Ethics and Information Technology (2023) 25:21
https://doi.org/10.1007/s10676-023-09699-6

ORIGINAL PAPER

The value of responsibility gaps in algorithmic decision-making

Lauritz Munch[1] · **Jakob Mainz**[1] · **Jens Christian Bjerring**[1]

Abstract
Many seem to think that AI-induced responsibility gaps are morally bad and therefore ought to be avoided. We argue, by contrast, that there is at least a *pro tanto* reason to welcome responsibility gaps. The central reason is that it can be bad for people to be responsible for wrongdoing. This, we argue, gives us one reason to prefer automated decision-making over human decision-making, especially in contexts where the risks of wrongdoing are high. While we are not the first to suggest that responsibility gaps should sometimes be welcomed, our argument is novel. Others have argued that responsibility gaps should sometimes be welcomed because they can reduce or eliminate the psychological burdens caused by tragic moral choice-situations. By contrast, our argument explains why responsibility gaps should sometimes be welcomed even in the absence of tragic moral choice-situations, and even in the absence of psychological burdens.

Keywords Responsibility Gaps · Algorithmic decision-making · Artificial intelligence

Introduction

Recent advances in machine learning suggest that AI systems in the near future could replace humans in carrying out many critical decision-making tasks (Bjerring & Busch, 2021; Baum et al., 2022).[1] Commonly cited examples involve assisting medical decision-making (Topol, 2019), screening job applicants (Langer et al., 2018), and operating cars (Levinson et al., 2011). A central worry about such deployments of AI systems concerns responsibility attributions.[2] If these systems are in the decision-theoretic driving seat, who can be said to be responsible in cases of erroneous decision-making? Some worry that the introduction of AI systems in vital stages of critical decision-making can result in so-called "responsibility gaps": roughly, outcomes for which no human agent can aptly be attributed responsibility.[3] To illustrate, consider the following case:

> **Decision-Procedure Designer.** The government intends to create a new law. According to this law, people with modest means can apply and receive monetary aid. Since incoming applications must be processed and decisions about eligibility must be made, the government faces a choice about which type of decision-maker to put in charge: (a) human decision-makers, or (b) an AI system capable of processing applications and making unilateral decisions.

Following the literature, this seems to be a paradigmatic example of how a responsibility gap might come to exist (Kiener, 2022; Danaher, 2016). Those who believe in the existence of responsibility gaps tend to motivate their belief by pointing to the autonomy and complexity of future (and

[1] Throughout the paper, we shall talk about 'AI systems', but we take this to include things like simple rule-based systems, machine learning systems, deep learning systems, etc.

[2] See Kraaijeveld (2020); Pagallo (2011); Tigard (2021); Matthias (2004); Sparrow (2007); Rubel et al. (2019).

[3] A burgeoning literature discusses whether artificial, non-human agents can be held responsible under certain conditions (see for instance Sebastián 2021; List 2021). We set this discussion aside here, since even if automatons could aptly be held responsible, this would change nothing from the perspective of our argument.

✉ Jakob Mainz
jtm@cas.au.dk

Lauritz Munch
laumu@cas.au.dk

Jens Christian Bjerring
filjcb@cas.au.dk

[1] Aarhus University, Aarhus, Denmark

Published online: 24 February 2023

even some currently conceivable) AI systems.[4] Let us go along with this idea and assume that the AI system in Decision-Procedure Designer is sufficiently complex and acts sufficiently autonomously, and that the government could not reasonably be expected to anticipate how the system will respond under all possible conditions of deployment. The government does therefore not satisfy commonly endorsed conditions for being responsible for its decisions (e.g., Matthias 2004). If the government opts for b), the thought goes, no one will therefore be fully responsible for all token decisions that the AI system will make down the line. And, as such, we create a responsibility gap by opting for b). By contrast, in opting for a), human decision-makers will be responsible for token decisions.

Of course, by deciding to implement the AI system in the first place, it seems plausible that the government is at least partially responsible for the actions of the AI system. This line of reasoning has led some to call into question the very possibility of responsibility gaps (Tigard, 2021; Hindriks & Veluwenkamp, 2023; Königs, 2022). Perhaps, for some salient class of decisions, there will always be a particular link in some causal chain of events leading up to the introduction of an alleged responsibility gap to which we can ascribe responsibility. Similarly, in the case of drunk driving, although the driver may have lacked control over the car in the moment of crashing, we normally still hold the driver responsible because we can trace the responsibility back to a significant action that the driver had control over, say, deciding to drink alcohol in the first place (Fischer & Tognazzini, 2009). If this "tracing back strategy" can be made to work for all apparent occurrences of a responsibility gap, then we might come to believe that such gaps do not in fact exist.[5] This aligns nicely with the intuition that many probably have about the case of drunk driving, namely that there is no responsibility gap exactly because the driver was responsible for driving in the first place. However, if the human agency involved in replacing human decision-makers with AI systems meant that someone was always morally responsible for the downstream decisions of the AI system, then it is puzzling that so many people have seemingly accepted that it is possible to create responsibility gaps in the first place. For if those picking the decision-procedure were always fully responsible for the decisions made by the AI system downstream, there would be no room left for the existence of responsibility gaps to exist in the first place.[6]

While there are important issues to address about the very possibility of responsibility gaps existing, we shall not dwell on these issues here but instead make our central line of argument conditional on their existence. Under the assumption that responsibility gaps do exist—something many seem to believe—we shall also bracket the question of whether they can be created *intentionally* as in cases where, for instance, a government deliberately decides to replace human decision-makers with an AI system.[7] Going forward, we will thus work on the assumption that if the government decides to opt for b) in Decision Procedure Designer, a responsibility gap is introduced.

Below we will offer two arguments for why responsibility gaps can sometimes be said to be desirable. The first argument focuses on preventing the consequences of being responsible for wrongdoing, whereas the second argument focuses on the badness of being responsible for wrongdoing as such. Hence, we argue that there is moral value in responsibility gaps—a claim which is of course consistent with the claim that responsibility gaps should, *all things considered*, be avoided in most real-world cases involving AI decision-making. While we are not the first to argue that responsibility gaps can sometimes be desirable (Danaher forthcoming), the broad focus of our argument on both the derivative as well as non-derivative badness of being responsible for wrongdoing is novel.

Now, suppose—perhaps unrealistically as things currently are—that the human decision-makers and the AI system in Decision-Procedure Designer are roughly identical in their accuracy levels and in the way that they distribute error-types, but that none of them are free of errors.[8] When each decision-system makes errors, it results in outcomes that harm people in the sense of not providing them with the monetary aid that they are duly owed.[9] In a case like this, it seems initially very natural to think that the government ought to choose option a). After all, since there are good

[4] For examples of theorists who believe that replacing humans with AI systems can create responsibility gaps, see (Matthias, 2004; Sparrow, 2007; Danaher, 2016; De Jong, 2020; Kiener, 2022; Danaher, forthcoming).

[5] See (Goetze, 2022) for discussion of the tracing back strategy.

[6] See (Simpson and Müller 2016) for discussion of this response.

[7] See also (Kiener, 2022) and (Hanson, 2009) for further discussions of how to "bridge" responsibility gaps.

[8] It has been argued that AI systems that make "social decisions" like the one in Decision-Procedure Designer are often extremely erroneous (Raji et al., 2022). We are not concerned with extant AI systems, but with AI systems that work as described above. We should also mention that many have argued that there are excellent reasons not to replace human decision-makers with AI systems for reasons unrelated to responsibility gaps. Finally, we should mention that it has been argued that AI systems are often no better than basic statistical techniques (Narayanan, 2019), making it less clear why we should be particularly concerned with replacing human decision-makers with AI systems per se. However, we shall set such worries aside for the sake of argument. We thank an anonymous reviewer for suggesting that we make these things explicit. observations and limpoint this out.

[9] Of course, the system will also distribute undeserved benefits to some. But since it is harder to see that such cases are wrongings of any particular individuals, we will focus on the other type of errors here.

reasons to ensure that responsibility can be placed in cases where critical decision-making goes wrong—for reasons to do with rectificatory justice for instance—the government should choose the human decision-makers over the artificial ones precisely because it is unclear who is responsible for the AI system's decisions.[10] If this reaction is correct, responsibility gaps are typically worrisome and should be avoided in contexts of critical decision-making. What "critical" decision-making amounts to exactly is of course a bit fuzzy. Factors that can render a decision more or less critical can include things like the badness of the outcomes when errors are made, the probability that errors will be made, the complexity of the decision, and so on and so forth. And arguably, the more critical the decision is, the more important it normally becomes to be able to hold *someone* responsible for the decision.

In this paper, however, we want to argue that the government at least has a *pro tanto* reason to opt for option b), even if doing so creates a responsibility gap. If our argument is sound, AI-induced responsibility gaps need not always be worrisome. Indeed, they might on occasion be desirable.

To flag our central line of reasoning, imagine that the government in Decision-Procedure Designer decides to opt for option a). As expected, a human decision-maker eventually makes an error that leads to some person being denied what they are properly owed. Not only is it regrettable that the citizen is being denied what they are properly owed, it is also regrettable that a person was directly responsible for such wrongdoing. As Victor Tadros has recently argued, being responsible for wrongdoing can both be *derivatively* and *non-derivatively* bad for the wrongdoer (Tadros, 2020). Being responsible for wrongdoing can be derivatively bad for the wrongdoer because it can lead to sanctions, blame, guilty conscience, etc., whereas it can be non-derivatively bad for the wrongdoer because it can make the wrongdoer's life go worse even in the absence of any derivative consequences of the wrongdoing. We motivate these claims later but notice first how they bear on the question at hand. If it is a bad thing to be responsible for wrongdoing, and if the government opts for a decision-process where we can expect some of the decision-makers to do wrong, then the government is opting for a decision-procedure that in one respect results in more moral badness than a salient alternative would do. There is a clear sense in which this is morally suboptimal. Or so we shall argue.

The paper is structured as follows. In section II, we make a few preliminary remarks about responsibility and responsibility gaps. In section III, we unfold our argument for the idea that there is *pro tanto* reason to create responsibility gaps, and we show what this claim implies for AI decision-making. In Section IV, we explain how our argument is situated in the current discussions on responsibility gaps. In Section V, we discuss and reject two objections to our argument. Finally, in Section VI, we make a few concluding remarks.

Responsibility and responsibility gaps

To prepare the ground for our argument, let us first be a bit more precise on the characterization of a responsibility gap. Following Johannes Himmelreich, let us say that:

> "[a] "responsibility gap" occurs whenever (i) an entity that is a "merely minimal agent" does X […], (ii) no-one else is (fully) responsible for X, and (iii) if X had been the action of a normal human person, then this person would have been responsible for it."[11] (Himmelreich 2019: 734).

Generally speaking, for a responsibility gap to be created, there needs to be an outcome that in some relevant sense is brought about by an agent—an entity that "acts"—to which we cannot attribute moral responsibility. This idea of an acting entity is reflected in Himmelreich's notion of a "merely minimal agent", which he characterizes as intentional agents who "can form beliefs, decide, and act but [who] cannot be responsible for their actions" (ibid.). As examples of merely minimal agents, Himmelreich cites "artificial agents such as AWS [autonomous weapons systems] or computer programs, but also group agents" (ibid.).

In light of Himmelreich's characterization, it is not hard to appreciate how a responsibility gap can occur in Decision-Procedure Designer. Given its stipulated ability to process applications and make unilateral decisions, the AI system in

[10] There are several different reasons for why it can be important to know who is responsible for erroneous decisions. First, it can be important because we sometimes need to know who to *punish* for erroneous decisions. Second, it can be important because it can help us avoid erroneous decisions in the future. Third, it can be important when we need to provide redress for the "victims" of erroneous decisions (Goetze, 2022); (Gotterbarn, 2001); (Nissenbaum, 1994). Fourth, it can be important when we need to provide explanations to victims of why errors were made (Coeckelbergh, 2021). Thanks to an anonymous reviewer for suggesting that we highlight these different reasons.

[11] Himmelreich only intends this to pick out one type of responsibility gap, notice, so it is not meant as a complete account. See (Hindriks and Veluwenkamp 2023) for discussion of Himmelreich's account and other ways of conceptualizing responsibility gaps. Hindriks and Veluwenkamp are skeptical of there being responsibility gaps, arguing that in the relevant range of cases, responsibility is always indirect or there is blameless harm (so there is no room for "gaps" in responsibility). As stated before, we remain neutral on the question of whether responsibility gaps exist, but notice that even if Hindriks and Veluwenkamp are right, our argument speaks to the desirability of cases of blameless harm.

Decision-Procedure Designer plausibly counts as a merely minimal agent. Moreover, since no one else in Decision-Procedure Designer is (fully) responsible when the AI system makes an erroneous decision, condition (ii) is satisfied as well. And, finally, since we would—at least under normal circumstances—have deemed humans responsible for harmful outcomes of the decision-making process, *had* they been in charge of administering the new law, condition (iii) is also satisfied for the AI system in Decision-Procedure Designer.

In the literature, it is not always clear what people have in mind when they raise concerns about responsibility gaps.[12] But conceptually, we can distinguish between

i) *being* responsible for some outcome O,
ii) being *fit* for being held responsible for O, and.
iii) being *held* responsible for O.

These distinctions can help us to better understand what is bad about responsibility gaps, but also what is good about them. When we say that someone *is* responsible for an outcome, we mean that a specific sort of connection obtains between the agent's agency and the outcome. This connection is invoked in claims such as "Steve was responsible for breaking the window by throwing a rock at it".[13] This is what is captured by i). When we say that someone is *fit* for being held responsible, we mean that it is appropriate to hold them responsible for some outcome given some standard of fittingness. Steve is fit to be held responsible for the broken window *because* it is fitting to hold him responsible for throwing a rock at it. This is what is captured by ii). Finally, if someone is *held* responsible for some outcome, this means that they are in fact held liable for rectificatory purposes—irrespective of whether it was fitting.[14] Steve is, for instance, held responsible (and fittingly so, it seems) when the owner of the window demands that Steve pays for a new window. This is what is captured by iii).

In many cases of critical decision-making, we need to know who *is* responsible for bringing about some erroneous outcome so that we know who is *fit* to be *held* responsible for bringing about that outcome. For example, when we need to decide who to fire because an erroneous decision was made, it is important that we know who was in fact responsible for the outcome, so that we do not end up firing someone who was in fact not fit for being held responsible. When responsibility gaps occur, no one *is* responsible for the outcome, and therefore no one is typically *fit* to be *held* responsible for the outcome. So the distinction between i), ii), and iii) is important when it comes to understanding the *badness* of responsibility gaps, because it helps us pin down the bad-making features of these gaps. Similar considerations apply when we focus on the *goodness* of responsibility gaps. After all, as we shall argue next, *being* responsible for wrongdoing is bad for the wrongdoer, but so is being *held* responsible for wrongdoing, regardless of whether one was in fact *fit* for being held responsible or not. If we can avoid these types of badness by introducing responsibility gaps, then responsibility gaps are at least *pro tanto* good.

Why responsibility gaps can be desirable

As mentioned, we think there are two distinct ways in which it can be bad for wrongdoers to be responsible for wrongdoing: it can be derivatively bad and non-derivatively bad (Tadros, 2020: 230). Let us begin with the former.

Being responsible for wrongdoing can be *derivatively* bad because it can bring about things that one should have reason to dislike such as sanctions, blame, guilty conscience, and so on. This derivative badness is primarily associated with the badness involved in being *held* responsible for wrongdoing. Of course, being responsible for wrongdoing can also be derivatively *good* for a wrongdoer. For example, if a person robs a bank and gets away with the loot, the wrongdoing can create many derivative goods for the person. When someone is responsible for wrongdoing—and the wrongdoing is discovered by others—the response is often to *hold* them responsible for their actions. They might get fired from their job, they might not be invited to social gatherings, they might be forced to pay compensation to the victims, and so on. It can be justified to punish people for certain types of wrongdoings. Committing sexual harassment in the workplace, for instance, may well warrant an employer to fire the offender. But surely, it is not always justified to punish people for their wrongdoings. Stealing a pencil from the workplace does not, for instance, seem to justify heavy sanctions such as being fired. Yet, following Tadros, "warranted or not, people have good reason to disvalue being responsible for wrongdoing because of these [harmful] effects" (Ibid.) That is, whether punishment is justified or not, punishment typically remains bad *for the*

[12] But see (Santoni de Sio and Mecacci 2021) who helpfully distinguish four interpretations of the term "responsibility gap"; see also (Goetze, 2022).

[13] There is a substantive question here about how "thick" the judgment that somebody is morally responsible for some outcome is. On the probably thinnest possible interpretation, A is morally responsible for some outcome O when A *caused* O. On a thicker notion, such as the one employed by Santoni de Sio and Mecassi (2021: 1062), responsibility for an outcome tracks blameworthiness (provided the outcome is one that warrants blame)—this they call "culpability".

[14] For our purposes we can understand the idea of being "held responsible" broadly. It may include activities such as blaming, punishing or harming.

wrongdoer. If we grant this idea, we have at least *some* reason to create responsibility gaps to remove people's responsibility for wrongdoings.

Why do we not just abstain from imposing costs on wrongdoers through punishment? The answer is that punishment, at least when fitting, seems to serve several important aims such as exemplifying a deterrent to wrongdoing, communicating blame and, perhaps, the idea that people get what they deserve. In other words, even though punishment might be regrettable according to some perspectives, we might nevertheless accept the corresponding disvalue because we are catering to more important aims that require us to punish people. In practice, then, we often accept the bad aspects of punishment as a necessary, but regrettable cost associated with promoting other moral aims. This tension is important to appreciate since it motivates us to look for ways to ensure that the regrettable trade-off of values does not arise to begin with—for instance, by working towards ensuring that people do not make themselves responsible for wrongdoing.

Admittedly, the thought that being responsible for wrongdoing is bad *for the wrongdoer* may seem odd. But note that the thought undergirds several uncontroversial policies. Think, for example, of the penal system and its deterrent function. While the deterrent function of the penal system is of course in place primarily for the sake of the *victims* of wrongdoing, it is also often argued that the function serves the *wrongdoer*. For wrongdoing makes peoples' lives go worse, whether or not a person is a victim or an offender. In this sense, part of the deterrent function of the penal system is to aid people in steering their agency away from wrongdoing (Tadros, 2020).

The *non-derivative* badness of being responsible for wrongdoing concerns the badness of *being* responsible for wrongdoing—whether or not anyone is in fact being *held* responsible by someone else. The thought that it can be non-derivatively bad to be responsible for wrongdoing is much more controversial than the thought that it can be derivatively bad. As such, it is also harder to motivate the non-derivative badness of wrongdoing since we cannot appeal to instrumental relations to other valuable things as a way of characterizing it.

Given that we cannot characterize the non-derivative badness of wrongdoing by appealing to downstream bad consequences of wrongdoing, we must do something else. We appeal to intuition:

> **Fight.** You are a parent to Anna. One day, you receive a phone call. Anna and another person, Dora, are both in the hospital with severe injuries from a fight. You do not know the details, but you are told that one of them attacked the other wrongfully, and that the other engaged in permissible self-defense. But you do not know who did what. Careful inspection of CCTV footage will hopefully reveal these details later.[15]

Imagine that Anna and Dora suffer from the same injuries. On the way to the hospital, you are wondering if your daughter wrongfully attacked Dora who then acted in permissible self-defense. What position would you hope that Anna occupies? We think that you should hope that Anna is *not* the one who wrongfully initiated the fight. How can we explain this intuition?

We might explain the intuition that Anna's parents should hope that it was Dora who wrongfully initiated the fight by reference to the derivative badness that Anna might experience if she is indeed the one responsible for the wrongdoing. If Anna is responsible for initiating the fight, she will likely be punished for her wrongdoing, and *this* is why—taking the perspective of Anna's parent—you should hope that it was Dora. But let us imagine that the CCTV footage is corrupted, and that neither Anna nor Dora can remember anything from the fight due to their injuries. Indeed, let us assume that it will never be known to anyone who initiated the fight and who merely engaged in permissible self-defense. But even in this case, it seems, you still have reason to want—*for the sake of Anna* (as well as for your own sake)—that she is not responsible for any wrongdoing. And this reason, plausibly, is explained by the fact we should generally hope for people that they do not end up directing their agency towards morally vile ends—even if their vile actions will not have any derivative bad consequences for them. If this is true, we also have motivation for believing that it can be non-derivatively bad to be responsible for wrongdoing.

Of course, you might still think that the reason you have for hoping that Anna was not responsible for any wrongdoing has to do with derivative badness, for example that a bad character trait might lead Anna into trouble in the future. But suppose that both Anna and Dora die during the fight, but that there is plenty of forensic evidence that one of them initiated the fight wrongfully while the other one merely defended herself. Suppose also that no one but you ever finds this forensic evidence. Even in that case, it seems, you should still hope that Anna was not responsible for wrongdoing. Accordingly, insofar as there is both derivative and non-derivative badness associated with wrongdoing, and insofar as we can remove these types of moral badness from the world by introducing responsibility gaps, responsibility gaps can be morally desirable.

However, even if it is true that it can be non-derivatively bad to be responsible for wrongdoing, one might now object

[15] This case is inspired by a similar case from (Tadros 2020).

that we should ignore this fact in practice.[16] The reason is, the thought goes, that wrongdoers *deserve* (some) punishment for their wrongdoings. While this might not necessarily make the punishment a good thing, perhaps it detracts from the badness of being responsible for wrongdoing. More generally, the idea that the punishment of wrongdoers detracts from the badness of being a wrongdoer can be attributed to (positive) retributivists, since they believe, more broadly, that it is intrinsically good that wrongdoers suffer (Walen, 2021; Tadros, 2011; Alexander & Ferzan, 2018). Transposed into the current context, retributivists might then deny that our reasons for introducing responsibility gaps are indeed good reasons. For if we can permissibly ignore the impact of at least some of the moral badness that wrongdoers suffer, responsibility gaps cannot do any significant moral work since the main attraction of introducing these gaps was exactly to remove the relevant moral badness. How should an adherent of the moral desirability of responsibility gaps respond to these retributivist ideas? Obviously, a satisfying answer to this question would involve a lengthy and detailed discussion about retributivism. But note that our argument is significant even if retributivism is true. The reason is that retributivism only entails that wrongdoers suffer in direct proportion to their desert base: the features in virtue of which suffering could be deserved. But in cases where the combined derivative and non-derivative badness of being a wrongdoer is likely to exceed the limits imposed by the desert base, we cannot permissibly ignore the suffering of wrongdoers. This is easiest to see when we consider the derivative badness of being responsible for wrongdoing. Consider, for example, a situation of decision-making on insufficient information grounds. If the government opts for a) in Decision-Procedure Designer, but the human decision-makers, at no fault of their own, only have very limited or misleading information at their disposal when making decisions, it is easy to imagine that the badness of being responsible for wrongdoing could exceed the limits of suffering imposed by the desert base. The suffering the decision-makers deserve in this case would probably be relatively limited, and yet the badness of being responsible for the wrongdoing of not granting applicants in need of monetary aid what they are duly owed can be very significant.

Recall that in Decision-Procedure Designer, neither type of decision-making will eliminate all errors from the decision-making processes. That is, both decision-procedures will entail that some people become victims of an injustice because they are denied what they are duly owed. But if the government opts for human decision-makers, there will be a further source of badness to take into consideration when these errors inevitably occur. For the errors are then caused by people's responsible exercise of their agency, either intentionally, or non-intentionally through recklessness or something similar. In other words, if the government opts for human decision-makers, there will be people who *are* responsible for wrongdoing. And, as we have argued, since it can be non-derivatively bad for people to be responsible for wrongdoing, being responsible for wrongdoing can still be morally bad even if no one is also *held* morally responsible for wrongdoing. By contrast, if the government opts for the AI system instead of human decision-makers, then although we will not—by assumption—eliminate more errors from the decision-making process, we will nevertheless ensure that no humans are responsible for wrongdoing when errors occur. Provided that we care about directing human agency away from wrongdoing, we thus have a *pro tanto* reason to prefer AI decision-makers over human decision-makers in cases like Decision-Procedure Designer. In a nutshell, why should we run the risk of implicating human agency in wrongdoing if we can avoid it by way of automating the relevant decision-making process? So when we ask whether human decision-makers should be replaced by AI systems, the mere possibility of creating responsibility gaps through the use of AI systems should no longer convince us to side with human decision-makers.

Situating our position

As mentioned, many believe that responsibility gaps are morally bad. Although there are several distinct explanations of what makes responsibility gaps morally bad, there is widespread agreement in the literature that they should be avoided. It is in this spirit that Christian List writes that

> "[s]ociety, via its regulatory authorities, should permit the use of autonomous AI systems in high-stakes settings only if structures are in place to ensure these systems'—or at least their legal representatives'—fitness to be held responsible for their actions." (List 2021: 17).[17]

.If we do not follow List's advice and strive to eliminate responsibility gaps wherever they occur, some worry that individuals can "hide" behind the decision of an AI system and ultimately reallocate responsibility from humans to technologies (List, 2021; Rubel et al., 2019; Feier et al., 2022). One common way to avoid responsibility gaps involves making sure that there is a "human in the loop" who, given suitable background conditions, is fit to be held responsible for the decisions made by the AI system (Baum

[16] Thanks to an anonymous reviewer for pressing us to discuss this objection.

[17] See also (Sparrow 2007); (Danaher 2016); (Felder 2021).

et al., 2022). On such a proposal, a human agent must oversee the AI system's decision-making to validate the correctness of its decisions and intervene if they prove erroneous. If and when errors occur, we can then typically be confident that the human in the loop is responsible for the outcome and maybe even hold them responsible too.

In contrast to common thinking, we have argued that there is a *pro tanto* reason to welcome responsibility gaps into automated decision-making. But just like there are different arguments *against* allowing responsibility gaps, so are there different arguments *in favor* of allowing them. To the best of our knowledge, only Danaher has so far endorsed the view that responsibility gaps should sometimes be welcomed. While we are in general agreement with Danaher in holding that responsibility gaps should sometimes be welcomed, we think that there are more reasons to welcome these gaps than Danaher recognizes. In this sense, our arguments complement his.

Danaher's argument proceeds from the observation that people often find themselves in genuine moral dilemmas where

> "two or more moral obligations or values compete with one another in such a way that they cannot be resolved or reconciled through decision-making. One of the obligations or values must be traded off against, or sacrificed in favour of, the other. This leads to a moral 'taint' or stain on our decision-making and makes moral decision-making a fraught and difficult business." (Danaher, forthcoming: 25).

.The "problem of tragic choice", as Danaher calls it, results from us having to resolve moral dilemmas in our lives. While not everyone is convinced that genuine moral dilemmas exist—or, if they do, how often they arise—Danaher appeals to them when motivating the thought that it can be psychologically burdensome to make choices when placed in such dilemmas.[18] For instance, it may be hard to live with the consequences of one's choice in a moral dilemma if one keeps second-guessing whether one did the right thing. It has, as Danaher puts it, a "phenomenological" impact on us that we must make tragic choices in the face of hard moral dilemmas. But automation may—via the creation of responsibility gaps—detach us from such dilemmas in a sense that may be psychologically beneficial to us.

So how does our argument complement Danaher's? As we have seen, Danaher's argument deals primarily with the *psychological* burdens of having to make decisions in *moral choice-situations* and with how responsibility gaps can alleviate such psychological burdens. In Danaher's own words:

> "I have defended an alternative perspective on techno-responsibility gaps. Although the prevailing wisdom seems to be against such gaps, and the policy proposals tend to try to find ways to plug or dissolve such gaps, I have argued that there may be reasons to welcome them. Tragic choices — moral conflicts that leave ineliminable moral remainders — are endemic in human life and there is no easy way to deal with them. […] That said, one potential advantage of advanced autonomous machines is that they enable a form of delegation with reduced moral and psychological costs." (Danaher, forthcoming: 23).

There are at least two ways in which our argument is broader in scope than Danaher's. First, our argument is not limited to cases of moral dilemmas. Our argument can thus explain why responsibility gaps can be morally good even in the absence of any such dilemmas. Second, our argument does not appeal to any psychological dispositions of wrongdoers but rather to the moral badness of being responsible for wrongdoing. Surely, it may *feel* bad to be responsible for wrongdoing, and especially so if you are morally conscientious and *held* responsible. But the badness associated with *being* responsible for wrongdoing is detached from any psychological facts about the decision-maker. By locating the badness associated with being responsible for wrongdoing outside the psychology of the decision-maker, we can explain why certain responsibility gaps can be desirable, even if the decision-maker who is replaced by an AI system experiences no psychological *burden* from being responsible for wrongdoing. To illustrate, consider the following example:

> **Co-Workers.** Smith and Jones both work in the HR department of a large company. Their primary task is to screen new job applications and decide on who should be called in for interviews. Both Smith and Jones are fanatic racists, and they consistently turn down job applications from black applicants simply because of their skin color. Smith gets a kick out of turning down black people's applications, while Jones has stopped caring many years ago, and now just turns them down out of old habit.

Suppose that we cannot, for whatever reason, replace Smith and Jones with other human decision-makers. In this case, it seems that Co-Workers is a paradigmatic example of a situation in which it would be morally desirable to replace human decision-makers with AI systems, even if the resulting process of automatization creates a responsibility gap.

But Danaher's account does not explain why. First, there is no moral dilemma in Co-Workers. Smith and Jones are

[18] See (Tessman, 2017) for discussion of moral dilemmas.

clearly morally blameworthy for their racist acts, and they ought to change their behavior and mindset to that of a non-racist. So there is no difficult moral choice to be made in this situation. But note, even if there was a moral dilemma, neither Smith nor Jonas would face any psychological burdens from being responsible for wrongdoing. Smith gets a kick out of turning down black applicants, whereas Jones' psychological response is indifferent. By contrast, since we are not limited to locating the badness of wrongdoing in people's psychology, we can straightforwardly get the right verdicts in cases like Co-Workers. Of course, it is consistent with Danaher's account to hold that it would be good both for the company and for the black applicants if Smith and Jones were replaced by an AI system. Yet, we can explain why it would be good to replace Smith and Jones directly by reference to the desirable properties of responsibility gaps. Whether Smith or Jones realize it or not, it would be good to replace their decision-making by an AI system because it would remove from their lives (as well as the lives of others) the badness that is associated with their wrongdoing. The lives of Smith and Jones go worse both for derivative and non-derivative reasons. Derivatively, their lives can go worse if their wrongdoings are discovered, or if they one day come to feel bad about their racist behavior. They can get fired, if their boss discovers that they systematically discriminate against black applicants. The black applicants might also find out what is going on and sue Smith and Jones for their wrongdoings, or their friends and family might decide that they do not want anything to do with them because of their racist attitudes. So it is clear that it can be derivatively very bad for Smith and Jones to be responsible for wrongdoing. But it is arguably also non-derivatively bad for Smith and Jones to *be*—and not merely *held*—responsible for wrongdoing. Even though Smith benefits psychologically from his racism by experiencing pleasure, and even though Jones is not psychologically burdened by his racism, both their lives would plausibly—from an objective point of view—go better if they were not wrongdoers. Following the Tadros-inspired argument we developed above, then, no matter if Smith and Jones realize it or not, and no matter if no one ever finds out what they did, there is a way in which their lives went worse because they were responsible for racist acts against so many black applicants.

Objections

Let us consider two objections to our argument. Although neither succeed, we believe, they point to interesting axiological nuances to our argument.

a. The Symmetry Objection.

According to the first objection, which we shall call the "Symmetry Objection", there is an axiological symmetry between the negative value of being responsible for a *wrong* decisional outcome, and the positive value of being responsible for a *right* decisional outcome. That is, if we think that being responsible for a wrong decisional outcome has negative value, then we should also think that being responsible for a right decisional outcome has some, or even corresponding positive value. This is an important observation, the objection goes, because in opting for automation we thereby eliminate not only the risk of people becoming responsible for wrongdoing, but also the possibility of people becoming responsible for doing something *right*, and hence something that makes their lives go better. So if the government decides to implement an AI system instead of human decision-makers in Decision-Procedure Designer, the government prevents not only humans from being responsible for wrongdoing, but also from being responsible for doing something good.

We concede—at least for the sake of argument—that doing the right thing has positive value, just like doing the wrong thing has negative value. Just as we think that a life goes worse if it is tainted by wrongdoing, so we think that a life goes better by doing good. But even if we grant this, it does not follow that the elimination of human agency from critical decision making is always unwelcome. For instance, it is not necessarily true that picking the right option in a particular choice situation comes with the same amount of positive value as picking the wrong option comes with negative value. Indeed, making the correct decision often gets its importance because it avoids the badness of making the wrong decision—and not because, say, the correct decision brings about a corresponding positive sum of goodness. To illustrate, imagine that you are a decision-maker tasked with screening job applications. You face the following two scenarios. In the first scenario, you will be processing two applications and you will classify both correctly as fitting candidates for an interview. In the second scenario, you will be processing four applications, three of which you will classify correctly and one of which you will classify wrongly because of a racial prejudice (you will be rejecting an otherwise qualified candidate because of their perceived race). Suppose making the correct decision and making the wrong decision were of equal value in the sense that correct classifications have + 10 value and incorrect ones have − 10 value. In that case, one should be indifferent about being in either scenario because the overall values in both scenarios add up to + 20. But that seems counterintuitive: it is natural to prefer being in the first scenario as opposed to the second, since each scenario adds up to + 20, but the second scenario involves acting out of racial prejudice. If so, it seems that it can be comparatively more important to avoid the badness

associated with doing wrong than harvesting the good associated with doing right. If this is true, we can then also be morally justified in omitting human agency from critical decision making—even if it means also removing *some* good from peoples' lives. Accordingly, it is not true, as the Symmetry Objection suggests, that if the government opts for b) in Decision-Procedure Designer, they will eliminate just as much potential for goodness as they would eliminate potential for badness.

Above we questioned if there is always symmetry between the absolute positive value of doing the right thing vs. the absolute negative value of doing the wrong thing. If we are right, the Symmetry Objection doesn't succeed. But we could also—for the sake of argument—assume that there is indeed such symmetry and yet show that our argument would be significant. This is because decision-procedure designers should not only be responsive to the value of possible outcomes, but also take into consideration what choices people would most likely make under a specific decision-scheme.

This point matters in the following way. Often, the possible options are not equally likely to be chosen. For example, imagine that a decision-designer is thinking about whether to implement an AI system or a set of human decision-makers and none of the available options are very good. Let's stipulate that each procedure on average makes the correct decisionin only 4 out of 10 cases. This is extremely bad, to be sure, and fortunately unlikely in many contexts, but we are making a principled point here. For simplicity, let's focus only on the non-derivative value of being responsible for the incorrect decision (−10), and the non-derivative disvalue of being responsible for the correct decision (+10), respectively. If we assume that 10 decisions will be made, the calculation will look as follows: The human decision-process will net us a total value of $(0.4 \times +10) + (0.6 \times -10) = -2$. That is, given the probabilities of making errors, the expected value of making the incorrect decisions exceeds the expected value of making the correct decisions.

When it comes to the AI system, on the other hand, there is neither non-derivative value nor disvalue from being responsible for outcomes, since no one is responsible for the outcomes. The AI system will therefore result in a net 0. Since $0 > -2$, this would be a case where we would have a reason to prefer the AI system over human decision-makers, even if we assume that the value of the correct decision and the disvalue of the incorrect decision are of equal magnitudes. Another way to put the same point is that if we strongly suspect that we are setting people up for (moral) "failure", we should feel the attraction of removing the option for people to do wrong entirely.

b. The Paternalistic Objection.

According to what we shall call the "Paternalistic Objection", our argument is objectionably paternalistic because it identifies a reason for replacing human decision-makers with AI systems for the sole sake of increasing the well-being of the human decision-makers. Generally, it is thought to be paternalistic—and perhaps objectionably so—when someone interferes with others' choices for the sake of improving the well-being of those interfered with (Dworkin, 2020). We have argued that we have a *pro tanto* reason to prefer AI systems over human decision-makers partly because of the potentially bad consequences that people will experience when the inevitable decisional wrongdoings will happen. Suppose, for vividness, that we choose to replace all medical doctors with AI systems to eliminate human wrongdoing. In such a case, it seems likely that the doctors would object that we are treating them paternalistically—depriving them of a preferred option of keeping their jobs—for the sake of their own well-being.

In response, note that the objection does not threaten our axiological claim. We have argued that there is a *pro tanto* reason to create responsibility gaps through automatization since they will lead to fewer people being responsible for wrongdoing. So even if the Paternalistic Objection is correct, it only shows that we have reason—*all things considered*—to avoid responsibility gaps in critical decision-making. While antecedent commitments to anti-paternalism dictate that we should not treat our identified *pro tanto* reason as decisive in debates about whether to automate a decision process, these commitments do not threaten our axiological claim. To wit, were we to opt for AI automation—perhaps for reasons to do with, say, decisional speed and accuracy—it would constitute an additional virtue if the resulting automatization process also created a responsibility gap: namely that fewer people would be responsible for wrongdoing.

What the objection points to, perhaps, is that we often place weight on aspects of a choice situation that goes beyond the disvalue associated with making the wrong decision.[19] As an example, consider again the medical profession. Perhaps there is a certain value involved in practicing medicine very skillfully, or perhaps it is important for creating and sustaining a certain set of professional norms and codes that people are allowed to make critical decisions (including the option to get things wrong). There is also a strong case to be made for the thought that many people find a sense of purpose and meaningfulness in being tasked with making high-stakes decisions where they can apply their skills to their fullest. For instance, a physician who takes practicing their profession as their conception of a good life could rightly feel insulted were someone to deny them an

[19] We thank an anonymous reviewer for asking us to elaborate on this point.

opportunity to pursue this conception for the sake of preventing them from making errors. Again, the ghost of paternalism seems to be lurking in the background.

But all this just goes to show that when thinking about how best to design critical decision-procedures, many factors must be considered. The arguments we have made here for automation are only a mere part of a much larger picture.

Concluding remarks

Many in the literature on AI-induced responsibility gaps seem to believe that such gaps are undesirable. To add more nuance to this debate, we have argued that AI-induced responsibility gaps can sometimes be desirable. The reason is that it is bad for people to be responsible for wrongdoing—something that makes people's lives go worse either derivatively or non-derivatively. This, we have argued, gives us a novel *pro tanto* reason for automating decision-making and removing human agency from critical decision-making processes. Accordingly, if responsibility gaps exist, and if they can be created intentionally, then we have shown that responsibility gaps need not always be morally problematic but rather something we should sometimes welcome in the decision-making process.

Our contribution is not only of theoretical importance, but also of practical importance. When we are faced with the task of deciding whether to replace human decision-makers with AI systems, it is crucial that we take all relevant considerations into account. While many have pointed out strong *pro tanto* reasons in favor of avoiding responsibility gaps, we have stressed the existence of *pro tanto* reasons in favor of creating them. Having all the relevant reasons on the table is important even if we ultimately decide—*all things considered*—that responsibility gaps are best avoided.

References

Alexander, L., & Ferzan, K. (2018). *Reflections on crime and culpability: problems and puzzles*. Cambridge: Cambridge University Press.

Baum, K., Mantel, S., & Schmidt, E., and Timo Speith (2022). From responsibility to reason-giving explainable Artificial Intelligence. *Philosophy & Technology*, *35*(1), 12.

Bjerring, J. C., & Busch, J. (2021). Artificial Intelligence and patient-centered decision-making. *Philosophy & Technology*, *34*, 349–371.

Coeckelbergh, M. (2021). *AI Ethics*. MIT Press.

Danaher, J. (2016). Robots, Law and the Retribution gap. *Ethics and Information Technology*, *18*(4), 299–309.

Danaher, J. Tragic Choices and the Virtue of Techno-Responsibility Gaps. *Philosophy and Technology*, forthcoming.

De Jong, R. (2020). The retribution-gap and responsibility-loci related to robots and automated technologies: a reply to Nyholm. *Science and Engineering Ethics*, *26*(2), 727–735.

Dworkin, G. (2020). Paternalism. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, https://plato.stanford.edu/archives/fall2020/entries/paternalism/.

Feier, T., Gogoll, J., & Uhl, M. (2022). Hiding Behind Machines: Artificial Agents May Help to Evade Punishment.Science and Engineering Ethics,28.

Felder, R. (2021). Coming to terms with the Black Box Problem: how to justify AI Systems in Health Care. *Hastings Center Report*, *51*(4), 38–45.

Fischer, J., & Tognazzini, N. A. (2009).The Truth about Tracing. Noûs, 43(3):531–56.

Goetze, T. (2022). Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement. *FAccT '22*

Gotterbarn, D. (2001). Informatics and professional responsibility. *Science and Engineering Ethics*, *7*, 221–230.

Hanson, F. A. (2009). Beyond the skin bag: on the moral responsibility of extended agencies. *Ethics and Information Technology*, *11*(1), 91–99.

Himmelreich, J. (2019). Responsibility for Killer Robots. *Ethical Theory and Moral Practice*, *22*(3), 731–747.

Hindriks, F., & Veluwenkamp, H. (2023). The risks of autonomous machines: from responsibility gaps to control gaps. *Synthese*, *201*, 21.

Kraaijeveld, S. R. (2020). Debunking (the) retribution (gap). *Science and Engineering Ethics*, *26*, 1315–1328.

Kiener, M. (2022). Can we Bridge AI's responsibility gap at Will? *Ethical Theory and Moral Practice*, *25*, 575–593.

Königs, P. (2022). Artificial intelligence and responsibility gaps: what is the problem?Ethics and Information Technology, 24(36).

Langer, M., Cornelius, J., & König, and Andromachi Fitili (2018). Information as a double-edged Sword: the role of computer experience and information on applicant reactions towards Novel Technologies for Personnel Selection. *Computers in Human Behavior*, *81*, 19–30.

Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., Zico, J., Kolter (2011). Towards Fully Autonomous Driving: Systems and Algorithms. *IEEE Intelligent Vehicles Symposium (IV)*, 163–68.

List, C. (2021). Group Agency and Artificial Intelligence. *Philosophy & Technology*, *34*, 1213–1242.

Matthias, A. (2004). The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*, 175–183.

Narayanan, A. (2019). How to Recognize AI Snake Oil. *Arthur Miller lecture on science and ethics, Massachusetts Institute of Technology*, http://www.cs.princeton.edu/~arvindn/talks.

Nissenbaum, H. (1994). Computing and accountability. *Communications of the ACM*, *37*(1), 72–80.

Pagallo, U. (2011). Killers, fridges, and slaves: a legal journey in robotics. *AI & Society*, *26*, 347–354.

Rubel, A., Castro, C., & Pham, A. (2019). Agency laundering and Information Technologies. *Ethical Theory and Moral Practice*, *22*(4), 1017–1041.

Raji, I., Elizabeth Kumar, A., Horowitz, & Selbst, A. (2022). The Fallacy of AI Functionality. *FAccT '22*.

Santoni de Sio, F., & Mecacci, G. Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. Philosophy & Technology, 34:1057–1084.

Sebastián, M. (2021). First-person representations and responsible Agency in AI. *Synthese*, *199*(3), 7061–7079.

Simpson, T., Vincent, C., & Müller. Just war and robot's killings.The Philosophical Quarterly, 66(263):302–22.

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77.

Tadros, V. (2020). Distributing responsibility. *Philosophy & Public Affairs*, *48*(3), 223–261.

Tadros, V. (2011). *The ends of harm: the Moral Foundations of Criminal Law*. Oxford: Oxford University Press.
Tessman, L. (2017). *When doing the right thing is impossible*. Oxford University Press.
Tigard, D. (2021). There is no techno-responsibility gap. *Philosophy & Technology*, *34*, 589–607.
Topol, E. (2019). High-performance medicine: the convergence of human and Artificial Intelligence. *Nature Medicine*, *25*(1), 44–56.
Walen, A. (2021). Retributive Justice, The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2021/entries/justice-retributive/.