Reasons for Meaningful Human Control

Veluwenkamp, H.M.

DOI
10.1007/s10676-022-09673-8

Publication date
2022

Document Version
Final published version

Published in
Ethics and Information Technology

Ethics and Information Technology (2022) 24:51
https://doi.org/10.1007/s10676-022-09673-8

ORIGINAL PAPER

Reasons for Meaningful Human Control

Herman Veluwenkamp[1]

Abstract

”Meaningful human control” is a term invented in the political and legal debate on autonomous weapons system, but it is nowadays also used in many other contexts. It is supposed to specify conditions under which an artificial system is under the right kind of control to avoid responsibility gaps: that is, situations in which no moral agent is responsible. Santoni de Sio and Van den Hoven have recently suggested a framework that can be used by system designers to operationalize this kind of control. It is the purpose of this paper to facilitate further operationalization of ”meaningful human control”.

This paper consists of two parts. In the first part I resolve an ambiguity that plagues current operationalizations of MHC. One of the design conditions says that the system should track the reasons of the relevant agents. This condition is ambiguous between the kind of reasons involved. On one interpretation it says that a system should track motivating reasons, while it is concerned with normative reasons on the other. Current participants in the debate interpret the framework as being concerned with (something in the vicinity of) motivating reasons. I argue against this interpretation by showing that meaningful human control requires that a system tracks normative reasons. Moreover, I maintain that an operationalization of meaningful human control that fails to track the right kind of reasons is morally problematic.

When this is properly understood, it can be shown that the framework of MHC is committed to the agent-relativity of reasons. More precisely, I argue in the second part of this paper that if the tracking condition of MHC plays an important role in responsibility attribution (as the proponents of the view maintain), then the framework is incompatible with first-order normative theories that hold that normative reasons are agent-neutral (such as many versions of consequentialism). In the final section I present three ways forward for the proponent of MHC as reason-responsiveness.

Keywords Meaningful Human Control · Normative reasons · Motivating reasons · Agent-relativity

Introduction

Rapid developments in Artificial Intelligence (AI) have changed many aspects of our daily lives. Moreover, AI generates much value in many different domains. The increased autonomy of the systems that are being used, however, also presents a potential problem. If a socio-technical system makes a mistake, it is not always clear who is responsible for that error. Situations in which responsibility is diminished because human beings give up control to autonomous systems are called responsibility gaps (Matthias, 2004; Sparrow, 2007, 2016) Responsibility gaps are taken to occur in many different domains, and have been discussed predominantly in relation to autonomous weapons and self-driving vehicles. In a much-cited paper, Sparrow (2007) argues that autonomous weapon systems should not be used in warfare because doing so would be in conflict with international humanitarian law. Humanitarian law requires that there is always someone who can be justly held responsible for deaths that occur in the course of the war. Sparrow takes this to show that autonomous weapons with responsibility gaps are impermissible.

Responsibility gaps are both discussed in relation to technological agents (e.g., (Coeckelbergh, 2019; Danaher, 2016; Himmelreich, 2019; Köhler, 2020) and collective agents (e.g., (Collins, 2019; Copp, 2007). In both contexts we see some authors denying that responsibility gaps exist (see for example (Köhler et al., 2017; Simpson & Müller, 2016; Tigard, 2020) for the claim that technological responsibility gaps do not exist and (Ludwig, 2007; Miller, 2018) for the claim that collective responsibility gaps do not exist).

✉ Herman Veluwenkamp
herman@veluwenkamp.net

[1] Delft University of Technology, Delft, Netherlands

Published online: 23 November 2022

There is also a lively debate on the question how to understand responsibility gaps (Himmelreich, 2019; Köhler et al., 2017; <reference removed>), and, relatedly, whether they are intrinsically problematic (e.g., (Königs, 2022; Robillard, 2018).

Authors who think that responsibility gaps are both problematic and real, have proposed different ways of dealing with those gaps. Hevelke and Nida-Rümelin (2015) have defended the claim that drivers of self-driving cars are collectively responsible, which, arguably, justifies a mandatory tax or insurance for those drivers. Alternatively, Nyholm (2018) maintains that we should think of human agents as supervisors of autonomous systems and draw on philosophical analyses of collaborative agency to deal with these collaborations.

Ideally, however, we would design autonomous systems in such a way that responsibility gaps do not even occur. That is, we should design these systems in such a way that there is always the kind of control that is needed to be able to talk about moral responsibility for the consequences of these systems.[1] In the literature, this kind of control has been labeled *meaningful human control*. An influential framework that has been developed to design for this kind of control is MHC as reason-responsiveness (Santoni de Sio & Van den Hoven, 2018). Santoni de Sio and Van den Hoven propose an account in which a system is under meaningful human control only if the system is responsive to the relevant moral reasons. They take meaningful human control as reason-responsiveness to require two conditions: the *tracking* and *tracing* condition. The tracking condition tells us that a socio-technical system should be able to respond to both the relevant (moral) reasons of the humans designing and deploying the system and the relevant facts in the environment in which the system operates. The tracing condition requires that (1) at least one human agent is present in the system design history or use context, who (2) has the right cognitive and physical capacities to fit their role; and (3) is adequately aware of such controlling role and their own active and passive responsibility. There is, however, an ambiguity in the kind of reasons involved: the account does not tell us if MHC requires a system to track the normative or the motivational reasons of the relevant agents.

In the first part of this paper, I resolve this ambiguity. I do this by first spelling out what the difference is between motivational and normative reasons, and why it is important for designers of socio-technical systems to make sure that these systems track the right kind of reasons. I then highlight the urgency of resolving the ambiguity further by showing (1) that MHC is currently interpreted as being involved with motivational reasons and (2) that, in contrast to what the current literature suggests, the framework should be understood in terms of normative reasons.

When this is properly understood, it becomes clear that this framework of MHC is committed to the *agent-relativity* of reasons. More precisely, I argue in the second part of this paper that if the tracking condition of MHC plays an important role in responsibility attribution (as the proponents of the view maintain), then the framework is incompatible with first-order normative theories that hold that normative reasons are agent-neutral (such as many versions of consequentialism). In the final section I present three ways forward for the proponent of MHC as reason-responsiveness.

Meaningful Human Control

Many autonomous systems can only be justifiably deployed if it is clear who is morally responsible when the system causes harm. Let us say that a system is under meaningful human control of an agent in context C if, in context C, this agent has that kind of control over the system to render the agent morally responsible. Many socio-technical systems are designed in such a way that they leave a human being “in the loop”. Sometimes these systems only recommend a certain course of action and a human agent must decide whether to adhere to this recommendation. Other examples of systems with a human ”in the loop” have a human being present to pause or stop a specific procedure. Although these human beings have some kind of control over the systems, being in the loop is not always enough to have meaningful human control in the sense defined above. If, for example, the system is not designed in a way that provides the human agent with enough information to assess the correctness of a recommendation made by the system, then she might lack the knowledge to judge when to intervene. And, this knowledge is required to appropriately hold her responsible for the decision to intervene or not. So if having a human agent ”in the loop” is not sufficient for meaningful human control, what exactly does MHC amount to?

In response to this question, Santoni de Sio & Van den Hoven (2018) have developed a theoretical framework that sees meaningful human control in terms of reason-responsiveness. In the philosophical literature on individual moral responsibility, reason-responsiveness is one of the dominant approaches. Reason-responsiveness grounds moral responsibility by reference to agents’ capacities for being appropriately sensitive to the reasons that are relevant for their actions (Brink & Nelkin, 2013; Fischer & Ravizza, 1998; Haji, 1998; McKenna, 2013; Wolf, 1990). Santoni de Sio and van den Hoven base their account for MHC and the

[1] Note that this is important even if responsibility gaps do not exist. There is a need for MHC, even if there are no responsibility gaps. In that case MHC should make sure that the *correct* human agent has the kind of control for responsibility.

two necessary conditions on these notions for individual moral responsibility. Giulio Mecacci and Filippo Santoni de Sio (2020) have operationalized this account further by elaborating on the reasons that are relevant for MHC. More specifically, they have introduced a proximity scale of the reasons that are relevant for the tracking condition. They observe that there are often a large number of reasons that bear on a single action, and that these reasons can be ordered on what they call a proximity scale. Mecacci and Santoni de Sio do not define this scale exactly, but notice that more proximal reasons are often closer in time to the action and also simpler than the more distal reasons. Let us look at an example of a self-driving system that is about to steer right.

> [...] the vehicle steering right could be explained by a driver's intention to exit the highway, as much as by her intention to go home, or even her broader plan to go to bed early to be well-rested the day after, which is part of her general goal of performing well in her profession... and the story may continue. These are all good and relevant reasons to want the system to steer right. In the terminology of our "scale of reasons" introduced earlier, we say that there are more distal reasons, e.g. the plan to safely go home, and more proximal reasons, e.g. the intention to steer right. (2020, p. 111)

Mecacci and Santoni de Sio focus on autonomous vehicles, but their account is supposed to apply to all autonomous systems. To avoid responsibility gaps, these systems should be responsive to both the proximal and the distal ”reasons” of the relevant agents. I use ”reasons” in quotation marks, because the entities that are identified in the text (intentions, plans, values, goals and norms) are strictly speaking not reasons at all. Take for example the intention to steer right. Of course, there is a reason in the vicinity here, but it is not the intention itself. And the fact that Mecacci and Santoni de Sio spell out their position in terms of non-reason entities is exactly the problem. If our goal is to specify design requirements for MHC, it is of crucial importance that designers know which entities to track. For, an agent can have intentions that do not correspond to that agent's goals, and one can form plans that are incompatible with the reasons one has. The question to which entities a system is supposed to responsive to, is therefore not merely theoretical, but also one of *crucial practical importance*.

Reasons

In the previous section we have explained what MHC is and noticed that not everyone interprets the tracking condition as concerned with reasons. Moreover, investigating the recent literature tells us that Mecacci and Santoni de Sio are not the only ones that spell out MHC in terms of entities that are different from reasons. Calvert et al., for example, remark that they'll use ”the term 'reasons' to denote any factor that can motivate and explain human behavior, such as intentions and plans.” (2018, p. 3321) and Steven Umbrello sometimes talks about reasons as ”intentions” (2020, p. 47) but also defines reasons as ”any element that can both prompt and demonstrate human behavior, such as objectives, programs and strategies.” (2020, p. 45).

To start resolving the ambiguity in the tracking condition, it is helpful to start distinguishing between two kinds of reasons: motivational and normative reasons.[2] A normative reason is, in Tim Scanlon's words, a "consideration that counts in favour of" someone acting in a particular way (1998). A motivating reason, on the other hand, is the reason that someone acts in a particular way. It is something that, according to the agent, counts in favour of acting. When I am driving a car, you can ask me what my reason is for turning left (motivating reason) and you can ask if there is a reason for turning left (normative reason).

Motivating and normative reasons are taken to be very different kinds of entities. A motivating reason is, on Donald Davidson's influential account, a belief-desire pair that both causes and rationalizes action (1963, p. 687). On this account, motivating reasons are mental states. This contrasts with the dominant view on normative reasons which takes these reasons to consist of facts (Raz, 1975; Scanlon, 1998).[3] E.g., the fact that my grandmother is ill, is a reason

[2] Some people distinguish a third kind of reason: explanatory reasons (Alvarez, 2010; Hieronymi, 2011). Explanatory reasons are, just as motivating reasons, usually taken to be mental states. When I say ”motivating reason” in this paper, this can often be understood as ”motivating reasons or explanatory reason”. For ease of exposition I won't discuss explanatory reasons in this paper.

[3] Factualism about normative reasons is indeed the dominant position. However, so-called reason internalists have argued that normative reasons are always related to motivations and/or intentions (Williams 1979; Schroeder 2007). One of the problems for reason internalists is that in some domains (i.e., the moral and the epistemic), factualism is very intuitive. Many find it, for example, difficult to accept that the paradigmatic bad guy only has a normative reason to better his life if he has the intention or motivation to do good. In the same spirit, it is difficult to accept that scientific beliefs should be responsive to something other than evidence: motivations or intentions do not seem to be relevant for rational belief. This has led some internalists to combine the claim that motivations relate to reasons with the position that all agents have the same motivations if they were in some idealized condition (Korsgaard, 1986; Smith, 1994). This, arguably, allows internalists to avoid the problems described

for me to visit her in the hospital.[4] This does not mean, however, that normative and motivating reasons are unrelated. When everything goes right, that is, no mistakes are made, normative reasons figure in the content of the mental states that are the motivating reasons for an action. For example, if my motivating reason for playing padel is my belief that playing padel is healthy and the fact that playing padel is healthy is indeed a normative reason to play it, then I have acted for a good reason.

The distinction between normative and motivating reasons is relevant for several reasons. One of the important ones is that we can be mistaken about the reasons we take ourselves to have. Let us look at an example. Alida might think that it is going to rain, and therefore bring an umbrella to work. We can say that Alida has a motivating reason to bring an umbrella. The motivating reason is the belief that it is going to rain and her desire not to get wet. But this doesn't mean that Alida also has a normative reason to bring an umbrella to work.

Of course, if Alida has a motivating reason to bring an umbrella, then so takes herself to have a normative reason as well. There are, however, two different ways in which Alida can be mistaken about this. Firstly, she can make a normative mistake: she can fail to realise that the fact that it is going to rain is a reason to bring an umbrella. And, secondly, she can make a non-normative mistake: she can be mistaken about the weather. In both cases, Alida has a motivating reason without the accompanying normative reason.

On first glance, one might think that the motivating/normative reasons distinction is similar to the distinction between proximal and distal reasons.[5] For example, one might think that proximal reasons are motivating reasons, while the distal reasons correspond to the normative reasons an agent has. And while I agree that this is sometimes the case, this is only so because we can have both normative and motivating reasons on all levels of the proximity scale, not because a specific level corresponds to one or the other type of reasons. In fact, you can have distal motivating reasons and proximal normative reasons. For example, the vehicle steering right could be explained by a driver's proximal intention to exit the highway, but also by her distal intention of performing well in her profession. And, the belief that steering right is a good way to exit the highway can be a motivating reason to steer right, just as the belief that steering right ultimately leads to an agent performing well in her profession. However, the agent can be mistaken about these reasons, so, for both levels she can have the motivating reason without the normative reason.

I have distinguished between normative reasons (facts) and motivational reasons (mental states). In the quotes at the beginning of this section, we also saw talk of intentions, goals, norms and values. Goals, norms and values are related to reasons, but I think that it is not helpful to discuss these entities. If they are helpful in understanding MHC, this is only so because of their relation to the right kind of reasons. Intentions are a slightly different matter. The reason for this is that they are mental states that are conceptually very close to motivating reasons: the intention to act is formed on the basis of one's motivating reason. However, as I will argue, MHC should track neither.

We have seen that MHC as reason-responsiveness is ambiguous between motivating and normative reasons. We have, furthermore, seen that participants in the literature interpret the account in terms of motivating reasons (or intentions). Finally, we have seen that the distinction between normative and motivating reasons is important, because one can be mistaken about the motivating reasons one has. In fact, one can have normative reasons that conflict with one's motivating reasons. One can be motivated to bring an umbrella to work because of a belief that it is going to rain, and a desire not to get wet. If, in fact, it is not going to rain, then one has a normative reason not to bring the umbrella. This is a rather insignificant example, but one can imagine that if the stakes are higher, it is of crucial importance that the correct kind of reasons is being tracked by a system. I will therefore, in the next section, argue how the tracking condition of MHC should be spelled out: i.e., in terms of normative reasons.

Normative reason-responsiveness

MHC derives its attraction partly from a philosophical tradition that sees responsibility as responsiveness to reasons. It is, therefore, important that MHC tracks the same kind of reasons as the individual accounts of moral responsibility. In this section I will explain the individual accounts and argue that they have to be concerned with normative reasons.

above. However, the consequence of such a move for MHC is that the tracking condition should refer to these ideal motivations, and not the actual motivations, of agents. On these versions of internalism it would still be a mistake to track the actual intentions and motivations of agents. Alternatively, internalists could bite the bullet and deny that it is a problem that the paradigmatic bad guy does not have a reason to be moral. In that case, the upshot of this paper is that it has made explicit an assumption for MHC about the nature of reasons.

[4] Much of the debate in metaethics concerns metaphysical and epistemic questions regarding normative reasons. Motivating reasons are usually taken to be relatively unproblematic. A metaethical error theorist, for example, typically does not deny that there are motivating reasons. Her position only commits her to denying that there are normative reasons. And the metaethical realist could hold that motivating reasons are mental states and agent-relative; while defending the claim that normative reasons are a fundamental part of reality.

[5] This option was suggested to me by one of the authors above.

Reason-responsive theories are introduced to distinguish between agents who can be held responsible for their actions, and those who cannot. The basic idea is that there is a certain level of control necessary in order to say that an agent is responsible for the outcomes of an action: if you don't have control over an action, you also can't be blamed for it. Reason-responsive theories of responsibility cash out this level of control in terms of a responsiveness to reasons: if you don't have the ability to act on (moral) reasons in a certain circumstance, then you are in that circumstance not in control.

The circumstances that theorists have in mind when developing these theories, are those in which an agent is hypnotized, has irresistible desires because of an addiction, or is some other relevant sense manipulated. In those circumstances, agents are not able to act for good reasons. The drug addict, for example, has a reason to stop taking drugs, go to work, etc., but does and cannot act on those reasons because of her irresistible addictive desires. Different theorists have developed the account in slightly different ways, but the core idea is that free action presupposes an ability to act on good reason (Brink & Nelkin, 2013; Fischer & Ravizza, 1998; Wolf, 1990).

Santoni de Sio and van den Hoven take Fischer and Ravizza's account as their starting point, so let us look at their account in more detail. Fischer and Ravizza ask us to consider an agent S and a mechanism, M, that leads to agent S's action. We can then say that the mechanism M is reason-responsive in the right sense if and only if M (1) in many circumstances leads to a recognition of the relevant reasons and (2) in some circumstances causes S to act on some reasons (Fischer & Ravizza, 1998, ch. 3).

Now let us consider the drug addict. She is not morally responsible, because there are no (or only a few) circumstances that cause her to act for some reason. The drug addict has several reasons, she has a normative reason to stop using drugs, to get a job, etc. She also has motivating reasons, most notably, she has irresistible addictive desires. The drug addict's failure to have the appropriate degree of control consists in her failure to respond to her normative reasons. This means that Fischer and Ravizza have to be concerned with normative reasons (instead of motivating reasons): it are *those* reasons that the drug addict is unable to act upon.

Wolf (1990) and Nelkin (2011) are much more explicit on this point, but I think it is also clear that Fischer and Ravizza cannot be talking about motivating reasons here (see also (McKenna, 2013; McKenna et al., 2017). Of course, even someone who is hypnotized performs her actions because of motivating reasons. Similarly, even someone who acts from an irresistible addictive desire acts on a desire for a temporary high combined with the belief that the drugs will cause that high. In those circumstances the agents have beliefs and specific desires that cause them to act. The manipulation only makes it the case that these beliefs and desires are unreliable. The good reasons that hold in those circumstances, that stopping would lead to a better life, more friends, etc., are not recognized and not acted upon. That is what makes it the case that these people are not in control, and therefore not responsible for their actions. Therefore, it is evident that these accounts have to spell out reason-responsiveness in terms of normative reasons.

This means that if *it is indeed the case that MHC derives its plausibility partly from the traditional accounts of individual responsibility, they need to spell out the tracking condition in terms of normative reasons as well*. This also means that we do have to make some changes to the conclusions that the authors from the previous section make. Mecacci and Santoni de Sio's proximity scale is still very useful. However, in engineering design, the tracking condition should be operationalized in terms of proximal *normative* reasons and distal *normative* reasons. This is crucial, because, as mentioned above, we can have distal normative reasons that conflict with our distal intentions (and the same holds for proximal intentions and reasons). Analogously, Umbrello's conclusion shouldn't be that ”a smart home climate control system […] should be designed in such a way that [it] is able to coherently co-vary with that agents' (moral) *reasoning*” (2020, p. 45, emphasis mine). Instead, it should co-vary with an agent's (moral) *reasons*. Again, this is crucial, because one's reasoning doesn't have to correspond to one's reasons (this corresponds only if an agent reasons well). With this result I conclude the first part of this paper.

The incompatibility argument

In the second part of the paper I argue that MHC as reason-responsiveness is incompatible with moral theories that hold that normative reasons are agent-neutral. This argument has three premises and is structured as follows:

P1 MHC should tell us who is responsible if a socio-technical system causes harm.

P2 The tracking condition of MHC plays an important role in responsibility attribution.

P3 If all reasons are agent-neutral, then the tracking condition of MHC cannot play an important role in responsibility attribution.

C MHC is incompatible with views that hold that all reasons are agent-neutral.

The first premise

This argument is valid, so let us investigate the two premises in turn. As MHC is a term of art, it is not uncontroversial what exactly the content of the term should be. However, we can specify success-conditions of a definition by looking at the problem the term is supposed to solve. The demand for MHC has primarily arisen in response to worries related to responsibility gaps. So it is plausible to evaluate an operationalization of the notion with respect to its ability to take away these worries (see also (Horowitz & Scharre, 2015). What exactly is the responsibility worry? Sparrow considers whether it is morally permissible to send AI robots into war, by asking who is responsible if these systems are involved in actions that we would normally consider war crimes. He discusses several potential loci of responsibility: the designers and programmers, the commanding officer and the machine itself. He argues that these options are all unattractive. Yet, he notices, the principle of jus in bellum (which specifies conditions for fighting a just war) dictates that we should be able to appropriately hold someone responsible for deaths that occur in war. From these premises, Sparrow concludes that we shouldn't send AI robots into war.

We can identify at least two different types of responsibility worries. The first is existential in nature: *is* the appropriate (group of) person(s) responsible when an autonomous system causes harm? The second worry relates to an epistemic question: do we *know* who is responsible when an autonomous system causes harm?[6] These questions are often not properly distinguished, but it is important to see that answers to these questions can come apart. There are, for example, contexts in which the appropriate agents *are* responsible, but in which it is impossible to hold these agents responsible because of epistemic uncertainties. These uncertainties can have their origin in the fact that many different people are involved in the development of an autonomous system (problem of many hands) (Poel et al., 2015) or the opaqueness of machine learning algorithms involved in the development (Burrell, 2016).

If we take MHC to be a response to Sparrow's worry, then it has to be an answer to both the existential and the epistemic responsibility worry. It both has to give the appropriate agents control of the autonomous system and it should give us conditions under which we can hold someone responsible when unjust harm occurs. Because if Sparrow is right, that is a necessary condition for sending drones into war. We fall short of meeting this requirement if we are only told that the drone was under meaningful human control, without giving us a way of determining from this more abstract fact who the proper bearer of responsibility is.

This demand can be made clearer by looking at an example. Suppose that an account of MHC tells us that a system is under MHC iff it always chooses that option that minimizes the number of casualties. Although a proper specification of the rules of engagement obviously needs to be more complex, this abstract objective would at least incorporate a lofty goal. But would it constitute an answer to the epistemic responsibility worry? It is clear that it does not. The reason for this is that if someone gets killed unjustly, the account doesn't give us any way of telling which agent is responsible for that action (in other words, it doesn't provide an answer to the epistemic question). It is therefore plausible that *MHC should tell us who is responsible if a socio-technical system causes harm.*

The second premise

The framework of MHC specifies two conditions for meaningful human control. The second, tracing condition tells us that the system must be designed in such a way that it is always possible to trace back the outcomes of the system to at least one moral agent. This means that it is involved in attributing responsibility to moral agents. However, the tracking condition is understood to play an important role in responsibility attribution as well. In the fragments below we can see some of the proponents of the MHC framework claiming that the tracking condition plays this role:

> Identifying reasons bearers is important because it allows to determine which agents, and to what extent, are or could be in control of the behaviour of a certain system, and what it takes for a given system to be under the control of given agents (Mecacci & Santoni de Sio, 2020, p. 110).

Depending on the extent that the vehicle fully responds to [the drivers'] intentions, they are fully responsible for what the vehicle does. (Calvert et al., 2018, p. 3324)[7]

The idea is that the agents whose reasons are being tracked have the kind of control over the system that makes them morally responsible for the actions of the system. And this idea is very attractive, as it establishes a connection between an agent and the action that is independent of any direct causal link. This shows that at least some of the proponents of MHC as reason-responsiveness hold that *the tracking condition of MHC plays an important role in determining who is responsible if a socio-technical system causes harm.*

[6] Note that a sceptic about responsibility gaps can rephrase these two questions as questions about control.

[7] Note that in this quote we can also find the assumption that a vehicle should be responsive to the *intentions* of the agent to be under meaningful human control. As I have argued in the first half of the paper, this claim is based on a mistake.

The third premise

Now let us turn to the third premise: if all reasons are agent-neutral, then MHC as normative reasons responsiveness cannot tell us who is responsible if a socio-technical system causes harm. To motivate this claim, I first have to distinguish agent-neutral from agent-relative reasons. This distinction was originally introduced by Thomas Nagel, and is nowadays considered as one of the most important distinctions in normative ethics.[8] Here is Nagel introducing the distinction:

> [An agent-relative] reason is one whose defining predicate *R* contains a free occurrence of the variable *p*. [...] All universal reasons and principles expressible in terms of the basic formula either contain a free-agent variable or they do not. The former are [agent-relative]; the later will be called [agent-neutral]. (Nagel 1970, p. 91)[9]

On a first reading, this formulation is difficult to grasp, but the idea behind it is simple and powerful. Let us look at a few different reasons to explain the notion. I have a reason to drive carefully so as to avoid accidents. As this description of the reason does not mention a specific agent, the reason is agent-neutral: if I have a reason to drive carefully in certain circumstances, then you have that reason as well if you are in the same circumstances. Let us now look at a slightly more interesting example. Suppose that I have a reason to drive to my mother. If I have this reason because it is *my* mother, then this reason is agent-relative. However, if the reason for visiting my mother is that it produces a certain amount of happiness, then this is an agent-neutral reason. Because, if I could have driven to someone else and his would generate as much happiness, then this would constitute as strong a reason to perform that action.

Traditional utilitarianism holds that we have a moral reason to perform that action that maximizes happiness. On this theory I have a moral reason to visit my mother, or to drive carefully, if and only if it maximizes happiness. Moreover, as every agent has the same reasons, traditional utilitarianism says that all reasons are agent-neutral. Not all theories deny that there are agent-relative reasons. Some (extreme) versions of deontology, for example, hold that we always have a moral reason to refrain from stealing. They do think that this is so, even if refraining from stealing would lead to more stealing (by other people). Although consequentialists have argued that this is irrational, deontologists typically respond by holding that the reason not to steal is agent-relative. So, I have a reason to make sure that *I* do not steal; it is not the case that I have a reason to make sure that there is no stealing. Deontologists often also hold that some other reasons are agent-neutral. What distinguishes these versions of deontology from traditional utilitarianism is that only the latter says that *all* reasons are agent-neutral.

So there are many views in normative ethics that hold that all reasons are agent-neutral. Classical utilitarianism is one of the best-known theories that is committed to the agent-neutrality of reasons, but many other versions of consequentialism do so as well. Most versions of consequentialism that deny that all reasons are agent-neutral take at least a large subset of normative reasons to be agent-neutral.[10] Moreover, even some deontologists hold that many reasons are agent-neutral. It is therefore safe to say that it is a widespread view in normative ethics that many reasons are *agent-neutral*.

We are now in a position to argue for the claim that if all reasons are agent-neutral, then the tracking condition of MHC cannot tell us who is responsible if a socio-technical system causes harm. For convenience, let us suppose that classical utilitarianism is true (I will later explain why this assumption is not a problem for the argument). Classical utilitarianism entails that all (moral) normative reasons are agent-neutral reasons: if something is a reason for you, it is also a reason for me. Now imagine an autonomous vehicle driving on the road, and different agents (operators, designers, etc.) who all have reasons for action in the circumstance the vehicle is in.

The tracking condition of MHC tells us that the vehicle is under meaningful human control only if it is responsive (it tracks) the normative reasons of the relevant agents. If this condition is supposed to tell us who is responsible in cases of harm, this means that we first have to find out who the relevant agents are, then determine the normative reasons they have, and then act on the basis of these reasons. So far, so good. However, if these reasons are agent-neutral reasons, then there is no set of reasons that are, for example, the operators' reasons. On this account of reasons, reasons are not uniquely tied to one agent (or a group of agents). If something is a reason for the operator, it is also a reason for the designer. This means that there is no set of normative reasons that ”belongs” to one agent, and not to another. So *if all reasons are agent-neutral, then the tracking condition*

[8] For expressions of this sentiment, see (Hurka, 2003, p. 628; Portmore 2013; Ridge, 2011).

[9] When Nagel first introduced this distinction, he used the terms 'objective reasons' and 'subjective reasons'. However, Derek Parfit soon convinced him that the terms 'agent-neutral' and 'agent-relative' were more apt (Parfit, 1984). I have modified the quote above to match the updated terminology.

[10] Some philosophers even hold that a normative theory shouldn't count as consequentialism unless it holds that all reasons are agent-neutral (Zong, 2000).

of MHC cannot play an important role in responsibility attribution.

Note that this is not the case for an account of MHC that tracks motivating reasons. Motivating reasons are mental states, and therefore by definition agent-relative (even if all normative reasons are agent-neutral). Given P1, P2, P3 and the validity of the argument, we have established that the way MHC was initially defended is incompatible with the claim that all reasons are agent-neutral.

Three ways forward

In this paper I have made two claims. Firstly, MHC should be understood in terms of normative reasons. In the first half of the paper I have indicated how our understanding of MHC should be changed in order to reflect this fact. The second claim I have defended in this paper is that the way MHC was initially defended is incompatible with agent-neutral reasons. In this final section, let me discuss what I believe are the available options for the advocate of MHC.

The first, and most obvious, way to resolve the incompatibility from the previous section is to deny the agent-neutrality of reasons. This would be rather costly however, as many normative theories are committed to agent-neutral reasons (what makes this even more costly is the fact that also many non-philosophers accept some version of consequentialism). However, many people who engage in this discussion write as if reasons are agent-relative. So perhaps some don't really consider this as a real problem.

The second option is to define MHC as motivational reason-responsiveness. The upshot of this approach is that it doesn't rule out first-order moral theories. This is so, because even if all normative reasons are agent-neutral, motivational reasons are by definition agent-relative (as they are characterized as the mental states of agents). As we have seen above, however, MHC as motivational reason-responsiveness cannot rely for its plausibility on theories of individual moral responsibility. Anyone who chooses this path will have to seek legitimacy elsewhere.

I consider the last option most plausible. It involves rejecting that the tracking condition plays an important role in assigning responsibility. On this understanding, the tracking condition tells us only that an autonomous system should track the relevant normative reasons and the relevant facts in the environment in which the system operates. The relevant normative reasons can be agent-neutral. The tracing condition is then the sole determinant of responsibility. The moral agents to which the actions can be traced back (as formulated in the original tracing condition) are the people who are in control and responsible for the outcome of the system.

An increasing number of philosophers, designers and policy-makers see the importance of meaningful human control over autonomous systems. I believe that with the recommendations from this paper, we are in a better position to further operationalize meaningful human control and design better autonomous systems.

Acknowledgements I would like to thank Jeroen van den Hoven and Filippo Santoni de Sio for helpful feedback on an early draft of the paper.

1 References

Alvarez, M. (2010). *Kinds of Reasons: An Essay in the Philosophy of Action*. Oxford University Press.

Brink, D. O., & Nelkin, D. K. (2013). Fairness and the Architecture of Responsibility. *Oxford Studies in Agency and Responsibility*, *1*, 284–313.

Burrell, J. (2016). How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data and Society*, *3*(1), https://doi.org/10.1177/2053951715622512.

Calvert, S. C., Mecacci, G., Heikoop, D. D., & De Sio, F. S. (2018). *Full platoon control in truck platooning: A meaningful human control perspective*. 3320–3326.

Coeckelbergh, M. (2019). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-019-00146-8.

Collins, S. (2019). Collective Responsibility Gaps.Journal of Business Ethics, *154*(4).

Copp, D. (2007). The collective moral autonomy thesis. *Journal of Social Philosophy*, *38*(3).

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

Davidson, D. (1963). Actions, Reasons, and Causes. *Journal of Philosophy*, *60*(23), 685. https://doi.org/10.2307/2023177.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility* (2Issue vol.). Cambridge University Press.

Haji, I. (1998). *Moral Appraisability: Puzzles, Proposals, and Perplexities*. Oxford University Press.

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. *Science and Engineering Ethics*, *21*(3), 619–630. https://doi.org/10.1007/s11948-014-9565-5.

Hieronymi, P. (2011). Reasons for Action. *Proceedings of the Aristotelian Society*, *111*(3pt3), 407–427. https://doi.org/10.1111/j.1467-9264.2011.00316.x

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

Horowitz, M., & Scharre, P. (2015). *Meaningful human control in weapon systems: A primer*. Washington: Center for a New American Security.

Hurka, T. (2003). Moore in the Middle. *Ethics*, *113*(3), 599–628. https://doi.org/10.1086/345624.

Köhler, S. (2020). Instrumental Robots. *Science and Engineering Ethics*, *26*(6).

Köhler, S., Roughley, N., & Sauer, H. (2017). Technologically blurred accountability?: Technology, responsibility gaps and the robustness of our everyday conceptual scheme. *Moral agency and the politics of responsibility* (pp. 51–68). Routledge.

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem?Ethics and Information Technology, *24*(3).

Korsgaard, C. M. (1986). Skepticism about practical reason.Journal of Philosophy, *83*(1).

Ludwig, K. (2007). The argument from normative autonomy for collective agents.Journal of Social Philosophy, *38*(3).

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183.

McKenna, M. (2013). Reasons-responsiveness, agents, and mechanisms. *Oxford Studies in Agency and Responsibility*, *1*, 151–183.

McKenna, M., Timpe, K., Griffith, M., & Levy, N. (2017). *Reasons-responsive theories of freedom*. The Routledge companion to free will.

Mecacci, G., & de Santoni, F. (2020). Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. *Ethics and Information Technology*, *22*(2), 103–115.

Miller, S. (2018). Collective responsibility. *Dual Use Science and Technology, Ethics and Weapons of Mass Destruction* (pp. 39–54). Springer.

Nagel, T. (1970). *The Possibility of Altruism*. Oxford Clarendon Press.

Nelkin, D. K. (2011). *Making Sense of Freedom and Responsibility*. Oxford University Press.

Nyholm, S. (2018). Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics*, *24*(4), 1201–1219. https://doi.org/10.1007/s11948-017-9943-x.

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

van de Poel, I., Royakkers, L., & Zwart, S. D. (2015). *Moral Responsibility and the Problem of Many Hands*. Routledge.

Portmore, D. W. (2013). Agent-Relative Vs. Agent-Neutral. In H. LaFollette (Ed.), *International Encyclopedia of Ethics*. Wiley-Blackwell.

Raz, J. (1975). Reasons for Action, Decisions and Norms. *Mind*, *84*(336), 481–499. https://doi.org/10.1093/mind/LXXXIV.1.481.

Ridge, M. (2011). Reasons for Action: Agent-Neutral Vs. Agent-Relative. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Robillard, M. (2018). No Such Thing as Killer Robots.Journal of Applied Philosophy, *35*(4).

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account.Frontiers in Robotics and AI, 15.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Belknap Press of Harvard University Press. http://books.google.nl/books?id=FwuZcwMdtzwC.

Simpson, T. W., & Müller, V. C. (2016). Just War and Robots' Killings. *The Philosophical Quarterly*, *66*(263), 302–322. https://doi.org/10.1093/pq/pqv075.

Smith, M. (1994). *The Moral Problem*. Blackwell. Issue 1.

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

Sparrow, R. (2016). Robots and Respect: Assessing the Case Against Autonomous Weapon Systems. *Ethics and International Affairs*, *30*(1), 93–116. https://doi.org/10.1017/s0892679415000647.

Tigard, D. W. (2020). There Is No Techno-Responsibility Gap.Philosophy and Technology, *34*(3).

Umbrello, S. (2020). Meaningful Human Control Over Smart Home Systems: A Value Sensitive Design Approach. *Humana Mente Journal of Philosophical Studies*, *13*(37), 40–65.

Wolf, S. (1990). *Freedom Within Reason*. Oup Usa.

Zong, D. (2000). Agent Neutrality is the Exclusive Feature of Consequentialism. *Southern Journal of Philosophy*, *38*(4), 676–693. https://doi.org/10.1111/j.2041-6962.2000.tb00921.x.

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