Responsibly Engineering CONTROL

Sebastian Köhler, Giulio Mecacci and Herman Veluwenkamp

This document is a pre-print version of a paper that has been accepted for publication in *American Philosophical Quarterly*. For citation purposes, please refer to it as follows

- Köhler, S., Mecacci, G. & **Veluwenkamp, H.** (forthcoming), Responsibly Engineering CONTROL, *American Philosophical Quarterly*.

Abstract

A number of concerns have been recently raised regarding the possibility of human agents to effectively maintain control over intelligent and (partially) autonomous artificial systems. These issues have been deemed to raise "responsibility gaps." To address these gaps, several scholars and other public and private stakeholders converged towards the idea that, in deploying intelligent technology, a meaningful form of human control (MHC) should be at all times exercised over autonomous intelligent technology. One of the main criticisms to the general idea of MHC is that it could be inherently problematic to have high degrees of control and high degrees of autonomy at the same time, as the two dimensions appear to be inversely related. Several ways to respond to this argument and deal with the dilemma between control and autonomy have been proposed in the literature.

In this paper, we further contribute to the philosophical effort to overcome the trade-off between automation and human control, and to open up some space for moral responsibility. We will use the instrument of conceptual engineering to investigate whether and to what extent removing the element of direct causal intervention from the concept of control can preserve the main functions of that concept, specifically focusing on the extent it can act as foundation of moral responsibility. We show that at least one philosophical account of MHC is indeed a conceptually viable theory to absolve the fundamental functions of control, even in the context of completely autonomous artificial systems.

Keywords: meaningful human control, responsibility gaps, conceptual engineering, autonomy

1. Introduction

A number of concerns have been recently raised regarding the possibility of human agents to effectively maintain control over intelligent and (partially) autonomous artificial systems (for brevity, we will call these systems, embedded and non-, generically "AI"). For example, AI's decision-making processes might be too fast for humans' monitoring capacities and, hence, their ability to timely stop and prevent action where necessary. AI is designed to process and make sense

of substantial amounts of information, sometimes in obscure and hardly explainable ways: human controllers may not be cognitively able to gain a clear understanding of what AI is doing and why, making it hard to intervene when necessary. AI's behavior might be also unpredictable: machine learning systems may be designed to develop novel behaviors (Matthias 2004). AI may generate confusions of agency: a human controller could be unsure about their role in joint action, sometimes attributing to themselves decisions and actions they have not initiated, some other times blaming AI for something they actually did themselves (Berberian et al. 2012; Norman 1990; Schwarz 2018).

These issues have been deemed to raise "responsibility gaps" (Matthias 2004; Santoni de Sio and Mecacci 2021): situations where automated systems display unwanted behavior for which there is a lack of responsibility or control, respectively (Di Nucci and Santoni de Sio 2016; Sparrow 2007).[1] To address these gaps, a number of scholars and other public and private stakeholders converged towards the idea that, in deploying intelligent technology, a meaningful form of human control should be at all times exercised over autonomous intelligent technology. Multiple accounts of meaningful human control (MHC thereafter) have been recently produced (see (Ekelhof 2019)). They mostly consist of sets of standards and normative requirements to promote a legally, ethically and socially acceptable form of human control. Originally proposed in the context of autonomous warfare (Amoroso and Tamburrini 2018; Article 36 2014; Chengeta 2016; Ekelhof 2019; Moyes 2016; Scharre and Horowitz 2015; Schwarz 2018), MHC has been recently investigated in the field of automated driving systems (Calvert et al. 2018, 2020; Calvert and Mecacci 2020; Heikoop et al. 2019; Mecacci and Santoni de Sio 2020; Santoni de Sio and Van den Hoven 2018) and medical automation (Braun et al. 2021; Ficuciello et al. 2019).

One of the main criticisms to the general idea of MHC is that it could be inherently problematic to have high degrees of control and high degrees of autonomy at the same time, as the two dimensions appear to be inversely related (Schwarz 2018). Highly automated systems have high degrees of autonomy. This is normally a desirable feature as the whole idea for those systems is that they should have the capacity to discharge humans from tasks that are for several reasons better suited to machines. As the argument goes, higher degrees of automation would entail lower degrees of human control, *by definition*, hence making any form of control harder as we progress towards increasingly autonomous artificial systems. We have to go for either autonomy or human control, but not for both at the same time: there seems to be no way to have the cake and eat it. The most immediate consequence for this, the argument goes, is that responsibility gaps are unavoidable.

There are several ways to respond to this argument and deal with the dilemma between control and autonomy. We will mention only some of them here, to provide just a glimpse of the possibilities. One way is to accept the argument as valid but claim that loss of control is inconsequential, or that its consequences can be dealt with by revising certain moral and legal practices. We should push innovation and AI, this approach claims, and get rid of human control where needed, as that is not necessary for the attribution of responsibility. There might be better, more efficient and more realistic ways to attribute responsibility even in absence of any control or controllability. We could for instance move away from classic forms of moral culpability and progressively shift towards easier to manage legal solutions and stipulations (Schellekens 2018; Santoni de Sio and Mecacci 2021). Additionally, one could think of providing some form of personhood to very advanced AIs, thereby making them suitable targets of moral and legal blame (Avila Negri 2021; Delvaux 2017).

Another way to deal with the dilemma between human control and automation is to have humans partially "in the loop" at all times, perhaps being very careful to avoid well known pitfalls in human-machine interface, such as automation complacency (Merritt et al. 2019). Some authors have suggested that full automation may not be the optimal aim in designing intelligent systems (Nyholm 2018; Nyholm and Smids 2020), and recommend finding the optimal type and amount of human contribution to an automated task. While this is a sensible and feasible response, it also prevents realizing the full potential of automation and AI, partially hindering technological innovation.

Finally, some authors reject the premise that human control requires the ability to *exercise* direct or indirect causal interventions. Santoni de Sio and van den Hoven (2018) and successively Mecacci and Santoni de Sio (2020) produced and refined an account of MHC that combines Fischer's and Ravizza's theory of responsibility and control (Fischer and Ravizza 1998), Nozick's counterfactual notion of tracking (Nozick 1981), and the classical theory of intention and action (Anscombe 1957; Bratman 1987; Davidson 2001). Amongst others, one of their conditions for meaningful human control prescribes an alignment between human reasons and an automated system's actions. This alignment, combined with a controller's sufficient knowledge of their moral responsibilities, would grant meaningful control even *in absence of a direct causal intervention*. This could in turn bridge (some) responsibility gaps. Despite such promises, however, this proposal remains highly philosophical and is rather challenging to operationalize into technical or institutional design solutions (Calvert et al. 2020; Mecacci and Santoni de Sio 2020).

In this paper, we further contribute to the philosophical effort to overcome the trade-off between automation and human control, and to open up some space for moral responsibility. We will use the instrument of conceptual engineering to investigate whether and to what extent removing the element of direct causal intervention from the concept of control can preserve the main functions of that concept, specifically focusing on the extent it can act as foundation of moral responsibility (this follows and substantiates a recent methodological turn in the ethics of technology, as several authors have argued that conceptual engineering has an important place in the ethics of technology. See for example Himmelreich and Köhler 2022, Löhr 2023, Löhr and Hopster *forthcoming*, or Veluwenkamp and van den Hoven 2023). Conceptual engineering as we understand it is the design, implementation and evaluation of concepts (Chalmers 2020, p. 2). This practice is justified, for example, when technology introduces new contexts that make the use of our old concept unsuitable for the new context (Veluwenkamp et al. 2022). For instance, we know that direct causal intervention plays an important function in the way we use the concept of control *in some contexts*. Yet, we can also argue that some other times this particular component plays a minor role, for example when we talk about political control. Ultimately, we aim at establishing, in the context of autonomous systems, whether and which conceptions of control can grant responsibility, and if that's the case, which kinds of.

To understand this paper's innovative potential, we should take a step back and observe the larger debate about meaningful human control. Most meaningful human control theories aim to preserve human responsibility and agency in highly automated systems by introducing a number of practical, technical, cognitive and moral constraints to integrate the classic notion of control (Article 36 2014; ICRAC n.d.; ICRC 2018; Kania 2017; USSB 2012). These theories are effective at addressing human control and responsibility in the context of high automation, but show their limits when applied in the context of full autonomy. A few markedly philosophical accounts of MHC (Himmelreich 2019; Santoni de Sio and Van den Hoven 2018) take a more radical step and

claim to be able to account for those latter situations. In order to do so, though, they completely revise the concept of control, questioning the fundamental intuition that all forms of control require a causal connection between controllers and their controlled systems. Rather than integrating them, they radically change the conditions on which common sense notions of control are based. These philosophical MHC theories, hence, seem to pay a steep price, and are prone to at least one important criticism: the concept of control they flesh out seems so removed from the one we commonly understand and use that it won't serve the same function(s), such as grounding responsibility attributions. This paper uses conceptual engineering to defuse that fundamental criticism and show that at least one philosophical account of MHC (Santoni de Sio and Van den Hoven 2018) is indeed a conceptually viable theory to absolve the fundamental functions of control, even in the context of completely autonomous artificial systems.

In section 2, we argue that how "control" is to be understood is a conceptual engineering problem. The correct understanding of "control," when it comes to thinking about autonomous systems, is the one suggested by conceptual engineering. In section 3, we will delve into what the object of our investigation is, that is, what it is that we are conceptually engineering, and will discuss the specific methodology. We will argue that we ought to engineer the content of expressions and defend a functionalist approach. In section 4 we apply our functionalist conceptual engineering approach to the case of "control" and show that the conception of control proposed by MHC theory fulfills the desired function better than other conceptions. Section 5 is dedicated to discussing two potential worries with our central thesis. The first one is that our engineering approach produces an unwanted proliferation of different conceptions of "control". The second one concerns circularity, as we may seem to be assuming what we need to prove. We aim to show that our concept of control can help preserve responsibility, but some may object that we already assume a particular concept of responsibility that fits the bill.

2. Control is a Conceptual Engineering Problem

As we've seen, when it comes to a certain range of automated systems, it is important that humans remain in control. However, it seems that there is a tension between humans remaining in control and the benefits from high degrees of automation. Specifically, it seems that control and automation of AI systems are inversely correlated, such that we can only have the one to a degree by lowering the degree of the other. It is important to note, however, that whether there really is a problem here depends significantly on how we understand the expression "control." The appearance of a paradox arises, because we operate with a certain idea of what is meant by "control" in this context in the background. Maybe, we associate "control" in this context with the idea that "a system is under the control (in general) of an agent if, and to the extent to which, its behavior responds to the agent's plans, maneuvers or operations" (Mecacci and Santoni de Sio 2020, p. 105; the conception they describe is the one developed by John Michon (1985)). Let us call this sort of control "operational control." If we understand "control" as operational control, it comes with an interventionist flavor: we have control to the extent that we can causally intervene to change the system's behavior. If we understood "control" this way, there will, indeed, be an inverse relationship between the degree of automation of systems and the degree of control we have over their behavior.

Given this, it is important to ask what the correct way to understand "control" in this context is. What, however, do we mean when we ask for the "correct" way to understand

"control?" Typically, when philosophers ask this question, this is understood in terms of *conceptual analysis*: the correct way of understanding "control" is that understanding that gives us the (best candidate for the) *actual* meaning of the term "control." One plausible way to cash this out, for example, is in terms of Frank Jackson's (1998) framework. In this framework, we determine the correct understanding of terms by considering what content would make best sense of our dispositions to apply them. Such a content can be mildly revisionary, as it is unlikely for any singular term that we can make sense of *all* our dispositions to apply it. Still, the correct understanding of a term is the one that makes most sense of people's dispositions to apply it. So, it will matter greatly that the content we assign to a term is not highly counterintuitive — not *too* far removed from our common understanding of the term.

However, conceptual analysis does not offer the only interpretation of what it means for the understanding of a term to be "correct" — nor the best for our context, as we will argue shortly. Another way to do so is in terms of what is called "conceptual engineering." Conceptual engineering is an approach to meaningful entities (such as expressions or concepts) that, likely, has always been a part of philosophy (though not under that name), but which has been getting increasing attention and systematic discussion only in recent years (e.g., Cappelen 2018; Isaac, Koch, und Nefdt 2022; Burgess, Cappelen, und Plunkett 2020; Eklund, 2021). According to conceptual engineering, the "correct" way for understanding an expression can only be determined through *normative* considerations. Specifically, the "correct" way to understand an expression is determined by how we *ought to* understand it. This view takes seriously that what an expression means and how we use it is, to some extent, up to us — the expression's users —, but that expressions also do important things for us, both epistemically and practically. Given these two observations, we should evaluate and assess our expressions and choose a way for understanding and using them that is normatively optimal. Specifically, conceptual engineers urge us to ask: what expressions should we use and how should we use them?

It is easiest to understand conceptual engineering's core suggestion by considering expressions that suffer from some conceptual defect. Some argue (e.g., Scharp 2013), for example, that our ordinary use of "true" is to some significant degree incoherent, as it leads to paradoxes (e.g., the liar paradox). If this is true, then conceptual analysis will fail to uncover a coherent content for "true." However, this does not mean that we should abandon "true." Rather, we should — according to conceptual engineers — look for ways of understanding "true" that are not defective, but preserve (some of) the important things we use "true" for. Note, though, that the defect in response to which we should change our concept need not be *epistemic* or *conceptual*. For example, Sally Haslanger (e.g., 2000) has argued that our current race and gender terms facilitate social oppression and that we should instead opt for an understanding of these terms that allows us to highlight and address such social problems. However, conceptual engineering is not just about addressing defects (e.g., Simion 2018). Rather, it comes into play whenever there is a potential (relevant) *improvement* that could be gained by introducing a certain expression or by changing the way we understand and use one.

So, conceptual engineering suggests that to determine an expression's "correct" understanding, we need to determine how we *ought to* understand it. Of course, this suggestion raises many further questions, some of which will depend on issues in the philosophy of language and mind. We will make some concrete suggestions on how we want to understand conceptual engineering in the next section. For now, though, we can go back to this paper's topic: "control." Above, we've highlighted that whether and to what extent there is a problem generated by the

relationship between control and automation depends on the correct way to understand "control." We can now argue the first point this paper wants to make: how "control" is to be understood is a conceptual engineering problem — the correct understanding of "control," when it comes to thinking about autonomous systems, is the one suggested by conceptual engineering.

We can argue this very straightforwardly: given the importance of "control" when thinking about autonomous systems, any understanding of "control" uncovered by conceptual analysis lacks the necessary normative significance. Suppose, for example, that there was a coherent content to "control" in the context of control over autonomous systems that satisfies the demands of conceptual analysis. In fact, assume that this content is what we've called "operational control." Even in this case we should ask: Is this how we *ought to* understand "control" in this context? Given the enormous practical dimensions and implications regarding the relation between control and automation, this is a very important question we need to ask: given these implications, we should have *good* reasons for understanding "control" in one way rather than another. However, given that conceptual analysis does not answer this question and there is no reason to assume that the content determined by conceptual analysis overlaps with the content we ought to assign, how we understand "control" in this context should, primarily, be understood as a conceptual engineering question. This leads us to the next question that is relevant for our paper: how, exactly, ought we to engineer "control" for the context of autonomous systems? To answer this question, we first need to unpack the specific assumptions we make about conceptual engineering and its methodology.

3. Conceptual Engineering: A Functional Approach

While conceptual engineering has always been a central business of philosophy, the systematic investigation of this methodological approach has only recently gotten traction. As such, the field is very diverse and philosophers disagree about several issues that significantly impact how we understand conceptual engineering's core suggestion. For example, there is disagreement on whether the target of conceptual engineering should just be expressions (e.g., Cappelen 2018; Thomasson 2022) or also concepts (e.g., Eklund 2015; Haslanger 2000; Plunkett 2015), whether we should engineer intensions and extensions (e.g., Cappelen 2018), use-patterns (e.g., Jorem 2021), commitment and entitlement structures (e.g., Löhr 2021), and so on. And, of course, any of these issues opens up further questions, such as what concepts are, what contents are, and so on. This paper is not the place to engage in these debates. Here, we will just make assumptions without arguing for them — assumptions we find independently attractive. However, most of the discussion should work, *mutatis mutandis*, on other feasible assumptions.

To approach an issue as a conceptual engineering problem, we need to settle at least two questions. First, what is it that we are engineering, when we are engaged in conceptual engineering? Second, what methodology should we use in conceptual engineering, that is, how do we approach a conceptual engineering problem? We take these questions in turn.

First, for the purposes of this paper we will focus on engineering the content of *expressions*, in our case "control" and "responsibility." We will assume that in ordinary English, the actual content of expressions is to some extent *indeterminate*, such that there are many different ways to make that content precise *without a change of topic*. What this means can be illustrated as follows: assume that we wanted to assign some determinate content to the expression "free." Some such assignments will, plausibly, count as changing the topic: for example, if we assigned a content to

"free," such that something is free if and only if it is green, we plausibly have changed the topic compared to how "free" is used in English. However, the same does not hold if we assigned any content that aligned with some prominent view in political philosophy. Here, different more determinate contents can still be said to preserve the expression's topic. For example, if we assigned a content to "free" such that an action is free if and only if it is free from outside interference, we would *not* think that we have changed the topic compared to how the expression is currently used in English (even if we might think that this is not the *best* content to assign to "free"). And similarly for other potential contents.

We will call any complete way of making an expression's content precise without changing that expression's topic *T* a *conception* of *T*. So, for example, the topic of "control" is control and Santoni de Sio's and van den Hoven's (2018) account of meaningful human control is one conception of control. Given this we have to say something about when two ways of making an expression's content more precise have the same topic. Roughly, we assume that two determinate contents are conceptions of the same topic if they are similar enough (Cappelen 2018; Sundell 2020). And, what counts as similar enough depends on our purposes. When Rawls, for example, proposed a new conception of justice, he was interested in conceptions that fulfilled a specific role: namely, providing "principles for assigning rights and duties" (1999, p. 5). However, in different contexts we might have different purposes, so what counts as similar enough differs too (see also Eklund 2021). Two conceptions are therefore of the same topic if they are similar enough, where "similar enough" is determined by contextual factors.

Talk of expressions' contents also needs to be unpacked a little, so that we understand what a conception consists in. It is common to understand this in terms of intensions and extensions. However, we will assume that conceptual engineering operates at a more fundamental level. Broadly speaking, we need to distinguish two approaches as to how expressions get their contents (e.g., Loar 2006). First, according to *representationalist* accounts, expressions have their contents in virtue of what they represent. On this approach, conceptual engineering aims to engineer expressions such that they *represent* a particular sort of thing (i.e., the thing they ought to represent). Second, according to *non-representational* accounts, expressions have their contents in virtue of certain patterns of use, such as their conceptual or inferential role. On this approach, conceptual engineering instead aims to engineer expressions such that they are characterized by a different conceptual or inferential role (i.e., the role it ought to play). Note that each of these approaches can be used to then fix contents in terms of intensions and extensions — the disagreement between these two approaches is, primarily, in virtue of what expressions have their contents and not what these contents are. Furthermore, note that even if one adopts a non-representationalist account, one can hold that *some* expressions have their content in virtue of a representational role. All that one denies is that this is true for all expressions. For independent reasons we are attracted to a non-representationalist account, so we will presuppose it here.

Cashed out in terms of the non-representational approach, we will now assume that a complete way of making the content of an expression precise gives us a determinate conceptual or inferential role for that expression. Two or more determinate conceptual or inferential roles provide us with *conceptions* of the same topic just in case, as suggested above, they are similar enough. This allows us to say what it is we think conceptual engineering should engineer: conceptual engineering should determine what determinate conception, understood in terms of topic-preserving conceptual or inferential role, ought to be associated with an expression. For our context this means: what conception of *control* ought to be associated with "control?"

Of course, before we can approach this question, we need to turn to the second question raised above: How should we *do* conceptual engineering? Answering this question will also answer what is meant by "ought" in the question asked by conceptual engineers. Here we will draw on Amie Thomasson's work (2020). Thomasson favors and has recently argued for a *functionalist approach* to conceptual engineering (see also Queloz 2022; Köhler and Veluwenkamp *forthcoming*; Simion and Kelp 2020; see Cappelen 2018 for a skeptic about the functional approach). This approach proceeds on the assumption that language does many different important things for us and that conceptual engineers should consider what the important functions are and engineer expressions so that they fulfill them.[2]

On the version of the approach that will assume here, we first look at the current function of an expression: why do we employ a specific expression, or, what couldn't we do if we didn't have this expression in our semantic repertoire? Such a function *might* be to represent reality (as in the case of important scientific concepts, for example), but on the functional approach it does not *have* to be: expressions can do many different important things for us. Once we have determined the actual function, we critically assess if this is a function we want the expression to have and what function the concept *ought* to possess. We might, for example, be critical about the function that traditional race and gender conceptions have (Burgess and Plunkett 2013) and argue that we need a conception with a function that accords better with our social and moral purposes. Finally, when we know what function our concept ought to serve, we select or engineer a conception that fulfills this function best.

Opting for the functionalist approach has another added benefit. Above, we've introduced the idea that two conceptions could be on the same *topic*. As it turns out, staying *on topic* is quite important in conceptual engineering, because when a revision changes topics, this can easily mean that we have failed to preserve why we were concerned with the expression in the first place. Suppose, for example, that we were worried about the compatibility of free will with determinism and someone suggested that we deal with this problem by understanding "free will" such that something has free will if and only if it is green. One can hardly expect this revision to address the original worry. We have just changed the topic, without dealing with the problem we started with.

Above we suggested that two determinate contents are conceptions of the same topic if they are similar enough, where similarity depends on our purposes. However, one quite attractive sort of similarity between two conceptions when it comes to dealing with the worry just sketched *is* in terms of the most important function they could perform (see also Sundell 2020; Thomasson 2020). After all, suppose that our revision preserved the most important point of the concept we revised: our revision preserves and maybe even advances why the concept matters in the first place. In this case, it is hard to take worries about "changing the subject" seriously. After all, the concept still does the thing in virtue of which we should care about it, so even if we might now strictly speaking be talking about something different, the topic of the concept is plausibly preserved in the most attractive sense of that phrase.

We have now cleared up how we will answer the two questions posed above: conceptual engineering ought to engineer the content of *expressions* and it should do so using a functionalist methodology. We can now turn to the core question of this paper: how should we engineer "control" in the context of autonomous systems?

4. Designing Autonomous Systems for CONTROL

As we saw above, the functionalist approach to conceptual engineering consists of three distinct steps. First, we determine what the current function of "control" is. What function does it serve to say of an agent that it has control over someone or something else? Second, we determine whether this function is appropriate: is this a function that we ought to have in our society? Finally, we determine which conception of control fulfills this function best.

To determine the current function of "control" we will use Miranda Fricker's Paradigm-Based approach.[3] In this approach, we first establish what the paradigmatic instance of applying a particular expression is. We then hypothesize about the function of this paradigmatic instance, and test our hypothesis by assessing whether it can explain other derivative practices.

Let us see if we can find a paradigmatic form of control. There are many different forms of control, for example, interpersonal control, self-control and political control. Our hypothesis is that control over the external environment is explanatorily basic. It is the kind of control that allows us to alter the world around us and conform it to our wishes and desires. Let us call this *environmental control*.

What is the point of indicating that someone has environmental control? In particular circumstances we can have different reasons to say that someone controls her surroundings. Typically, however, when we indicate that someone has control over the environment, we want to signal that she has the ability to make sure that she can change or modify the world around her. That is, an agent has certain goals, and we want to indicate the agent's ability to realize some of these goals by changing the world around her. So, we take the function of indicating that X has control over the environment to be that of indicating X's ability to influence the environment to align it with X's goals.

To test our hypothesis that environmental control is a paradigmatic form of control we have to determine whether it can explain derivative forms of control. Let us look at interpersonal control first. Interpersonal control is the kind of control one has over another person. One is not changing the external environment directly. Instead, one is changing the external world by changing another agent's actions. So, when we say that X has control over Y, we indicate that X is able to realize her goals by changing the actions of Y. Interpersonal control is therefore an indirect form of environmental control.

Self-control is structurally similar to interpersonal control, except that in the former there is no *other* person to influence. We think a good way to understand this is to distinguish first between fleeting and entrenched goals.[4] We can have fleeting goals, such as eating that bar of chocolate, and entrenched goals, such as staying healthy. In our weaker moments we have a tendency to align our actions with fleeting goals. When we have self-control, however, we make sure that our behavior aligns with deeper, entrenched goals. In this way we can see that self-control can be explained by interpersonal control. Both interpersonal control and self-control can therefore readily be understood as derivative from environmental control. So, environmental control is, plausibly, a paradigmatic form of control.

The next step is to determine whether we want, or need, a conception of "control" in the context of autonomous systems. That is, is it good to indicate that an agent has the ability to influence the actions of an autonomous system in order to align its action to the agent's goals? Yes, for at least two reasons. The first relates to responsibility attributions. It is often important to determine the right locus of responsibility when autonomous systems cause harm (Nyholm 2018).

A conception with the current function of "control" would contribute to this in the right way. For suppose that the human Harry is in control of an autonomous weapon system (AWS). This would require that Harry can ensure that the AWS aligns its behavior with Harry's goals. Hence, in cases where the AWS causes harm, Harry can be held appropriately responsible for this harm. After all, Harry was able to align the AWS's actions with his goals. Moreover, suppose that the AWS caused harm, but Harry failed to intervene. In this case Harry can also be held responsible. Indicating that he had control, given the function we identified, indicates that he could have prevented the harm if his goal were to do so.

Secondly, we care about control in the context of autonomous systems because we want to minimize the amount of harm we expose people to. Indicating that someone has the ability to direct the actions of the autonomous systems tells us that someone is able to minimize harm (although it is of course a further question whether this person acts on the ability). This is also why it is important that self-driving cars and planes are designed semi-autonomously. At the current state of technology, it is too risky for these machines to function entirely on their own. We say that a driver is at least partly in control of a semi-autonomous car to indicate that they are in some circumstances able to align the action of the system to their goals.

Having justified this function of "control," let us see which conception fulfills this function best. The conception presupposed in arguments for responsibility gaps (Matthias 2004; Sparrow 2007) is operational control. We can specify this conception with the following inferential role:[5]

Operational control

Agent A is responsible for outcome O → A is in control of O

Agent A is in control of O → A is (or has been) able to causally influence O

This conception of control cannot fulfill the function of "control" in the context of autonomous systems. This can be seen by considering two scenarios: one where people who are capable of causally intervening in the operation of an autonomous system are unable to align the actions of said system with their goals, and one where people who can align the actions of the system with their goals are unable to causally intervene.

In the first scenario, someone in the driver seat of a semi-autonomous vehicle is physically capable of causally intervening in the car's operation. According to operational control, this person is in control. However, under many circumstances the driver is unable to causally influence the car's behavior in such a way that the car's actions are aligned with the goals of the driver. Consider for example Hanah. Hanah is unaware of the fact that the semi-autonomous car she is sitting in is unable to function in extreme weather conditions. One day she is driving the car to work and has the car in semi-autonomous mode. It is raining very hard, and because of this the car crashes into a lorry. Hanah could have causally influenced the car's operation. However, because the marketing the car manufacturer used to sell the car, she (blamelessly) did not know that this was expected of her. She was therefore unable to align the car with her goal of getting to work safely.

But operational control also fails in the other direction. To see this, consider fully autonomous cars. No-one can directly casually influence such a car; therefore, no one has operational control over the car. However, if the car is designed and regulated well, then there are operators (programmers, etc.) who can make sure that the car's behavior aligns with their goals.

So given the function of "control", we have reason to abandon operational control as the conception of "control" in the context of autonomous systems.

Johannes Himmelreich (2019) has recently defended a rival conception of control. Inspired by the aforementioned problems with operational control, he introduces "Robust Tracking Control" as conception of control in the context of AWS. For Himmelreich, a human agent A has Robust Tracking Control over an outcome O performed by a system S if the following three conditions are met. There is (1) a directive that A can give S such that, (2) if A gives this directive, then X occurs, and (3) if A were not to give directive X then X would not occur (2019, p. 736). We can specify this conception of control as follows:

Robust tracking control

Agent A is responsible for O → A is in control of O

Agent A is in control of O → O tracks (in a modally robust way) A's directives.

This conception of control has clear advantages over operational control. It entails, for example, that one can be in control of a fully autonomous car, even if one has no way of physically intervening in the system. Fully autonomous cars (usually) operate in response to directives given by one of the human drivers in the car (albeit by providing voice commands or in some other way). If the human gives this directive, the car travels to the destination provided in the directive, and, if the human were not to give this directive, the car would not commence the journey.

However, this conception shares some of the limitations of "operational control." To see this, let us consider Hanah again. If she would have given the directive to switch to manual control, then she would have avoided crashing into the lorry. Moreover, she did not give the directive and for that reason did cause the incident. She therefore seems to have robust tracking control over the crash, rendering her morally responsible. And this problem generalizes. Human directives can be wrong, misinformed, or result from akrasia. In this case, there is an intuitive sense in which Hanah was not in control of the vehicle, because crashing is not what she *really* intended to do. With increasing automation, we want AI to avoid responding to a user's mistakes.

Above we have discussed two conceptions of control that combine the link to responsibility attribution with an ability to causally intervene in a system's operation. We determined that the function of judging that X exercises control over Y is to indicate that X had the ability to influence Y's behavior in order to align Y's actions with X's goals. Moreover, this is an important function because it retains the necessary connection with proper responsibility attributions. We concluded that operational control and robust tracking control do not always fulfill this function of "control." We will now show that the main function of "control" can be preserved by a conception of control that removes the element of direct causal intervention.

Meaningful human control (MHC), in the philosophical account developed by Santoni de Sio and Van Den Hoven (2018) and further substantiated by (Mecacci and Santoni de Sio 2020) and (Cavalcante Siebert et al. 2022), relies on two conditions and some additional assumptions. One key assumption is worth being specified immediately. The authors often refer to human control over a system. In their account, "system" does not refer only to the mere technological artifact or infrastructure. The term is rather to be interpreted to denote a larger *sociotechnical* system whose boundaries can move depending on what is relevant in different contexts, where relevance is normatively established on a case-by-case basis (Mumford 2006). With that out of the way,

MHC's theory of control is based on two distinct conditions. The first condition for a system to be deemed under control of a human agent, also called "tracking," is similar to Himmelreich's robust tracking control, with the important difference that *it is agnostic with regard to the specific nature of a user's intervention*. Where Himmelreich requires that the system *receives a directive* from some human agent(s), MHC requires a system *to be responsive* to the relevant human agents' relevant reasons for actions (Mecacci and Santoni de Sio 2020; Veluwenkamp 2022). Controllers and their controlled system are required to consistently share the same goals and values. This is something that can be strived for and maximized during, for instance, the design phase. Explicit directives instantiating those goals may or may not be provided at any point in time.

The second condition for control, according to this conception, is called "tracing." It states that (1) one or more human controllers should exist and be identifiable; that (2) they should have the right cognitive and physical capacities, where required, to fit their role; and that (3) they should be adequately aware of such controlling role, especially what it means in terms of active and passive responsibility bearing. We can depict this conception of control as follows:

Meaningful human control
- Agent A is responsible for sociotechnical system S → A is in control of S
- Agent A is in control of S → A's reasons are being tracked by S
- Agent A is in control of S → S is designed such that the tracing condition applies to A.

The extent to which the two conditions of tracking and tracing are satisfied depends on whether and the extent to which–respectively–a system and its controllers possess certain properties. However, the definition of system is such that in certain cases a controller might be usefully considered part of the system themselves. Hanah's case, presented above, shows why this is relevant. Hanah's control over the car satisfies Himmelreich's robust tracking condition, but we are still unhappy to deem her in control and responsible for her actions, since that's not what she really intended to do. Moreover, increasing automation offers the opportunity to make up for human mistakes, and we ideally should reap that benefit. MHC's conception of control helps make sense of this aspect in several ways.

Let us, therefore, consider to what extent MHC fulfills the function of "control" in the context of autonomous systems. That is, does it allow us to indicate that a human has the kind of ability to influence the system's action to render them responsible for the outcome of the system? First consider Hanah again. She is unable to influence the car's behavior in order to align its actions with Hanah's goals, because she is unaware of the car's capabilities. A proper conception of control for the context of autonomous systems would take this into account. In contrast to the other conceptions, MHC is sensitive to the fact that Hanah did not *really* have a reason to crash into the lorry. Mecacci and Santoni de Sio (2020) make use of the notion of "proximity" to identify differences among different orders of volitions. In that way, one could say that, despite the car being fully responsive to Hanah's proximal intention not to provide a directive, the vehicle was not equally responsive to her distal, more general, more *relevant* goal of getting home safely. Where the robust tracking condition in Himmelreich's conception is satisfied, this is not the case for Santoni de Sio and Van den Hoven's. By seeing Hanah as part of the sociotechnical system she is meant to control, one can additionally argue that she fails to exercise control over herself, in the

sense that she let her fleeting goals (proximal intentions) take over her entrenched goals (distal intentions).

Second, MHC is sensitive to the fact that Hanah may or may not have lacked the necessary capacities to understand that a certain situation demanded an explicit directive from her, or she may have lacked the capacity to act at the right time or in the right way. This is expressed in the second condition for MHC, tracing: diminished controller's capacities mean diminished control.

5. Objections

We will close this paper by responding to two potential objections that might be raised to our arguments. Discussing these objections should, hopefully, help to clarify the commitments that come with our account.

The first worry is a worry about the implications of engaging in conceptual engineering for a concept for a specific purpose. The worry is this: what we are suggesting here is *not* to introduce a new concept to capture a phenomenon we do not yet have a concept to pick out. Instead, we are suggesting to revise "control" such that it fits a specific use case, namely the case of human-AI interactions and responsibility attributions. However, one might worry that this strategy suggests that we introduce very many different conceptions of "control" each suited for different kinds of circumstances. Specifically, one might worry that our way of doing conceptual engineering commits us to a very fine-grained proliferation of conceptions of "control" for specific contexts as long as a different conception fits that context better. This is a worry, because it might appear as if this approach introduces huge amounts of ambiguity as to what "control" means, ambiguity that would be detrimental to the ability of ordinary speakers to understand what "control" means in any specific context.

To this worry we have the following responses. First, we find it questionable to what extent the approach we suggest requires us to introduce ambiguities. One important thing to note is that the conception for control offered by MHC can actually cover a whole range of circumstances. For example, in the case where you are in control of your own actions, MHC would be just as applicable as in the case where you control an AI and the conception would have the same implications about whether you are in control, as more orthodox conceptions, such as operational control.

Of course, the requirements that need to be satisfied for one to be in control might be different depending on the circumstances. However, it is unclear why this would pose a distinctive issue for our account. After all, the way we think about responsibility already seems to indicate that we take different kinds of control to be required for responsibility in different kinds of situations. For example, both a sober and a drunk driver can be responsible for any accidents they cause. However, the drunk driver only has *indirect* control over their actions, while the sober driver is responsible in virtue of their *direct* control. This supports the second response we want to give: even if our approach introduces and requires *some* ambiguity, this should not be a problem, given that we already are familiar with shifting standards for control.

Of course, this presupposes that our approach only requires some ambiguity. This brings us to our third response, which connects to and bolsters the second response: it is unlikely that we will have to proliferate conceptions of control beyond reasonable limits. Nothing in our ordinary responsibility practice speaks for the fact that for "control" to play its proper function in responsibility attributions we need extremely fine-grained ideas about control for different kinds

of circumstances. Of course, as the case of autonomous artificial systems shows, novel circumstances might require us to introduce new or to revise old conceptions to deal with challenges our responsibility practice might face, but there is no reason to expect developments that suddenly require a huge proliferation of conceptions of "control."

Our last response is this: assume that it was indeed the case that we must proliferate conceptions of control on our account. For this to be true, it would have to be the case that "control" can only play its role vis-a-vis responsibility, if control is sensitive to very peculiar, fine-grained, and minute details of situations. But it seems like this implies that our current practice of responsibility, which is *not* sensitive to these details, is deeply wrong. After all, if what we have stipulated was true, then what we currently mean by "control," which does not track such fine-grained details would be deeply insufficient for playing the role of "control" in responsibility attributions, meaning that many of our attributions would be off-track. In this case, it seems that we would be morally *required* to introduce very fine-grained conceptions of control for different kinds of situations. So, it seems like if the objection is correct about what our account requires, then what our account requires is what we actually should do.

Let us turn to the second objection, which raises a worry about *circularity*. Our starting point in the paper are worries about the impact that increasing automation has on responsibility due to its incompatibility with control. What we suggest to address such worries, is to *revise* our conception of control, such that automation and control can be aligned. We argue that such a conception is satisfactory, because it is able to fulfill a valuable function, namely to properly attribute responsibility in the context of human-AI interactions. However, one might worry here that this argumentation is circular, because the argument presupposes an idea of *responsibility* that comes with normative standards as to what responsibility requires and that those who are worried about automation reject this idea. Specifically, for our argument to be successful, it must be the case that the conception of control we suggest allows us to *properly* attribute responsibility in the relevant context. However, to determine whether the attributions it enables are *proper* responsibility-attributions, we need to make assumptions about what responsibility requires. This, then, raises the worry that our argumentation presupposes a view about responsibility that requires only a "weaker" notion of control, and in this way builds in the result of our argument into our starting point - a starting point those who do think there is a worry reject, because they hold that only a conception of control that is such as to be incompatible with control can be fitting for responsibility attributions.

We have two responses. First, our argumentation does not presuppose any specific conception of responsibility. The way we formulate what the function of "control" is in the context of responsibility attributions is broad and compatible with many different views on what responsibility requires. Our actual arguments themselves also just rely on general considerations and intuitions that have to do with control and responsibility. Of course, maybe the opponent here thinks that we are missing something important, something which only, for example, operational control can deliver. However, without knowing what this is, it is going to be difficult to engage with the worry. Of course, anyone skeptical of our approach here should feel free to investigate further what conception of control best allows "control" to perform its distinctive function and we gladly welcome further debate on this question.

Second, assume that the objection is correct in that our argumentation only works with a specific set of views about responsibility and that some, hence, will reject our arguments because they have a different view. This then just moves the discussion one level up, as we now have to

ask what conception of "responsibility" we should accept. This, again, is best seen as a conceptual engineering problem for similar reasons as the ones we've presented above. And, there is little reason to assume that a conception of responsibility that coheres well with our arguments here would not best fit the most important functions of "responsibility" (Himmelreich and Köhler 2022). In any case, engaging in this discussion would take us too far out of the scope of this paper.

6. Conclusion

In this paper we have shown a few important things. We claimed that understanding "control" requires a normative approach. This is needed to preserve its function in the context of attributing responsibility to human agents in sociotechnical systems where highly automated or autonomous artificial agents are deployed. A fruitful methodology to do so, we argued, is conceptual engineering, and specifically when approached from a functionalist stance. We therefore showed that "environmental control" is a paradigmatic, explanatorily basic form of control that also supports the function of responsibility attribution. Zooming in to the context of automation, we discussed a few context-specific conceptions of control and observed that the one that better fulfills the paradigmatic functions absolved by "environmental control" is "meaningful human control," and specifically in the philosophical account of Santoni de Sio and Van den Hoven. Finally, we defended our thesis from two main possible objections.

All in all, our aim was to show, aided by the methodology of conceptual engineering, that it is possible to conceive, in a meaningful and purposeful sense, human control–and responsibility–over the actions of automated and autonomous artificial agents. It is indeed necessary, to do so, to normatively revise the conceptions we intuitively deploy when talking and thinking about control in the context of AI. We indicated "meaningful human control" as the conception of control that ought to be used in that context. What we also hope to have shown is that such a conception is not as far-fetched as it could prima facie seem to be. Neither is it a deep revision over what we called "environmental control", which is a fundamental and (at least implicitly) widely adopted conception of control. "Meaningful human control" simply instantiates in the specific context of automated systems those functions that "environmental control" and, derivatively, other commonly accepted conceptions of control, already successfully absolve. We hope our work has further paved the way towards a future where AI and human control and responsibility are reconciled.

Sebastian Köhler
Frankfurt School of Finance & Management
Computational Science & Philosophy Department
Adickesallee 32-34, 60322 Frankfurt am Main, Germany
s.koehler@fs.de

Giulio Mecacci
Radboud University Nijmegen
Donders Institute for Brain, Cognition and Behaviour
Department of Cognitive Artificial Intelligence
Houtlaan 4, 6525 XZ Nijmegen, The Netherlands
giulio.mecacci@donders.ru.nl

Herman Veluwenkamp
University of Groningen
Faculty of Philosophy
Department of Ethics, Social and Political Philosophy
Oude Boteringestraat 52, 9712 GL Groningen, The Netherlands
h.m.veluwenkamp@rug.nl

Notes

We would like to thank Hein Duijf, Thomas Grote, Sven Nyholm, Kai Spiekermann, Philippe van Bashuysen, as well as other members of the Digital Ethics workshop at the Leibniz University Hannover for valuable comments. GM initiated the paper. All authors contributed equally to the writing and revisions.

[1] Some have argued that the notion of a responsibility gap is inconsistent and unmotivated, and that we should focus on gaps in control instead (Hindriks and Veluwenkamp 2023). For the purpose of this paper, this difference does not matter.
[2] Veluwenkamp et al. (2022) have recently investigated different projects in which philosophers of technology engage in conceptual work to see what methodological choices they made. What they found is that, insofar a distinct methodology could be recognized, these philosophers also opted for the functionalist approach (albeit implicitly). So, the assumption here should fit with common methodological assumptions in the philosophy of technology.
[3] There are several other methodologies in the literature that are also developed to determine what the actual function of our conceptions is, such as the genealogical approaches of Catarina Dutilh Novaes (2015) and Matthieu Queloz (2020). We have opted for the simpler approach championed by Fricker, which she describes 'as a more straightforward and transparent way of achieving the very same explanatory pay-off' (Fricker 2016).
[4] Using Frankfurt's (1988) terminology, we can say that fleeting goals are our first order goals, while the entrenched goals are the goals we would have if our second order goals were satisfied.
[5] Let X → Y mean that if someone utters that X, then she ought to accept that Y. For a discussion of conceptions of control in the context of Value Sensitive Design, see (Veluwenkamp and van den Hoven 2023).

References

ICRAC. n.d. About ICRAC. ICRAC. Retrieved January 29, 2023, from https://www.icrac.net/about-icrac/

ICRC. 2018. Treaties, States parties and Commentaries: General Protection of Civilian Objects. Retrieved April 24, 2023, from https://ihl-databases.icrc.org/en/ihl-treaties/api-1977/title/commentary/1987

Kania, E. B. 2017. "Battlefield Singularity," Artificial Intelligence, Military Revolution, and China's Future Military Power, CNAS.

USSB. 2012. Defense Science Board Task Force Report: The Role of Autonomy in DoD Systems. Washington, DC. https://doi.org/ADA566864

Amoroso, D., and Tamburrini, G. 2018. "The ethical and legal case against autonomy in weapons systems," Global Jurist, *18*(1).

Anscombe, G. E. M. 1957. Intention (Vol. 57, Issue 40, pp. 321–332). Harvard University Press.

Article 36. 2014. Autonomous weapons, meaningful human control and the CCW.

Avila Negri, S. 2021. "Robot as legal person: Electronic personhood in robotics and artificial intelligence," Frontiers in Robotics and AI, 419.

Berberian, B., Sarrazin, J.-C., Le Blaye, P., and Haggard, P. 2012. "Automation technology and sense of control: A window on human agency," PloS One, *7*(3), e34075.

Bratman, M. 1987. Intention, plans, and practical reason.

Braun, M., Hummel, P., Beck, S., and Dabrock, P. 2021. "Primer on an ethics of AI-based decision support systems in the clinic," Journal of Medical Ethics, *47*(12), 3–3.

Burgess, A., and Plunkett, D. 2013. "Conceptual Ethics I," Philosophy Compass, *8*(12), 1091–1101.

Burgess, Alexis, Herman Cappelen, and David Plunkett, 2020. Conceptual Engineering and Conceptual Ethics. Oxford University Press

Calvert, S. C., Heikoop, D. D., Mecacci, G., and Van Arem, B. 2020. "A human centric framework for the analysis of automated driving systems based on meaningful human control," Theoretical Issues in Ergonomics Science, *21*(4), 478–506.

Calvert, S. C., and Mecacci, G. 2020. "A conceptual control system description of Cooperative and Automated Driving in mixed urban traffic with Meaningful Human Control for design and evaluation," IEEE Open Journal of Intelligent Transportation Systems, 1, 147–158.

Calvert, S. C., Mecacci, G., Heikoop, D. D., and De Sio, F. S. 2018. Full platoon control in truck

platooning: A meaningful human control perspective. 3320–3326.

Cappelen, H. 2018. Fixing language: An essay on conceptual engineering. Oxford University Press.

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., and Jonker, C. M. 2022. "Meaningful human control: Actionable properties for AI system development," AI and Ethics, 1–15.

Chalmers, D. J. 2020. "What is conceptual engineering and what should it be?" Inquiry, 1–18.

Chengeta, T. 2016. "Defining the Emerging Notion of Meaningful Human Control in Weapon Systems," NYUJ Int'l L. & Pol., *49*, 833.

Davidson, D. 2001. Essays on Actions and Events: Philosophical Essays Volume 1. Clarendon Press.

Delvaux, M. 2017. "Report with recommendations to the commission on civil law rules on robotics," European Parliament, *A8-0005/2017*.

Di Nucci, E., and Santoni de Sio, F. 2016. "Drones and responsibility," Legal, Philosophical and Sociotechnical Perspectives on Remotely Controlled Weapons. London and New York: Routledge.

Dutilh Novaes, C. 2015. "Conceptual genealogy for analytic philosophy," in Beyond the analytic-continental divide (pp. 83–116). Routledge.

Ekelhof, M. 2019. "Moving beyond semantics on autonomous weapons: Meaningful human control in operation," Global Policy, *10*(3), 343–348.

Eklund, M. 2015. "Intuitions, conceptual engineering, and conceptual fixed points," in The Palgrave handbook of philosophical methods (pp. 363–385). Springer.

Eklund, M. 2021. "Conceptual Engineering in Philosophy," in J. Khoo and R. Sterken (Eds.), The Routledge Handbook of Social and Political Philosophy of Language.

Ficuciello, F., Tamburrini, G., Arezzo, A., Villani, L., and Siciliano, B. 2019. "Autonomy in surgical robots and its meaningful human control," Paladyn, Journal of Behavioral Robotics, *10*(1), 30–43.

Fischer, J. M., and Ravizza, M. 1998. Responsibility and control: A theory of moral responsibility. Cambridge university press.

Frankfurt, H. G. 1988. "Freedom of the Will and the Concept of a Person," in What is a person? (pp. 127–144). Springer.

Fricker, M. 2016. "What's the Point of Blame? A Paradigm Based Explanation," Noûs, *50*(1), 165–183.

Haslanger, S. 2000. "Gender and race:(What) are they?(What) do we want them to be?" Noûs, *34*(1), 31–55.

Heikoop, D. D., Hagenzieker, M., Mecacci, G., Calvert, S., Santoni De Sio, F., and van Arem, B. 2019. "Human behaviour with automated driving systems: A quantitative framework for meaningful human control," Theoretical Issues in Ergonomics Science, *20*(6), 711–730.

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

Himmelreich, J., and Köhler, S. 2022. "Responsible AI Through Conceptual Engineering," Philosophy and Technology, *35*(3), 1–30.

Hindriks, F., and Veluwenkamp, H. 2023. "The risks of autonomous machines: From responsibility gaps to control gaps," Synthese, *201*(1), 21.

Hopster, J. and Löhr, G. *forthcoming.* "Conceptual Engineering and Philosophy of Technology: Amelioration or Adaption?" Philosophy & Technology

Isaac, M.G., Koch, S. and Ryan Nefdt, R. 2022. "Conceptual Engineering: A Road Map to Practice," Philosophy Compass 17 (10).

Jackson, F. 1998. From Metaphysics to Ethics: A Defence of Conceptual Analysis (Vol. 1, pp. 13–28). Oxford University Press.

Jorem, S. 2021. "Conceptual engineering and the implementation problem," Inquiry: An Interdisciplinary Journal of Philosophy, *64*(1–2), 186–211.

Sebastian Köhler and Herman Veluwenkamp. forthcoming. "Conceptual Engineering: For What

Matters," Mind.

Loar, B. 2006. "Language, Thought, and Meaning," in M. Devitt and R. Hanley (Eds.), The Blackwell Guide to the Philosophy of Language. Blackwell Publishing.

Löhr, G. 2021. "Commitment engineering: Conceptual engineering without representations," Synthese, *199*(5), 13035–13052. https://doi.org/10.1007/s11229-021-03365-4

Löhr, G. 2023. "If Conceptual Engineering is a new Method in the Ethics of AI, what Method is it exactly?" AI and Ethics

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

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

Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., and Shirase, L. 2019. "Automation-induced complacency potential: Development and validation of a new scale," Frontiers in Psychology, *10*, 225.

Michon, John A. 1985. "A critical view of driver behavior models: what do we know, what should we do?." in Human behavior and traffic safety, pp. 485-524. Boston, MA: Springer US.

Moyes, R. 2016. Key elements of meaningful human control. Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons.

Mumford, E. 2006. "The story of socio-technical design: Reflections on its successes, failures and potential," Information Systems Journal, *16*(4), 317–342.

Norman, D. A. 1990. "The 'problem' with automation: Inappropriate feedback and interaction, not 'over-automation.'," Philosophical Transactions of the Royal Society of London. B, Biological Sciences, *327*(1241), 585–593.

Nozick, R. 1981. Philosophical Explanations (Issue 1, pp. 81–88). Harvard 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.

Nyholm, S., and Smids, J. 2020. "Automated cars meet human drivers: Responsible human-robot coordination and the ethics of mixed traffic," Ethics and Information Technology, *22*(4), 335–344.

Plunkett, D. 2015. "Which concepts should we use?: Metalinguistic negotiations and the methodology of philosophy," Inquiry, *58*(7–8), 828–874.

Queloz, M. 2020. "From Paradigm-Based Explanation to Pragmatic Genealogy," Mind, *129*(515), 683–714.

Queloz, M. 2022. "Function-Based Conceptual Engineering and the Authority Problem," Mind.

Rawls, J. 1999. A theory of justice (revised edition) *Oxford: Oxford University Press*.

Santoni de Sio, F., and Mecacci, G. 2021. "Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them," Philosophy and Technology, *34*(4), 1057–1084.

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

Scharp, K. 2013. Replacing Truth. Oxford University Press UK.

Scharre, P., and Horowitz, M. C. 2015. "Meaningful Human Control in Weapon Systems: A Primer," Center for a New American Security, *16*.

Schellekens, M. 2018. "No-fault compensation schemes for self-driving vehicles," Law, Innovation and Technology, *10*(2), 314–333.

Schwarz, E. 2018. The (im)possibility of Meaningful Human Control for Lethal Autonomous Weapon Systems. https://blogs.icrc.org/law-and-policy/2018/08/29/im-possibility-meaningful-human-control-lethal-autonomous-weapon-systems/

Simion, M. 2018. "The 'should' in conceptual engineering," Inquiry: An Interdisciplinary Journal of Philosophy, *61*(8), 914–928.

Simion, M., and Kelp, C. 2020. "Conceptual Innovation, Function First," Noûs, *54*(4), 985–1002.

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

Sundell, T. 2020. "Changing the subject," Canadian Journal of Philosophy, *50*(5), 580–593.

Thomasson, A. L. 2020. "Pragmatic Method for Normative Conceptual Work," in A. Burgess, H. Cappelen, and D. Plunkett (Eds.), Conceptual Engineering and Conceptual Ethics. Oxford University Press. https://doi.org/10.1093/oso/9780198801856.003.0021

Thomasson, A. L. 2022. "How should we think about linguistic function?" Inquiry, 1–32.

Veluwenkamp, H. 2022. "Reasons for Meaningful Human Control," Ethics and Information Technology, *24*(4), 51. https://doi.org/10.1007/s10676-022-09673-8

Veluwenkamp, H., Capasso, M., Maas, J., and Marin, L. 2022. "Technology as Driver for Morally Motivated Conceptual Engineering," Philosophy & Technology, *35*(3), 71.

Veluwenkamp, H., and van den Hoven, J. 2023. "Design for Values and Conceptual Engineering," Ethics & Information Technology