Ethics and Information Technology (2025) 27:21
https://doi.org/10.1007/s10676-025-09830-9

ORIGINAL PAPER

Taking responsibility for the outcomes of autonomous technologies

Niël H. Conradie[1] · **Saskia K. Nagel**[1]

Accepted: 20 March 2025 / Published online: 6 April 2025
© The Author(s) 2025

Abstract
It has been extensively argued that emerging autonomous technologies can represent a challenge for our traditional responsibility practices. Though these challenges differ in a variety of ways, at the center of these challenges is the worrying possibility that there may be outcomes of autonomous technologies for which there are legitimate demands for responsibility but no legitimate target to bear this responsibility. This is well exemplified by the possibility of techno-responsibility gaps. These challenges have elicited a number of responses, including dismissals of the legitimacy of these demands, attempts to find proximate agents that can be legitimately held responsible, and arguments for prohibiting the use of technologies that may open such gaps. In this piece we present a general argument that an overlooked but valuable option lies in adopting a strategy of *taking* responsibility for the outcomes of autonomous technologies even when the conditions for being legitimately held responsible are not met. We develop a general argument that the adoption of such a strategy is often justified not only by the demands of being responsible, but by practical considerations rooted in our relationships: the need to preserve of the quality of our relationships and the trustworthiness of the socio-technical system of which the autonomous technology is both a result of and embedded in.

Keywords Autonomous technologies · Taking responsibility · Techno-responsibility gap · Collingridge Dilemma · Dual-use · Trustworthiness

Introduction

There is a vast and growing discussion on the ways in which the introduction of autonomous technologies stands to impact our responsibility practices (Danaher, 2016, 2022; Johnson & Verdicchio, 2017; Matthias, 2004; Santoni de Sio & Mecacci, 2021; Sparrow, 2007; Tigard et al., 2020; Hindriks & Veluwenkamp, 2023; Coecklebergh 2023; Oimann, 2023; Vaseashta, 2023; Yazdanpanah et al., 2023; Sienknecht, 2024; Heine & Quintavalla, 2024). These include considerations of the Problem of Many Hands, the Collingridge Dilemma, and the issue of dual-use. Perhaps most paradigmatic of all is the debate surrounding the possibility of, and possible solutions to, so-called (techno-)responsibility gaps. Our aim in this piece is to contribute to this discussion by arguing for greater recognition of the role of *taking responsibility* as an often overlooked but valuable part of dealing with the various challenges to responsibility raised by autonomous technologies. Though we think our argument applies to at least some examples of all these responsibility challenges, we will focus on the techno-responsibility gap as an exemplar. This is for two reasons: (i) the techno-responsibility gap is a challenge uniquely raised by autonomous technology, and (ii) each of the other primary concerns involves, at least in part, the worry of opening a responsibility gap in an analogous sense to the techno-responsibility gap though not necessarily for the same reasons. Reason (i) gives relevance to our discussion as we are focusing here on the impact of autonomous technologies, and reason (ii) explains how insights garnered from how taking responsibility might help in techno-responsibility gap cases might generalise out to be of benefit in dealing with the other challenges as well.

For gaps, the threat rests on the possibility of there being harms resulting from our interactions with autonomous technologies that create a legitimate call for responsibility but for which there is no legitimate target—i.e. no-one to legitimately bear the responsibility. Our contention is that if

✉ Niël H. Conradie
niel.conradie@humtec.rwth-aachen.de

Saskia K. Nagel
saskia.nagel@humtec.rwth-aachen.de

[1] RWTH Aachen University, Aachen, Germany

or when such instances occur, we (variously, the developers, users, and regulators of these technologies) are confronted with an important choice. This choice is between taking responsibility for our interactions with these technologies and for the outcomes of these interactions or dismissing this responsibility. We argue that in many such cases there are good reasons to adopt a strategy of taking responsibility. In particular, we contend that there are at least some cases, what we will call *extended* cases, where the reasons for taking responsibility are practical ones from outside of the legitimacy conditions for being held responsible—a line of argument we will explain and develop below. This is to say that these are cases where it would be wholly fitting, according to the demands of responsibility, for us to ignore or dismiss claims that we are responsible, but where we have compelling practical reasons to adopt a strategy of *taking* such responsibility anyway. We argue that these practical reasons include the preservation of the quality of our relationships and the trustworthiness of the socio-technical system of which the autonomous technology is both a result of and embedded in.

We all take responsibility for a variety of things every day. Some of these are prospective or forward-looking, as when a person takes responsibility for watering their plants or for calling their ailing mother. Some of these things are retrospective or backward-looking, as when a person takes responsibility for stepping on another's foot in a crowd or for forgetting their partner's birthday. In many of these cases, there is an overlap between the things that we take (prospective or retrospective) responsibility for and the things for which we meet the conditions to be legitimately held responsible. However, this is not always the case. There are instances where we take responsibility beyond that which could be legitimately demanded of us by the strict definitions of being responsible (Enoch, 2012, 2015; Kiener, 2022; Mason, 2019; van de Poel & Sand, 2021). What is more, there can be compelling practical reasons for taking responsibility in this, what Mason (2019: 194) calls "extended" way—a term for describing taking responsibility beyond what we can be legitimately held responsible for—as a strategy depending upon the external considerations at work. Our argument is that the development and use of autonomous technology can give rise to precisely such contexts.

We begin by briefly discussing the relationships between autonomous technologies and our responsibility practices, with a focus on the possibility and impact of techno-responsibility gaps. Though not amounting to a full account of this dense and ongoing debate, we will identify a number of important responses to the challenge of techno-responsibility gaps as a means of positioning our own contribution. We take our proposal to be supplementary—perhaps even complementary—rather than a replacement to a number of the existing approaches, and that recognising an important role for a strategy of taking extended responsibility can help to alleviate moral costs that the existing approaches cannot easily deal with.

Turning to the details of our proposal, we will outline the distinction between holding responsible and taking responsibility. This will involve explaining what we mean by "taking responsibility", which is indebted to the account developed by Elinor Mason of what she calls "extended blameworthiness" (see 2019: 179–207 for her full discussion of the notion). This will initially guide us in determining the conditions under which taking responsibility is fitting. As will be shown, these conditions are weaker than those for holding responsible, although they remain important. There is such a thing as taking too much responsibility—taking responsibility when it is irrational or unreasonable to do so. The most immediate example being to take responsibility for an outcome that is clearly impossible to influence. Our aim is to provide an explanation for why and when taking responsibility in cases involving autonomous technologies, even when holding responsible would otherwise be illegitimate, can be reasonable and justified.

In our final step we argue that taking extended responsibility in these situations might not only be justified but that there may be an obligation to do so, though this relationship is not straightforward. We contend, in agreement with van de Poel and Sand (2021) and Mason (2019). That taking extended responsibility cannot be obligatory in any single particular case—that it's being supererogatory is an integral feature of taking extended responsibility. However, we argue that the practical considerations of maintaining our relationships and preserving the trustworthiness of our socio-technical environment can make it obligatory for us to adopt a general strategy of taking responsibility under the right conditions. This strategy requires a sufficient degree of taking extended responsibility but does not compel it in every case or any particular case.

Autonomous technologies and responsibility

The category of "autonomous technology" is very difficult to pin down with precision. For our purposes here we take autonomous technologies to include any technological system that exhibits what Liu (2016) labels *discretionary autonomy*. This is the ability of the system to receive data, work upon it through its own internal processes, and based on this work make a discretion between potential responses.[1]

[1] What is left aside here is the question of whether any of this constitutes the technology "making a decision", though we might at times use language like this for the sake of convenience even if it is not actually fitting (Johnson & Verdicchio 2017).

It is precisely the possession of this feature by an assortment of related new technologies (often loosely grouped under the label "AI", and including machine learning, deep learning, and neural networks) that has sparked a concern that they represent a potentially insuperable threat to our responsibility practices. In many cases the worries raised are familiar ones, made more acute by the features of these new technologies: the Problem of Many Hands (Thompson, 2014, 2017; van de Poel et al., 2012; van der Poel et al., 2015), the Collingridge Dilemma (Collingridge, 1980; Genus & Stirling, 2018; Levy, 2011; Rosen, 2004; Zimmerman, 2022), and harmful dual-use (Forge 2010; Vaseashta, 2023) being the foremost examples. However, we will primarily focus on the relatively new worry of techno-responsibility gaps. This is for the two reasons mentioned in the introduction: (i) the techno-responsibility gap is a challenge uniquely raised by autonomous technology, and (ii) each of the other primary concerns involves, at least in part, the worry of opening a responsibility gap in an analogous sense to the techno-responsibility gap though not necessarily for the same reasons. As such we will bring in discussion of the other issues where relevant, but they will not be our focus.

Techno-responsibility gaps are only ever possible when sufficiently autonomous technology is present (Matthias, 2004; Sparrow, 2007): When we employ autonomous technologies such as AI—either generally (Matthias, 2004) or at least in certain morally significant fields such as the military (Sparrow, 2007) – we run the risk of opening a *techno-responsibility gap*.[2] Such a gap is the result of the concurrence between two states of affairs: (1) there is some outcome that results from the discretion of the autonomous technology on the basis of which there is a legitimate call for responsibility, and (2) there is no suitable entity to which this responsibility can legitimately attach.[3] By a "legitimate call for responsibility" what is meant here is that those harmed by the output of these technologies might be justified in calling for or demanding that some entity bear responsibility for the harm that has befallen them. This call must be justified in order for the techno-responsibility gap to reveal a challenge to our responsibility practices, as if it is unjustified then it can (presumably) be dismissed.

To return for a moment to our reasons for focusing on techno-responsibility gaps, it is important for us to be clear that we do think that our argument extends (or at least has the potential to extend) beyond such cases, and encompasses at least some cases of Many Hands, Collingridge Dilemmas, and dual-use. Specifically, those that present a challenge to responsibility—in whole or in part—by threatening to open a "gap", a situation where there is a legitimate call for responsibility but no legitimate target to bear it. Thus, though the techno-responsibility gap is the only challenge unique to autonomous technologies—and so particularly relevant when discussing the relationship between responsibility and such technologies—solutions that might help resolve these techno-responsibility gaps can, and we would contend do, provide us with tools for confronting other sorts of responsibility gaps.

Two of the most prominent early examples of AI applications where the threat of responsibility gaps was raised were those of autonomous vehicles and in (lethal) autonomous weapon systems. In either of these cases there is the risk that if the technology could respond with discretion in such a way as to cause harms—a collision with a pedestrian, or the killing of an unarmed civilian, for example. The technologies themselves could not be legitimate targets for blame as they lack the relevant agential features, so the question is then whether there are proximate agents to bear the responsibility. Here, the epistemic opacity of the inner workings and the related unpredictability of the outputs of AI technologies means that it is possible that all the developers, users, regulators, etc. involved with the technologies might have been fully scrupulous—i.e. might have been as conscientious as could be reasonably demanded of them—and yet still the technologies evince the harmful malfunction. And this concern generalises out to any AI technology that is: (a) capable of making discretions between responses, (b) where these responses can result in harms, and (c) where the technologies themselves are sufficiently epistemically opaque to allow for the harmful response to be outside of reasonable foreseeability for those involved with the development, regulation, and use of the technology. Given that the ability to make discretions as well as great complexity are both key features of AI as a (albeit loosely defined) class of technology, we assume that most if not all such technologies at least have the potential to trigger these concerns, perhaps excepting those where the output cannot do any harm (if such cases exist). However, where this assumption does not hold, we are happy to exempt that example of AI technologies from our concerns here and reiterate that our focus will be on such technologies that do instantiate (a)-(c) listed above.

The challenge's introduction has given rise to a dense literature grappling with the possibility of

[2] We follow Tigard (2021) in preferring the term "techno-responsibility gap" over the term "responsibility gap" employed by Matthias and Sparrow. Our reason for this is that since we can plausibly talk about responsibility gaps emerging without any reference to autonomous technology (or indeed any technology) as in the case of the Problem of Many Hands (van de Poel et al., 2015), a more precise term should be used to pick out the phenomenon we have in mind.

[3] Although the discussion surrounding this topic has seen a proliferation of concern with different types of techno-responsibility gaps or other gaps in the vicinity—with talk of retribution gaps (Danaher 2016), control gaps (Hindriks & Veluwenkamp 2023), and Santoni de Sio & Mecacci's (2021) four interrelated types of responsibility gaps—we will focus here on the techno-responsibility gap as it is most generally understood.

techno-responsibility gaps in cases involving AI of different sorts. Some have attempted to dismiss the possibility of these gaps by either denying that the calls for responsibility raised by the outcomes of autonomous technologies are legitimate (Königs, 2022), or by contending that there are nearby entities to which the responsibility could attach (Nyholm & Smids, 2020). Others have sought to "bridge" the gap by stressing the versatility and adaptability of our responsibility practices, recommending different revisions or amendments to our usual responsibility practices that will allow us to retain targets for responsibility (Gunkel, 2020; Danaher, 2022; Santoni de Sio & Mecacci, 2021; Tigard, 2021). Yet others contend that the threat of the gap is real and unbridgeable, and that a prohibition on the use of autonomous technologies in situations where a gap might be opened is our best response in at least some cases (e.g. in context of military AI) (Sparrow, 2007). Attempts to insert collective agents into the role of responsibility bearers in these cases have been put forward (Conradie, 2023; Robillard, 2018) but also strongly challenged (Taddeo & Blanchard 2022; Taylor, 2024). Each of these approaches remain live options, though each does face its own serious hurdles.[4]

The strategy of taking extended responsibility that we develop and defend should not be thought of as an attempt at providing a full standalone answer to the challenge of techno-responsibility gaps. We take various attempts at bridging the gaps to have merit, be these by locating legitimate bearers of responsibility based on existing responsibility practices or by revising some of our practices to accommodate our developing needs. Prohibition should only be a last resort, turned to only if alternative strategies fail. As to the approach of dismissing the legitimacy of the demands for responsibility, our proposal takes seriously that at least some (perhaps many) cases of apparent techno-responsibility gaps may instead be correctly thought of as (tragic) accidents, and that the legitimacy of holding agents responsible for them may be dubious. We differ from this approach as we suggest that this is not the end of the responsibility story: it ignores the role that taking responsibility can and should play in mitigating the otherwise significant costs incurred by this way of resolving the techno-responsibility gap.

Holding responsible and taking responsibility

Determining the conditions under which it is legitimate to hold another agent retrospectively responsible, including accounting for exemptions and excuses, has been the primary focus of the philosophical discussion around moral responsibility in general and techno-responsibility gaps in particular. Though this focus is shifting somewhat, as there is an increasing interest in prospective responsibility (Ronca 2022) as well as in taking responsibility (Mason 2018, 2019; van de Poel & Sand 2021; Danaher 2022). To briefly summarise a nuanced discussion: holding an agent retrospectively responsible involves attributing to them moral praise- or blameworthiness on the basis of some slice of their conduct.[5] A vital feature of our practices of holding each other responsible is that we should not do so *unfairly*. To think through how this works, it is useful to follow McKenna (2012) in thinking of our responsibility practices as paradigmatically conversational.

McKenna (2012) argues that there are three usual steps in the responsibility conversation: firstly, when an agent acts, and her acts are morally charged, she appreciates that she might be introducing a meaningful contribution to a kind of conversational exchange with others. This initial stage McKenna calls *Moral Contribution*. The second stage, in which that agent is blamed (or praised) by a respondent on the basis of the morally charged action, he calls *Moral Address*. In the third stage, *Moral Account*, the blamed agent extends the conversation by offering an excuse, a justification, or an apology, for example. The respondent might at this point continue the conversation, perhaps by forgiving or punishing. Crucially, one of the moves available during *Moral Account* is to raise what has been called the Unfairness Objection—the objection that to target the provider of *Moral Account* with responsibility would in this case be unfair as they lacked the appropriate relationship to the slice of conduct. The strength of the Unfairness Objection stems from the intuitive idea that no person should be held blameworthy for that which falls outside of their control, suitably understood (Watson, 1996; Levy, 2005; Himmelreich & Lawford-Smith 2019).[6]

[4] For example, dismissing the legitimacy of the responsibility claims of victims of harm is fraught with moral peril. As long as possible ways of meeting these demands remain viable, we should be reluctant to adopt dismissal. To view these harms as mere accidents risks incentivising negligence and recklessness among the developers, users, and regulators of these technologies, and it risks doubly victimizing those harmed through the dismissal (Smith 2020[1759]). The contention that we can dismiss the claim that there is no nearby entity to which the responsibility can legitimately attach raises the concern of scapegoating the individuals involved – that is, blaming agents who are not in fact blameworthy (Himmelreich & Lawford-Smith 2019; Reiff 2008; Sverdlik 1987).

[5] The use of "slice of conduct" here is to be neutral between actions, omissions, and perhaps even mental action or attitudes, all of which have been thought to be potential targets of praise or blame in the discourse.

[6] This concern has, to our knowledge, usually been discussed in relation to blame, though it is not difficult to imagine that this is because the cost of misplaced blame is usually higher than that of misplaced praise. But even then, it does seem that an objection to misplaced praise on the basis of fairness is an available move during *Moral Account*, though is practically more likely to be brought forth by a third party than the provider of the account.

In the case of autonomous technologies, we are not assuming that it is possible for the technologies themselves to engage in the Moral Conversation. We take them to be incapable of making a Moral Contribution, as they lack the moral agency necessary for such. However, the developers, users, and regulators involved with the technology are certainly open to such contribution and engagement. If self-driving car ploughs into a pedestrian, and the reason for malfunction is that one of the developers of its software failed to run a requires safety check, then this developer would be an appropriate target for the Moral Conversation. Similarly, the decision makers at a company who choose to continue using a hiring algorithm after discovering that it is racially biased would also be open to such. In cases not involving a techno-responsibility gap, it is the humans involved with the autonomous technology who acts as an interlocutor—who sees their conduct involving the technology as a Moral Contribution, who would be the target of Moral Address, and who would deliver a Moral Account.

Techno-responsibility gaps, if they exist and function as paradigmatically assumed, are cases where this pattern is problematised, as anyone held responsible would be justified in invoking the Unfairness Objection. This is usually thought to be because the developer or user in question lacked the appropriate control over the conduct of the autonomous technology—*including a lack of ability to reasonably foresee the technology's conduct*. Thus, the lack of control in these cases can often be understood as a failure of an epistemic condition necessary for responsibility. These technologies are infamous for being epistemically opaque and for the possible unpredictability of their outputs (Hindriks & Veluwenkamp, 2023). If neither the passenger nor developer of a self-driving car could reasonably foresee that it would, under certain conditions, veer into a pedestrian, it would be unfair to *hold* either responsible for this outcome. Regardless of the possible success of the various bridging strategies aimed at avoiding this conclusion, which we mentioned in the previous section, we assume it likely that in at least a substantial number of cases things will play out as in this example. Given that any target for responsibility can justifiably provide the Unfairness Objection during *Moral Account*, and based on the norms that govern holding responsible, the victim's attempt to hold anyone responsible lacks legitimacy and can be dismissed. This does, however, leave us with the very real concern that without anyone being held responsible, negligence and recklessness might be incentivized, and that the victims of harms in these cases will not receive reparations.

Here we turn to what we hold to be the vital role of *taking responsibility* in these contexts. Taking responsibility occurs when an agent either (i) *recognises* the legitimacy of some responsibility demand in relation to their own conduct or (ii) when an agent *adopts* responsibility for some future or past conduct for which they would not previously have been responsible. The former of these is usually an important and necessary part of the Moral Conversation: to make proper amends and bring closure to a victim, the responsible party should recognise their responsibility by recognising the legitimacy of the victim's demands, and in doing so should *take* responsibility. An actual demand need not be made for this to happen, an agent can recognise themselves as a legitimate target and then respond accordingly. Vital as this is, it is the latter occurrence that is of particular interest to us: *when an agent takes on responsibilities that were not (previously) legitimately attributable to them*. Following and building on language employed by Elinor Mason (2019) in her discussion of taking responsibility, we call instances of this latter type *taking extended responsibility*.

An important difference between holding responsible and taking responsibility that makes such extended responsibility possible is that the norms governing when it is legitimate to hold someone responsible are stricter and more limiting than those governing when it is legitimate for an agent to take responsibility (Calhoun, 2019, 2022; Enoch, 2012, 2015; Mason, 2019; van de Poel & Sand, 2021, 2024). As explained by van de Poel and Sand (2021: S4779):

> What is typical for all these situations [those of holding responsible] is the existence of another agent j that ascribes responsibility to agent i, which naturally raises the question whether such attribution is fair. This is different in the case of taking responsibility for some ϕ as the voluntary act of an agent i. Nevertheless, there is a pendant to the fairness issue in the realm of 'taking responsibility'. Obviously, there are events for which an agent cannot reasonably take responsibility, for example, because she is unable to affect them. The conditions for holding someone responsible in a fair manner are equally relevant when assessing whether it is reasonable for someone to take responsibility. Nevertheless, there is an important difference. There is more room for reasonably taking responsibility than there would be for someone else attributing the same responsibility in a fair way. In other words, for some ϕ for which it may be unfair to hold an agent i responsible, that same agent i may still reasonably take responsibility.

Those cases where an agent takes reasonable responsibility for something that it would be unfair for them to be held responsible for, this is taking extended responsibility.

Though our approach here owes much to Mason's work, our aim does not perfectly overlap with hers. Her explicit concern is with taking extended *blameworthiness*, whereas our interests concern the wider taking extended *responsibility*. This broadens our endeavor in two ways: first, we are not only concerned with taking retrospective responsibility but

also taking prospective responsibility. Second, we are not concerned only with blame but also with praise.

What we take to be one of Mason's most crucial insights is that an agent's possible justifications for taking extended responsibility—regardless of the responsibility's valence or direction—are not wholly determined by the relationships between the agent and the act but can include richer considerations originating in our social embeddedness (Mason, 2019: 200–205). Whereas for unextended responsibility, the justifications for blame or praise are conditioned upon their being the right relationship between agent and act (i.e. is the person blameworthy or praiseworthy), the justifications for taking extended responsibility can involve a consideration of reasons external to this agent-act relationship. If the correct relationship between agent and act is present, then that agent *is* blameworthy or praiseworthy, and in the case of blameworthy would, ceteris paribus, be under an obligation to take this responsibility. On the other hand, even an agent lacking such a relationship to an act might have good reason to adopt a strategy of taking extended responsibility, not because that is what is *demanded* by the relationship between agent and action, but because doing so allows for the fulfilment of other practical and desirable aims rooted in their web of relationships.

Though an agent violates no "rules" of responsibility in dismissing extended responsibility, doing so—and most crucially doing so *consistently*—comes at an important cost. It threatens to undermine the basis for which our responsibility practices operate: the preservation of our relationships and the maintenance of the moral community. The relationship between the internal legitimacy conditions of a practice and the etiology of the self-same practice can be complex or highly attenuated. This certainly appears the case when we consider the myriad responsibility practices that form a ubiquitous part of our lives. Psychological evidence points at the enablement of co-operation, the maintenance of relationships, and the reinforcement of self-regulation as important functions performed by these practices. As Mason (2019) points out, an individual who systemically dismisses extended responsibility will find themselves alienated from others and untrusted by those with whom they might seek to cooperate. This is most apparent in personal relationships—within families, friendships, romantic relationships, etc.—where the taking on of extended responsibility is way to reflect our respect and for the other person and our investment in the relationship. In Mason's (2019: 199–200) words:

> Insofar as we owe respect to our fellows, it is something that requires investment. It is not mechanical, or legalistic. We cannot respect our fellow people without attitudinal back up. Taking responsibility for inadvertent disrespect is part of the required attitudinal back up. It is a way to show the person who has been slighted that they are respected, and a way to demonstrate investment in a shared community.

Refusals to exhibit this respect and care, to properly invest in our relationships, creates a situation where our relationship partners will (perhaps gradually) withdraw from the relationship. Mason's (2019: 201) preferred example here is the romantic partner who "gradually stops seeing their partner as a proper interlocutor, stops holding them accountable…a sign that the relationship has dissolved." One was of describing this could be to say that the (largely tacit) norms within these personal relationships include an expectation of taking extended responsibility to some sufficient degree.

It is less clear how this is to be understood when the relationships in question are increasingly impersonal. In these cases, the norms of interaction could be unclear or might not include an expectation for taking extended responsibility. We will turn to discussing this concern in Section "Practical considerations for adopting a strategy of taking extended responsibility" but will note here that though the levels of expectations may differ we do contend that even in professional, impersonal, relationships there are expectations of at least certain levels of taking extended responsibility.

It could also be asked how Mason's account, which rests so much focus on relationships, applies to cases involving autonomous technologies. To clarify our position: as in our discussion of McKenna's conversational account, we are not claiming that it is the relationship with the autonomous technologies themselves that matters here. What does matter are the relationships between the different humans involved. Depending on the case in question, developers, purveyors of technologies, regulators (including policymakers and executive regulators), users, and victims are all embedded in relationships with each other—impersonal though they may often be. When we talk of the role of taking extended responsibility, we are talking, for example, about developers of an autonomous technology taking responsibility for harm it has caused to a victim, or regulators of this technology taking responsibility for their failure to prevent it. The value of her account is that it provides insights into how these instances of taking responsibility should be understood, and how we should understand the deontic status of doing so.[7]

Though we do take on many elements of Mason's account, it should be noted that her focus is on retrospective responsibility, whereas our argument extends to considerations of prospective responsibility. Our motivation for this expansion can be summed up as "prevention is better than cure"—it is often the case that a potential responsibility gap can be prevented by a given agent prospectively taking

[7] Our thanks to an anonymous reviewer for pushing us to clarify this point.

responsibility for an outcome. As our overall aim is to argue for the usefulness of taking extended responsibility in dealing with techno-responsibility gaps, it is important that we recognise this useful role that prospective responsibility can play in diffusing gaps before they even arise. Though still leaning on important insights from Mason's account, these are our own attempts to add to the understanding of a dimension of taking extended responsibility with which she has not (to our knowledge) engaged.

Responsibility strategies

Importantly, taking extended responsibility is not an all or nothing situation. Here we wish to introduce talk of *responsibility strategies*. By a responsibility strategy we mean a rule scheme for how and when to take responsibility. These strategies are often (indeed usually) built up through a process of unintentional accrual over time, but sometimes they are developed or more often modified through intentional endeavour. Importantly, there might be a number of acceptable strategies for taking responsibility that differ quite a lot in how much, and for what, responsibility will be taken. How we approach extended responsibility (as with other facets of responsibility) is dynamic and permitting of some variety. In part this is due to the fact that there are several sorts of cases where we have deeply equivocal and ambiguous intuitions and responses as to whether or not responsibility can or should be attributed. We see this most vividly in how humans engage with what Shoemaker (2015) might call cases of responsibility at the margins. The most common examples of these being persons with certain psychological impairments, or persons suffering from the consequences of a traumatic childhood and further life. In these cases, it is not clear that there is an unequivocal correct response, there may be a number of acceptable responses. In addition to this, our limited cognitive and epistemic resources means that in practice we must adopt generalised strategies of responsibility (for both holding and taking) despite being aware that at times we might get it wrong.

Very importantly, a strategy can permit of rules with a dispositional rather than a categorical character. We will *usually* take responsibility in situation X, rather than we will always take responsibility in situation X. I do not always take responsibility for changing the diapers, but I will *usually* do so—and this works if my partner shares the same (or similar) strategy, and so the diapers get changed but neither of us has to see this as an exclusive duty of ours, nor do we need to arrange some careful balance of when one or the other of us takes the task. For this to work, it is necessary that the *usually* in both my and my partner's strategies be sufficiently regular. Of course this is not an example of taking *unextended* responsibility, as presumably my partner and I would be open to being held responsible if we fail to do so, but it illustrates how responsibility strategies, on our understanding of them, can permit of dispositional rules, and that whether these rules work can depend on the frequency with which it commits me to taking responsibility. This will be an important element in developing our final argument.

A responsibility strategy can be imagined as a list of commitments, some categorical and some dispositional, held by an agent toward when to take responsibility. Since part of the usefulness of such strategies lies in flexibility, there is no single stipulative form that this list should take, but rather any given list should be assessed in terms of how well or poorly it adheres to the constraints on fittingness for taking responsibility – which we discuss in the rest of this section – and how well it allows the agent to meet the practical aims of preserving their relationships – which will be discussed in Section "Practical considerations for adopting a strategy of taking extended responsibility".

There are important constraints in the case of taking responsibility, and there are two dimensions to this: first, taking *unextended* responsibility is fitting in cases where being held responsible would be legitimate. Any strategy that involves refusals to take responsibility for that for which an agent meets the conditions for being legitimately held responsible would be unacceptable. Second, and more important for our arguments, when we turn to taking extended responsibility, the constraints are far less defined. It would be prima facie unacceptable for an agent to take extended *prospective* responsibility for things that they can have no control over (to take responsibility for the rising and setting of the sun, for example). In a case like this the taking of responsibility for the future outcome would mean that the agent is now accountably committed to bringing it about, and if they have no ability to do this then their commitment is unreasonable—we would not, and should not, take it seriously. Thus, a strategy for taking responsibility should avoid such instances. On the other hand, it can be wholly plausible for a strategy to call for taking retrospective responsibility for already occurred outcomes over which the agent did not *have* control, as in such a case the taking of responsibility commits the agent to participation in the Moral Conversation and to undertake relevant measures of redress—apologies, emotional or social reparations, etc.[8] This does require that the agent taking responsibility be in a position to provide such participation and redress—so it would be unreasonable for a senior citizen in Scotland watching news of a brutal killing in South Sudan to take extended retrospective

[8] Redress here is not taken to include what we take to be meant by "punishment", the punitive infliction of adverse treatment. Questions of punishment we leave to the sphere of legal responsibility, which is not considered here.

responsibility for it, but would be reasonable to take responsibility for losing a friend's necklace due to forgetfulness. The participation in the Moral Conversation will also be different to what we would expect in cases of unextended responsibility. As per Mason's (2019: 191–196) account, the first-personal reactive attitude when taking extended blame would be remorse, rather than the guilt characteristic of unextended blame. Whereas guilt involves the self-attribution of a substandard Quality of Will, remorse does not. Rather remorse involves the thought: "that injurious conduct was mine and I wish I had acted differently."

Though "weaker" than guilt in this sense, two things should be noted: (i) this still means that the responsibility at stake here is accountability rather than only answerability as argued by Kiener (2022). It is not enough that the taker be answerable for the outcome and open to redress, but that they take ownership and have the correct attitudinal responses to properly engage in the Moral Conversation. And (ii) this applies to taking extended retrospective responsibility, but not to taking extended prospective responsibility. Taking prospective responsibility at a time, t_1, does not immediately entail any particular attitudinal responses at t_1—as the success or failure of that for which responsibility has been prospectively taken at t_1 has not yet been determined. Instead, what taking responsibility does commit the taker to is to have the appropriate responses once this determination has occurred at some later time, t_2: if x takes prospective responsibility for bringing about some y at t_1 (then they have now taken ownership of the eventual outcome, be this y or not-y), then if y does not occur at t_2 then they are prima facie open to being held responsible—and so guilt (not remorse) would be an apt response. However, this is further complicated by the fact that an agent can take responsibility for bringing about some y but then fail to do so in ways or for reasons that we would think renders them an illegitimate target for being held responsible.

An example of this might look like the following: a city is suffering from low water quality in its water supply network. Though various attempts to remedy the issue have been made by local and national politicians (who presumably should take unextended responsibility in cases like this), no solution has worked. Stepping into this context, Margaret, the head of a small tech company organises and leads a team of her developers in order to create an algorithmic system that is able to identify impurities in the water and dynamically combat these by automatically introducing countermeasures. By taking this role, Margaret is taking extended responsibility for helping to resolve the water crisis—i.e. it is not something that could be legitimately demanded of her, but she has chosen to take it on. By doing so she is now open to responsibility for how her attempt plays out. If the program is successful and the water crisis resolved then she, and her team, would be open to praise. If the attempt fails, and especially if it results in any additional harms or indignities, due to a lack of conscientiousness on Margaret or her team's part—e.g. they do not take proper precautions, are lax in their testing, etc.—then she and they would be open to blame. This blame would not be for the water crisis itself, but for failing to meet the commitment to help resolve the crisis that the taking of responsibility represented.[9] However, if the program fails for reasons that are outside of Margaret and her team's control then they would not be legitimate targets for being held blameworthy. This could occur because despite all reasonable effort the system developed a critical error, one that no-one could be expected to have foreseen. Such a case seems to still leave us with a gap. To bridge this, we contend that the agent in question (in this case Margaret and her team) can still take *extended retrospective* responsibility for these outcomes, with the accompanying remorse.

This notion of taking ownership of the conduct or outcome in question is a central element to taking responsibility, as this seems necessary if we are to engage adequately in the Moral Conversation with victims. It is also part of why we cannot take responsibility for that which can in no way be claimed in this manner. Mason (2019: 183–184) makes sense of this relationship through her idea of "ambiguous agency" in her words:

> It is not completely clear that they [certain actions] are connected to the agent in the right way. On the other hand, it is not completely clear that they are not connected to the agent. It is this ambiguous agency that is puzzling. We have acts that are plausibly the agent's acts, and yet the badness of them does not seems traceable to the agent in any meaningful way

Her argument is that when a morally bad outcome is the result of such ambiguous agency, it is appropriate for us to take extended responsibility for it. In a similar, though different fashion, Kiener (2022: 586) stipulates requirements that the agent taking responsibility's involvement must be sufficiently proximate and sufficiently relevant to the outcome in question. He describes these as follows:

> [I]nvolvement can be usefully analysed in terms of two important metrics. The first is proximity, i.e. the temporal or spatial distance between an agent's conduct and the AI-related outcome. End-users normally show greater proximity to an outcome than the computer scientists who developed the system. The second metric is relevance, which concerns the importance of

[9] In some ways, this is analogous to the structure of a promise (Keiner 2022; van de Poel & Sand 2024). Though there may have been no obligation to make one, if it is made and then not lived up to this is grounds for criticism.

> a person's involvement. Here, the computer scientists could sometimes have made a more important contribution to the AI's outcome than the end-user, at least when the latter uses the system only for its previously determined purpose.

Our contention is that these conditions make sense in the case of taking retrospective responsibility, though there is likely some unavoidable vagueness in establishing their precise boundaries. What is less clear is whether these are useful in establishing limits on strategies of taking prospective responsibility. As such, we move beyond Mason's account here and present our own account—though informed by Mason's—of the conditions for taking extended prospective responsibility.

Here it seems that the issue at stake is not establishing if proper ownership is possible, but rather about predictions of control. This should not be surprising as the taker of prospective responsibility is claiming ownership over outcomes which have not yet occurred. The most immediate thought would be that an agent can only take prospective responsibility for that over which they will have adequate control, i.e. over what they could reasonably assume they would be able to bring about. This is an admittedly vague standard, as appeals to reasonableness often are, and it is inevitable that there will be dispute on the fringes as to when the control is adequate. The most apt test would be to ask whether the control that the taker would have over the outcome amounts to the level of control that would be demanded for them to be held responsible for that outcome—a test that depends on which of the competing control conditions on responsibility one endorses.[10] We will not be taking a stand between these competing control conditions. Though this expectation of control preceding the taking of responsibility is often the case, we follow van de Poel and Sand (2024: 32) part of the way when they claim that:

> [A]t least under some conditions it can be permissible and reasonable to take on new forward-looking responsibilities, even if the object of such responsibilities is initially beyond our control. This is not to say that control is irrelevant in these situations, quite the contrary. We will suggest that in these cases, the typical relation between control and responsibility is reversed. Rather than being responsible for what is already under our control (and, perhaps, because it is under our control), we are sometimes moved by the call of responsibility, and as a result of taking responsibility, we aim to increase control.

We concur that taking responsibility for an outcome can be done without the taker yet having control over the outcome, and that by taking this responsibility the taker commits themselves toward increasing their control. However, for it to be reasonable for an agent to take on responsibilities that bring such commitments they must have sufficient control over being able to gain this further control over the outcome. For ease of discussion, we can speak here of *proximate control*—the control the agent has over bringing about the outcome in particular (the control that van de Poel and Sand claim is not necessary at the point responsibility is taken)—and *distal control*—the control that the agent has over being able to gain proximate control. If it were the case that the agent did not have distal control, and was aware of this, then it would be unreasonable for them to take responsibility. Thus, even though taking responsibility can precede proximate control, considerations of distal control still condition how reasonable it is to take responsibility in any particular case.

Suffice it to say that it is reasonable for an agent to take such responsibility if they have good reason to believe that they would have control under some understanding. Importantly, it might be reasonable for them to do so even if this turns out to be untrue, provided that their belief it was true was well-founded. This gives us three categories: (i) those who take prospective responsibility for that which they cannot reasonably expect to control. In this case the taker is being unreasonable, likely criticisably so, and the taking of responsibility is unfitting—no legitimate responsibility practices follow from it, the taker acquires no new legitimate responsibilities and would not be a legitimate target for being held responsible regardless of the outcome. (ii) Those who take prospective responsibility over what they can reasonably expect to control and where in fact they do have this control. These are the paradigmatic cases, the taking of responsibility is fitting and our responsibility practices play out as we would anticipate. (iii) Those who take prospective responsibility over what they can reasonably expect to control but where in fact they do *not* have this control. Here the taker is not being unreasonable, and the taking is fitting, but the taker might well be *excused* for failure to bring about the relevant outcomes. This would mean that it would be unfitting to hold the agent taking responsibility in this case responsible, though provided that the outcome was still the result of ambiguous agency it would still be fitting for

[10] Though there are many competing control conditions, the one recommended by Fischer and Ravizza (1998) has perhaps been the most influential, and is adopted by van de Poel and Sand (2024) in an important recent engagement with the relationship between control and taking responsibility. On this understanding, what Fischer and Ravizza call guidance control, what is required for an agent to have appropriate control over some outcome is that the outcome be the result of that agent's own reasons-responsive actual-sequence mechanism. There is much more that can be said about the strengths and weaknesses of this understanding of control, and van de Poel and Sand discuss these in their work, but as we are not committing to any particular understanding of control, we leave issues to the side here.

the agent to take extended responsibility. Thus, if the taker could reasonably expect to have control, when the outcome is determined and it turns out that the taker had control, then they would be a legitimate target for being held responsible. If they did not, then it seems that they might be in a position to take extended retrospective responsibility.

If the agent has taken prospective responsibility but not *extended prospective responsibility*, what will determine the case in the latter is whether or not the outcome still falls within the bounds of ambiguous agency, proximity, and relevance. But taking extended prospective responsibility involves taking prospective responsibility for that for which one could not be required to take responsibility by the rules of our responsibility practices. It extends the range of that for which responsibility can be taken. Our contention is that in such a case the agent need only have *ambiguous control* over the possible future outcome, analogous to Mason's ambiguous agency. Such ambiguous control still requires that the agent be suitably proximate and relevant to the activity in order for the control to be ambiguous rather than merely lacking. However, the benefit of this is that by taking such extended prospective responsibility the agent is prima facie committed to taking retrospective responsibility for these ambiguous outcomes even if they end up falling outside of their control strictly speaking, though this retrospective responsibility for the outcomes will be in the extended sense. Of course, if the agent believes that they have ambiguous control but in fact they do not—they have less control than this—then they may well take extended prospective responsibility (and perhaps reasonably so, if they had good reason for their mistaken assumption). However, on examination it would in fact have been inappropriate in this case and so should be dismissed in the course of the Moral Conversation.

With all this said, it can still be asked why agents should adopt strategies of taking extended responsibility when the strict rules of being responsibility do not require it. We suggest that there are practical demands outside relationship between agent and act that justify demanding the adoption of such strategies. In the next section we unpack these.

Practical considerations for adopting a strategy of taking extended responsibility

The maintenance of our relationships and the preservation of trust are served by a strategy of taking extended responsibility for the outcomes of autonomous technologies. That this is the case provides a practical justification—in line with the similar practical justification we saw from Mason—for adopting a general strategy of taking extended responsibility for the outcomes of interactions involving autonomous technologies.

Our relationships are a deeply valuable and perhaps unavoidable feature of our shared lives. A part of partaking in these relationships, in this shared life, is to take part in our shared responsibility practices (Calhoun, 2022; Mason, 2019; Strawson, 1962; Wolf, 2001). If we choose to systematically cut ourselves off from taking extended responsibility and engaging in the necessary Moral Conversation, our fellows are likely to stop approaching their interactions with us from within the participant stance, but rather adopt the objective stance toward us. Consider the following description from Mason (2019: 201):

> In those cases it is true that, in a sense, you can avoid blameworthiness by refusing to take it on [in the case of extended responsibility]. My response to that worry is that in this context, we should see blameworthiness as a privilege, not a burden. Think of a romantic relationship in which one person gradually stops seeing their partner as a proper interlocutor, stops holding them accountable. That is a sign that the relationship has dissolved. Although being blameworthy, and being blamed, sting in the moment, over the long term it is the better way to be. Taking responsibility is part of being in relationships, part of being a member of a community. We can let go of all that, but estrangement and alienation are a high price to pay for avoiding blameworthiness.

Part of avoiding these deeply undesirable outcomes is achieved by at least sometimes taking extended responsibility. In many ways it is precisely because an agent is not obligated to do so but does so anyway that their commitment to their relationships, to their community, is affirmed. Making these choices obligatory would in some ways reduce this affirming role.

When we are dealing with autonomous technology, as a developer, purveyor or in some other role, how we respond to the outcomes generated by these technologies stands to impact our relationships with others. Relevantly here, if we refuse to take extended responsibility for the outcomes (especially the harmful ones) of these technologies then our relationships with others can be radically damaged. As per Mason's account (2019: 204–205), this can hold for both personal and impersonal relationships, though not in precisely the same way. In personal relationships taking extended responsibility serves to affirm that you are taking the other seriously as a person. In the case of impersonal relationships "where an inadvertent act [an instance of ambiguous agency] expresses disrespect, an agent who cares about the broad social relationship she has with her peers has a reason to take on extended blameworthiness, to show her fellow people that she does, after all, respect them, and to ask for respect in return". Here we diverge from Mason's view in that it does not seem that the commitment

to "the broad social relationship she has with her peers" is limited to cases where disrespect is expressed. We take it to be more plausible that there is good reason to take extended responsibility in an impersonal context as an affirmation of the value placed on our broad social relationships. Doing so signals that we recognise the value of the other even if we are not in a personal relationship, and that this valuing goes beyond merely the avoidance of disrespect.

It is very likely that most cases dealing with techno-responsibility gaps resulting from autonomous technologies will involve impersonal relationships—those between victims and the users or purveyors of the technology, between victims and the developers of the technology, and between victims and the regulators tasked with overseeing the technology. To illustrate this with an example: consider a team of developers who develops a machine learning-enabled system capable of ranking job applicants in terms of predicted job success. This system is then purchased by an insurance company to use in their hiring process. This entire process, or at least parts of it, are overseen by a regulatory body. As it turns out, the algorithm's outputs reveal a systemic bias against female candidates, a number of such candidates are rejected due to this bias, become aware of this and call for blame to be apportioned to those responsible. However, it turns out that the source of the problem was such an unpredictable one that no-one in the socio-technical system could have reasonably foreseen it. This is despite there being tests conducted at several stages, and scrutiny applied by the regulators. This is then a usual example of a techno-responsibility gap, and it would seem to be unfair to extend blame to any of entity involved. However, though none of them would be a legitimate target for holding responsible, they have good, practical, reasons for taking responsibility in a case like this.

Developers or purveyors (in this example, the insurance company who purchased and used the system) who adopt responsibility strategies that systematically dismiss calls for responsibility for outcomes in cases like this will face increasing alienation and isolation from potential customers and partners—even if a dismissal in any single case would not do so. This is a threat to successful future co-operation, including, likely most pertinently to the developers and purveyors, future economic co-operation. This is so even though these are impersonal relationships—they still require affirmation on the part of both parties that mutual respect and care is present, affirmation displayed through taking responsibility. And this is known, developers and purveyors looking to foster good relations with their customers and partners will sometimes make a big show of taking responsibility—though the sincerity of such displays can be dubious. Importantly, if this taking of responsibility is to be effective (as discussed in Section "Responsibility strategies") the developers or purveyors must be able to engage appropriately in the Moral Conversation, which requires that their motivation for doing so cannot be solely self-interested—they must be capable of expressing genuine remorse.[11] With regulators, at least those appointed by the state in a democracy, this is in many ways an even clearer dynamic. Regulators, and the state structures that appoint and govern them, that do not have a strategy that includes taking extended responsibility in at least a sufficient number of cases such as this are likely to alienate the citizens that they are tasked with protecting, a degenerating relationship that can undermine the credibility of their social role as regulators. In the example case, where the agency is ambiguous and there is sufficient proximity and relevance, the developers and regulators could reach out to apologise and express remorse to the victims, providing explanations for the failure if this is possible, and affirm commitments to future improvement. The insurance company could similarly apologies, express remorse, and affirm future improvement, but also perhaps include reparatory measure for some sort—perhaps giving the unjustly dismissed applicants a new chance at an improved hiring process. Importantly, we are not arguing that in each and every case such as this all of the above should take place, but that each of the above-discussed groups have reason to adopt a strategy of taking responsibility in some sufficient number of such cases. Also, we are hesitant to be too precise in our specifications of reparations as these would have to be ultimately sensitive to a myriad contextual features that this toy example simply cannot capture. Our arguments here are very general and are intended to be. We mean them to be a starting point, a defense of the general usefulness of adopting strategies of taking extended responsibility in dealing with techno-responsibility gaps, rather than an attempt to concretise this approach in any particular case.

A potentially even better response that the actors in this case could have had would have been to adopt a strategy of taking extended prospective responsibility prior to the occurrence of the harm in the first place. To take one group of actors as an example, if the regulators adopted a strategy of taking extended responsibility to do tests *beyond what is reasonably demandable* (recall that we are talking about taking on *extended* responsibility), then it is quite possible that they identify the error before harm occurs, and there is no gap at all. Such a preemptive strategy could have the further advantage, if made known, of communicating their investment in their relationship to the users and potential victims of these technologies.

[11] We leave aside here whether or not this taking of extended responsibility can be undertaken by a collective (e.g. the insurance company as a whole), or whether it will reduce to instances of individual responsibility. Though we are sympathetic to the former possibility, we do not defend this here but elsewhere [Redacted].

A closely related consideration in favour of taking extended responsibility is that it allows us to preserve the trustworthiness of our socio-technical systems. Despite sometimes unclear talk of trustworthy AI, which can seem to suggest that certain autonomous technologies might be trustworthy in themselves, we adopt the view that only agents or systems with the correct agential properties can be legitimately trustworthy (Conradie & Nagel, 2024). When it comes to a given autonomous technology, this means we should have as our goal the preservation of the trustworthiness of the socio-technical environment within which this technology is embedded, as the technology itself is not a candidate for legitimate trust. This is a matter of the trustworthiness of the agents who constitute the environment, as well as the structures and policies that bind the environment as a collective whole. Though this is a multifaceted task, one important element thereof is ensuring that responsibility is properly accounted for. Trust, insofar as it is related to vulnerability (Hall et al., 2001; Kerasidou et al., 2021) and betrayal (Baier, 1986; Vanneste & Puranam, 2024), are inextricably linked to responsibility concerns. To overgeneralise somewhat, where there is no ability to be responsible, there can be no betrayal of trust, only of reliance.

Responsibility gaps then, be they of the techno- kind or raised by one of the other challenges, threatens to undermine the trustworthiness of the socio-technical system. To trust is to make oneself vulnerable, and vulnerable in a very particular way. Trusting makes us vulnerable to the emotional harm of the betrayal itself, above and beyond any direct harms that would be present in cases of mere reliance (Vanneste & Puranam, 2024: 26). But we open ourselves to this vulnerability with the expectation that the other is an agent who not only will be trustworthy in seeking to avoid failing our trust, but is also trustworthy in that if things go wrong they will be there to make amends—to meet our call of betrayal appropriately. If harmful outcomes might result where such amends are not possible, then it ceases to be reasonable to open ourselves to that vulnerability. And this is possibly (if not probably) costly for all concerned: the developer and purveyor of the technology might lose interest and clients, and the users might withdraw from using services that could bring them benefit. Taking extended responsibility would help to reduce this as it assures the potential truster that even if failures occur responsibility will be taken, improvements made, betrayals recognised and responded to.

Let us consider an example in which a group of software engineers have developed the AI system for a new autonomous vehicle, which is being distributed by a major car company. The development and sale of the AI and the car as a whole is overseen by a regulatory body who have to certify it before it is released. Adding in the legal rules around intellectual property, the purchase of automobiles, and the company's own internal quality control policy on autonomous vehicles, this—loosely speaking—is the socio-technical context within which the new autonomous vehicle is embedded.[12] As we have already mentioned, all the above-mentioned actors: developers (the software engineers), the company, as well as the regulators, have a reason to want to ensure that this socio-technical system is trustworthy. If it is deemed to be so, then potential customers, as well as the wider public impacted by this system, will be more prepared to make themselves vulnerable to it by purchasing their products and by being prepared to accept these products into general society. Ensuring this trustworthiness is not achieved through any single measure, but ensuring a sufficient frequency of taking extended responsibility is vital. If autonomous vehicles cause sufficiently frequent harms to which the developers, companies, and regulators merely throw up their hands and (legitimately) dismiss these as unpredictable accidents, this will reduce the trustworthiness of the socio-technical system. On the other hand, if the developers, companies, and regulators adopt strategies of taking extended responsibility for these cases with sufficient regularity, then the trustworthiness can be preserved as if things go wrong they will be there to make amends and to meet our call of betrayal appropriately. For the socio-technical system as a whole to preserve its trustworthiness, what matters is that extended responsibility is taken, exactly how is not of particular importance provided they meet the constraints of ambiguous agency, proximity, and relevance if this taking is retrospective, and the constraints of distal and proximate control if the taking is prospective.

Though these two considerations might speak strongly in favour of taking extended responsibility, it remains unclear what follows from this. Is taking extended responsibility laudable but supererogatory or is it obligatory—perhaps a duty in some sense. In favour of the first option, van de Poel and Sand (2021: S4779) say that taking responsibility in the extended sense is virtuous, it exhibits an excellence of character on the part of the taker, but it is not obligatory. In their words:

> Such acts of taking responsibility are morally supererogatory, i.e. they are not morally required, but if agents take responsibility for them it is morally praiseworthy. This also suggests an association of the taking of responsibility and the possession of certain virtues. In fact, as we have argued before, responsibility itself may be seen as a kind of virtue that is associated with

[12] There are, undoubtedly, different ways to circumscribe the socio-technical system in this case – elements that some would add and some would subtract. However, as our argument here will deal with what are pretty central elements of this system, we do not think that these variations should undermine our central point.

the assuming, or taking, of specific backward-looking and forward-looking responsibilities.

In contrast, Kiener (2022) contends that there can be an obligation to take extended responsibility. This can be differently motivated, either by consequential reasons, deontic reasons or agent-affirmative reasons. For the first he contends that in "cases in which the harm to another party is severe and there is no one else who could take responsibility in a meaningful way" (Kiener, 2022: 588) we have an obligation to take responsibility. The deontic reasons stem from the possibility, already mentioned earlier, that refusing to take extended responsibility can be disrespectful or offensive to the victim—something we have an assumed duty to avoid. In terms of agent-affirmation, these reasons follow from the idea that taking part in our responsibility practices affirms our roles and practical identities as "real-world" agents. Much of this later is bound up with the discussion of the role that taking responsibility plays in preserving our relationships.

Both of these approaches have merits and costs, and though we will not deal with either account exhaustively here, it is worth considering both. To consign taking extended responsibility wholly to the realm of the supererogatory would mean that an agent who never took extended responsibility—who perhaps adopts this as a policy—would not be criticisable on any level for this lifelong dismissal. This is strikingly counterintuitive, and certainly revisionary to our practices. We absolutely do judge and criticize agents who never take extended responsibility, and so some explanation for this seems necessary. That said, and contra Kiener, it seems correct that any individual instance of taking extended responsibility will be supererogatory. The thought is that if taking extended responsibility was obligatory in a given case then it would not be clear that this could still be understood as a case of *extended* responsibility. There is a tension here between keeping extended responsibility extended, while at the same time reconciling the criticisability of those who never take extended responsibility.

Our solution is to argue that it is obligatory to adopt a general strategy of taking extended responsibility, but not obligatory to do so in any single particular instance. In this way, the obligation is akin to the duty of charity. The targeted context of autonomous technologies, maintaining our relationships and preserving the trustworthiness of the socio-technical environment—though reasons in favour of taking extended responsibility—does not oblige us to do so in any possible case (though they may, of course, motivate us to do so). However, if our relationships and the environment's trustworthiness are to be maintained and preserved, we must take such responsibility with sufficient frequency and competency. It is not incidental that the deterioration of our relationships or of trustworthiness is a gradual and systemic process. Throughout our discussion of these concerns, it has always been the case that the worry arises when a failure to take extended responsibility is systemic, or the failure is sufficiently reliable. This does mean that, when assessing an agent, we should be evaluating their overall approach to taking extended responsibility, not obsessing over much about any particular given case. Such general strategies do not have only one form, but can be instantiated in various ways, the unifying element being that extended responsibility will be taken with enough reliability that the costs associated with not doing so are staved off. In this sense, perhaps more than any other, the core consideration at work here is practical.

Conclusion

We have discussed how the introduction of autonomous technologies into our workplaces and into our lives pose a number of challenges for our responsibility practices. Though the mechanisms differ, a commonality among these challenges is that due to various features of autonomous technologies and the socio-technical context in which they are developed and embedded there can be cases where these technologies bring about harmful outcomes but where it is unclear who—if anybody—should or even can bear responsibility for them. Without dismissing other extant attempts at resolving these challenges we have argued that an underdiscussed tool in our responsibility toolbox has been the role of taking responsibility.

In agreement with prevailing work on the topic we contend that the conditions for taking responsibility are wider and more permissive than those for being held responsible. This allows for the taking of extended responsibility, i.e. the taking of responsibility (prospective or retrospective) for that which it would be illegitimate to be held responsible for. The benefit for the topic at hand is then shown to be that even if the responsibility concerns raised by autonomous technologies due result in cases where there is no apt target to be held responsible, these gaps can still be bridged since agents can take extended responsibility. For this to be a useful bridge it must manifest as the adoption of a responsibility strategies to take responsibility for the outcomes of autonomous technologies. Such strategies should be both prospective (being committed to taking responsibility for possible future outcomes) and retrospective (being committed to taking responsibility when outcomes have or do occur). We then consider how, given that extended responsibility is not strictly compelled by the rules of our responsibility practices themselves, the adoption of such strategies can be justified. Our contention is that there are usually practical considerations beyond the strict rules of responsibility that can make the adoption of such strategies obligatory. In the case of

autonomous technologies, we highlight considerations of retaining the quality of our relationships and preserving the trustworthiness of our socio-technical environment.

Finally, we contended that though any individual instance of taking extended responsibility must retain its supererogatory character in order to remain an instance of extended responsibility, at the same time there is clearly something criticisable about an agent who never takes extended responsibility. Our answer to this tension is to hold that agents are obligated to adopt general strategies of taking extended responsibility, where these strategies might vary in precisely which individual cases responsibility will be taken. The concern here remains overridingly practical: that the strategies adopted should be sufficiently reliable so as to avoid the possibility of a degeneration in our relationships or in the trustworthiness of the socio-technical context.

Acknowledgements Sincere thanks to the members of the RWTH Aachen Applied Ethics Group for the many discussions and suggestions that have helped to give this work the shape it has.

Author contributions N.C. and S.N. contributed equally on each section of this manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) under the Excellence Strategy as part of the cluster of excellence Internet of Production (EXC 2345/1), German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung) as part of the project BridgingAI (16DHBKI023).

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Baier, A. (1986). Trust and antitrust. *Ethics, 96*(2), 231–260. https://doi.org/10.1086/292745

Calhoun, C. (2019). XI—responsibilities and taking on responsibility. *Proc Aristotel Soc, 119*(3), 231–251. https://doi.org/10.1093/arisoc/aoz017

Calhoun, C. 2022. *A Social Practice Account of Responsible Persons*. Descartes Lectures.

Coeckelbergh, M. (2023). Narrative responsibility and artificial intelligence: How AI challenges human responsibility and sense-making. *AI & Society, 38*(6), 2437–2450.

Collingridge, D. (1980). The dilemma of control. *The Social Control of Technology*, 13–22.

Conradie, N. H. (2023). Autonomous Military Systems: Collective responsibility and distributed burdens. *Ethics and Information Technology, 25*(1), 20.

Conradie, N. H., & Nagel, S. K. (2024). No agent in the machine: Being trustworthy and responsible about AI. *Philosoph Technol, 37*(2), 72.

Danaher, J. (2016). Robots, law and the retribution gap. *Ethic Inf Technol, 18*(4), 299–309.

Danaher, J. (2022). Tragic choices and the virtue of techno-responsibility gaps. *Philosoph Technol, 35*(2), 1–26.

Enoch, D. (2012). Being Responsible, Taking Responsibility, and Penumbral Agency. In U. Heuer & G. Lang (Eds.), *Luck, value, and commitment: Themes from the ethics of Bernard Williams* (pp. 95–132). Oxford University Press.

Enoch, D. (2015). Tort liability and taking responsibility. In J. Oberdeek (Ed.), *Philosophical foundations of the Law of Torts*. Oxford University Press.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Forge, J. (2010). A note on the definition of "dual use." *Sci Eng Ethic, 16*, 111–118.

Genus, A., & Stirling, A. (2018). Collingridge and the dilemma of control: Towards responsible and accountable innovation. *Res Pol, 47*(1), 61–69.

Gunkel, D. (2020). Mind the gap: Responsible robotics and the problem of responsibility. *Ethic Inf Technol, 22*(4), 307–320.

Hall, M. A., Dugan, E., Zheng, B., & Mishra, A. K. (2001). Trust in physicians and medical institutions: What is it, can it be measured, and does it matter? *The Milbank Quarterly, 79*(4), 613–639.

Heine, K., & Quintavalla, A. (2024). Bridging the accountability gap of artificial intelligence – what can be learned from Roman law? *Legal Studies, 44*(1), 65–80. https://doi.org/10.1017/lst.2022.51

Himmelreich, J., & Lawford-Smith, H. (2019). Punishing groups: When external justice takes priority over internal justice. *The Monist, 102*(2), 134–150.

Hindriks, F., & Veluwenkamp, H. (2023). The risks of autonomous machines: From responsibility gaps to control gaps. *Synthese, 201*(1), 1–17.

Johnson, D. G., & Verdicchio, M. (2017). AI, agency and responsibility: The VW fraud case and beyond. *AI & SOCIETY, 32*(4), 595–604.

Kerasidou, C. X., Kerasidou, A., Buscher, M., & Wilkinson, S. (2021). Before and beyond trust: Reliance in medical AI. *J. Med Ethic, 48*(11), 852–856.

Kiener, M. (2022). Can we Bridge AI's responsibility gap at Will? *Ethic Theor Moral Pract, 25*(4), 575–593.

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology, 24*(3), 36. https://doi.org/10.1007/s10676-022-09643-0

Levy, N. (2005). The good, the bad and the Blameworthy. *J Ethic Soc Philosoph, 1*(2), 2–16.

Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford University Press.

Liu, H.-Y. (2016). Refining responsibility: Differentiating two types of responsibility issues raised by autonomous weapon systems. In N. Bhuta, S. Beck, & R. Geiß (Eds.), *Autonomous weapon systems: Law, ethics, policy* (pp. 123–140). Cambridge University Press.

Mason, E. (2018). Respecting each other and taking responsibility for our biases. In M. Oshana, K. Hutchison, & C. Mackenzie (Eds.), *Social dimensions of moral responsibility* (pp. 163–184). Oxford University Press.

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

Matthias, A. (2004). The responsibility gap in ascribing responsibility for the actions of automata. *Ethic Inf Technol, 6*(3), 175–183.

McKenna, M. (2012). *Conversation & Responsibility*. Oxford University Press.

Nyholm, S., & Smids, J. (2020). Automated cars meet human drivers: Responsible human-robot coordination and the ethics of mixed traffic. *Ethic Inf Technol, 22*(4), 335–344.

Oimann, A. K. (2023). The responsibility gap and LAWS: A critical mapping of the debate. *Philosoph Technol*. https://doi.org/10.1007/s13347-022-00602-7

Reiff, M. (2008). Terrorism, retribution, and collective responsibility. *Soc Theor Pract, 28*(3), 442–455.

Robillard, M. (2018). No such thing as killer robots. *J Appl Philosoph, 35*, 705–717. https://doi.org/10.1111/japp.12274

Ronca, G. (2022). Collective guilt, individual and prospective responsibility. *Inscriptions, 5*(2), 1–20.

Rosen, G. (2004). Skepticism about Moral Responsibility. *Philosoph Perspect, 18*, 295–313.

Shoemaker, D. (2015). *Responsibility from the Margins*. Oxford University Press.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosoph Technol, 34*(4), 1057–1084.

Sienknecht, M. (2024). Proxy responsibility: Addressing responsibility gaps in human-machine decision making on the resort to force. *Aus J Int Affairs, 78*(2), 191–199. https://doi.org/10.1080/10357718.2024.2327384

Smith, A. (2020[1759]). *The Theory of Moral Sentiments*. Sanage Publishing House.

Sparrow, R. (2007). Killer Robots. *J Appl Philosoph, 24*(1), 62–77.

Strawson, P. (1962). Freedom and Resentment. In *Proceedings of the British Academy* (Vol. 48).

Sverdlik, S. (1987). Collective Responsibility. *Philosoph Stud, 51*(1), 61–76. https://doi.org/10.1007/bf00353963

Taddeo, M., & Alexander B. (2022). 'Accepting Moral Responsibility for the Actions of Autonomous Weapons Systems—a Moral Gambit'. *Philosophy & Technology 35*(3), 78. https://doi.org/10.1007/s13347-022-00571-x.

Taylor, I. (2024). Collective responsibility and artificial intelligence. *Philosoph Technol, 37*(1), 27.

Thompson, D. F. (2014). Responsibility for failures of government: The problem of many hands. *Am Rev Pub Administrat, 44*(3), 259–273. https://doi.org/10.1177/0275074014524013

Thompson, D. F. (2017). Designing responsibility: The problem of many hands in complex organizations. In J. van den Hoven, S. Miller, & T. Pogge (Eds.), *Designing in Ethics* (pp. 32–56). Cambridge University Press.

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

Tigard, D. W., Conradie, N. H., & Nagel, S. K. (2020). Socially responsive technologies: Toward a co-developmental path. *AI & Society, 35*(4), 885–893.

van de Poel, I., Nihlén Fahlquist, J., Doorn, N., Zwart, S., & Royakkers, L. (2012). The problem of many hands: Climate change as an example. *Science and Engineering Ethics, 18*, 49–67. https://doi.org/10.1007/s11948-011-9276-0

van de Poel, I., & Sand, M. (2021). Varieties of responsibility: Two problems of responsible innovation. *Synthese, 198*(19), 4769–4787.

van de Poel, I., & Sand, M. (2024). Responsibility beyond control. In A. Placani & S. Broadhead (Eds.), *Risk and responsibility in context* (pp. 31–50). Routledge.

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

Vanneste, B. S., & Puranam, P. (2024). Artificial Intelligence, Trust, and Perceptions of Agency. *Academy of Management Review*, (ja), amr-2022.

Vaseashta, A. (2023). Existential risks associated with dual-use technologies. *Proc Stanford Existent Risk Conf*. https://doi.org/10.25740/zy474yf0050

Watson, G. (1996). Two faces of responsibility. *Philosoph Top, 24*(2), 227–248.

Wolf, S. (2001). The moral of moral luck. *Philosoph Exchang, 31*(1), 2–16.

Yazdanpanah, V., Gerding, E. H., Stein, S., Dastani, M., Jonker, C. M., Norman, T. J., & Ramchurn, S. D. (2023). Reasoning about responsibility in autonomous systems: Challenges and opportunities. *AI & Society, 38*, 1453–1464. https://doi.org/10.1007/s00146-022-01607-8

Zimmerman, M. J. (2022). *Ignorance and moral responsibility*. Oxford University Press.

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