of this, it seems very plausible that the protest function of blame could be fulfilled if the blamee is an AI system. For illustration, let us take up the case of the self-driving car again, which prioritizes driving customers home quickly over the safety of pedestrians. Such a hierarchy of goals is objectionable and the behavior that expresses it can thus be an appropriate target of protest. As pedestrians, it makes complete sense to stand up for our safety and make clear that the goal structure that manifests itself in the car's conduct is unacceptable. We, thereby, show to ourselves and those around us that our safety matters to us. Whether or not the car can understand our protest, or modify its behavior in reaction to our protest, is irrelevant for whether it makes sense to protest.

The protest function thus seems to be a clear example of an important function that our blaming practices can still fulfill when the blamee is an AI system.

3.5 Signaling

The same holds true for what has recently been argued to be another important function of blame, namely, *to signal one's commitment to certain norms and values* (Shoemaker and Vargas 2021), or, more specifically, to signal that one is "a member of a particular moral tribe, someone who cares about a set of norms and their breaches, someone who is disposed to police the norms, and more" (Shoemaker and Vargas 2021, 587).

For illustration, imagine you witness your colleague telling a racist joke about another colleague.[18] In responding to this with blame (e.g., by telling the joke-teller angrily that their

[18] The following is inspired by Shoemaker and Vargas' discussion of the case of Sarah (2021, 589–90).

joke is inappropriate and deeply hurtful to the victim), one is sending the signal that one is committed to the norm that racist behavior is not okay. Importantly, one is not merely sending this signal to the blamee, but also to bystanders as well as to the victim. To the latter, one is also sending a signal of solidarity ("I know that what x is saying is wrong and I've got your back!"). Finally, one is sending information about one's "agential qualities", i.e., roughly speaking, about one's character or regard for others. Thus, a single blaming response may send "many different signals far and wide" (Shoemaker and Vargas 2021, 590) and hence fulfill its signaling function *through many different channels*.

The latter point is important because it suggests that there can be instances of blame which are, again, more about the blamer and those who witness the blaming response than about the blamee. This point is also highlighted by Shoemaker and Vargas:

> Given its multichannel nature, in some cases blame's signal may even *exclude the blamed agent altogether*. This is a significant and underappreciated point, for it makes clear just how distinct blame may be from harsh treatment, sanctions, and punishment of the blamed agent. In such cases of 'gossipy' blaming, the blamed agent is oftentimes beside the point. Yet the moral signal can remain crucial for the reputation of the blamer and an important data point for social cooperation. (Shoemaker and Vargas 2021, 590 emphasis in original)

To briefly expand on the last point, note that blaming responses are often *quasi-automatic* reactions to perceived breaches of norms and, in view of their quasi-automaticity, difficult to fake. Hence, there is a high likelihood that observers of a blaming response will be able to gather accurate information from it, making such responses indeed "an important data point for social cooperation" (Shoemaker and Vargas 2021, 590).

If we apply the considerations detailed in the preceding to the question we're interested in, namely, whether the signaling function can be fulfilled when the blamee is an AI system, we arrive at the same affirmative answer as we did in the case of the protest function and the conversation function—and for parallel reasons. For illustration, take, again, our example of the self-driving car. When we blame the car for prioritizing driving its customers quickly to their destination over the safety of pedestrians, we signal that we are committed to certain moral norms (e.g., about the importance of not putting other people's lives at risk for trivial reasons).

This in turn allows others who observe our response to gather valuable information about our normative stance towards certain types of traffic behavior, about how we would behave in traffic, and, more generally, about certain general agential qualities we possess. For instance, our caring about the safety of pedestrians shows that we possess some amount of regard for our fellow human beings (at least if we additionally assume that the car's conduct presents no immediate danger to ourselves). And just as before, the signaling function can be fulfilled in this case, *even if* we assume that the target itself does not understand our signaling nor modifies its behavior in response to it. This is because the signaling function, just like the protest and conversation function, can be more about the blamer and those who witness the blaming response than about the blamee and, due to its "multi-channel nature", can be fulfilled even if the channel from blamer to blamee is "closed".[19]

The signaling function is thus another example of an important function of blame that could be fulfilled when the blamee is an AI system. On a final note, we believe that this function might even become increasingly important to us (i) the more AI systems become part of our daily social interactions and (ii) the more such systems perform activities that we could also perform ourselves (such as driving cars, waiting tables, taking care of the elderly, etc.). After all, assuming that we will increasingly face situations in which AI systems display problematic conduct in the course of performing actions that *we* could also perform, the following further assumption seems plausible, too: we will increasingly feel the need to signal our commitment to certain norms and values in order to reassure each other that we belong to the same "moral tribe" and our solidarity with potential victims.[20]

---

[19] Some readers may still feel uncomfortable with the idea that our blaming practices can be more about the blamer and those who witness an instance of blame than about the blamee. Here is a further consideration in support of this point: even when we focus exclusively on instances of blame where all parties involved are human beings, so-called dyadic cases of blame, where the victim of a transgression overtly blames the transgressor face to face, "are actually not all that frequent" (Shoemaker and Vargas 2021, 590). While they certainly occur, they seem to be far outnumbered by non-dyadic cases and, more specifically, cases in which *we blame a transgressor to others in the absence of the transgressor*.

[20] To illustrate this point with a concrete example, take the (imagined) case of a waiter robot that prioritizes serving customers with white skin over customers with a different skin color. On witnessing this, many of us would presumably feel the need to signal our commitment to the norm that racist behavior is not okay, as well as our solidarity with potential victims.

3.6 Relationship Management

Tim Scanlon has argued that blame should be understood in terms of relationship modification. According to him, to blame is, roughly, to register impairments in relationships—for example, between friends—and to modify one's attitudes accordingly (see Scanlon 2008, 128–29). In this paper, we remain agnostic about how, exactly, to spell out the nature of blame (see the beginning of sec. 3). However, it seems plausible to us that Scanlon has identified a further valuable function of blame: by blaming people, we can manage our relationships with them. In what follows, we will argue that, somewhat surprisingly perhaps, this function can be fulfilled to an important extent when the blamee is an AI system.

Let us begin with Scanlon's account of relationships that we will presuppose in the following. His view starts with paradigmatic intimate relationships like friendship. But it is also meant to make sense of less intimate relationships, for example between colleagues, and even of people's relationships with countries, companies, and other entities, as we will spell out in more detail below. The core idea is that relationships consist in attitudes and dispositions that the parties have towards each other (see Scanlon 2008, 131). For our purposes, we can think of representational states about, for example, what to expect from one another and motivational states about how to act towards each other. Take the relationship between colleagues as an example. The relationship-specific standards tell us what we, as colleagues, can be expected to believe and desire in our roles as colleagues. These standards also tell us what an entity needs to be able to be a party in a relationship. In particular, Scanlon argues that being able to make decisions and to regularly and non-accidentally conform to the standards that govern a relationship is sufficient for being able to be a party in the relevant kind of relationship (see Scanlon 2008, 161–62, 165).

Very briefly, our main argument is this: many AI systems can make decisions in the sense of interacting with their environments based on their representational and motivational states (see sec. 2). Moreover, they can non-accidentally conform to certain standards. Therefore, they can be parties in some of the relationships Scanlon is concerned with. They

can also breach these standards and, thus, we need ways to register these breaches and to revise our relationships accordingly. Blaming these systems can fulfill this important function. This is the skeleton of our view. Let us now flesh it out.

Consider, first, an asymmetrical, non-close relationship between humans. In Kazuo Ishiguro's novel *Remains of the Day,* the butler Stevens reflects on the issue of what makes a great butler. Especially important is the duty "to devote the utmost care in the devising of the staff plan" (Ishiguro 1989, 5). Imagine that the new employer, Mr. Farraday, expects from Stevens utmost care, realizes Stevens' "slovenliness at the stage of drawing up the staff plan" (Ishiguro 1989, 5), and responds by placing this responsibility on another employee. Thereby, Mr. Farraday would revise their relationship as a response to Stevens' not having the attitudes he expects from his butler. A response of this kind is important in a non-ideal world because we need ways to revise our professional relationships in accordance with whether others exercise the care we can reasonably expect from them. Human responses to AI systems can play very similar roles. Imagine that Stevens is replaced by an AI system. The users train it such that when devising a good staff plan comes in conflict with other jobs, say, searching the Internet for deals, devising the staff plan is prioritized. Imagine that this works well for a long time, but then the system autonomously prioritizes searching for deals which results in faulty staff plans and "many quarrels, false accusations, unnecessary dismissals" (Ishiguro 1989, 5). The users' response would be very similar to the one we imagined from Mr. Farraday: they would register that an expectation regarding the program's priorities has been breached, they would revise their attitudes to it by deciding to not rely on the system anymore and express this by, for example, ordering a new one. It is important for us to be able to respond in this way. If some entity does not have the priorities we can reasonably expect it to have, then we need to be able to change our attitudes towards it. Thus, blaming AI systems in this way fulfills a valuable function.

Some may reply that Stevens is a human being, but an AI system is not, which is, they may say, a crucial difference for whether revising relationships makes sense. We think that *being human* is not an important feature for the relevant kind of relationship management. To

see this, consider, second, relationships between individual humans and non-human entities, such as collective agents. Scanlon, for instance, discusses the case of a ferry accident with many casualties. He argues that we sometimes "have grounds to suspend our trust of the ferry company (say, by revoking its license to operate ferries)" (Scanlon 2008, 163). He explains that this "presupposes trust as the (…) default relationship against [which] a given relationship is measured" (Scanlon 2008, 164). Therefore, suspending our trust is a response to the company's impairing the default relationship and hence a form of blame, on Scanlon's account. For another case, consider NGOs and their donors. They are parties in a relationship that is partly constituted by the NGOs' expectation to be financially supported and the donors' expectation that the money is used in accordance with certain values. Sometimes NGOs fail on this. A Greenpeace activist injured two spectators of a Euro 2020 soccer game and risked harming many more when parachuting into the Munich Olympic Stadium to protest diesel and petrol cars (Guardian 2021). Plausibly, the donations of donors were not used in adequate ways in this case. For a donor, it would have been appropriate to revise their relationship with Greenpeace, for example, by sending critical emails or donating less for a certain period. Such responses would play the important role of re-shaping the relationship that the NGO has impaired. AI systems can, in the relevant ways, be like NGOs. Imagine an AI system that calculates how to use donations in the most efficient way to support human well-being and decides to invest in a certain program, but this turns out to be a very inefficient way to achieve the goal. Then, it would be appropriate for the users to revise their reliance on the system, to give negative feedback, and to look for a better alternative. This response is very similar to the donors' blaming Greenpeace in the parachuting case and it fulfills the same important functions. Thus, blaming AI systems can be an important way to manage our relationships with non-human agents (just like blaming NGOs can). [21]

[21] For a defense of the view that there are important parallels between the "regulatory interactions" we can have with collective agents on the one hand and with AI systems on the other, see also List (2021).

Some may reply that companies and NGOs, in contrast to AI systems, are constituted by human beings and that this makes an important difference for whether revising relationships with them makes sense. Again, we think that *being constituted by humans* is not a relevant factor here. To see this, consider, third, relationships between humans and their pets. Scanlon argues that for many humans the point of having pets is to have close relationships with them (see Scanlon 2008, 166). This relationship includes the expectation that the other party will not harm you or, depending on the kind of pet, that it does what you order it to do. If our pets do not live up to these expectations, it makes sense to revise our attitudes and relationships, for example, by modifying our desire to spend time and play with them. However, the same, we would argue, holds for some near-future or even current AI systems, like care, toy, or sex robots. For some people, one important point of having them is to have a relationship with them (see, e.g., Nyholm 2020, 105–9 for examples; see Ishiguro 2022 for a vivid fictional example). Such a relationship is governed by, for example, the standard not to harm the owners, and, in some cases, the standard that the robots do what the owners order them to do. If the systems breach these standards, their owners can appropriately revise their attitudes towards them, for example, by modifying their desire to spend time with them.

To sum up, many of us have important relationships with employees, companies, NGOs, or pets. These asymmetrical relationships differ in many respects from paradigmatic intimate relationships like close friendship or romantic love. However, what they share with the latter is that they are governed by standards that the parties involved in the relationships can (fail to) live up to. If the other party breaches the standard and, thereby, impairs the relationship, we can register this and revise our attitudes accordingly. This form of blame enables us to manage our relationships with these entities, which is important in the non-ideal world we live in. The same, we have argued, holds true for AI systems. We can have asymmetrical relationships with them that are governed by standards that these systems can (fail to) live up to. If they breach these standards, we can understand this as impairing the relationship we can have with them. It is important for us to be able to manage these

relationships. Thus, blaming AI systems within relationships of these kinds plays a valuable role.[22]

3.7 Taking stock

In the preceding, we took a closer look at the various valuable functions of our blaming practices and discussed which of these functions could still be fulfilled when the blamee is an AI system. We began with a negative claim: the retribution function can plausibly no longer be fulfilled. However, as we furthermore argued, it is also doubtful whether this function is valuable. Regarding the behavior-modification function, we contended that there are no in-principle obstacles to its fulfillment, but that the degree to which this function could be fulfilled would ultimately depend on whether AI systems will be equipped with the relevant learning mechanisms. When we turned to the conversation, protest, and signaling function of blame, such empirical contingencies became less important. These functions, we argued, could still be fulfilled surprisingly well (even if, e.g., AI systems never reach the status of "conversable entities"). The same held true for the relationship-modification function. We've thus arrived at the conclusion that our blaming practices could fulfill several valuable functions when targeting AI systems. If correct, this result would ensure that they would still "have a point" and give us a *pro tanto* reason to extend them to these systems (see the beginning of sec. 3).

4. Blaming AI and AI-blameworthiness

The result of the last section, however, may not seem enough to make such an extension fully appropriate. This is because, intuitively, it is *fully* appropriate to blame an entity for its conduct only if that entity is *blameworthy*, i.e., morally responsible for that conduct.[23] Hence, it seems

[22] Some authors, inspired by Peter Strawson's "Freedom and Resentment" (1962), claim that another important function of blame is to enable close, personal, symmetrical relationships: without blame responses like resentment, the idea is, there would be no such thing as real friendship or love (see, e.g., Shabo 2012). However, we are skeptical about whether this Strawsonian picture is correct (see, e.g., Milam 2016) and hence will not pursue this line of thought any further.

[23] To clarify, we presuppose that there are different senses in which it can be appropriate to blame a target. When we say that blaming a certain target is *fully* appropriate, we mean that blaming the target is appropriate in *all* (relevant) senses, i.e., that blaming the target would not merely be *all-things-*

that in order to show that it can be fully appropriate to blame AI systems, one would also have to show that AI systems can be morally responsible agents.

List, who suggests that certain forms of holding responsible other than blame should be extended to AI systems (see sec. 2), also discusses the issue of AI responsibility. His general stance on this issue is quite optimistic:

> (…) while there are significant technical challenges here, conceptually, there is no reason why an AI system could not qualify as a moral agent and, in addition, satisfy the knowledge and control conditions I have stated. Even if existing AI-systems do not yet meet these requirements, there is no reason to think that having an electronic or otherwise engineered hardware is an in-principle barrier to their satisfaction. (List 2021, 1229)

Thus, according to List, there are no in-principle obstacles to (future) AI systems fulfilling the conditions for blameworthiness (see also List 2021, 1227–31). Assuming List's optimistic stance on this point is correct, this would enable us to arrive at the following conclusion: we have reason to assume that it will be fully appropriate to extend our blaming practices to some future AI systems, since (i) we have reason to assume that some future AI systems will be blameworthy for their conduct and (ii) our blaming practices would still fulfill several valuable functions in targeting AI systems (as was argued previously).

However, not everyone will share this optimistic stance on the point of AI-blameworthiness (see, e.g., Hakli and Mäkelä 2019). Unfortunately, this is an issue too big to be settled within the scope of this paper.[24] So let us suppose that there are in-principle obstacles to AI systems fulfilling the conditions for blameworthiness. It may then seem to follow

---

considered permissible, but also *fitting* and *deserved*. An anonymous referee urged us to address the important issue of whether the practice of blaming children may be an everyday counterexample to our claim that, intuitively, only blameworthy agents are fully appropriate targets of blame. Here is a brief sketch of how we would respond to this: the practice of blaming children may show that it can sometimes be *all-things-considered permissible* to blame those who are not fully blameworthy (perhaps because it may sometimes have good consequences to blame children). But we do not think that the practice of blaming children shows that it can sometimes be *fully appropriate* (fitting, deserved, etc.) to blame those who are not (fully) blameworthy.

[24] For more on this topic, see also our unpublished manuscript "How Robots Can Be Blameworthy" (co-authored with Peter Schulte).

that our above reasoning would at best be of merely theoretical interest. However, this conclusion may be premature.

One common way to frame discussions about blameworthiness is in *moral* terms. The general idea is that the "worthiness" in "blameworthiness" should be understood in terms of fairness (e.g., Wallace 1994), justice (e.g., G. Strawson 1994), or desert (e.g., McKenna 2019). Despite important differences, these views share the following core assumption: if an agent fails to fulfill the conditions for blameworthiness, then it would be, in some sense, *morally* inappropriate to blame her (e.g., unjust, unfair, or undeserved), since blame, and, in particular, "open blame", is (at least somewhat) harmful for the blamee. However, as we've argued before (sect. 3.1), blame seems to *lose* its harmful character when the blamee is an AI system. Now, suppose that we are right about this. Then, it seems to follow that one key motive for avoiding "blame without blameworthiness"—namely, its being morally inappropriate in the way just articulated—no longer seems to apply when the blamee is an AI system.[25]

This, in turn, enables us to arrive at the following result: even if we combine our above reasoning with the assumption that no future AI system will fulfill the conditions for blameworthiness, we might still have good reason to extend our blaming practices to these systems. This is because one key type of moral concern for avoiding "blame without blameworthiness" no longer seems to apply when the blamee is an AI system. And this consideration, combined with the consideration that blame could still fulfill several valuable functions when targeting AI systems, might seem enough for an extension to these systems to be justified.

Against this, though, one might object that blaming a non-blameworthy AI system might still be problematic, especially if there is a blameworthy agent in the vicinity.[26] In particular, one

[25] To clarify, we do *not* want to claim that AI systems lack moral status (or lack moral rights). Our point is a much weaker one: unlike in the case of human beings, a certain prominent class of moral concerns about displaying blaming responses toward non-blameworthy entities seems to become irrelevant when the blamee is an *AI system*.

[26] Many thanks to an anonymous reviewer for raising this important objection.

might worry that it may deflect attention away from the real culprit (e.g., the designer or the company) and enable them to get off the hook too easily.

We agree that this is a valid worry. In reply, let us make three points. First, according to the account we've defended, the fact that blaming AI systems would fulfill several valuable functions merely gives us a *pro tanto* reason to blame these systems. This reason may very well be outweighed by considerations of the kind just articulated. Thus, our account is perfectly compatible with the claim that we should sometimes only blame the designer or the company, even if it would also make sense to blame the AI system.

Secondly, sometimes there will be no other agent (either individual or collective) who is blameworthy if an AI system causes (unjustified) harm. In fact, the assumption that we should expect such cases to arise is one key driving force for discussions about responsibility gaps (see sec. 2). In these cases, blaming a non-blameworthy AI system would not have the problematic consequences mentioned above.

We would maintain, though, that sometimes there will be another agent who is blameworthy and it will also be true that blaming the non-blameworthy AI system will have some undesirable consequences, but we may still have sufficient reason to blame the AI system. For instance, sometimes it may be very important to respond directly, i.e., in the given situation, to harmful behavior displayed by an AI system, but the real culprit may not be available. For illustration, think, once more, of the signaling function of blame (sec. 3.5). We can imagine cases in which we have strong reason to send a signal of solidarity to the victim and it may be that we can only achieve this by responding directly (and in a negative manner) to the AI system that caused the harm in that situation. In sum, we think that there may also be cases in which we will have sufficient reason to blame a non-blameworthy AI system even if this could, in a sense, be said to amount to an act of "misfired" blame and even if doing so had the undesirable consequences described above.

5. Conclusion

A common and important part of our everyday moral lives is to blame ourselves and others for bad conduct. The arrival of powerful AI systems that operate autonomously in high-stakes contexts raises the question of whether it makes sense to target these systems with blame when they make bad decisions. We have argued for the admittedly surprising claim that it indeed makes sense to include these systems in our blaming practices, since many of the important functions that are fulfilled by blaming humans can also be served by blaming AI systems. We concluded that this gives us good *pro tanto* reason to extend our blaming practices to AI systems.

It does not follow from this that we are obliged to include AI systems in our blaming practices or that there are no important differences between blaming humans and blaming AI systems. Still, the conclusion is important. For even if the arrival of powerful AI systems should require that we re-shape some of our moral theories and regulatory practices, our blaming practices do not need a fundamental revision and are in this sense "future-proofed": we can hold onto them and have good reason to include more players on the field.[27]

[27] The paper was accepted by JESP on April 27, 2023. We would like to thank two anonymous referees for JESP for engaging deeply with the paper and for providing two sets of very helpful comments, which greatly improved it. Furthermore, we would like to thank the participants of an online workshop on practical philosophy in January 2023, Susanne Burri, Max Kiener, Sebastian Köhler, Peter Königs, and Sven Nyholm, for highly constructive oral and written feedback. We are also grateful to Leonie Eichhorn and Shawn Wang for very helpful written comments, to Peter Schulte for very helpful advice, and to Dorothea Debus, Damiano Ranzenigo, Fabian Stöhr, as well as to the students of Leonhard Menges' Advanced Seminar in Practical Philosophy in December 2022 for very helpful discussions. Finally, we would like to thank Claire Davis for proofreading. Leonhard Menges' work on the paper was supported by the Austrian Science Fund (FWF) P34851-G and is part of the research project *The Sense of Responsibility Worth Worrying About*.

6. References

Bathaee, Yavar. 2018. “The Artificial Intelligence Black Box and the Failure of Intent and Causation.” *Harvard Journal of Law & Technology*, 889-938, 31 (2).

Baum, Kevin, Susanne Mantel, Eva Schmidt, and Timo Speith. 2022. “From Responsibility to Reason-Giving Explainable Artificial Intelligence.” *Philosophy & Technology* 35 (1): 12. https://doi.org/10.1007/s13347-022-00510-w.

Burri, Susanne. 2018. “What Is the Moral Problem with Killer Robots?” In *Who Should Die? The Ethics of Killing in War*, edited by Jay Strawser, Ryan Jenkins, and Michael Robillard, 163–83.

Coates, D. Justin, and Neal A. Tognazzini. 2012. “The Nature and Ethics of Blame.” *Philosophy Compass* 7 (3): 197–207.

———. 2013. “The Contours of Blame.” In *Blame: Its Nature and Norms*, edited by D. Justin Coates and Neal A. Tognazzini, 3–26. New York: Oxford University Press.

Danaher, John. 2016. “Robots, Law and the Retribution Gap.” *Ethics and Information Technology* 18 (4): 299–309. https://doi.org/10.1007/s10676-016-9403-3.

———. 2022. “Tragic Choices and the Virtue of Techno-Responsibility Gaps.” *Philosophy & Technology* 35 (2): 26. https://doi.org/10.1007/s13347-022-00519-1.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. 2018. “AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.” *Minds and Machines* 28 (4): 689–707. https://doi.org/10.1007/s11023-018-9482-5.

Fricker, Miranda. 2016. “What’s the Point of Blame? A Paradigm Based Explanation.” *Noûs* 50 (1): 165–83.

Gogoshin, Dane Leigh. 2020. “Robots as Ideal Moral Agents per the Moral Responsibility System.” *Culturally Sustainable Social Robotics*, 525–34. https://doi.org/10.3233/FAIA200952.

———. 2021. “Robot Responsibility and Moral Community.” *Frontiers in Robotics and AI* 8. https://www.frontiersin.org/articles/10.3389/frobt.2021.768092.

Guardian. 2021. “Greenpeace Apologises for Injuries Caused by Parachuting Protester at Euro 2020.” *The Guardian*, 0616 2021, sec. Football. https://www.theguardian.com/football/2021/jun/15/greanpeace-protester-avoids-accident-after-parachuting-into-germany-v-france.

Hakli, Raul, and Pekka Mäkelä. 2019. “Moral Responsibility of Robots and Hybrid Agents.” *The Monist* 102 (2): 259–75. https://doi.org/10.1093/monist/onz009.

Himmelreich, Johannes. 2019. “Responsibility for Killer Robots.” *Ethical Theory and Moral Practice* 22 (3): 731–47. https://doi.org/10.1007/s10677-019-10007-9.

Ishiguro, Kazuo. 1989. *The Remains of the Day*. 1st ed. London: Faber & Faber, 1999.

———. 2022. *Klara and the Sun*. London: Faber.

Kiener, Maximilian. 2022. “Can We Bridge AI’s Responsibility Gap at Will?” *Ethical Theory and Moral Practice*, July. https://doi.org/10.1007/s10677-022-10313-9.

Köhler, Sebastian. 2020. “Instrumental Robots.” *Science and Engineering Ethics* 26 (6): 3121–41. https://doi.org/10.1007/s11948-020-00259-5.

Königs, Peter. 2022. “Artificial Intelligence and Responsibility Gaps: What Is the Problem?” *Ethics and Information Technology* 24 (3): 36. https://doi.org/10.1007/s10676-022-09643-0.

Langer, Markus, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. “What Do We Want from Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research.” *Artificial Intelligence* 296 (July): 103473. https://doi.org/10.1016/j.artint.2021.103473.

List, Christian. 2021. “Group Agency and Artificial Intelligence.” *Philosophy & Technology* 34 (4): 1213–42. https://doi.org/10.1007/s13347-021-00454-7.

London, Alex John. 2019. “Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability.” *Hastings Center Report* 49 (1): 15–21. https://doi.org/10.1002/hast.973.

Macnamara, Coleen. 2015a. “Blame, Communication, and Morally Responsible Agency.” In *The Nature of Moral Responsibility: New Essays*, edited by Randolph Clarke, Michael McKenna, and Angela Smith, 211–36. New York: Oxford University Press.

———. 2015b. “Reactive Attitudes as Communicative Entities.” *Philosophy and Phenomenological Research* 90 (3): 546–69.

Mason, Elinor. 2019. *Ways to Be Blameworthy: Rightness, Wrongness, and Responsibility*. New York: Oxford University Press.

Matthias, Andreas. 2004. “The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata.” *Ethics and Information Technology* 6 (3): 175–83. https://doi.org/10.1007/s10676-004-3422-1.

McGeer, Victoria. 2013. “Civilizing Blame.” In *Blame: Its Nature and Norms*, edited by D. Justin Coates and Neal A. Tognazzini, 162–88. New York: Oxford University Press.

———. 2019. “Scaffolding Agency: A Proleptic Account of the Reactive Attitudes.” *European Journal of Philosophy* 27 (2): 301–23. https://doi.org/10.1111/ejop.12408.

McKenna, Michael. 2013. “Directed Blame and Conversation.” In *Blame: Its Nature and Norms*, edited by D. Justin Coates and Neal A. Tognazzini, 119–40. New York: Oxford University Press.

———. 2019. “Basically Deserved Blame and Its Value.” *Journal of Ethics and Social Philosophy* 15 (3). https://doi.org/10.26556/jesp.v15i3.547.

Menges, Leonhard. forthcoming. “Blaming.” In *The Routledge Handbook of Responsibility*, edited by Maximilian Kiener. New York: Routledge.

Milam, Per-Erik. 2016. “Reactive Attitudes and Personal Relationships.” *Canadian Journal of Philosophy* 46 (1): 102–22. https://doi.org/10.1080/00455091.2016.1146032.

Müller, Vincent C. 2020. “Ethics of Artificial Intelligence and Robotics.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/ethics-ai/.

Noorman, Merel. 2020. “Computing and Moral Responsibility.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/.

Nyholm, Sven. 2018. “Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci.” *Science and Engineering Ethics* 24 (4): 1201–19. https://doi.org/10.1007/s11948-017-9943-x.

———. 2020. *Humans and Robots: Ethics, Agency, and Anthropomorphism*. Rowman & Littlefield.

Pereboom, Derk. 2014. *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press.

———. 2021. *Wrongdoing and the Moral Emotions*. New York: Oxford University Press. http://dx.doi.org/10.1093/oso/9780192846006.001.0001.

Pettigrove, Glen. 2012. “Meekness and ‘Moral’ Anger.” *Ethics* 122 (2): 341–70.

Scanlon, T.M. 2008. *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, Mass.: Harvard University Press.

Shabo, Seth. 2012. “Where Love and Resentment Meet: Strawson’s Intrapersonal Defense of Compatibilism.” *Philosophical Review* 121 (1): 95–124.

Shoemaker, David, and Manuel Vargas. 2021. “Moral Torch Fishing: A Signaling Theory of Blame.” *Noûs* 55 (3): 581–602. https://doi.org/10.1111/nous.12316.

Smith, Angela M. 2013. “Moral Blame and Moral Protest.” In *Blame: Its Nature and Norms*, edited by D. Justin Coates and Neal A. Tognazzini, 27–48. New York: Oxford University Press.

———. 2022. “Blame and Holding Responsible.” In *The Oxford Handbook of Moral Responsibility*, edited by Dana Kay Nelkin and Derk Pereboom. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190679309.013.15.

Solum, Lawrence. 1992. “Legal Personhood for Artificial Intelligences.” *North Carolina Law Review* 70 (4): 1231.
Sommers, Tamler. 2012. *Relative Justice*. Princeton, N.J: Princeton University Press.
Sparrow, Robert. 2007. “Killer Robots.” *Journal of Applied Philosophy* 24 (1): 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.
Strawson, Galen. 1994. “The Impossibility of Moral Responsibility.” In *Free Will*, edited by Gary Watson, 212–28. New York: Oxford University Press, 2003.
Strawson, Peter F. 1962. “Freedom and Resentment.” In *Free Will*, edited by Gary Watson, 72–93. New York: Oxford University Press, 2003.
Talbert, Matthew. 2012. “Moral Competence, Moral Blame, and Protest.” *The Journal of Ethics* 16 (1): 89–109.
Tigard, Daniel W. 2021a. “Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible.” *Cambridge Quarterly of Healthcare Ethics* 30 (3): 435–47. https://doi.org/10.1017/S0963180120000985.
———. 2021b. “Technological Answerability and the Severance Problem: Staying Connected by Demanding Answers.” *Science and Engineering Ethics* 27 (5): 59. https://doi.org/10.1007/s11948-021-00334-5.
Tognazzini, Neal, and D. Justin Coates. 2018. “Blame.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/blame/.
Vargas, Manuel. 2013. *Building Better Beings: A Theory of Moral Responsibility*. New York: Oxford University Press.
Walen, Alec. 2021. “Retributive Justice.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2021. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/justice-retributive/.
Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge, Mass.: Harvard University Press.
Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195374049.001.0001.
Wang, Shawn Tinghao. 2021. “The Communication Argument and the Pluralist Challenge.” *Canadian Journal of Philosophy* 51 (5): 384–99. https://doi.org/10.1017/can.2021.30.
Watson, Gary. 1987. “Responsibility and the Limits of Evil: Variations on a Strawsonian Theme.” In *Agency and Answerability: Selected Essays*, 219–60. New York: Oxford University Press, 2004.
Winfield, Alan F. T., and Marina Jirotka. 2017. “The Case for an Ethical Black Box.” In *Towards Autonomous Robotic Systems*, edited by Yang Gao, Saber Fallah, Yaochu Jin, and Constantina Lekakou, 262–73. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-64107-2_21.