Meaningful human control as reason-responsiveness the case of dual-mode vehicles

Mecacci, Giulio; Santoni de Sio, Filippo

DOI
10.1007/s10676-019-09519-w

Publication date
2019

Document Version
Final published version

Published in
Ethics and Information Technology

Citation (APA)
Mecacci, G., & Santoni de Sio, F. (2019). Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics and Information Technology*, *22*(2), 103-115. https://doi.org/10.1007/s10676-019-09519-w

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Ethics and Information Technology
https://doi.org/10.1007/s10676-019-09519-w

ORIGINAL PAPER

Meaningful human control as reason-responsiveness: the case of dual-mode vehicles

Giulio Mecacci[1] · **Filippo Santoni de Sio**[1]

Abstract
In this paper, in line with the general framework of value-sensitive design, we aim to operationalize the general concept of "Meaningful Human Control" (MHC) in order to pave the way for its translation into more specific design requirements. In particular, we focus on the operationalization of the first of the two conditions (Santoni de Sio and Van den Hoven 2018) investigated: the so-called 'tracking' condition. Our investigation is led in relation to one specific subcase of automated system: dual-mode driving systems (e.g. Tesla 'autopilot'). First, we connect and compare meaningful human control with a concept of control very popular in engineering and traffic psychology (Michon 1985), and we explain to what extent tracking resembles and differs from it. This will help clarifying the extent to which the idea of meaningful human control is connected to, but also goes beyond, current notions of control in engineering and psychology. Second, we take the systematic analysis of practical reasoning as traditionally presented in the philosophy of human action (Anscombe, Bratman, Mele) and we adapt it to offer a general framework where different types of reasons and agents are identified according to their relation to an automated system's behaviour. This framework is meant to help explaining what reasons and what agents (should) play a role in controlling a given system, thereby enabling policy makers to produce usable guidelines and engineers to design systems that properly respond to selected human reasons. In the final part, we discuss a practical example of how our framework could be employed in designing automated driving systems.

Keywords Meaningful human control · Ethics of self-driving cars · Accountability for autonomous systems · Proximity scale of reasons · Responsible innovation in self-driving cars · Ethics of human–robot interaction

Introduction

Automation is increasingly becoming part of even the most common technological solutions. A number of technological devices, from smartphone to cars to war drones, are increasingly acquiring a certain degree of intelligence and, consequently, autonomy. These technological devices have the capacity to plan and initiate actions autonomously, urging us to reflect on the extent a human subject, capable of moral reasoning and of carrying responsibility for their actions, can still be fully responsible for their behaviour. The deployment of automated solutions can obscure and conceal the role of a human agent in technologically mediated action, and conceal the exact relation that links the controller to their automation-aided action. This is not only an interesting philosophical question but also an urgent practical issue. When an agent gives up part of the control over a certain action by delegating part of the activity to an autonomous device, this can lead to unwanted and unpredictable results, while also creating so-called responsibility or accountability gaps (Matthias 2004) (Sparrow 2007) (Heyns 2013), situations where the behaviour of the device leads to unwanted outcomes or even lethal accidents, but it is not clear whether any human agent can legitimately be deemed accountable for that. In this paper we address these questions in relation to one specific case-study: autonomous vehicles. In particular, we investigate which kind of control over intelligent vehicles is required to maintain high levels of safety and accountability (Sparrow and Howard 2017). However, we believe that many aspects of our discussion may well apply to other kinds of intelligent machines.

✉ Giulio Mecacci
g.mecacci@tudelft.nl

Filippo Santoni de Sio
f.santonidesio@tudelft.nl

[1] Delft University of Technology, Jaffalaan 5, 2628BX Delft, The Netherlands

Published online: 06 December 2019

Concerns for reliability and accountability of intelligent systems have previously been voiced in relation to autonomous weapon systems (stopkillerrobots.org). In order to address these issues, the notion of Meaningful Human Control (MHC hereafter) has been recently gaining popularity. The concept of MHC appeals to the intuition that when autonomous systems are deployed in unstructured, dynamic and potentially unpredictable environments, simply having a human agent involved at some point in the decisional chain (sometimes called 'the kill chain'[1]) may not be sufficient to prevent unwanted mistakes and so-called accountability gaps; human persons must maintain a role that is as prominent as possible (Article 36 2014) (Human Right Watch 2015).

The concept of meaningful human control has remained under-defined in the political debate on autonomous weapon systems; however, Santoni de Sio and Van den Hoven recently provided a comprehensive philosophical account of it (Santoni de Sio and van den Hoven 2018). In that paper, the authors suggested a possible application of the notion of MHC to automated driving systems. In this paper, we pick up that suggestion and start developing a full account of MHC over automated driving systems.

Santoni de Sio and van den Hoven's approach is original in two ways. First, the authors produced an encompassing notion of control, one that applies not just to intelligent artefacts, but also to the entire "socio-technical system" of which these are part. In their notion of intelligent system, devices themselves play an important role but cannot be considered without accounting for the numerous human agents, their physical environment, and the social, political and legal infrastructures in which they are embedded. Second, in line with the so-called Value-Sensitive Design approach (van den Hoven 2013), Santoni de Sio and Van den Hoven's work is meant to ultimately provide not just political and legal regulation but also *general design guidelines*—applicable to devices and (social) infrastructures alike—to achieve and maintain a meaningful form of control over autonomous systems in the military domain as well as in civilian domains like transportation. Their claim is that, in order to achieve meaningful human control over intelligent systems, two conditions should be jointly satisfied; they termed them the *tracking* and *tracing* conditions. The tracking condition requires a system to be responsive to the relevant human reasons to act; tracing requires instead the presence of one or more human agents in the system design history or use context who can at the same time appreciate the capabilities of the system and their own responsibility for the system's behaviour. The joint satisfaction of these two conditions grants human controllers, designers, programmers, regulators and others a more meaningful kind of control over automated systems, thereby maximizing safety and eliminating unwanted accountability gaps.

Whereas we think that Santoni de Sio and Van den Hoven account is very promising, we also believe that it needs to be further developed in order to fulfil the general function the authors attribute to it. In this paper, in line with the general framework of value-sensitive design proposed by the authors, we aim to *operationalize* the general concept of MHC in order to pave the way for its translation into more specific design requirements.[2] In particular, we will focus on the operationalization of the *first* of the two conditions the authors investigated: the so-called 'tracking' condition. Our investigation will be led in relation to one specific subcase of automated system: dual-mode driving systems (e.g. Tesla 'autopilot').

The operationalization of the tracking condition of MHC will be done in two ways. First we will connect and compare meaningful human control with a concept of control very popular in engineering and traffic psychology: (Michon 1985), and we will explain to what extent tracking resembles and differs from it. This will help clarifying the extent to which the idea of meaningful human control is connected to, but also goes beyond current notions of control in engineering and psychology. Second, we take the systematic analysis of practical reasoning as traditionally presented in the philosophy of human action (Anscombe, Bratman, Mele) and we adapt it to offer a general framework where different types of reasons and agents are identified according to their relation to an automated system's behaviour. This framework is meant to help explaining what reasons and what agents (should) play a role in controlling a given system. This can enable policy makers to produce usable guidelines and engineers to design systems that properly respond to selected human reasons.

This paper has four different goals: first, contributing to the discussion on Responsible Innovation and Value Sensitive Design, by showing how MHC can be operationalized and embedded into automated systems by design; second, contributing to the philosophical debate on autonomous systems and human responsibility by connecting the theory of MHC to engineering notion of controls on the one hand and to the philosophical theory of action on the other; third, contributing to the ethics of autonomous driving systems by starting a systematic application of the theory of MHC to the case study of dual-mode vehicles; fourth, contributing to the philosophy of action by proposing an original application

[1] https://web.archive.org/web/20130613233413/http://cno.navylive.dodlive.mil/2013/04/23/kill-chain-approach-4/

[2] On the operationalization of values in value-sensitive design see van de Poel (2013).

of some of its basic notions to the new case of (partially) autonomous systems.

The remainder proceeds as follows: we first recapitulate the general notion of control developed by Michon and widely applied in traffic engineering and psychology. Then, we delve into the notions of "meaningful human control" and "tracking" as introduced by Santoni de Sio and Van den Hoven, and we consider their advantages over a more traditional notion of control in engineering and behavioural psychology; we then introduce the analysis of practical reasoning as presented in the philosophy of action and explain why it is relevant for our goals; based on this, we introduce and present what we call the proximity of reasons scale; we show how our framework can help solve some issues left open by Santoni de Sio and van den Hoven (2018).

The driver's tasks: strategical, tactical, operational control

John Michon (1985) describes three tasks that a driver must perform: strategical planning, tactical manoeuvring and operational control (Fig. 1). These three tasks are layered on top of each other and are meant to specify three functional levels of control. Higher levels coordinate and constrain lower ones. This is a well-known theory in traffic psychology, and we believe it well represents a classic notion of control as used in engineering more generally. It is unclear to which extent these three functional levels of performance should be correlated with a general notion of control. We think it is reasonable to take all the three of them as being different modes of control. Michon's notion of control could be then paraphrased by stating that a system is under the control (in general) of an agent if, and to the extent to which, its behaviour responds to the agent's plans, manoeuvres or operations. Correspondingly, an agent loses control of a vehicle as soon as this does not respond anymore to any of those levels of control. This notion of control can bring about some interesting implications, especially when applied to intelligent systems. One of them regards the fact that this notion can apply to human and non-human agents alike. An automated driving system, for instance, can be deemed in operational control of a car for as long as there is a correspondence between the software operations and the car behaviour. Another implication is that certain intelligent systems might be deemed under human control just because they are responding to a human agent's very general strategical planning.

On the one hand, the notion of control that Michon promotes is well suited to model the interaction between drivers and autonomous systems. Partially automated vehicles, for instance, might be modelled as taking over lower levels of control (i.e. operational and tactical), while leaving the driver with strategical control. On the other hand, we believe that this model, *as is*, has also some limits. Namely, it might make up for a 'lowly demanding' notion of control; one that, if utilized to model control over autonomous driving systems, can potentially generate safety concerns and accountability gaps. It could generate safety concerns, because it might mislead into deeming humans in control even when they are holding very loose reins (humans' general plans being released through processes they don't really understand or control). This creates the possibility of misalignments between human relevant decisions and the system's actions. This notion of control can also generate responsibility gaps, because it doesn't stress the differences between human controllers and artificial controllers, whereas this is a key difference when it comes to moral and legal responsibility attribution, as only humans can be held responsible for unwanted actions or mistakes of a technical system. In the next paragraph, we will see how the notion of meaningful human control tries to provide stronger conditions for control, in the attempt to avoid the above-mentioned unwanted implications.

Meaningful human control, tracking and the role of reasons

Santoni de Sio and Van den Hoven (2018) produced a philosophical theory of "meaningful human control" that might avoid the potential issues described in the last section. In their account, the adjectives 'meaningful' and 'human' should be read as indicating that the notion promotes a stronger and clearer connection between human agents and intelligent systems, thereby resulting in better safety and clearer accountability. A system that is under meaningful human control is less likely to cause accidents, as the relationship between human controller—be it a designer, a programmer, a driver—and the controlled system is more robust and resilient. This is achieved by satisfying two conditions, termed tracking and tracing.

The tracking condition requires the system to be able to respond to—i.e. to track—the relevant reasons of the relevant human agents to act, or refrain from acting. There can be meaningful human control over a system to the extent such system is able to seamlessly co-vary its behaviour according to certain human patterns of (moral) reasoning. According to the second condition, named *tracing*, meaningful human control can be achieved only if it is possible to identify one or more human agents within the design and use chain that have the capacity to (i) understand the capabilities of the system while at the same time (ii) appreciating their own moral responsibility for its behaviour.

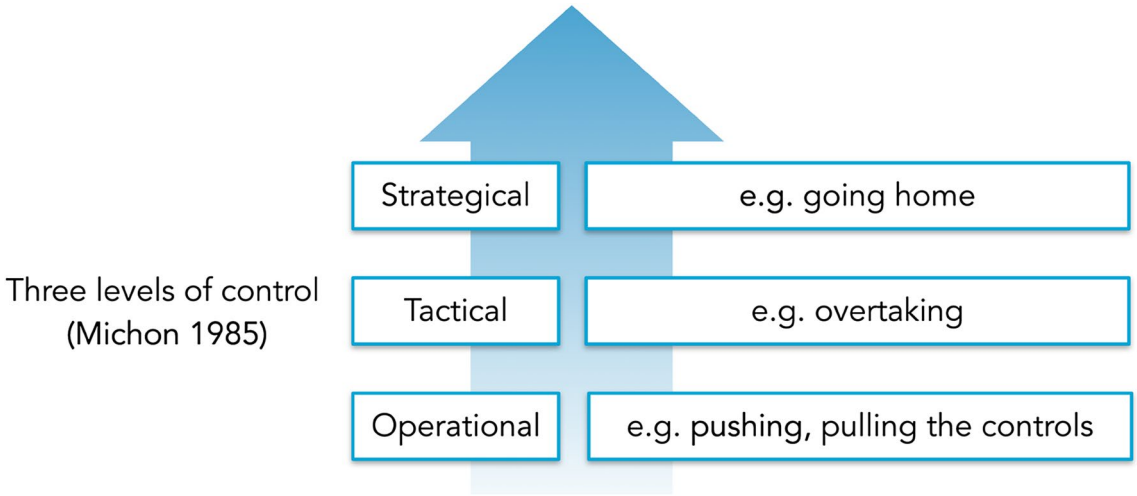

Fig. 1 Michon's three levels of control (simplified)

The role of the tracing condition in the prevention of accidents with dual-mode vehicles has been already highlighted and discussed in (Santoni de Sio 2016) and (Santoni de Sio and van den Hoven 2018). In a nutshell, their claim is that a dual-mode vehicle is far more prone to accidents if the tracing condition does not realize, that is: if the appreciation of the real limits of the driving capacities of the vehicle and its driver, and the appreciation of one's moral responsibility for the (mis)behaviour of the system are not owned by the same (group of) persons. This may for instance happen when: the car manufacturers are well-aware of the technical limits of the driving system they produce and/or the (current) mental limits of a human drivers for whom it's produced, but they shift all responsibility for accidents related to these limits to the human drivers, by having them accepting certain terms and conditions. On the other hand, the human driver is well-aware of her responsibility but badly overestimates the driving capacities of the vehicle and/or her own capacities to appropriately retake control when necessary. However, we believe that the role of the tracking condition for the design of safer and more just dual-mode vehicles hasn't been equally appreciated. In this paper, we will therefore delve into the tracking condition.

Classical engineering accounts of control such as Michon's (1985), as said, mainly focus on the controller and its capacity to interact with a vehicle's behavior. The tracking condition for meaningful human control has a wider focus, in at least two senses. First, it explicitly employs a more encompassing notion of *system*. A driving system includes human agents and vehicles as well as the whole traffic environment and the social, legal and political infrastructures. Second, a larger and more diverse number of potential agents involved in control tasks is considered. These agents can all be potential controllers of the vehicles insofar as their reasons can be reflected in the system in multiple ways. These two features have interesting implications.

By adopting a wider notion of a system, the tracking condition suggests that in order for the system to be under meaningful human control, each of its element, including the human agents themselves, should be maximally responsive to reasons. This implies that, while humans should be capable (e.g. appropriately trained or skilled) to behave according to certain reasons, all the other elements of the system should be designed to do the same. Not only the numerous components of the system should be in that sense 'responsive' to reasons, but could in turn offer relevant reasons for action, reasons that the other components of the system should be able to recognize and respond to. For instance, according to an example from Santoni de Sio and van den Hoven, an automated driving system should not only appropriately respond to the plans of its individual driver but also to some relevant features of road infrastructures—e.g. signs, traffic lights—as well as to some formal and informal traffic norms present in a given society. This might seem odd at first, given that the definition of the tracking condition explicitly mentions responsiveness to human reasons to act, and not to other features of a system. However, as we will explain in more detail below, infrastructures and traffic norms can be said to reflect in turn the intentions of designers, policy makers or even the society in which they are embedded.

In synthesis, the MHC approach adopts a notion of a (driving) system made of diverse and numerous elements, and suggests that all of them should be optimized to respond to the intentions and plans of its driver as well as to some intentions and plans of the traffic system's designers, the policy makers or even to some general norms of a society. Admittedly, this may be seen as something making control

more demanding than the traditional engineering notion of control, insofar as more design requirements potentially enter the picture. However, this approach also allows for an original combination of higher level of autonomy (i.e. less human driving) with a higher human control on a driving system (via technical and institutional infrastructures); in fact, according to MHC, in principle control can be achieved also via agents that are not directly related to the driving task as drivers or supervisors, provided the vehicle is designed to respond to the relevant intentions and plans of these other relevant agents: designers, policy-makers, and the society as a whole.

Santoni de Sio and van den Hoven explore the ideal conditions to achieve meaningful human control. In this paper, we take a step toward the practical operationalization of those conditions, and therefore we prefer to interpret them as criteria for meaningful human control. The conceptual shift is subtle but substantial. From a system design perspective, tracking and tracing can be interpreted as evaluation criteria to assess the extent to which meaningful human control is reached—and reachable—for each given system. They can also serve as instrumental values the design process should strive for, in order to optimize systems for meaningful human control. That said, work can be done to clarify those criteria and make them more usable. First of all, in order to design for tracking, it should first be established *which reasons of which humans* are relevant in any given context; this requires an appropriate *unit of measure* to identify and potentially categorize different reasons, and determine which ones a certain system can, in general, be designed to respond to; and, second, in order to implement specific engineering design solutions that promote meaningful human control by maximizing tracking, they also need a general *reference framework* to represent (i) how different reasons stand in reciprocal relation and (ii) how they stand in relation to a system's behaviour. From the perspective of designing systems that realize tracking, identifying different reasons in a vaguely specified space of "relevance" might open the door to arbitrary, not well-grounded, and thus morally problematic design choices. That's why, without denying that normative decisions have to be taken in any design process including designing for tracking, and therefore some disagreement may always emerge in this respect,[3] we also believe that a general theory of MHC should at least provide some objective reference framework within which to identify and prioritize different reasons of different agents.

We will argue that reasons to act can be represented in a two-dimensional space, a 'proximity scale' that identifies them according to how closely they may influence a system's behaviour. In order to do that, we will draw inspiration from the proximal/distal distinction which has been used multiple times in philosophy of action to characterize intentions (Bratman 1984a). This is why in the next section we will provide an overview of the philosophical background that will serve to give solid theoretical foundation to our proximity scale of reasons.

Intentions, reasons, and practical reasoning

The concept of meaningful human control crucially relies on the idea of reasons tracking. In order to better understand and operationalize this concept, we propose to look at the philosophical analysis of reasons and actions as developed in philosophy in the so-called theory of action. In her 1957 seminal book *Intention*, Elizabeth Anscombe made four points which heavily influenced the theory of action of the decades to come. First, she distinguished three kinds of intention: *intentions-in-action*, for instance opening a door intentionally, *intentions with which it is acted*, for example entering an apartment with the intention of stealing, and *simple intentions*, for instance intending to go to the movies tonight (while not taking yet any action). So, intentions can be a different "distance" from action—they can coincide with the action, accompanying, or anticipating it. Secondly, the three kinds of intentions are conceptually connected: we as humans are able to recognize an intentional action as such because we have the concept of further intentions; and we understand further intentions because we know what a simple intention is. Thirdly, what crucially characterizes human intentions is not their causal role in the production of behaviour but rather their capacity to provide complex rational explanations of it. We need the concept of intention not as much to causally explain human behaviour—physical explanations can (better) do this job—but mainly to *make sense of* it. Fourthly, and relatedly, the task of a philosophical theory of action is not investigating the mental causes of human behaviour but rather creating a toolbox of concepts to make sense of human actions, by identifying and classifying reasons and intentions according to their relationship with each other and with the behaviour they are supposed to explain; in Anscombe's own words—borrowed from Aristotle—to elaborate a logic of *practical reasoning*; much in the same sense in which traditional logic provides the tools to conceptualize (sound) theoretical reasoning (Anscombe 1957).

More recently, Michael Bratman (Bratman 1987; Bratman 1984b) has developed Anscombe's project by explaining the relationship between actions, intentions and plans.

As summarized by Mele (1992, p. 137)

among intentions there are intentions for the specious present and intentions for the nonimmediate future, or what I

[3] We offer some example of these different normative options in the final section.

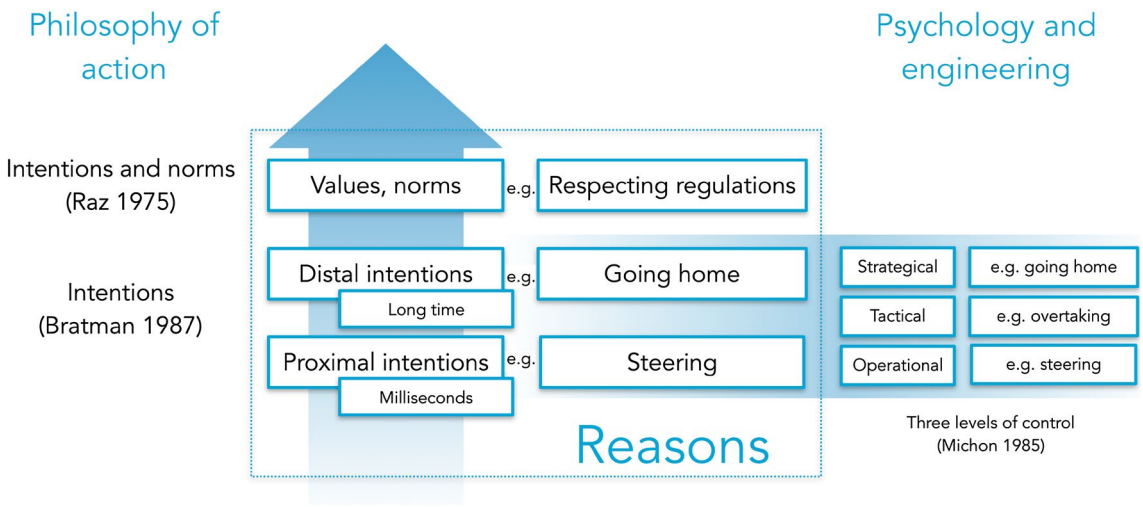

Fig. 2 Mapping Michon's model of levels of control to the distinctions made in philosophy of action. To be noticed that by considering norms and values we can expand the field of reasons to act beyond Michon's model

shall call, respectively, proximal and distal intentions. Distal intentions, Michael Bratman has argued, "are typically elements in larger plans," plans that "help me to coordinate my activities over time, and my activities with yours" (1984a; p. 379; cf. 1987). For Bratman, the coordinative roles of distal intentions rest on several features of these intentions: they have the capacity to control behaviour; they "resist (to some extent) revision and reconsideration"; and they involve dispositions to reason with a view to intention satisfaction and "to constrain one's intentions in the direction of consistency" (1987, pp. 108–109).

Philosophy of action theorists have put forward frameworks that identify reasons and intentions even beyond the individual agent's plans. These theories include reasons to act that are quite removed from the action they are related to, up to including norms. Raz (1975) suggested that rules and norms can also be considered important reasons for action insofar as, similarly to plans, they coordinate and structure more proximal intentions and ultimately behaviour.[4]

To recap, one main idea in the philosophy of action of the last sixty years is that human behaviour is open to different kinds of rational explanations; some of these refer to intentions which are very close in time or even coincide with the behaviour they explain; others refer to reasons which can be further away from the behaviour, and also shared with other agents, like plans or norms. This phenomenon was famously dubbed the "accordion effect" by Davidson (2001); human action can be legitimately described and explained with reference to many different reasons that are nested into each other; and different valid descriptions and explanations of the same actions are possible, depending on how broad or narrow is the set of reasons included in the explanation.

The proximity scale of reasons for tracking

Whereas in the traditional philosophy of action, practical reasoning was meant as an explanatory aid to make sense of the relationship between human reasons to act and human action, we propose to use the structure of practical reasoning to make sense of the relationship between human reasons and *the behaviour of systems which include human and non-human agents*. It is precisely on the nature of this relationship that the tracking criterion for MHC crucially depends.

Michon's model (Michon 1985) already employs concepts that are very close to those of philosophy of action, and he does that to explicitly describe control over vehicles. His tripartition of driving tasks can be easily correlated to Bratman's dichotomy of proximal and distal intentions. Strategical tasks would map to Bratman's distal intentions. Significantly, both authors make use of the word 'plans' while describing their respective notions. Michon's operational tasks would typically connect to Bratman's proximal intentions (Fig. 2). The fact that Michon's model is slightly more granular (three vs. two classes) is only due to the different aims of the respective models. For what matters to us, both models can be seen as describing and categorizing an ultimately continuous space.

The philosophical tradition is wider and more generally applicable if compared to Michon's model of control, that

[4] This opens the door to the idea of "shared" or "collective intentionality" which is itself the topic of a whole strand of literature which, for reasons of space, will not be discussed in this paper. See for instance (Schweikard and Schmid 2013).

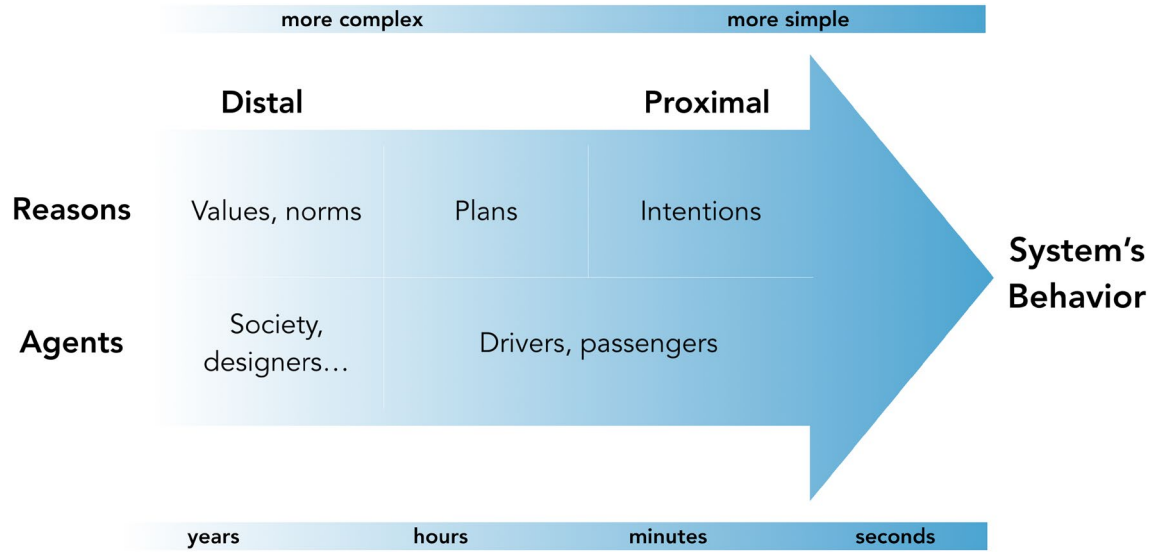

Fig. 3 The proximity scale. Reasons can be classified according to their proximity value. Bratman's proximal and distal intentions (plans) are typically temporally closer to a system's behaviour than Raz's values and norms. They are also simpler, in the sense that more complex reasons explain and affect a system's behaviour only through more proximal ones. Different agents can also be identified as typical endorsers of certain kinds of reasons

mostly concerns an individual driver's internal reasoning and the resulting actions upon a vehicle. For that reason, philosophy of action can provide elements that are not considered by traditional models of control, and produce a theory with broader scopes. This is especially true when we consider a theory of meaningful human control that strives to model control in terms of a relationship between human reasons in general and autonomous systems–at–large (not just vehicles then). As explained in "Meaningful human control, tracking and the role of reasons" section, meaningful human control potentially concerns multiple agents, and multiple elements of a system. A system can be deemed to be under meaningful human control by more than one agent, or even by supraindividual agents, such as a company, a given society or a state. To model this complex relation, and substantiate the tracking criteria, we propose a model, inspired by both psychological and philosophical accounts, where human reasons are ordered in a scale with respect to how closely they influence a system's behaviour (Fig. 3). This scale is meant to be a reference framework that is configured as a continuous space. Although the only relevant aspect is the reasons' relative position, a few classes of reasons can be identified. This, although not necessary, is useful in two ways. First, it allows us to understand how the scale maps to the models it is inspired by (traffic psychology and philosophy of action). Second, classifying reasons might contribute to the identification of different classes of agents to whom those reasons can be typically attributed. Having a few discrete classes of reasons makes it easier to draw connections with discrete classes of agents. It should be noted that one could draw different numbers of distinctions to adapt to different contexts and needs.

It is important, before proceeding, to clarify what kind of magnitude the proximity value represents. Compatibly with what we have seen regarding both psychological and philosophical models, *time* is an important factor in determining proximity, together with *complexity*.

There is an important caveat for what concerns the time factor in our proximity scale, that does not apply to the other models we have discussed. In fact, whereas traditional models were meant to explain the relationship between human intention and human action, the tracking criterion represents the relation between human intention and an intelligent system's action. Within those theories, operational tasks and proximal intentions could, respectively, precede or coincide in time with the action itself. In those cases where the tracking criterion is meant to be applied, even the most proximal intentions might be arbitrarily distant in time from the behaviour itself. For instance, an automated system might include a vehicle driving on mars that, though responding extremely seamlessly to the most proximal reasons—i.e. somebody on earth intending to steer away from a crater–, would do that with a delay of around 14 min. What matters, however, is again the *relative* distance between two or more reasons and the system's behavioural response. Introducing an *absolute* temporal delay will not affect the scale (Fig. 4).

Also, distal reasons are usually more *complex* than proximal ones, and the latter reasons might figure as simpler components of the former ones. A very proximal reason will not just be close(r) in time to the execution of a system's

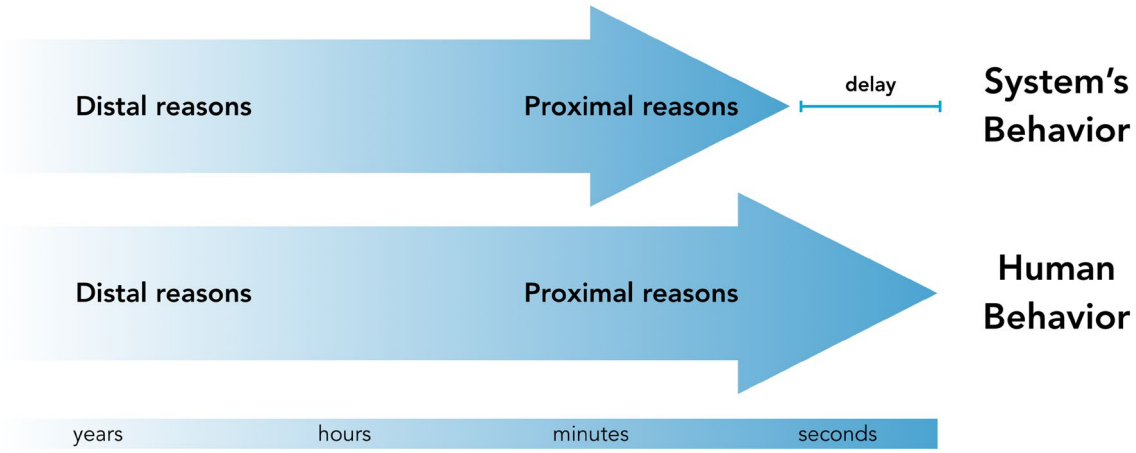

Fig. 4 Between human reasons and systems' behaviour there can be a temporal gap which does not compromise the scale

behaviour, then, but will also explain such behaviour at a very detailed, lowly abstract level. Typically, it will not be possible to further decompose proximal reasons into simpler reasons. Typical proximal reasons clearly map to Michon's operational tasks and, in the case of a driving system, are the intention to steer left or right, to brake or to accelerate, and so on. Proximal reasons can be assembled into more general, abstract ones. Distal reasons are reasons that are meant to explain why a certain system adopted a certain strategy, or made a certain plan of action. For instance, a driving system might respond to an agent's general intention to drive to a bar rather than to their workplace. A distal reason, e.g. that of driving home, can be described by appealing to the smaller, simpler elements that compose them, down to the most proximal intentions related to every single turn, lane change or speed adjustment. One can have the general plan to go home *by* intending to take a certain road with a certain speed, speeding a couple times, and so on. In turn, one can take a certain road by intending to brake, steer right at the crossroad, and accelerate accordingly.

It should be noticed that an agent might endorse certain distal reasons while not explicitly expressing any proximal one. This is most times the case. One does not usually plan whether and how to accelerate or brake at each of the turns that the trip involves. It is more likely that a driver who just wants to go home, will not commit to any route in particular, or to the average speed of the car, let alone to the numerous, finely grained actions to be taken by the driving system in order to realize the plan. And this is exactly what we do when we take a public transport service like a train, a bus or a taxi; we plan to get to a certain place at a certain time and we realize our plan without committing to (in the case of trains and buses: without even being able to) realizing more proximal intentions to take a certain route or speed etc.

The proximity scale helps us to also identify typical agents playing a role in a system. Different types of agents can typically bear different types of reasons. These reasons can range from those of drivers and final users, typically more proximal to the system's behaviour, to those of the government, which expresses its reasons through laws and regulations that in turn constrain and coordinate more proximal reasons, such as individual plans and intentions. 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. In the next two sections, we will see how the concepts we have discussed above can be applied, respectively, to evaluate the presence and extent of meaningful human control in case scenarios and to design systems that are optimized for tracking and meaningful human control.

Tracking and the scale of reasons: assessing meaningful human control in dual-mode driving systems

An automated driving system can assist a driver in different ways and to widely different extents. A lowly automated system will assist drivers in controlling the car, implementing a number of functions aimed at facilitating driving tasks (e.g. cruise control) and maximizing safety in potentially dangerous situations. A highly automated system might entirely replace the driver, rendering them a de facto passenger. Driving automation hold the promise of a safer, environmental friendlier, more efficient traffic system; but it is an open question how the development and introduction of these systems should be realized from a technical, regulatory, and socio-psychological point of view, in order to achieve the desired results and prevent unwanted risks. Recent fatal accidents like the Tesla (Shepardson 2018) and the Uber (Bellon 2018) have given a vivid representation of the risks involved

in a non-responsible introduction of automated driving systems on the public road. In this paper we have looked at the ethics of the introduction of automated driving systems from the angle of meaningful human control, and of the tracking criterion more in particular. We have introduced the scale of reasons as a tool for better understanding meaningful human control and actively designing for it. In this and the next section we will show how tracking can help better understand and design for meaningful human control in the use case of "dual-mode" vehicles such as for instance the Tesla model S.

In 2016, a lethal traffic accident involved a Tesla model S that was at that time making use of the "auto-pilot" feature. As the on-board sensors failed to recognize a lorry crossing the street, the vehicle crashed into it causing the death of its driver (Yadron and Tynan 2016). Tesla "auto-pilot" is a level 2 assistive driving system that provides partial driving autonomy in certain circumstances, such as while driving in a highway. As for every level 2 system, it is mandatory for the driver to constantly remain vigilant and ready to regain operational (manual) control at any time. The company's defence line was built around the claim that, while the vehicle behaved according to its (limited) capabilities, the driver was not properly monitoring the car's behaviour. Specifically, though he had been warned multiple times to regain control of the wheel, he was not ready to correct the car's behaviour when the assistive systems ceased to be able to control the trajectory.

The theory of meaningful human control may provide an interesting angle on this case. It requires us to look, first, at the system's responsiveness to reasons or its lack thereof (tracking criterion) and, second, at the presence of at least one human agent in the system design and use that can: (a) appreciate the capabilities of the system and (b) understand their own role as morally responsible for the consequences of the system's actions (tracing criterion). In their brief discussion of the Tesla accident, Santoni de Sio (2016) and Santoni de Sio and van den Hoven (2018) focus on the "tracing condition" and wonder to what extent it was satisfied by the driver. In fact, the driver might not have been properly trained for this special mode of interaction with a partially automated car; he might not have been trained to realize the requests formulated by the company, namely not to get distracted while supervising the car's behaviour; most importantly he may have not even been aware of his own limited capacities in driving this new kind of car; furthermore, his appreciation that he was fully responsible for the behaviour of the vehicle might have been impaired by e.g. bad communication on the company's part (the system was advertised as "autopilot") or simply by him lacking sufficient experience in the use of partially automated systems. All the above reasons might indicate poor tracing, leading to conclude that meaningful human control might have been hardly achievable in that case, making in turn the driver not (fully) morally responsible for the accident. As a matter of fact, the authors themselves seem to come to this conclusion on these bases. However, they do not return on how and the extent to which the "tracking condition" may or may not be achieved in such cases, and how to design to achieve it.

Let's then consider the dual-mode vehicle from the perspective of tracking. As we are interested in reflecting on some general design principles to achieve more safety and clearer responsibility, rather than just assessing individual responsibilities in specific accidents, we will go beyond the analysis of the 2016 Tesla accident and will frame the discussion in relation to dual-mode vehicles more generally. In order to assess whether and to what extent the tracking criterion is also satisfied, we have to assess the extent to which the semi-automated car is responsive to the relevant reasons of the relevant agents. Let's consider the reasons of the designated controller, i.e. the driver. Lowly automated driving systems, typically driving assistance systems, are usually meant to be controllable by the driver sitting behind the wheel. More highly automated systems can be designed to be remotely controlled by e.g. aggregated control facilities. We notice that the Tesla of the example is a SAE level 2 automated vehicle. According to this classification, it is designed to engage in automated behaviour under certain circumstances (e.g. highways). The driver, however, is requested to keep the hands on the wheel and be ready to intervene at all times. From the perspective of the tracking criterion, the question is whether and to what extent the system's behaviour is responsive to the controller's relevant reasons to act.

In line with Davidson's "accordion effect", multiple concurring reasons of the driver can be identified as potentially explaining the car's behaviour for each given instance. For example, 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. These seem to be both identifiable, amongst others, as explaining the vehicle's behaviour. As a matter of fact, *if* we only consider responsiveness to the proximal reasons of the vehicle's driver, a dual-mode vehicle could largely satisfy the tracking criterion for MHC. Whenever a competent driver acts with the intention of steering right, the vehicle will steer right, whenever she acts with the intention of braking, the vehicle will slow down etc. whenever the driver will (intentionally) set the vehicle the autonomous mode, the vehicle will switch to that mode, whenever the driver will

touch the steering wheel or the pedals, the system will return under her direct control. In this limited sense, the vehicle is (designed to be) under human control. In fact, going back to the 2016 accident, *if* the driver had pushed the brakes in time, the vehicle would have slowed and stopped in time to avoid the crash.

However, that is what we would conclude if the car were an old-timer rather than a level 2 automated vehicle. Based on the analysis of this paper, we argue that this is not sufficient to establish that the system is designed to maintain *meaningful* human control. In fact, the driver's proximal intentions are certainly not the only relevant reasons a safe automated system should respond to. Other, potentially conflicting, more distal reasons of hers should be reflected in the functioning of the system. In the 2016 accident, the Tesla was not designed to detect a relatively simple conflict between two basic reasons of the driver: he did intend to relinquish the operational control of the vehicle—as showed by the fact that he hasn't touched at all the controls for a long time before the fatal crash—but he also intended to safely get to his destination (and, a fortiori, he did not intend to crash into any other vehicle). The Tesla was not designed to respond to this latter reasons of the driver: first, and most obviously, because right before the crash the vehicle was not able to perceive the lorry crossing the road in front of it, and so to avoid that crash on its own; second, because at a previous time the system was not able to perceive that the driver was not able to intervene either, and so the system couldn't adopt any alternative strategy to realise the driver's (and the other road users') general intention to not incur in any accident, for instance by gradually slowing down and eventually stop in the emergency lane, after the driver didn't react to the vehicle's multiple requests to remain alert and ready to intervene.

Therefore, not only was the system arguably designed to leave room for (moral) *responsibility gaps*, due to issues in the distribution of knowledge about the functioning of the system and perception of one's responsibility between driver and manufacturers (Santoni de Sio 2016, Santoni de Sio and van den Hoven 2018)[5]; the system was also designed to leave room to *control gaps*, insofar as it was designed to have the driver (sometimes) ceasing to realise his general intentions on her own (by relinquishing operational control of the vehicle), while at the same time not being designed to (always) respond to these reasons on its own (Calvert et al. 2019).

Admittedly, designing automated driving systems for more distal reasons, (i.e. realizing a general plan, such as safely driving home) is far more complex and full of variables than responsiveness to more proximal ones. This is not to say the current dual-mode vehicles do not respond to that class of reasons at all. On the contrary, they can drive quite efficiently over long distances and display complex behaviour and decision making. However, having a car that is responsive to individual plans *as reliably* as it currently responds to (proximal) intentions, requires *better* automation. This is an interesting conclusion that can be drawn by utilizing the model we propose. It is not the case that more human intervention will always grant more meaningful human control; more automation can promote better satisfaction of the tracking criterion, and hence more meaningful human control, provided the automation of the system is designed to grant a better responsiveness to the relevant human reasons. Automation, in turn, requires us to be able to design systems that can easily track—that is: *recognize, navigate and prioritize*—the numerous reasons and agents that can co-occur in every given situation. The conceptual tool that we have called 'proximity scale' seems to allow us to more thoroughly reason about designing for meaningful human control. We will see how design solutions are concretely enabled by our findings in the next section.

Designing for tracking and meaningful human control

The proximity scale informs us on the generality of a reason and how closely it influences a certain system's action. Proximity is not meant to be an absolute value, but mainly intended to give an idea of how different reasons relate to each other and with respect to the influenced action. Identifying reasons according to their proximity value allows an engineer with imaginary—yet quite imaginable—technology, to design for meaningful human control. For instance, a highly automated driving system might be designed to satisfy the tracking criterion by obeying the following simple algorithm made of two rules that use the proximity value as main variable:

(i) respond to a proximal reason IFF it does not conflict with a more distal reason
(ii) respond to the most proximal reason allowed by (i)

Of course, different algorithms could be produced to implement different policies. The above example presents a simple pyramid-like scheme where, given some proximity value, and some understanding of reasons, the system knows that it should respond to the most proximal reasons available unless they conflict with any other reason which is relatively more distal. As a result, the policy the above algorithm describes is one that seems to privilege safety over individual freedom and flexibility, and therefore prioritizes responsiveness to societal rules and regulations. Different policies might privilege individual freedom and

[5] See the discussion of the tracing criterion above.

independence, therefore prioritizing (some) proximal reasons over distal ones, granting more control to final users.

An important objection to this idea is that recognizing reasons and potential conflicts between them requires a certain degree of semantical intelligence, something that might not be achievable for our technical solutions in the near future. However, this results from a rather common misunderstanding of the notion of tracking and reason responsiveness We should not forget that we are talking about systems-at-large, and not just intelligent devices. A system's reason responsiveness might well result from the human agents that are part of the system or be embedded in the technology by smart design solutions. Current and future limitations of artificial intelligence are therefore not necessarily an issue for the realization of systems that track human reasons to a satisfactory extent. Smart design solutions should strive to harmonize systems' behaviour and our—potentially changing—reasons and moral values, while minimizing the need for constant active causal contribution from intelligent human controllers.

Let us now see how our exemplificatory algorithm might be implemented in a practical case of dual mode driving. The following are two fictional stories that are meant to illustrate a system that is designed for tracking. Unlike the Tesla case discussed above and in line with our analysis of this paper, we will here broaden the scope of the analysis, as to include not only the different reasons of the driver, but also some reasons of some other agents potentially involved in road traffic.

The first case sees Lucy as the protagonist. Lucy is driving home in a winter night. Visibility, due to a dense fog, is very low, but she is a little technophobic and, despite having a very expensive car, she does not want to use the provided autopilot. She grabs the wheel and starts driving herself. At some point, near a dark alley, a fastidious beeping signal breaks the otherwise surprisingly smooth ride. The vehicle cuts the engine and swerves gently, dodging what seems to be a wrecked car, still smoking on the asphalt. Right past the accident site, Lucy slows down and pulls off in a safe spot to check the situation.

The second case is that of John. John lives near Lucy. He is heading back home around the same time of Lucy. He is using the autopilot feature and relaxing with some jazz after working till very late that day. As he gets to the location of the accident, his dashboard starts beeping and the car slowly drives around the site. However, he's not as willing as Lucy to stop and check. He knows it is his duty to pull off and offer help, but he really does not feel like it that night. He notices that his car slows down dramatically, and seems to be pulling off. He immediately grabs the wheel and push on the accelerator, trying to avoid that annoyance. However, his dashboard warns him that an emergency procedure is about to be deployed. In a matter of seconds, the steering wheel becomes loose, spinning freely on its axis, while the pedals seem to have lost control of the engine. The vehicle slows down and pulls off automatically, with its hazard lights on. A voice compels John to wear his safety vest, leave the car and offer his help, while an emergency call is being automatically dispatched. Left with no choice, and starting to understand the entity of the situation, he leaves the car to offer his help.

That night, Lucy wanted to drive home. She wanted to do it herself though. The system allowed that. It allowed her reasons to influence the system, allowing her to drive manually. However, if the car would have kept realizing those proximal intentions, she would probably have intended to keep driving straight, unaware of the wreck that occupied her lane. Fortunately, her intentions were conflicting with her more general plan, which was to get home safely, and the car was programmed to prioritize those kinds of more distal reasons. It can be observed how in certain situations driving could just be safer with an automated driving system permanently in control. However, this becomes relevant only if this is in some way specified within the system, i.e. there is some reason of some agent, such as the government, that is more distal than Lucy's reasons, and conflicts with them. That would have been the case if, for instance, a regulation was in place establishing that manual driving is unsafe and therefore not allowed.

John's plan was also to go home that night, but his case is different. His car not only denied his intentions to push the accelerator and keep driving, but also refused to comply with his more general plan to go home, pulling off instead. The car was prioritized the interest of the victims of an accident over the driver's will. Rescue (legal) obligations are one of the possible reasons that lead driving behaviour, and it is on the proximity scale more general and further away in time than John's plan to leave the site of the accident without checking. No matter what John's plans were, the system was designed for cars to comply with this obligation.

This is just one example, and, to be clear, we are not necessarily campaigning for vehicles that track specific reasons, like the interests of victims of accidents to be rescued. Our general claim is that vehicles that are under MHC should also respond to some distal reasons of their owners/drivers as well as to some (distal) reasons of other agents in society, as reflected in some moral and legal norms. Which of these reasons specific systems should track remains a normative question on which reasonable persons and policy-makers may disagree.

Conclusions

In this paper, we have offered a framework to systematically reflect on the reasons we want autonomous systems to respond to, and a tool to map these reasons and design systems which track them. By providing conceptual scaffolding to fill the gap between normative values and technical implementations, we contributed to the discussion on Responsible Innovation and Value Sensitive Design. In particular, we started from Santoni de Sio and Van den Hoven notion of meaningful human control and we made their tracking condition more easily operationalizable in engineering design. The general 'proximity' value, instantiated by the proximity scale, has contributed to provide a further, more precise conceptual tool to assess how well MHC is, or can be, expressed, for each given scenario (as exemplified in the last section).

This paper contributed to the philosophical debate on MHC, by connecting this theory to the more technical field of human-vehicle interaction, on the one hand, and to philosophy of action, on the other hand. We believe that such connection has the potential to enable a great deal of further, interdisciplinary research on the subject of MHC and on the ethical design of automated driving systems. Indeed, our study can be interesting to experts in traffic psychology, traffic engineering and engineering more generally looking for conceptual tools to help making value driven design choices. It could also interest philosophers of action, especially those concerned with exploring to what extent philosophical theory of action may apply to the case of human interaction with artificial autonomous systems.

The theory of MHC is gaining increasing attention in both political and technical environments. One of the most important objectives for the future of this theory is to work out insights on how to implement its normative indications in real world technical and institutional design solutions that are often constrained by a number of contingent factors. We believe that with this work we have moved one further step towards this direction, and we hope that others will welcome our insights to develop the theory further.

Acknowledgements We would like to thank Simeon Calvert for his precious insights on the final draft of our manuscript.

Funding This work is part of the research project "Meaningful Human Control over Automated Driving Systems" with project number MVI.16.044, which is (partly) financed by the Dutch Research Council (NWO).

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

References

Anscombe, G. E. M. (1957). *Intention*. Oxford: Basil Blackwell.

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

Bellon, T. (2018). Fatal U.S. self-driving auto accident raises novel legal questions [WWW Document]. Reuters. Retrieved September 18, 2018 from https://www.reuters.com/article/us-autos-selfdriving-uber-liability-anal/fatal-u-s-self-driving-auto-accident-raises-novel-legal-questions-idUSKBN1GW2SP.

Bratman, M. (1984a). Two faces of intention. *Philosophical Review, 93,* 375–405.

Bratman, M. E. (1984b). Two faces of intention. *Philosophical Review, 93,* 375–405.

Bratman, M. E. (1987). *Intentions, plans and practical reason*. Cambridge, MA: Harvard University Press.

Calvert, S., Mecacci, G., Van Arem, B., Santoni de Sio, F., Heikoop, D., Hagenzieker, M. (2019). Gaps in the control of automated vehicles on roads. IEEE intelligent transportation systems magazine

Davidson, D. (2001). Actions, reasons, and causes. In D. Davidson (Ed.), *Essays on actions and events* (pp. 3–20). Oxford: Oxford University Press. https://doi.org/10.1093/0199246270.003.0001.

Heyns, C. (2013). Report of the special rapporteur on extra-judicial, summary or arbitrary executions.

Human Right Watch. (2015). *Mind the gap: The lack of accountability for killer robots*. New York: Human Right Watch.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6,* 175–183. https://doi.org/10.1007/s10676-004-3422-1.

Mele, A. R. (1992). *Springs of action: Understanding intentional behavior*. New York: Oxford University Press Inc.

Michon, J. A. (1985). *Human behavior and traffic safety, human behavior and traffic safety*. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4613-2173-6.

Raz, J. (1975). Reasons for action, decisions and norms. *Mind, 84,* 481–499.

Santoni de Sio, F. (2016). Ethics and self-driving cars. A white paper on responsible innovation in automated driving systems. Dutch Minist. Infrastruct. Water Manag. Rijkswaterstaat.

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers Robotics AI, 5,* 15. https://doi.org/10.3389/frobt.2018.00015.

Schweikard, D.P., Schmid, H.B. (2013). Collective intentionality. Stanford Encyclopedia of Philosophy

Shepardson, D. (2018). Tesla, NTSB clash over autopilot investigation | Reuters [WWW Document]. Retrieved September 18, 2018, from https://www.reuters.com/article/us-tesla-crash-autopilot/tesla-ntsb-clash-over-autopilot-investigation-idUSKBN1HJ2JS.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24,* 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

Sparrow, R., & Howard, M. (2017). When human beings are like drunk robots: Driverless vehicles, ethics, and the future of

transport. *Transportation Research Part C: Emerging Technologies, 80,* 206–215. https://doi.org/10.1016/j.trc.2017.04.014.

Van De Poel, I. (2013). *Philosophy and engineering: Reflections on practice, principles and process, philosophy of engineering and technology*. Dordrecht: Springer. https://doi.org/10.1007/978-94-007-7762-0.

van den Hoven, J. (2013). Value sensitive design and responsible innovation. In R. Owen, J. Bessant, & M. Heintz (Eds.), *Responsible innovation* (pp. 75–83). Chichester: Wiley. https://doi.org/10.1002/9781118551424.ch4.

Yadron, D., Tynan, D. (2016). Tesla driver dies in first fatal crash while using autopilot mode | Technology | The Guardian [WWW Document]. Retrieved September 28, 2018, from https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk.