

Inquiry

An Interdisciplinary Journal of Philosophy

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/sinq20

A way forward for responsibility in the age of AI

Dane Leigh Gogoshin

To cite this article: Dane Leigh Gogoshin (12 Feb 2024): A way forward for responsibility in the age of AI, *Inquiry*, DOI: [10.1080/0020174X.2024.2312455](https://doi.org/10.1080/0020174X.2024.2312455)

To link to this article: <https://doi.org/10.1080/0020174X.2024.2312455>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 12 Feb 2024.



Submit your article to this journal [↗](#)



Article views: 996



View related articles [↗](#)



View Crossmark data [↗](#)

A way forward for responsibility in the age of AI

Dane Leigh Gogoshin 

Department of Practical Philosophy, University of Helsinki, Helsinki, Finland

ABSTRACT



Whatever one makes of the relationship between free will and moral responsibility – e.g. whether it's the case that we can have the latter without the former and, if so, what conditions must be met; whatever one thinks about whether artificially intelligent agents might ever meet such conditions, one still faces the following questions. What is the value of moral responsibility? If we take moral responsibility to be a matter of being a fitting target of moral blame or praise, what are the goods attached to them? The debate concerning 'machine morality' is often hinged on whether artificial agents are or could ever be morally responsible, and it is generally taken for granted (following Matthias 2004) that if they cannot, they pose a threat to the moral responsibility system and associated goods. In this paper, I challenge this assumption by asking what the goods of this system, if any, are, and what happens to them in the face of artificially intelligent agents. I will argue that they neither introduce new problems for the moral responsibility system nor do they threaten what we really (ought to) care about. I conclude the paper with a proposal for how to secure this objective.

ARTICLE HISTORY Received 2 May 2023; Accepted 7 January 2024

KEYWORDS Moral responsibility; responsibility gaps; artificial intelligence; artificial moral agency; AI ethics; responsible AI

1. Introduction

It is generally taken for granted that moral responsibility is vital both for a functional society as well as for a valuable and meaningful life (Dennett 1984; Vargas 2007; Smilansky 2000; Strawson 2008; Pereboom 2014). Without it, no one would ever truly deserve moral blame, punishment, praise, or reward, and our interpersonal relationships would take on a hollow ring. When faced with a tragedy involving artificial intelligence (AI), such as the death of a pedestrian by a self-driving vehicle, or the massacre of multiple civilians by a lethal, autonomous military drone, where

CONTACT Dane Leigh Gogoshin  dane.gogoshin@helsinki.fi  Department of Practical Philosophy, University of Helsinki, Helsinki, Finland

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

the assignment of moral responsibility is far from straightforward (giving rise to possible responsibility gaps), we naturally conclude that we have landed in new and considerably worse territory. Responsibility gaps, it is claimed, threaten ‘the moral framework of society and the foundation of the liability concept in law’ (Matthias 2004, 176), preclude (where lethal autonomous weapons are deployed) morality and legality in war (Sparrow 2007), undermine public trust in the rule of law (Danaher 2016), and interfere with healthcare providers’ ability to fulfil moral duties and uphold patient rights (Lang, Nyholm, and Blumenthal-Barby 2023, 152).

In this paper (see also Gogoshin 2023a), I challenge the assumption underlying these claims. It is incorrect, or so I will argue, to view AI-inhabited moral contexts, from the standpoint of moral responsibility, as either significantly new or worse. This is because (1) the moral responsibility system (responsibility attributions and the practices associated with holding responsible) is fundamentally flawed and (2) the true goods which are associated with this system are not threatened by the introduction of artificially intelligent agents (AIs). Hence, worries about threats to the responsibility system and efforts to salvage it are seriously misguided. Identifying and dealing adequately with AI-related threats will require, at a minimum, acknowledging this. I will also propose a way forward for responsibility in the age of AI which gets to the heart of what we really (ought to) care about with respect to responsibility.

Not only does the techno-responsibility gap debate¹ ignore the well-founded, long-standing scepticism about moral responsibility (Spinoza and Curley 1985; Honderich 2002; Strawson 1994; Waller 2011; Pereboom 2014; Levy 2011; Caruso 2017, 2021²), it takes the benefits associated with moral responsibility for granted and disregards the empirically confirmed harms of the responsibility practices (Pickard 2017; Snoek et al. 2021; Waller 2011, 2014; Holroyd 2021; Jeppsson and Brandenburg 2022).³ It is my first task in this paper to rectify these oversights. To this end, I adopt the sceptics’ critical stance toward the responsibility system and

¹This is Tigard’s (2021) term for responsibility gaps arising in the context of autonomous AI.

²The first two of these are ‘hard determinists.’ They assume that determinism is true and that free will and moral responsibility are incompatible with determinism. The others are ‘hard incompatibilists.’ They argue that moral responsibility is incompatible with determinism and indeterminism.

³This is to say that, with rare exceptions (e.g. Munch, Mainz, and Bjerring 2023; Danaher 2022), the techno-responsibility gap debate is predicated on the belief that moral responsibility is vital to a meaningful life and a functional society – irrespective of whether one argues for or against the existence of responsibility gaps, or whether or not we can solve them. There are others (e.g. Himmelreich 2019; Robillard 2018; Königs 2022) who argue/assume that responsibility gaps exist and that they are unproblematic, but for different reasons than I do.

their optimism about a world without basic desert responsibility, but I do so without their theoretical commitments (according to which moral responsibility is strictly incompatible with a naturalistic-scientific system).⁴ I assume in this paper that AIs cannot be responsible in the traditional (basic desert⁵) sense. I also assume that insofar as responsibility just is a matter of adopting certain blame or praise attitudes in response to wrong or right action or to certain qualities of will, AIs cannot be morally responsible. Accordingly, I assume that responsibility gaps arise in AI-inhabited contexts. However, because I challenge the value and legitimacy of responsibility attributions and practices, I do not think these gaps are unique to AI or a result of AI itself, and that there exist preferable alternatives to our existing practices.

I begin, in §2, with a comprehensive overview of the functions and aims of the responsibility system as per various compatibilist accounts. I then identify the overarching ones and situate them within the broader scope of ethics. In §3, I present criticisms of the responsibility system and alternatives to it. These tasks (and the thoroughness with which they are undertaken) are crucial for identifying what is at stake in the techno-responsibility gap debate – for getting clear, assuming there are such gaps, about what precisely is at risk and whether that is worth protecting. In §4, I consider what introducing AIs into societal roles does to the previously identified aims. I focus on the threats identified in the responsibility gap literature (control and transparency) and consider their relationship to these aims. Two insights surface in this discussion – that what is of chief moral importance regarding AI has more to do with the nature of moral decision-making than responsibility, and with the human actors and agendas behind AI than AI's (hard-to-control, opaque) nature. In §5, I suggest a way forward for the responsibility worth wanting in the age of AI.

2. Responsibility: aims and functions

2.1. Prefatory remarks

It is widely accepted that forward-looking notions of responsibility (Schlick 1962; Nowell-Smith 1948; Smart 1961; Dennett 2003, 2015;

⁴By 'optimistic sceptics,' I have in mind especially Waller, Pereboom, and Caruso. They are optimistic that justice, a robust sense of our own agency and achievement, and meaningful interpersonal relationships are all possible in the absence of basic desert responsibility.

⁵To be defined in §2.1.

Pereboom 2014; McGeer 2019; Milam 2021), as well as the ‘answerability’ version defended by Scanlon (1998) and Bok (1998) (per Caruso and Morris 2017, 839), are immune to the metaphysical threats posed to the traditional concept of moral responsibility, that of ‘basic desert responsibility.’⁶ Per Pereboom (2014, 2), an agent who is morally responsible in the basic desert sense deserves ‘to be blamed or praised just because she has performed the action, given an understanding of its moral status, and not, for example, merely by virtue of consequentialist or contractualist considerations.’ It is this basic desert conception which compatibilists are charged with defending. There are a variety of compatibilist strategies which I do not undertake to elaborate. In what follows, I take a neutral stance toward them and simply assume, for the sake of argument, that either they succeed or that their lack of success does not undermine the responsibility function they attach to. My present concern is to identify and categorize the aims or functions that are tied to the responsibility practices (acts, expressions, and attitudes of moral praise and blame), with the goal of assessing their value. Even if basic desert responsibility turned out to be fully compatible with a scientific world view and our practices justified in that respect, it might still be the case that they do not get us closer to what really matters and/or that what really matters does not depend on basic desert responsibility.

2.2. Inculcating responsible behaviour

Consequentialist accounts (those of Schlick, Nowell-Smith, Smart, e.g.) focus on the undisputed socially regulative effects of the responsibility practices.⁷ I will focus here on Daniel Dennett’s argument (Dennett 2003, 2015). He says that in order to understand the concept of responsibility, we must look at the *point* of holding responsible (Dennett 2015, 172). For simplicity’s sake, he takes up the ‘most explicit (public, codified, instituted) responsibility practice,’ punishment, which he looks at through the lens of the institution of criminal law. Understanding its rationale, Dennett says, is key to understanding our status as responsible moral agents (2015, 173). Via punishment, we successfully minimize the frequency of wrongdoings in society. This is because even just

⁶Pereboom (2014) suggests that there are additional compatibilist accounts which, with certain qualifications, can be accepted by the free will sceptic. Notably, consequentialist accounts like those of Schlick and Smart can be seen as revisionary compatibilist accounts (Pereboom 2014, 131), but the kind of free will and moral responsibility they defend (see especially Dennett 2003) is sufficiently revisionist to sidestep the free will debate entirely.

⁷As acknowledged by Hobbes, Hume, Mill, Adam Smith, P.F. Strawson, etc.

approximately rational agents, which most of us are, are deterred by the negative legal and/or social consequences tied to prohibited behaviour. We need the system of threatened punishments to keep us on track because though rational, we are not angels, and our world is less than ideal (2015, 172). Staying on track, being part of the 'moral agents club,' Dennett says, in turn gives one access to society's goods (goods which are made possible by responsible behaviour and cooperation). Our practices of holding responsible are simply 'the best game in town' when it comes to the goal of minimizing wrongdoing 'while also trying to minimize the costs of enforcement and punishment' (2015, 175-177).

Dennett's naturalistic conception of free will, while revisionary, is sufficient, he claims, to ground feeling legitimately proud or ashamed, grateful or guilty, and to take responsibility for our actions. It is neither possible nor necessary to determine whether a particular trait is of someone's making, 'instead of trying to assay exactly to what degree a particular self is self-made – we simply hold people responsible for their conduct (within limits we take care not to examine too closely). And we are rewarded for adopting this strategy by the higher proportion of "responsible" behavior we thereby inculcate' (Dennett 2015, 179).

2.3. Moral agency cultivation

Whereas consequentialist accounts are satisfied with a low bar for responsible agency (the capacity to be regulated by the sanctions and rewards of our responsibility practices), they have been criticized, among other things, for incorrectly reducing our practices to carrots and sticks aimed only at behavioural regulation (Wallace 1994). Instrumentalist accounts (e.g. Vargas 2013; McGeer 2019), while maintaining the metaphysical advantage of the consequentialists, overcome this criticism by interpreting our practices as cultivating moral agency, by way of enhancing our sensitivity to moral reasons. Via a sensitivity to positive or negative social feedback in the form of reactive attitudes (blaming and praising attitudes like resentment, indignation, gratitude, admiration, etc.), 'reactive exchanges,' we come to internalize the moral reasons being targeted by these attitudes. This is because 'the capacity to recognize and respond to *moral* reasons is an essentially social skill, requiring social feedback to develop and maintain' (McGeer 2019, 313). Due to our particular psychological disposition to learn from the corrective interventions of others and to correct them, this feedback is distinctly reactive in nature (McGeer 2015).

2.4. Moral address and its outcomes

Communicative accounts (Watson 2004; Darwall 2006; Shoemaker 2007; McKenna 2012; Macnamara 2015; Fricker 2016) claim that the function of our practices (especially blame) is the communication of 'demand for reasonable regard' (Watson 2004, 229). Blame, in this light, amounts to a kind of message meant to elicit 'a response in a recipient, where that response amounts to uptake of the message sent' (Macnamara 2015, 553). As with the instrumentalist exchange, the moral address of the communicative account is a two-way street, with a 'call and response' structure. According to Fricker, blame has the 'illocutionary point' of 'inspiring remorse in the wrongdoer,' which is 'a pained moral perception of the wrong one has done' (Fricker 2016, 167). Its aim is to bring the moral understanding of the wrongdoer and the wronged into alignment.

2.5. Protest and relationship modification

Protest accounts (Hieronymi 2004; Smith 2012; Talbert 2012) explain the function of blame in terms of registering and challenging the moral claim implicit in an agent's behaviour. In effect, when we blame, by way of modifying our attitudes and intentions toward, or expectations of wrongdoers, we protest their moral commitments (or negative quality of will) and seek some kind of moral acknowledgement from them or from the moral community (Smith 2012, 43). The protest account shares the call-and-response structure of the communicative account as well as the relationship modification aspect of Scanlon's account.

In line with his view of morality as essentially interpersonal, Scanlon (2008) holds that our blame practices serve to modify relationships in accordance with the relationship impairment caused by the wrongdoing. When we blame our partner for some action, we take our partner to have impaired the relationship, and we, in turn, modify our attitudes toward and expectations of our partner in a way that reflects this impairment. Restoration of the relationship hinges on the proper uptake and reparative response from the wrongdoer.

2.6. Normative landscaping

Like communicative, protest, and Scanlonian accounts, where our responses signal and enforce/reinforce normative commitments, the remaining accounts also highlight the (passive and active) relationship

between our practices and the normative landscape.⁸ First, according to McGeer (2019), the normative landscape is dynamic and subject to ever-changing material circumstances. It is through the responsibility practices, she says, that we discover, negotiate, and communicate the new contours as they arise. Although this view meshes with a constructivist metaethics, it isn't limited to it. Naturally, material conditions change over time and aspects of our normative obligations inevitably change with them. Reactive exchanges serve to surface disagreements about these normative obligations and then to 'fine-tune' our normative expectations in light of these new conditions (Sie 2018).

Irrespective of one's metaethics, the set of obligations we incur via morally relevant actions like wrongdoing (our 'normative footprint'; Sliwa 2019), may well be an object of construction, the method of construction being our reactive exchanges. Sliwa (2019) argues that blame has an epistemic function; it communicates how an act of wrongdoing has reshaped the normative landscape, incurring reparative rights and duties (e.g. to apologize). Other authors attribute constructivist or negotiatory powers to our practices. Bagnoli (2021) suggests that our responses to blame (by way of disclaiming it) constitute 'modes of exercising normative powers, whose main functions are demanding recognition, responding to wrongs, voicing disagreement, exiting alienating conditions, and calling for a fair redistribution of specific responsibilities.' Bicchieri (2017) describes how responsibility practices can (re)shape the normative landscape.

2.7. Retribution

I acknowledge that there are many proponents in the legal and philosophical literature of retributivism.⁹ 'Retributivism,' per Caruso and Morris (2017, 841), 'refers roughly to the justification for treatment whereby an individual is either rewarded or punished as payback for the moral rights and wrongs he has committed.' Along with Waller (2011) and Caruso (2021), I dispute the normative basis for retributivism and maintain that the retributive urge should be resisted. Whatever one's normative position, the fact of the matter, Danaher (2016) says, is that the public looks to the law to manage this urge. Responsibility gaps may make this impossible, resulting in an erosion of public trust in the rule of law.

⁸See Shoemaker and Vargas (2021) for the costly-signalling functionalist account of blame.

⁹See Moore (1997) for a legal scholar's defence of retributivism, Morris (1968) and McKenna (2020) for philosophical defences.

In response, Kraaijeveld (2020) offers a debunking argument of retributivist intuitions and thus of retributivism and retribution gaps. While it is not entirely clear that Kraaijeveld speaks to Danaher's worry, Königs (2022, 35) does. He points out that there are non-retributivist, strictly forward-looking justifications for legal responses (e.g. sanctions) which can satisfy the public's desire for justice.¹⁰

I take seriously the thought that, when faced with grave, malicious harm, the wrongdoer cannot simply be allowed to 'get away with it,' even while rejecting the claim that punishing wrongdoers is intrinsically good. Like the consequentialists and the optimistic sceptics, I hold that justice can be pursued in meaningful, non-retributivist, and even non-punitive ways.

2.8. Taking responsibility

Though accounts of taking responsibility (TR) are not functionalist accounts, the active conception they advance captures what I view as the second of two high-level aims of the responsibility system, the first being the cultivation of responsible behaviour (and coinciding minimization of harms). There are roughly two conceptions of TR, one more general and passive, and the other, more specific and active. The first one is taken by its authors to constitute responsible agency. For Frankfurt, it is in identifying oneself with the springs of one's actions that one takes responsibility and becomes morally responsible for them (Frankfurt 1975, 122). For Kane (2007, 41–42), it is by TR that we resolve the indeterminacy underlying all human action, making it our *own* choice. For Dennett (2015), it is by TR that we become responsible agents. In seeing ourselves as in control – even if at a certain level of abstraction we are not – and by TR, even for accidents, we increase our control and make ourselves 'less likely to be "accident" victims in the future' (2015, 157). In what follows, I focus on the more specific, active notions of TR in both traditional and AI contexts.

2.8.1. Taking responsibility in the traditional context

Mason (2018), targeting biases for which we are traditionally deemed to lack control, proposes that we can and should take responsibility (also 'TR') for them. 'I argue that the zone of responsibility can be extended to include acts that we are not fully in control of, and acts whose moral

¹⁰Whether AIs pose a threat to even forward-looking responsibility will be discussed in §4.2.2.

status we are nonculpably ignorant about at the time of acting. This extension of responsibility happens through a voluntary *taking* of responsibility' (2018, 164). By TR for a belief or action, she argues, we can become responsible for it (2018, 179). TR is not, on Mason's view, a moral duty and it extends only to those things which are genuinely an agent's actions. TR is valuable within the context of interpersonal relationships and it is only insofar as one values a good relationship that one has a normative reason to TR. Enoch (2012), by contrast, argues that the things for which we can TR extend as far as the 'penumbra' of our agency, and for these, we have a moral duty to TR. Sliwa (2023) defines TR as a matter of properly 'owning' one's normative footprint: recognizing and committing to abide by the obligations incurred by it – to, e.g. apologize, repair, feel remorse, etc.

2.8.2. *Taking responsibility in the AI context*

Champagne and Tonkens (2015) argue for TR *qua* prospective liability. They suggest, in the case of military AI, that a sufficiently high-ranking person or group of persons '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' (2015, 126). Kiener (2022, 582–283) proposes instead TR *qua* answerability (cf. Duff 2009), by which he means the obligation to answer and explain one's conduct and to appropriately respond to those affected (with the standing to demand such an explanation). These responses may include 'the obligation to apologise, to follow up on the well-being of those who have been harmed, to take precautionary measures so that similar harm will not recur, and so on' (Ibid, 586). The taking of answerability in AI-related harms is restricted, however, 'to those who have been involved in the development and use of that AI' (Ibid). Finally, Goetze (2022) argues that computer professionals are morally obliged to TR for the actions of the systems they design, despite not being *prima facie* morally responsible for them.

2.8.3. *The value of taking responsibility*

TR, in the active sense, constitutes what I take to be a primary function of the responsibility system. It enables us to 'own,' mitigate, or contain the consequences of our normative footprint – to explain our actions, make reparations, apologize, commit to reforming ourselves, etc. It seems to be a way of closing the responsibility gap inherent in human life – with or without advanced technologies. As all TR proponents implicitly or

explicitly acknowledge, there are strong limits to what we are legitimately or automatically responsible for. The solution, in their view, is to TR.

2.9. Responsibility: the big picture

The preceding discussion gives rise to two primary functions of the responsibility system to which the other functions can be described as subservient: (1) the cultivation of responsible behaviour and the coinciding minimization (deterrence) of harms (which depend on communicating and shaping the normative landscape), at societal and interpersonal levels, and (2) the fulfilment of duties incurred by our normative footprints, aimed at the reparation and restoration of harms done (which may include reformation of the wrongdoer, contributing in effect to (1) and/or (2)). The aim of this brief sub-section is to situate these aims with respect to ethics more broadly.

In his counterfactual genealogical story of ethics, Pettit and Hoekstra (2018) suggests that responsibility concepts arise out of a need to build trust and reliability within groups whose members are wholly interdependent. Trust, in turn, is built on the good reputations of those members which, in turn, are built on their following through on their avowed desires (cum values). They commit to these by way of making pledges. Fulfilling these pledges often requires overcoming competing desires, something which is motivated by the desire to bolster one's credibility. Subject to social influence – in the form of exhortations (e.g. 'you can do it!'), blame arises as a retrospective exhortation, 'you could have done otherwise.'

Tomasello (2015), an evolutionary psychologist, in his reconstruction of the evolution of human morality, argues that morality is a form of cooperation (see also Curry, Mullins, and Whitehouse 2019), and that the shared intentionality of pre-humans led from strategic cooperation to morality. This involved a transformation of role-specific normative standards into idealized, impartial standards to which the group made a joint commitment. Collaboration according to these standards gave rise to mutual respect and trust. Individuals self-regulated according to these standards, to which they felt a sense of obligation and, correspondingly, a need to justify themselves to their community. Breaches of group standards led to censure in the form of blame and punishment.

Taking these two stories about the genesis of responsibility for granted, we can say that responsibility plays a central role in establishing the necessary conditions of cooperation – trust, reliability, and mutual

respect. Our practices track and foster norm adherence. When we can trust and rely upon one another to adhere to the standards we collectively endorse, we can cooperate more effectively. We can thus append to (1) above, that the responsibility system promotes cooperation.

3. The responsibility system: criticisms and alternatives

The goal of this section is to counter the view of our practices as straightforwardly positive. I do this by way of briefly conveying some criticisms and then by presenting superior alternatives. The strongest critical voices are those of the responsibility sceptics (Waller 2011, 2015; Pereboom 2014; Caruso 2017, 2021), but there are significant (and growing) whistle-blowers among compatibilists as well (Springer 2013; Pickard 2017; Snoek et al. 2021; Stichter 2020; Holroyd 2021; Jeppsson and Brandenburg 2022). The former argue that our existing practices should be substituted with non-desert based practices whereas the latter criticize specific aspects or applications of them, in recognition of their Janus-faced nature.

3.1. Criticisms

Criticisms of the moral responsibility system target both unintended effects of our practices (counter to the aims) as well as the aims themselves. Pickard (2017) points out that behavioural modification (particularly among substance abusers) is negatively affected by blame, in particular (see also Snoek et al. 2021). The sceptics claim that basic desert responsibility leads to harsh retributive practices which perpetuate harm and injustice (Waller 2011; Pereboom 2014; Caruso 2021).

Waller (2011), Springer (2013), Holroyd (2021), and Snoek et al. (2021), reference psychological research which shows that blame and praise can undermine moral motivation and reformation. Stichter (2020) shows that, without the right skills for uptake, blame stands to promote moral disengagement in the blamee. *Contra* the instrumentalists, per Waller (2023, 372), '[t]he "blame and shame" of moral responsibility are not effective tools for accomplishing the vitally important goal of "developing and sustaining our capacity to recognize and respond to moral reasons" (McGeer 2019, 313); to the contrary, they are more likely to limit and impede moral development.'

This is because, as I have argued (Gogoshin 2023b), it is possible that our practices, via providing prudential incentives for complying with

the moral reason (responding with blame to norm transgressions, praise for compliance), foster extrinsic motivation and prioritize norm-compliance, sometimes at the cost of moral reasons-sensitivity.¹¹ We might, e.g. drive the speed limit to avoid sanction or keep promises to avoid blame rather than for the (moral) reasons that would justify these responses (e.g. avoid causing harm, treat others respectfully).¹² We might internalize blame responses to certain actions (develop an aversion to performing them, e.g.) rather than the moral reasons (Harland 2020). Finally, we might come to prioritize the avoidance/pursuit of blame/praise responses over and above the moral reasons which justify them (caring more to avoid censure than harm, e.g.).

In order to show that our practices stand a chance at sensitizing agents to moral reasons, the instrumentalist must idealize our practices, in the form of the 'ideal reactive exchange.' It must involve the right audience (with the right moral views, the right moral standing) and the right feedback (the right amount of blame, delivered in the right way), and depends on the blamee having the right disposition for positive uptake of the blame response, e.g. Our *actual* practices are far from ideal. Second, even in meeting the conditions of the ideal reactive exchange, (enhanced) moral reasons-sensitivity is not a guaranteed outcome (as suggested above). The same criticism can be said to apply to the communicative model, the desired outcome of which is the improved moral understanding or moral competence of the blamee. This outcome is dependent on factors about both agents in the moral conversation – capacities for effective and proportional reactive response and for its proper uptake.

Because blame and praise are powerful tools of behavioural conditioning (Bicchieri 2017; Gogoshin 2021), we also risk thereby inducing compliance with bad norms (Fricker 2016, 25–26) and further entrenching existing injustices (Holroyd 2021; Harland 2020; Mackenzie 2021; Jeppsson and Brandenburg 2022), thereby shaping the normative landscape for the worse. As Coggins and Steinert (2023) have pointed out, building norm-compliant robots is problematic for these and other reasons.¹³ We also risk perpetuating individual harms and increasing injustice *vis-à-vis* TR. The duties incurred by wrongdoing – to apologize, repair, etc., are

¹¹See Duff (2001) for the view of our responsibility responses as providing prudential incentives for moral action. Moral appeals, he says, are weakly persuasive and moral agents, weakly responsive, and so we require this incentivization.

¹²By "moral reasons," I have in mind the concept at work in Arpaly's (2003) "right-making features" or Sliwa's (2016) "rightness" views of moral worth.

¹³Reasons pertaining to further sources of injustice not addressed in the previously cited literature.

in principle more difficult for certain (perhaps even a majority of) wrongdoers. Wrongdoers, especially those who end up in the prison system, are often disenfranchised members of society, victims of corrosive disadvantages (Waller 2011; Wolff and de-Shalit 2007). Many come from troubled homes and disadvantaged socioeconomic classes. Despite being viewed as responsible, they are much less likely to be able to fulfil these incurred duties and, in this way, can never own or repair their normative footprint, be forgiven, or move forward. Worse, by failing to fulfil these new duties, they are taken to commit further wrongs and are thus open to a vicious cycle of blame and punishment.

It is likely that it takes a certain privilege to TR for things (Waller 2011) – especially over which we initially lacked sufficient control. *Contra* Dennett, therefore (who claims it is in TR for actions which may even be accidents that we are less likely to be accident victims in the future), if it is the case that we lack control over something, being blamed or praised for it does not magically impart that control. On the contrary, it often leads to learned helplessness and apathy (Waller 2011, 77, 136-137). At the very least, in order for Dennett’s desired outcome to occur, blame and praise should target things over which we actually have control. While Waller (2011, 108) concurs that the ability to TR is vital and positive, it is not the same as moral responsibility – that which would justify blame, praise, punishment, or reward. We cannot legitimately *take* moral responsibility.

3.2. Alternatives

Pickard (2017), Pereboom (2014), and Caruso (2017) offer promising alternatives to our current practices, which I will only briefly mention here. Insofar as the techno-responsibility gap is a result of AIs’ lack of basic desert responsibility, these provide ready-made desert-free alternatives. If nothing else, they undercut the perceived threat to our current practices by showing that they’re replaceable.

Pickard argues (and shows empirically) that responsibility without blame is a much more effective method for behavioural rehabilitation. Pereboom reorients the aims of the responsibility system to the aims of protection, reconciliation, and moral reformation, and argues that they’re all attainable by way of non-desert-based emotional attitudes and responses (e.g. moral sorrow and disappointment rather than resentment) which preserve meaning in interpersonal contexts. Caruso offers a quarantine model as an alternative to (especially the American) crimin

al system, which he argues is plagued by basic desert-linked retributive practices. Here, I focus on Waller's alternatives to TR and to the 'blame and shame game': 'take-charge responsibility' (Waller 1998; 2011, 106–114) and the 'no-blame systems approach' Waller (2020; 2023), respectively.

According to Waller, Dennett confuses TR with 'take-charge responsibility' (Waller 2011, 146). Take-charge responsibility, inspired by Hart's (2008) 'role responsibility,' 'designates the broader taking of responsibility – including taking charge of one's own plans and projects and life – that must be distinguished from moral responsibility. Just as a captain may have role responsibility for a ship, so you may have take-charge responsibility for your projects, your values, your goals, your life' (Ibid, 107). Acquiring this power to exercise effective control over our life choices is not itself something we control, since it depends on lucky circumstances. Whether agents with take-charge responsibility are also morally responsible is a separate question, and something that a responsibility sceptic like Waller would deny.

In order to make the case for a blame-free alternative, Waller provides two real world examples. He first describes Toyota's disastrous early auto manufacturing process which, he suggests, was plagued by the destructive effects of blame and shame.

Its severely top-down control model demanded that workers follow orders without thinking or questioning. Mistakes on the assembly line were blamed on individual workers and severely punished: the problem was 'solved' when the individual worker was fired. Workers hid mistakes when possible and tried to shift the blame to others when mistakes could not be concealed. Small problems were covered up until they became big problems, cars rolled off the assembly line with multiple defects, and Toyota became notorious for poor workmanship. (Waller 2023, 372–373)

Eventually, Waller says, Toyota replaced this 'blame and shame *control* model' with a 'no-blame *systems* approach' (Bodek 2011) based on three radically different basic principles.

First, rather than blaming workers for mistakes and problems, the detection of problems was regarded as a valuable part of improving the manufacturing process: small problems that would have been hidden by worried workers were instead exposed, examined, and fixed before they evolved into disasters; and workers who reported problems were treated as valuable contributors to improving the process. Second, problems were not treated as the fault of individual workers to be solved by firing an individual at the immediate problem source. Problems are *systemic* and solving them requires careful examination

of the deeper causes. And finally, workers were treated as valuable contributors to accomplishing shared goals, and their expertise and insights were welcomed and treated with respect. (Waller 2023, 373)

According to Waller, these changes led to Toyota's transformation from a manufacturer of low quality automobiles to 'a company with an earned reputation for high quality workmanship with remarkably few manufacturing flaws' (Ibid).

Waller goes on to provide another example of adopting the systems model in the domain of air traffic control, after which the error rate decreased dramatically.

Instead of treating errors as evidence of individual negligence, the inevitable errors were now viewed as vitally important indications of deep systemic problems that required cooperative shared efforts to resolve (Sabatini 2008; Harris and Muir 2005). Small errors were reported and fixed before they became disasters, and cooperative efforts resulted in effective ways of radically reducing errors and preventing the inevitable errors from becoming disasters. (Waller 2023, 373)

As a responsibility sceptic, Waller wholly rejects basic desert and related practices. Abolishing them, he argues, is a necessary condition of a fairer and more just world. Applying the principles behind the systems model to social relations more broadly, he argues, will engender respect for abilities and persons and encourage a deeper understanding of systemic social problems, enabling meaningful solutions. Moral responsibility, as a concept applied to individuals, gives rise to a myopic, distorted view of problems and their solutions.

3.3. Interim conclusion

Whether these criticisms and alternatives, taken together, entail that we're better off without moral responsibility full stop, just the responsibility practices, or just certain aspects of these practices, is not immediately obvious. Whether we need the concept of moral responsibility and whether the concept can be wholly stripped of its backward-looking nature while yet having the forward-looking benefits we associate with responsibility practices, are also far from clearcut matters. It is clear enough that we cannot take our practices or their outcomes as straightforwardly positive. This alone warrants questioning the fear driving the techno-responsibility gap debate and motivates alternative practices.

If what we're after is acquiring greater degrees of control over self and the environment, opportunities to make our own decisions and exercise

effective control, as both Dennett and Waller suggest we are (despite disagreeing about what that amounts to), increasing responsible behaviour, decreasing harms, shaping and communicating the right normative landscape, building trust, mutual respect, cultivating positive relationships, increasing collaboration and cooperation, enhancing our sensitivity to moral reasons, and successfully managing the consequences of our actions, then these should be our driving concerns – not whether the responsibility system can continue *ad perpetuum* or whether AIs can be morally responsible.

4. AIs and responsibility

The aim of this section is to determine what introducing AIs into societal roles does to the above-stated aims. I focus on the responsibility-related threats identified in the techno-responsibility gap literature and consider their relationship to these aims. I begin by situating my argument with respect to the techno-responsibility gap debate.

4.1. The techno-responsibility gap debate

According to Tigard (2021), the techno-responsibility gap debate divides into two camps: the techno-pessimists and the techno-optimists. Both camps agree that AI creates a gap within the framework of our existing responsibility practices. The former argues that this gap is insurmountable, the latter argues the opposite and offers workarounds. Here I distinguish my argument from three others which, like mine, challenge the premises underlying this debate.

Tigard (2021) argues that machines can be responsible in ways consistent with our already diverse and flexible practices, and so concludes that there is no techno-responsibility gap. Santoni de Sio and Mecacci (2021) argue that discussants in the debate mistakenly lump different types of responsibility together and so make unwarranted assumptions about responsibility gaps. They identify four types of responsibility (culpability, moral accountability, public accountability, and active responsibility) and connect specific uses of AI to each type. Königs (2022) similarly complains that the debate fails to sufficiently qualify the existence or nature of responsibility gaps and, after qualifying them, argues that they are not cause for great concern. While I share these authors' aim of getting explicit about the premises of the debate and their non-alarmist approach, I take a step further back from the debate and question the value of its

subject, the responsibility system, and thereby question the concern driving the debate.

4.2. Control

4.2.1. Responsibility and control

A central claim of techno-responsibility gap discussants is that because AIs are hard to control and their processes largely non-transparent, given that control and foreseeability are crucial to moral responsibility, the use of AIs gives rise to responsibility gaps (Sparrow 2007; Gunkel 2020; Bathaee 2018; Wang, Kaushal, and Khullar 2020; Coeckelbergh 2020). Indeed, while most philosophers reject the possibility of absolute control over our characters or actions, a control condition of some sort is standard for compatibilist accounts (e.g. ‘guidance control,’ per Fischer and Ravizza 2000). This notwithstanding, the essence of moral responsibility is the relationship between an agent and her action, control being but one potential aspect of this relationship. There are alternatives. Attributionists (e.g. Scanlon 1998) are concerned with whether an action can be properly attributed to an agent for the sake of moral assessment on another basis than control, e.g. the agent’s judgment-sensitive attitudes. The sceptics reject the idea that there is any relationship between an agent and her action which would make her morally responsible for it. For Dennett, TR is a means of overcoming our inherent lack of control; increased control is the freedom of which we are capable and which we (ought to) care about.

Interestingly, despite being tied to beliefs about agents’ control, our responsibility practices – via conditioning and prudentially incentivizing norm-compliance, are clearly regulative (hence their social utility). While the ability to self-regulate according to externally imposed norms is a matter of control and may be a necessary condition of responsible agency (McGeer 2015), autonomy requires (minimally) full integration of these norms (Deci and Ryan 2013). This is not a guaranteed outcome of our practices (Harland 2020; Brandenburg 2021; Gogoshin 2023b) and even virtuous agents are dependent on their ongoing motivational scaffolding (McGeer 2019). The resultant picture of responsible agency is one of heteronomous rather than autonomous agency.

So, while we might think that blame and punishment are, in principle, only justified when an agent has sufficient control over her action, our actual practices appear to prioritize something else. This is further reflected in the phenomenon of moral luck as well as in numerous

empirical studies. As to the first, as Williams and Nagel (1976) pointed out, much of moral assessment, assessment which we take to be correct and justified (Nagel 1976), factors in things over which we have little to no control. One type of moral luck, resultant luck, suggests that outcomes often trump control in our moral judgments. Numerous empirical studies (starting with Nahmias et al. 2005), confirmed many times since (per Knobe 2014), reveal our responsibility judgments in specific cases to be independent of agential control. When specific agents engage in concrete, vicious behaviour, we take them to be responsible even when we believe their behaviour was causally determined. There is thus an apparent tension between our theoretical reasoning about control and responsibility on the one hand, and our actual assessments and practices on the other. The techno-responsibility gap debate overlooks this tension.

4.2.2. AI-specific control worries

The responsibility relevant control issue with AIs is that because we lack control over them (insofar as they are operationally autonomous) and they lack the right kind of control for moral responsibility, when harm involving them occurs, there will be no obvious blameworthy party. This gives rise to two kinds of worry. The first kind arises from a perceived threat to the prospective control-related dimension of our practices, for which the function is inculcating responsible behaviour, and the security founded on the reliability and predictability it provides. While complex and somewhat unpredictable, most of us (the set of responsible agents) are nonetheless rather predictable and reliable. We are moved by the norms themselves, by the actual or predictable reactive response directed at norm-relevant behaviour, or a combination of both. When we err, absent an acceptable excuse, we are held to account, by others if not also ourselves. This constitutes the retrospective control-related dimension of our practices. The second kind of worry arises from a threat to the security it provides. Naturally, when we introduce agents who are not predictably moved by shared normative reasons into contexts in which, further, we are unable to pinpoint an obvious blameworthy party, both senses of security appear to be threatened. I argue that the second sense is partly false and can be addressed else how. I address the first kind first.

In order to address the first kind of worry, that of AIs not being regulable (internally or externally) by our norms and thus unreliable, we have two options: (1) to restrict and/or diminish AIs' capacities and/or their uses/contexts, or (2) to equip them with the ability to govern themselves according to our standards. While some theorists support (2) (e.g. Wallach

and Allen 2009; Malle 2016; Riaz et al. 2018), others (e.g. Sparrow 2021) insist that it is a hopeless endeavour (due, per Sparrow, to the nature of normativity). Even supposing the pessimist is right, there are other criteria, e.g. safety – protocols for which could significantly increase predictability (see Hendrycks et al. 2019 for an example¹⁴). While (1) is a conservative approach that risks undermining progress dependent on AIs due precisely to their complexity (assuming their abilities arise from it¹⁵), given a certain amount of well-founded tolerance for uncertainty, following it need not extinguish progress. After all, despite humans' general reliability, we are notoriously morally fallible (Batson 2016). Machines, by contrast, are not plagued by competing, selfish impulses, or the kind of intense, righteous moral emotions which lead to atrocities. Despite their complexity as well as the complexity of the normative landscape, we have reason to think that AIs might be more reliably norm-compliant than humans.¹⁶

This notwithstanding, AIs still lack the ability to self-regulate or self-correct. Despite the 'I' in AI, AIs do not have minds of their own. They are trained according to data sets provided by their engineer-trainers. And they are only as good as the data they get. They do not, despite recent advances, break the 'barrier of meaning' (Mitchell 2019). They cannot judge whether their input or output is sensible or extract meaning from it. This notwithstanding, AIs need not be wildly unpredictable. The error margin and its range is, in principle, knowable (to a degree) and usage restrictions can be made accordingly. More data and further training can be administered until the AI is fit for deployment (i.e. until it has attained a morally acceptable level of risk, as per Hindriks and Veluwenkamp 2023). It is when AIs are deployed prematurely, when they are given authority by human decision-makers, or when their output is taken for granted and used for high stakes decisions, that we have a clearcut threat. This is where strong safety standards and protocols are crucial. I address the remaining uncertainty in §5.

The second kind of security which comes from holding responsible, while it cannot be attained with AIs directly, or so I assume, is partly false. First, we have means in interpersonal and institutional contexts of dealing with wrongdoers that do not depend on their being responsible. With children, the mentally insane or incompetent, and animals, we

¹⁴See the CAIS database of AI safety research at <https://www.safe.ai/research>.

¹⁵See DARPA (2016, 7) for the claim that AI abilities and opaqueness go hand-in-hand.

¹⁶Though there are problems with norm-compliant machines (as pointed out by Coggins and Steinert 2023), norm-compliance promotes cooperation and it might solve the predictability issue.

interact according to certain clearcut boundaries. With strangers, whose reliability is unknown, we also take certain precautions. When members of the first group commit a harm, we do not (at least not legitimately) hold them responsible. We must make peace with this harm in other ways. Second, in holding the responsible responsible, we assume, often falsely, that we can adequately deal with their normative footprint. As discussed, it isn't clear that the demands we make of wrongdoers to TR for their harms are fulfilable, either by them or in principle, due to the agents' specific capacities, or to the nature of the harms themselves. Can we fully repair or meaningfully mitigate harm via an apology, e.g.? By demanding apologies, promises, and reparations, can we reform characters? The experience of guilt and remorse may play a role in moral reformation, but genuine guilt and remorse cannot be demanded. Moreover, someone might comply in future with our normative expectations strictly in order to maintain friendship, a successful business relationship, or an important position. They can be relied upon, in this light, but they have not been morally reformed. It is possible, as previously discussed (in §3), to promote moral disengagement via censure and sanction as well as via praise.

When apologies and promises of reformation do serve to appease the wronged, which I suspect they often do not, we can say that a psychological burden has been lifted from the latter which, in turn, prevents the initial wrongdoing from causing further harm. Absent a fully satisfying response from wrongdoers, we might still derive psychological satisfaction from simply not letting them 'get away with it.' Insofar as this satisfaction relates to retribution, as discussed, it should be resisted. If it comes from signalling and enforcing our normative commitments, then this, as per the consequentialist, does not depend on an agent's deserving a blame response in any deep sense. As Pereboom has shown, we can get the forward-looking benefits without basic desert responsibility. Still, one might think, the problem is not that AIs lack basic desert responsibility; it is that we cannot achieve even the forward-looking benefits of holding responsible with them.

Pereboom's (2014) desiderata of reconciliation, moral (re)formation, and future protection are tied to non-reactive emotional responses like moral sorrow and disappointment, and they depend, for uptake, on an agent's reasons-responsiveness. Supposing AIs can be equipped with the relevant affective faculties (see Daily et al. 2017 for related research) and, via reinforcement learning (see Wallach and Allen 2009 for an example), be functionally influenceable on this basis, the non-

functionalist will presumably remain dissatisfied. One way or the other, we might assume that a psychological gap will persist. So, assuming that holding an AI responsible for forward-looking purposes does not give us the satisfaction we may derive through holding the responsible responsible (assuming that satisfaction is worth preserving), there is no *a priori* reason to think we must get it from the AI or else go without. Via scapegoating, blaming involved (however indirectly) human parties (Champagne and Tonkens 2015), or, more innocuously (per Kiener 2022), having involved human parties TR *qua* answerability, we might less or more legitimately pursue this satisfaction. I will propose a different kind of solution in §5.

4.3. Transparency and moral agency

The issue of transparency and AI is closely related to control and, moreover, there are different kinds of AI transparency related issues (Wadden 2022). Here I will focus on the issue as pertains to moral decision-making. Responsibility judgments are concerned with the relationship between an agent and her action. Motives matter, especially in theory. Our practices, as discussed, have trouble discriminating among motives. Our own minds, we might say, are black boxes; we are excellent post-hoc rationalizers (Haidt 2001), and yet, explanations matter greatly for quotidian to high-stakes matters, especially when things go wrong. We want to know why our friend was late to our meeting, why the doctor gave this vs. that diagnosis, why the emergency clinic neglected a patient's crucial symptom, etc. Despite the possibility of a false explanation, it matters to us that people answer for their actions. With at least some AIs, due to their complexity, it is generally thought that these answers are not even in principle possible (Burrell 2016). How can we trust the output of something whose reasoning process is a black box? How can we trust that it will make a moral decision? It is this second question I'm primarily concerned with here.

Given the opacity of our own minds and hearts, what is it that allows us to trust human military drone operators, e.g. and not autonomous AI-operated military drones? Empirical studies strongly suggest that we perceive human behaviour as the result of choice based on reasons, whereas we perceive computer behaviour as the result of causal processes, not reasons (Knobe 2014). But why should that fact give rise to less trust? In a compelling news article, psychologists Crockett, Everett, and Pizarro (2017), provide a possible explanation. We tend to trust, they

say the research suggests, people who are guided by social-emotional commitments and priorities, rather than cost-benefit analyses. AIs, they say, lack the features on which we base trust. 'In our fellow humans [...] we prefer those whose moral decisions are guided by social emotions like guilt and empathy.' AIs, on the other hand, or so we believe, do not act for (their own) reasons; their actions are caused by their programming (Knobe 2014). 'Even if machines were able to perfectly mimic human moral judgments, we would know that the computer did not arrive at its judgments for the same reasons we would' (Crockett, Everett, and Pizarro 2017).

Interestingly, this thesis suggests what the empirical studies on responsibility judgments do (what matters most is the moral weight of the action or outcome rather than the agent's control over it); it is a distinctly moral issue. We want these decisions to be made in the same way that we make them, or at least we want that possibility – by way of distinctly moral, emotional reactions and intuitive responses (per Crockett et al.). If AIs were acting according to their programming, though, we wouldn't have the above worries about control and transparency. There is a seeming contradiction here. On the one hand, autonomous AIs are hard to control, predict, and explain, and this gives rise to fears about how they will behave. On the other hand, we are dissatisfied with the fact that their actions are determined rather than (moral) reasons-sensitive.

One hypothesis that could explain both is the following. What we fear about AIs is that they do not have an autonomous capacity for moral reasoning; they can, at best, mimic ours.¹⁷ Supposing they could mimic ours, would this not be good enough? From a functionalist or consequentialist perspective, it surely would. It wouldn't be good enough for someone who cares about why agents do what they do (is it because of empathy or programming?) or that what they do reflects their own identity, manifests *their own* intentional agency. This is the issue fuelling the unending debate about moral responsibility and free will. In order to ever be truly praise- or blameworthy for something, it must be the product of our own agency, in light of reflectively endorsed reasons (c.f. Dworkin 2011). AIs, unless they have the capacity for autonomous moral reasoning, are never praise- or blameworthy. They are heteronomous moral agents at best. We could of course invoke the sceptical

¹⁷Some might fear that if AIs become autonomous in this sense, they might endorse reasons which are at odds with human well-being, but I think this fear is misguided.

argument here to say that humans are also never legitimately praise- or blameworthy. I will take another route.

4.3.1. *Moral decision-making*

I have argued that, from the standpoint of our practices, responsible agents need not be morally autonomous agents (Gogoshin 2021; Gogoshin 2023b).¹⁸ Responsible agents need only respond to the reasons presented by our reactive responses (e.g. avoidance of blame, pursuit of praise) to comply with normative demands rather than for the (moral) reasons which justify these reactive responses. It is hoped, even assumed (Vargas 2013; McGeer 2019) that via reactive exchanges, we tend to internalize these moral reasons but, as I have argued, this outcome is not guaranteed. When it comes to the societal function of the responsibility system, moreover – enabling cooperation – norm-compliance appears to be (taken as) primary and possibly sufficient.

Importantly however, responsible agents, unlike AIs, can and often do act morally autonomously. Furthermore, there might be a visceral connection between human agency and moral reasons (an emotional basis, e.g. for moral reasons-responsiveness) which can only be functionally replicated, at best, in AIs. The fact remains, however, that we are often not moved by the relevant moral reasons and are, to repeat the earlier point, highly morally fallible. If our chief concern is having the capacity to be moved directly by moral reasons, I think it is mostly for reasons relating to the value of moral worth and moral autonomy. As they take us beyond the scope of responsible agency (Gogoshin 2023b; Fischer 2022), they render the ultimate worry about AI and morality about something other than responsibility.

This concern owes, I suggest, to the relationship between moral autonomy and trustworthiness. Ultimate trustworthiness (vs. mere reliability, e.g.) is, I assume, tied to a capacity for robust moral success. Robust moral success, in turn, depends, in first part, on the ability to overcome both internal and external competing interests. It also depends on the ability to act well in new and various contexts (“moral flexibility”) rather than the ability to adhere to a rigid set of rules and principles (Bartels 2008). An independent and robust capacity for responding directly to moral reasons (moral autonomy) supports both abilities. So, moral autonomy is worth striving for, but it is not a reasonable bar to set for AI.

¹⁸I understand moral autonomy as being ultimately responsive to the moral reasons directly. Korsgaard (1996, 22) takes it to be a matter of acting on a law one gives to oneself. See Fischer (2022) for the distinction between responsible agency and autonomy.

5. A way forward

Returning now to the unresolved worry that we cannot get the same degree or kind of satisfaction in morally relevant situations involving AI, my first response is that we often don't get it from putatively clearcut responsible human wrongdoers either, and so very often, we have to try to make peace else how. In addition to reasons already discussed, this is because even when we perceive responsibility as clearcut, we are never seeing the whole picture (causal chain) and so falsely believe that the problem is limited to the individual. Since it is not, addressing the individual is never a complete solution. Second, AIs, unlike humans, can be readily reformed, and many of their concrete harms – *to no less a degree* than anthropogenic harms – repaired. Making peace with the remaining psychological gap relating to holding responsible, as well as the first kind of worry relating to prospective responsibility (reliability) previously discussed, is conceivable, I think, under certain conditions.

The way forward is, after all, a matter of control and transparency, but of a different sort than that which is often taken to be at stake in the 'blame game.' It is a matter, I suggest, of the legitimate stakeholders¹⁹ taking front-end control. To do so, these stakeholders must fully endorse the AI's deployment. This endorsement will require transparency concerning, minimally, its functionality, its success vs. error rate and range, as well as moral consent. This consent is, in turn, a function of the degree to which the AI's role is necessary²⁰ and contributes to the public interest (along the lines proposed by, e.g. Floridi et al. 2020; Züger and Asghari 2023). With these conditions in place, we can manage the uncertainty (indeterminacy) attached to the nature of AI and address the possibly more pressing threats posed by potentially negligent or otherwise bad actors who might otherwise deploy AIs prematurely, in the wrong contexts, to the wrong ends, with undue freedom or authority, etc.

By taking the right kind of front-end control, involving the right actors and the right agenda, we are in effect exercising Waller's 'take-charge responsibility,' the kind of control which is intrinsically rewarding and for which we are ready and willing to manage the consequences. Rather than exercising this responsibility as individuals, however, we do

¹⁹Legitimate stakeholders' must be carefully specified beyond mere standard usage – e.g. so as to include or represent all who are all in any significant way involved on the front or back ends of the tech, though I do not undertake to further qualify this herein. Current AI-related decisions do not consistently involve the legitimate stakeholders (Bélisle-Pipon et al. 2023).

²⁰This term also needs qualification that goes beyond the scope of this paper to provide.

so at a group level: shared aim, shared consent, shared control, shared responsibility. Exercising take-charge responsibility does not preclude the kind of mistakes for which we may wish to hold one another to account, but it corresponds to the kind of prospective control we're after when it comes to harm minimization, trust, and reliability. Insofar as blame and shame inhibit take-charge responsibility and moral reasons-responsiveness and block an understanding of the systemic factors involved in harm and wrongdoing, the small gain on retrospective security they may provide is outweighed by the losses. Moreover, the true goods of holding responsible – reparations, reformation, reconciliation, protection, and positive normative landscaping – are not clearly thereby enhanced and may be achieved else how.

Waller's previously cited 'no-blame systems approach' examples shed light onto how collaboration can be enhanced in the absence of individual moral responsibility attributions and reactive responses. Notably, this approach is not just a matter of not blaming individuals, it amounts to transforming the nature of the game. In Waller's examples, by implementing the systems approach rather than blaming and sanctioning individual workers for problems, the deeper causes of these problems surfaced and were successfully addressed. Under the threat of sanction (even just in the form of negative social feedback), people are often motivated to conceal their mistakes and/or to redirect blame to scapegoats. This is highly counterproductive to the end goal of progress and cooperation. It fuels internal, possibly ruthless competition instead, giving rise to mistrust and ill will. Recall that the aim of responsibility in ethics and in society writ large is cooperation, not just deserts.

Cooperation is built on trust, reliability, and respect. If AIs are helpful tools (or, potentially, collaborative partners) which support our goals, we have an overwhelming, positive reason to engage with them and to make them better.²¹ Mistakes and negative outcomes need not be blamed on individuals, but rather taken as opportunities for improvement. Tolerance for negative outcomes is only possible when we view them as inevitable and necessary, within careful limits, for progress. It is only when technologies are aimed at progress – in accordance with human values and the public interest – that we can view them in this light, maintaining caution without overwhelming fear.

²¹See Nyholm (2018) for a proposal on maintaining front-end control by way of restricting AI development/deployment to human-robot collaborations.

5.1. Objections and clarifications

The responsibility system alternative proposed herein excludes both just deserts and forward-looking blame. Here I address three possible objections. The first is that of the consequentialist. Insofar as our reactive practices generate favourable outcomes, their justification does not rely on basic desert moral responsibility; so why remove them? My reply is that (a) the favourable outcomes the consequentialists care about (social regulation) are either not necessarily the right aim (meaningful collaboration and moral reasons-sensitivity being better) or can be achieved without reactive practices and (b) since I assume we cannot legitimately hold AIs responsible, these practices are ineffectual (which, because of (a), I see as unproblematic).

Another objection is that we should retain a place for praise and that AIs' lack of responsibility precludes it. 'Social psychological evidence suggests that [...] specific expressions of praise positively contribute to agents' non-instrumental motivation to pursue the praised activity' (Telech 2022, 2). As suggested earlier in the paper, however, praise can misfire too and so, like blame, does not warrant blind rescue efforts. Still, provided the above conditions regarding front-end control and transparency are met, I do not see why the sense of achievement or promotion of right values and outcomes that praise stands to promote would be strictly precluded in AI-human contexts. A collective praise/credit system might nevertheless be more feasible than an individualistic one. The guideline for our practices I propose is simply the conscientious promotion of the right values and outcomes rather than the blind perpetuation of a flawed system.

The third objection is that my positive proposal is an idealization which, although we have evidence of the no-blame systems approach's effectiveness, requires excessive revision to current ways of thinking and doing. In reply, I clarify that my principal aim in this paper has been to shed light on the ineffectiveness of these ways for what we currently (and ideally) care about, which is not threatened by AI directly. I hope thereby to open the door to proposals which could get us closer to what we care about rather than to argue for a single way forward. Moreover, the no-blame systems approach is an example of an alternative to the current system, whereas my positive proposal is a matter of identifying the location and kind of control needed to steer AI responsibly. I should also clarify, however, that in light of the limitations of the responsibility system, the existence of alternatives for achieving the true goods

of this system, and our interest in creating and using AI to improve the human condition (which will continue to result in drastic societal changes), I consider drastic revisions to our current ways, assuming they're possible, to be well warranted.²²

6. Conclusion

In this paper, I have argued that the techno-responsibility gap debate is founded on shaky ground – on implicit and unwarranted beliefs about the positivity and necessity of the responsibility system. I have attempted to shine a light on the functions and outcomes, both good and bad, of this system, and to consider the relationship between them and the presence of AIs. The true goods associated with the responsibility system – harm minimization, reliability and cooperation, addressing one's normative footprint, communicating and shaping the normative landscape – as well as the things the responsibility system does not (reliably) foster – enhanced control, increased sensitivity to moral reasons, mutual respect, well-founded trust, meaningful collaboration, solutions to the wide-reaching roots of wrongdoing – are not threatened by AIs themselves. By taking the right kind of front-end control of AI development and deployment, we can tolerate inevitable mistakes, jointly take charge of the consequences of these mistakes, and avoid the avoidable mistakes originating in the very human actors and agendas behind AI.

Acknowledgements

This paper is dedicated to the memory of Bruce N. Waller (1946-2023), a philosopher with a heart as expansive as his mind, in appreciation of his mentorship, influence on the field of moral responsibility and free will, and on this paper in particular, and his efforts to make the world a better place. I would also like to thank an anonymous reviewer for this journal for their generous and helpful suggestions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Kone Foundation [grant no 201906341].

²²We have further evidence that they are. Pickard's (2017) revisionist approach to blame, e.g. has proven to be both empirically possible and effective.

ORCID

Dane Leigh Gogoshin  <http://orcid.org/0000-0002-6534-0714>

References

- Arpaly, Nomy. 2003. *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford: Oxford University Press.
- Bagnoli, Carla. 2021. "Disclaiming Responsibility, Voicing Disagreements, and Negotiating Boundaries." In *Oxford Studies in Agency and Responsibility*, edited by David Shoemaker. Oxford, New York: Oxford University Press.
- Bartels, Daniel M. 2008. "Principled Moral Sentiment and the Flexibility of Moral Judgment and Decision Making." *Cognition* 108 (2): 381–417. doi:10.1016/j.cognition.2008.03.001.
- Bathae, Yavar. 2018. "The Artificial Intelligence Black Box and the Failure of Intent and Causation." *Harvard Journal of Law and Technology* 31 (2018): 889.
- Batson, C. Daniel. 2016. *What's Wrong with Morality? A Social-Psychological Perspective*. Oxford, New York: Oxford University Press.
- Bélisle-Pipon, Jean-Christophe, Erica Monteferrante, Marie-Christine Roy, and Vincent Couture. 2023. "Artificial Intelligence Ethics Has a Black Box Problem." *AI & SOCIETY* 38 (4): 1507–1522. doi:10.1007/s00146-021-01380-0.
- Bicchieri, Cristina. 2017. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. New York, NY: Oxford University Press.
- Bodek, Norman. 2011. "Solving Toyota's Quality Problems." *Quality Digest*. <http://www.qualitydigest.com/inside/management-article/zenjidoka-solving-toyotas-quality-problems-021411.html>.
- Bok, Hilary. 1998. *Freedom and Responsibility*. Princeton, N.J.: Princeton University Press.
- Brandenburg, Daphne. 2021. "Consequentialism and the Responsibility of Children: A Forward-Looking Distinction between the Responsibility of Children and Adults." *The Monist* 104 (4): 471–483. <http://doi.org/10.1093/monist/onab013>.
- Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1): 205395171562251. doi:10.1177/2053951715622512.
- Caruso, Gregg D. 2017. "The Public Health-Quarantine Model." *SSRN Electronic Journal*, doi:10.2139/ssrn.3068021.
- Caruso, Gregg D. 2021. *Rejecting Retributivism: Free Will, Punishment, and Criminal Justice*. Law and the Cognitive Sciences. Cambridge, United Kingdom; New York, NY: Cambridge University Press.
- Caruso, Gregg D., and Stephen G. Morris. 2017. "Compatibilism and Retributivist Desert Moral Responsibility: On What Is of Central Philosophical and Practical Importance." *Erkenntnis* 82 (4): 837–855. doi:10.1007/s10670-016-9846-2.
- Champagne, Marc, and Ryan Tonkens. 2015. "Bridging the Responsibility Gap in Automated Warfare." *Philosophy & Technology* 28 (1): 125–137. doi:10.1007/s13347-013-0138-3.
- Coeckelbergh, Mark. 2020. "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability." *Science and Engineering Ethics* 26 (4): 2051–2068. doi:10.1007/s11948-019-00146-8.

- Coggins, Tom N., and Steffen Steinert. 2023. "The Seven Troubles with Norm-Compliant Robots." *Ethics and Information Technology* 25 (2): 29. doi:10.1007/s10676-023-09701-1.
- Crockett, Molly, Jim Everett, and David Pizarro. 2017. "Why Are We Reluctant to Trust Robots?" *The Guardian*, April 24, sec. Science. <https://www.theguardian.com/science/head-quarters/2017/apr/24/why-are-we-reluctant-to-trust-robots>.
- Curry, Oliver Scott, Daniel Austin Mullins, and Harvey Whitehouse. 2019. "Is It Good to Cooperate? Testing the Theory of Morality-as-Cooperation in 60 Societies." *Current Anthropology* 60 (1): 47–69. doi:10.1086/701478.
- Daily, Shaundra B., Melva T. James, David Cherry, John J. Porter, Shelby S. Darnell, Joseph Isaac, and Tania Roy. 2017. "Affective Computing: Historical Foundations, Current Applications, and Future Trends." In *Emotions and Affect in Human Factors and Human-Computer Interaction*, edited by Jeon, Myounghoon, 213–231. London: Elsevier. doi:10.1016/B978-0-12-801851-4.00009-4.
- Danaher, John. 2016. "Robots, Law and the Retribution Gap." *Ethics and Information Technology* 18 (4): 299–309. doi:10.1007/s10676-016-9403-3.
- Danaher, John. 2022. "Tragic Choices and the Virtue of Techno-Responsibility Gaps." *Philosophy & Technology* 35 (2): 26. doi:10.1007/s13347-022-00519-1.
- DARPA (Defense Advanced Research Projects Agency). 2016. "Explainable Artificial Intelligence (XAI)." *DARPA-BAA-16-53*. <https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf>.
- Darwall, Stephen L. 2006. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, Mass: Harvard University Press.
- Deci, Edward L., and Richard M. Ryan. 2013. "The Importance of Autonomy for Development and Well-Being." In *Self-Regulation and Autonomy*, 1st ed., edited by Bryan W. Sokol, Frederick M. E. Grouzet, and Ulrich Müller, 19–46. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139152198.005.
- Dennett, D. C. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, Mass: MIT Press.
- Dennett, D. C. 2003. *Freedom Evolves*. New York: Viking.
- Dennett, D. C. 2015. *Elbow Room: The Varieties of Free Will Worth Wanting*, New edition. Cambridge: MIT Press.
- Duff, Antony. 2001. *Punishment, Communication, and Community*. Studies in Crime and Public Policy. Oxford ; New York: Oxford University Press.
- Duff, Antony. 2009. "Legal and Moral Responsibility." *Philosophy Compass* 4 (6): 978–986. doi:10.1111/j.1747-9991.2009.00257.x.
- Dworkin, Ronald. 2011. *Justice for Hedgehogs*. Harvard Univ. Press paperback ed. Cambridge, Mass.: Belknap Press of Harvard University Press.
- Enoch, David. 2012. "Being Responsible, Taking Responsibility, and Penumbral Agency." In *Luck, Value, and Commitment*, edited by Ulrike Heuer, and Gerald Lang, 95–132. New York: Oxford University Press. doi:10.1093/acprof:oso/9780199599325.003.0005.
- Fischer, John Martin. 2022. "What Moral Responsibility Is Not." In *Thick (Concepts of) Autonomy* Vol. 146, edited by James F. Childress, and Michael Quante, 1–16. Cham: Springer International Publishing. doi:10.1007/978-3-030-80991-1_1.

- Fischer, John Mark, and Martin Ravizza. 2000. *Responsibility and Control: A Theory of Moral Responsibility*. First paperback ed. *Cambridge Studies in Philosophy and Law*. Cambridge: Cambridge University Press.
- Floridi, Luciano, Josh COWLS, Thomas C. King, and Mariarosaria Taddeo. 2020. "How to Design AI for Social Good: Seven Essential Factors." *Science and Engineering Ethics* 26 (3): 1771–1796. doi:10.1007/s11948-020-00213-5.
- Frankfurt, Harry G. 1975. "Three Concepts of Free Action: Part 2." *Aristotelian Society: Supplementary Volume* 49: 113–125.
- Fricker, Miranda. 2016. "What's the Point of Blame? A Paradigm Based Explanation." *Noûs* 50 (1): 165–183. doi:10.1111/nous.12067.
- Goetze, Trystan S. 2022. "Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement." In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 390–400. Seoul Republic of Korea: ACM. doi:10.1145/3531146.3533106.
- Gogoshin, Dane Leigh. 2021. "Robot Responsibility and Moral Community." *Frontiers in Robotics and AI* 8. <https://www.frontiersin.org/articles/10.3389/frobt.2021.768092>.
- Gogoshin, Dane Leigh. 2023a. "Challenging the Premises of the Techno-Responsibility Gap." In *Social Robots in Social Institutions*, edited by Raul Hakli, Pekka Mäkelä, and Johanna Seibt, 560–567. Amsterdam: IOS Press. doi:10.3233/FAIA220659.
- Gogoshin, Dane Leigh. 2023b. "A Challenge for the Scaffolding View of Responsibility." *Ethical Theory and Moral Practice* 26 (1): 73–90. doi:10.1007/s10677-022-10340-6.
- Gunkel, David J. 2020. "Mind the Gap: Responsible Robotics and the Problem of Responsibility." *Ethics and Information Technology* 22 (4): 307–320. doi:10.1007/s10676-017-9428-2.
- Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108 (4): 814–834. doi:10.1037/0033-295X.108.4.814.
- Harland, Harry. 2020. "Beyond the Moral Influence Theory? A Critical Examination of Vargas's Agency Cultivation Model of Responsibility." *The Journal of Ethics* 24 (4): 401–425. doi:10.1007/s10892-020-09328-0.
- Harris, Don, and Helen C. Muir, eds. 2005. *Contemporary Issues in Human Factors and Aviation Safety*. Aldershot, Hants, England ; Burlington, VT, USA: Ashgate.
- Hart, H. L. A. 2008. *Punishment and Responsibility: Essays in the Philosophy of Law*. 2nd ed. Oxford, New York: Oxford University Press.
- Hendrycks, Dan, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2019. *AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty*. <https://doi.org/10.48550/ARXIV.1912.02781>.
- Hieronymi, Pamela. 2004. "The Force and Fairness of Blame." *Philosophical Perspectives* 18: 115–148. <https://www.jstor.org/stable/3840930>.
- Himmelreich, Johannes. 2019. "Responsibility for Killer Robots." *Ethical Theory and Moral Practice* 22 (3): 731–747. doi:10.1007/s10677-019-10007-9.
- Hindriks, Frank, and Herman Veluwenkamp. 2023. "The Risks of Autonomous Machines: From Responsibility Gaps to Control Gaps." *Synthese* 201 (1): 21. doi:10.1007/s11229-022-04001-5.
- Holroyd, Jules. 2021. "Oppressive Praise." *Feminist Philosophy Quarterly* 7 (4), <https://ojs.lib.uwo.ca/index.php/fpq/article/view/13967>.

- Honderich, Ted. 2002. *How Free Are You? The Determinism Problem*. 2nd ed. Oxford, New York: Oxford University Press.
- Jeppsson, Sofia, and Daphne Brandenburg. 2022. "Patronizing Praise." *The Journal of Ethics* 26 (4): 663–682. doi:10.1007/s10892-022-09409-2.
- Kane, Robert. 2007. "Libertarianism." In *Four Views on Free Will*, edited by Robert Kane, John Martin Fischer, Derk Pereboom, and Manuel Vargas, 5–43. Malden, MA: Blackwell Publishing.
- Kiener, Maximilian. 2022. "Can We Bridge AI's Responsibility Gap at Will?" *Ethical Theory and Moral Practice* 25 (4): 575–593. doi:10.1007/s10677-022-10313-9.
- Knobe, Joshua. 2014. "Free Will and the Scientific Vision." In *Current Controversies in Experimental Philosophy*, edited by Edouard Machery and Elizabeth O'Neill. New York: Routledge.
- Königs, Peter. 2022. "Artificial Intelligence and Responsibility Gaps: What Is the Problem?" *Ethics and Information Technology* 24 (3): 36. doi:10.1007/s10676-022-09643-0.
- Korsgaard, Christine M. 1996. *Creating the Kingdom of Ends*. Cambridge; New York, NY, USA: Cambridge University Press.
- Kraaijeveld, Steven R. 2020. "Debunking (the) Retribution (Gap)." *Science and Engineering Ethics* 26 (3): 1315–1328. doi:10.1007/s11948-019-00148-6.
- Lang, Benjamin H., Sven Nyholm, and Jennifer Blumenthal-Barby. 2023. "Responsibility Gaps and Black Box Healthcare AI: Shared Responsibilization as a Solution." *Digital Society* 2 (3): 52. doi:10.1007/s44206-023-00073-z.
- Levy, Neil. 2011. *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. 1. publ. Oxford: Oxford Univ. Press.
- Mackenzie, C. 2021. "Culpability, Blame, and the Moral Dynamics of Social Power." *Proceedings of Aristotelian Society Supplementary* 95 (1): 163–182.
- Macnamara, Coleen. 2015. "Blame, Communication, and Morally Responsible Agency." In *The Nature of Moral Responsibility*, edited by Randolph Clarke, Michael McKenna, and Angela M. Smith, 211–236. New York: Oxford University Press. doi:10.1093/acprof:oso/9780199998074.003.0010.
- Malle, Bertram F. 2016. "Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots." *Ethics and Information Technology* 18 (4): 243–256. doi:10.1007/s10676-015-9367-8.
- Mason, Elinor. 2018. *Respecting Each Other and Taking Responsibility for Our Biases*. Vol. 1. New York: Oxford University Press. doi:10.1093/oso/9780190609610.003.0007.
- Matthias, Andreas. 2004. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6 (3): 175–183. doi:10.1007/s10676-004-3422-1.
- McGeer, Victoria. 2015. "Mind-Making Practices: The Social Infrastructure of Self-Knowing Agency and Responsibility." *Philosophical Explorations* 18 (2): 259–281. doi:10.1080/13869795.2015.1032331.
- McGeer, Victoria. 2019. "Scaffolding Agency: A Proleptic Account of the Reactive Attitudes." *European Journal of Philosophy* 27 (2): 301–323. doi:10.1111/ejop.12408.
- McKenna, Michael. 2012. *Conversation and Responsibility*. New York: Oxford University Press.

- McKenna, Michael. 2020. "Punishment and the Value of Deserved Suffering." *Public Affairs Quarterly* 34 (2): 97–123. doi:10.2307/26921122.
- Milam, Per-Erik. 2021. "Get Smart: Outcomes, Influence, and Responsibility." *The Monist* 104 (4): 443–457. doi:10.1093/monist/onab011.
- Mitchell, Melanie. 2019. "Artificial Intelligence Hits the Barrier of Meaning." *Information* 10 (2): 51. doi:10.3390/info10020051.
- Moore, Michael S. 1997. *Placing Blame: A General Theory of the Criminal Law*. Oxford: New York: Clarendon Press; Oxford University Press.
- Morris, Herbert, and The Hegeler Institute. 1968. "Persons and Punishment: Edited by Sherwood J. B. Sugden." *The Monist* 52 (4): 475–501. doi:10.5840/monist196852436.
- Munch, Lauritz, Jakob Mainz, and Jens Christian Bjerring. 2023. "The Value of Responsibility Gaps in Algorithmic Decision-Making." *Ethics and Information Technology* 25 (1): 21. doi:10.1007/s10676-023-09699-6.
- Nagel, Thomas. 1976. "Moral Luck." in *Proceedings of the Aristotelian Society Supplementary* 50: 137–151.
- Nahmias, Eddy, Stephen Morris, Thomas Nadelhoffer, and Jason Turner. 2005. "Surveying Freedom: Folk Intuitions About Free Will and Moral Responsibility." *Philosophical Psychology* 18 (5): 561–584. doi:10.1080/09515080500264180.
- Nowell-Smith, P. 1948. "IV.—Freewill and Moral Responsibility." *Mind; A Quarterly Review of Psychology and Philosophy* LVII (225): 45–61. doi:10.1093/mind/LVII.225.45.
- Nyholm, Sven. 2018. "Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci." *Science and Engineering Ethics* 24 (4): 1201–1219. doi:10.1007/s11948-017-9943-x.
- Pereboom, Derk. 2014. *Free Will, Agency, and Meaning in Life*. First edition. Oxford: Oxford University Press.
- Pettit, Philip, and Kinch Hoekstra. 2018. *The Birth of Ethics: Reconstructing the Role and Nature of Morality*. New York, NY: Oxford University Press.
- Pickard, Hanna. 2017. "Responsibility Without Blame for Addiction." *Neuroethics* 10 (1): 169–180. doi:10.1007/s12152-016-9295-2.
- Riaz, Faisal, Sohail Jabbar, Muhammad Sajid, Mudassar Ahmad, Kashif Naseer, and Nouman Ali. 2018. "A Collision Avoidance Scheme for Autonomous Vehicles Inspired by Human Social Norms." *Computers & Electrical Engineering* 69 (July): 690–704. doi:10.1016/j.compeleceng.2018.02.011.
- Robillard, Michael. 2018. "No Such Thing as Killer Robots." *Journal of Applied Philosophy* 35 (4): 705–717. doi:10.1111/japp.12274.
- Sabatini, Nicholas A. 2008. *Reaching the Next Level of Aviation Safety*. FAA Team News (Federal Aviation Administration – FAATeam – FAASafety.gov).
- Santoni de Sio, Filippo, and Giulio Mecacci. 2021. "Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them." *Philosophy & Technology* 34 (4): 1057–1084. doi:10.1007/s13347-021-00450-x.
- Scanlon, Thomas. 1998. *What We Owe to Each Other*. Cambridge, Mass: Belknap Press of Harvard University Press.
- Scanlon, Thomas. 2008. *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, Mass: Belknap Press of Harvard University Press.
- Schlick, Moritz. 1962. *Problems of Ethics*. New York: Dover Publications.

- Shoemaker, David. 2007. "Moral Address, Moral Responsibility, and the Boundaries of the Moral Community." *Ethics* 118 (1): 70–108. doi:[10.1086/521280](https://doi.org/10.1086/521280).
- Shoemaker, David, and Manuel Vargas. 2021. "Moral Torch Fishing: A Signaling Theory of Blame." *Noûs* 55 (3): 581–602. doi:[10.1111/nous.12316](https://doi.org/10.1111/nous.12316).
- Sie, Maureen. 2018. *Sharing Responsibility*. Vol. 1. New York: Oxford University Press. doi:[10.1093/oso/9780190609610.003.0013](https://doi.org/10.1093/oso/9780190609610.003.0013).
- Sliwa, Paulina. 2016. "Moral Worth and Moral Knowledge." *Philosophy and Phenomenological Research* 93 (2): 393–418. <https://www.jstor.org/stable/48578736>.
- Sliwa, Paulina. 2019. "Reverse-Engineering Blame." *Philosophical Perspectives* 33 (1): 200–219. doi:[10.1111/phpe.12131](https://doi.org/10.1111/phpe.12131).
- Sliwa, Paulina. 2023. "Taking Responsibility." In *New Conversations in Philosophy, Law, and Politics*, edited by Ruth Chang, and Amia Srinivasan. New York: Oxford University Press.
- Smart, J. J. C. 1961. "I-Free-Will, Praise and Blame." *Mind; A Quarterly Review of Psychology and Philosophy* LXX (279): 291–306. doi:[10.1093/mind/LXX.279.291](https://doi.org/10.1093/mind/LXX.279.291).
- Smilansky, Saul. 2000. *Free Will and Illusion*. Oxford, New York: Clarendon Press ; Oxford University Press.
- Smith, Angela M. 2012. "Moral Blame and Moral Protest." In *In Blame*, edited by D. Justin Coates, and Neal A. Tognazzini, 27–48. New York: Oxford University Press. doi:[10.1093/acprof:oso/9780199860821.003.0002](https://doi.org/10.1093/acprof:oso/9780199860821.003.0002).
- Snoek, Anke, Victoria McGeer, Daphne Brandenburg, and Jeanette Kennett. 2021. "Managing Shame and Guilt in Addiction: A Pathway to Recovery." *Addictive Behaviors* 120 (September): 106954. doi:[10.1016/j.addbeh.2021.106954](https://doi.org/10.1016/j.addbeh.2021.106954).
- Sparrow, Robert. 2007. "Killer Robots." *Journal of Applied Philosophy* 24 (1): 62–77. doi:[10.1111/j.1468-5930.2007.00346.x](https://doi.org/10.1111/j.1468-5930.2007.00346.x).
- Sparrow, Robert. 2021. "Why Machines Cannot Be Moral." *AI & Society* 36 (3): 685–693. doi:[10.1007/s00146-020-01132-6](https://doi.org/10.1007/s00146-020-01132-6).
- Spinoza, Benedictus de, and E. M. Curley. 1985. *The Collected Works of Spinoza*. Princeton, NJ: Princeton University Press.
- Springer, Elise. 2013. *Communicating Moral Concern: An Ethics of Critical Responsiveness*. Cambridge, Mass: MIT Press.
- Stichter, Matt. 2020. "Learning from Failure: Shame and Emotion Regulation in Virtue as Skill." *Ethical Theory and Moral Practice* 23 (2): 341–354. <http://doi.org/10.1007/s10677-020-10079-y>.
- Strawson, Galen. 1994. "The Impossibility of Moral Responsibility." *Philosophical Studies* 75 (1–2): 5–24. doi:[10.1007/BF00989879](https://doi.org/10.1007/BF00989879).
- Strawson, P. F. 2008. *Freedom and Resentment and Other Essays*. London, New York: Routledge.
- Talbert, Matthew. 2012. "Moral Competence, Moral Blame, and Protest." *The Journal of Ethics* 16 (1): 89–109. doi:[10.1007/s10892-011-9112-4](https://doi.org/10.1007/s10892-011-9112-4).
- Telech, Daniel. 2022. "Praise." *Philosophy Compass* 17 (10): e12876. doi:[10.1111/phc3.12876](https://doi.org/10.1111/phc3.12876).
- Tigard, Daniel W. 2021. "Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible." *Cambridge Quarterly of Healthcare Ethics* 30 (3): 435–447.

- Tomasello, Michael. 2015. *A Natural History of Human Morality*. Cambridge, Massachusetts: Harvard University Press.
- Vargas, Manuel. 2007. "Revisionism." In *Four Views on Free Will*, edited by John Martin Fischer, Robert Kane, Derk Pereboom, and Manuel Vargas, 126–165. Malden, MA: Blackwell Publishing.
- Vargas, Manuel. 2013. *Building Better Beings: A Theory of Moral Responsibility*. 1st ed. Oxford: Oxford University Press.
- Wadden, Jordan Joseph. 2022. "Defining the Undefinable: The Black Box Problem in Healthcare Artificial Intelligence." *Journal of Medical Ethics* 48 (10): 764–768. doi:[10.1136/medethics-2021-107529](https://doi.org/10.1136/medethics-2021-107529).
- Wallace, R. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.
- Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. Oxford, New York: Oxford University Press.
- Waller, Bruce N. 1998. *The Natural Selection of Autonomy*. SUNY Series, Philosophy and Biology. Albany, NY: State University of New York Press.
- Waller, Bruce N. 2011. *Against Moral Responsibility*. Cambridge, Mass: MIT Press.
- Waller, Bruce N. 2014. *The Stubborn System of Moral Responsibility*. Cambridge, Massachusetts: MIT Press.
- Waller, Bruce N. 2020. "Beyond Moral Responsibility to a System That Works." *Neuroethics* 13 (1): 5–12. doi:[10.1007/s12152-017-9351-6](https://doi.org/10.1007/s12152-017-9351-6).
- Waller, Bruce N. 2023. "Responsibility Without Blame." In *The Routledge Handbook of Philosophy of Responsibility*, edited by Maximilian Kiener. Abingdon: Routledge.
- Wang, Fei, Rainu Kaushal, and Dhruv Khullar. 2020. "Should Health Care Demand Interpretable Artificial Intelligence or Accept "Black Box" Medicine?" *Annals of Internal Medicine* 172 (1): 59. doi:[10.7326/M19-2548](https://doi.org/10.7326/M19-2548).
- Watson, Gary. 2004. *Agency and Answerability: Selected Essays*. 1st ed. Oxford: Oxford University Press. doi:[10.1093/acprof:oso/9780199272273.001.0001](https://doi.org/10.1093/acprof:oso/9780199272273.001.0001).
- Williams, B. A. O., and T. Nagel. 1976. "Moral Luck." *Aristotelian Society Supplementary Volume* 50 (1): 115–152. doi:[10.1093/aristoteliansupp/50.1.115](https://doi.org/10.1093/aristoteliansupp/50.1.115).
- Wolff, Jonathan, and Avner de-Shalit. 2007. *Disadvantage*. Oxford: Oxford University Press.
- Züger, Theresa, and Hadi Asghari. 2023. "AI for the Public. How Public Interest Theory Shifts the Discourse on AI." *AI & SOCIETY* 38 (2): 815–828. doi:[10.1007/s00146-022-01480-5](https://doi.org/10.1007/s00146-022-01480-5).