Humanistic Management Journal (2021) 6:249–272
https://doi.org/10.1007/s41463-020-00101-x

ORIGINAL RESEARCH

Between Real World and Thought Experiment: Framing Moral Decision-Making in Self-Driving Car Dilemmas

Vanessa Schäffner[1]

Received: 7 July 2020 / Accepted: 10 November 2020 / Published online: 23 November 2020

Abstract

How should driverless vehicles respond to situations of unavoidable personal harm? This paper takes up the case of self-driving cars as a prominent example of algorithmic moral decision-making, an emergent type of morality that is evolving at a high pace in a digitised business world. As its main contribution, it juxtaposes dilemma decision situations relating to ethical crash algorithms for autonomous cars to two edge cases: the case of manually driven cars facing real-life, mundane accidents, on the one hand, and the dilemmatic situation in theoretically constructed trolley cases, on the other. The paper identifies analogies and disanalogies between the three cases with regard to decision makers, decision design, and decision outcomes. The findings are discussed from the angle of three perspectives: aspects where analogies could be found, those where the case of self-driving cars has turned out to lie in between both edge cases, and those where it entirely departs from either edge case. As a main result, the paper argues that manual driving as well as trolley cases are suitable points of reference for the issue of designing ethical crash algorithms only to a limited extent. Instead, a fundamental epistemic and conceptual divergence of dilemma decision situations in the context of self-driving cars and the used edge cases is substantiated. Finally, the areas of specific need for regulation on the road to introducing autonomous cars are pointed out and related thoughts are sketched through the lens of the humanistic paradigm.

Keywords Ethics of algorithmic decision-making · Autonomous driving ethics · Ethics for self-driving cars · Crash algorithms · Moral dilemma · Trolley cases

✉ Vanessa Schäffner
vanessa.schaeffner@gmail.com

[1] Philosophische Fakultät S.J., Munich School of Philosophy, Kaulbachstraße 31a, D-80539 Munich, Germany

Introduction

Humanistic Management: State of Research and Research Gap

Although humanistic management is a relatively young field of research that has emerged mainly during the last ten years, its body of literature covers a broad range of subjects. Among the key topics are practical issues arising with management practices in organizational contexts as well as theoretical investigations into corresponding terms and concepts. The question of what humanistic management actually means has continuously been developed in seminal works (e.g. Bachmann et al. 2018; Melé 2016; Von Kimakowitz et al. 2011). Aspects that are given particular attention are – above all – the notion of human dignity in both conceptual and practical terms (e.g. Bal 2017; Kostera and Pirson 2017) and issues of responsibility (e.g. Glauner 2019; Pirson 2019), sustainability (e.g. Ferguson et al. 2019; Sasse-Werhahn 2019), and freedom (e.g. Dierksmeier 2018a; Pirson 2018). Another widely discussed issue is the role of humanistic management in the context of business education and how it can help to transform educational programmes towards a more holistic understanding of management (e.g. Amann et al. 2011; Dierksmeier 2016; Lepeley et al. 2016). Furthermore, a considerable number of articles deal with the driving force that humanist elements present for management practices in a global perspective (e.g. Dierksmeier et al. 2011; Lupton and Pirson 2014) referring especially to the notion of a world ethos (e.g. Gohl 2018; Pirson and Keir 2018). Apart from this thematic focus, humanistic management literature is enriched with single exploratory articles on emergent phenomena that are related to humanistic management in particular and business themes in general. Articles of this kind are meant to stimulate further debates on pertinent developments in the business domain; recent ones deal, for example, with the issue of cryptocurrencies (Dierksmeier 2018b; Seele 2018).

Humanistic management sees itself as a perspective on management devoted to the protection of human dignity as well as the promotion of (societal) well-being (e.g. Pirson 2017). Since making decisions is a central part of management practice, the humanistic management perspective is concerned with the question of how decisions can be made in a way that particularly takes this humanistic paradigm into account. It is interesting to note that the approaches suggested throughout humanistic management literature relate to decision situations on different levels. Firstly, decisions and coherent practical actions of individuals (micro level) in organizational contexts are discussed. Hormann (2018), for example, argues for resilience as a quality for leaders to deal with organizational trauma and Leisinger (2018) uses the concept of the world ethos to show that managers need to reflect on values and resulting actions. Secondly, several articles deal with decision-making on the corporate level of organizations and institutions (meso level). Kabadayi et al. (2019) identify respect, trust, fairness, and inclusion as values that should guide organizational processes in order to achieve positive outcomes of social innovation in service. Gonstead and Chhin (2019) suggest the Catholic Social Teaching concepts of subsidiarity and solidarity as suitable principles for participatory leadership and establishing shared decision-making in businesses. Thirdly, situations of decision-making on the level of society, culture, and civilization (macro level) are reflected. Clark et al. (2019) present an empirical study which investigates key factors considered to promote human flourishing and their dependence on geographic and contextual aspects. By bringing together humanistic management and social innovation, Fisk et al. (2019) conceptualise a notion of systemic social innovation that is able to promote the flourishing of every human being and all life on earth.

In spite of its wide thematic coverage, humanistic management literature has hitherto devoted little effort to a novel type of decision-making that is evolving at a high pace in a digitised business world: since more and more organizational processes are transformed into automated routines running on complex technological systems, algorithmic decisions become an integral part of the business domain. They present a fundamental change to the realm of decision-making that is increasingly finding its way into an ever-increasing number of businesses, and they will also become a central issue for management practices in the medium term. In catching up with this development, humanistic management might be able to develop its own academic scope further by expanding its expertise on algorithmic decisions, on the one hand, and to make valuable contributions at an early stage of algorithmic automation research, on the other. To put it in different words, the engagement of humanistic management research with algorithmic moral decision-making does not only bear the potential to strengthen its own field, but at the same time to encourage relevant research in other fields as well.

Algorithmic Decision-Making: The Case of Self-Driving Cars

The emerging phenomenon of machine algorithms taking decisions that have hitherto been exclusively dealt with by humans can be observed in several practical contexts. One of the most controversially debated technological innovations on the research and development agenda of the current decade is autonomous driving. Basically, the vision of fully automated cars encompasses a comprehensive software code that controls the behaviour of the car in every conceivable driving situation. Naturally, this entails situations where the driverless car might find itself confronted with the need to ‘take a decision’, i.e. to ‘choose’ between possible alternative courses of action. Among other issues, ethical solutions to situations where collisions and resulting personal harm are unavoidable – so-called moral dilemmas – are often at the centre of contentious debates in literature on self-driving cars (e.g. Gerdes and Thornton 2016; Goodall 2014; Lin 2013; Millar 2017). They represent a specific type of moral decision-making problems in which incompatible action alternatives are confronted with each other, all of which are both right and wrong at the same time for moral reasons. However, the phenomenon of dilemmas occurring in driving contexts is as old as road traffic itself. What makes moral dilemmas a matter of such controversial ethical interest just now? The German Ethics Commission on Automated and Connected Driving appointed by the Federal Ministry of Transport and Digital Infrastructure states: “Genuine dilemmatic decisions, such as a decision between one human life and another, depend on the actual specific situation, incorporating ‘unpredictable’ behaviour by parties affected. They can thus not be clearly standardised, nor can they be programmed such that they are ethically unquestionable.” (2017: 7, Rule No. 8) Why? It seems plausible that a human decision taken within fractions of a second before an imminent collision is far from being comparable to a sophistically implemented algorithm taking effect in emergency situations.

This paper takes up the case of self-driving cars as a prominent and vigorously debated example of applied algorithmic decision-making. Its core idea is to compare it to two edge cases involving decision situations characterised by dilemma structures. In recent years, several research papers have ascertained a striking resemblance between so-called trolley cases[1] and moral dilemmas in the context of autonomous driving. However, a growing number of authors have adopted a critical stance towards this view, insisting on manifold disanalogies.[2] This paper seizes their ideas and argues that for gaining a comprehensive picture of what makes up the decision situation related to driverless cars, it is not enough to rely on trolley cases alone. The given analysis contributes to the ongoing controversy on trolley cases, but at the same time, it goes beyond it. Tying in with the often-raised criticism that trolley cases are constructed and unrealistic, manual driving is proposed as an additional, real-life edge case. Mundane accidents have been a widespread everyday experience for decades and the way human drivers solve dilemma situations can be interpreted as being the status quo in this respect. On these grounds the paper asserts that a close look at moral decisions taken by human drivers deserves more attention. Besides, empirical studies (e.g. Bonnefon et al. 2015, 2016) have found evidence for the claim that there is a substantial divergence of the moral dimensions that are touched by dilemma reactions in the case of human drivers, on the one hand, and driverless cars, on the other. Why is that? It is argued in this paper that the conceptual and epistemic factors framing the respective situations of decision-making are principally vital for the way self-driving cars are morally perceived. A closer investigation into decision-making situations in cases of manual driving is therefore recognised as a reasonable means to enrich the discussion on ethics for crash algorithms.

The arguments presented in this paper are subject to two major limitations resulting from the assumptions that are made regarding particular implementation approaches. It is obvious that the definition of the decision situation about crash algorithms depends crucially on the nature of admissible decision criteria. This paper is concerned with the so-called top-down approach (e.g. Etzioni and Etzioni 2017) stating that behavioural responses to dilemma situations rely on the programming of specific ethical principles or rules into the system's software code. The basic idea of the alternative approach – the so-called bottom-up approach – refers to the concept of machine learning. Machines learn about ethical decisions from analysing data that has been collected from human driver behaviour in real-life driving situations. If either the bottom-up alternative or a hybrid approach combining elements of top-down and bottom-up (e.g. Misselhorn 2018a) was chosen, the decision situations of human drivers and decision makers on algorithms would be intertwined. However, the argumentation of this paper is essentially based on contrasting clearly separable situations of moral decision-making. A similar problem arises with regard to the degree to which the implementation of particular algorithms for dilemma cases is binding. Pertinent literature on self-driving car ethics contemplates two approaches: the first one – which this paper adopts – insists that all ethical behaviour of the vehicle is part of a mandatory software, while the second one implies an option for customised solutions. These so-called personalised settings

[1] In pertinent literature, there seems to be some confusion about the usage of related terms. Some authors use *trolley problem* or *trolley experiment* where they actually refer to *trolley cases*. These are constructed scenarios involving single-agent decision-making in moral dilemmas with unavoidable harm. Contrasting many trolley cases against one another to examine how moral intuitions change between them, Philippa Foot (1978) was the first to introduce the famous thought experiment called the *trolley problem*. It was later elaborated by Judith Jarvis Thomson (1976, 1985), who framed the variant that is most often referred to.

[2] Among others, Nyholm and Smids (2016) offer a comprehensive overview on the disanalogies between the trolley experiment and ethics for self-driving cars. Himmelreich (2018) continues the debate by revealing further insights into conceptual, epistemic, and practical difficulties inherent in trolley cases. Wolkenstein (2018) reflects on the role trolley dilemmas have played in recent debates on driverless car ethics. Wu (2020) examines potential analogies from the perspective of law. Keeling (2020) argues for the usefulness of the trolley problem by responding to common objections.

allow passengers to choose a specific set of predefined responses to potentially occurring dilemmas according to their own moral preferences (e.g. Applin 2017; Gogoll and Müller 2017; Millar 2017). For example, passengers may choose if their car should always prioritise their safety or whether they might be willing to sacrifice themselves in certain scenarios. If a personalised ethics setting was possible, there would be several layers of decision-making that ultimately lead to the behaviour of the driverless car; programmers as well as passengers would be involved. The findings presented in this paper must be interpreted in light of these assumptions; at the same time, they encourage further research that covers alternative implementation approaches.

This paper is arguably the first attempt that builds upon both manual driving and trolley cases as edge cases for ethical decision-making in the context of self-driving car dilemmas. The original contribution of the present paper is that it brings together the practical real-life perspective of manual driving and the theoretical concepts behind trolley cases for the purpose of creating valuable insights and contributing novel arguments to the ethical discussion on crash algorithms. The paper begins with some arguments substantiating the choice of the two edge cases, followed by a brief outline to the discipline of machine ethics which specifies that the issue of designing crash algorithms is ultimately a matter of human – not a machine's – decision-making (section 2). Then the three cases of self-driving cars, manual driving, and trolley cases are juxtaposed with respect to decision situations under dilemma conditions (section 3). The analysis is guided by a conceptual framework that covers essential aspects of decision situations: *who* takes the moral decision, *how* is it taken, and *what results* from it. This framework brings together a broad range of attributes that are suited for characterizing decision situations and for precisely working out similarities and differences between them. Subsequently, results are discussed from the angle of three perspectives (section 4). Firstly, identified analogies between driverless car dilemmas and either of the edge cases are reflected upon. It is explained why both edge cases could be helpful, only to a limited extent, to inform driverless car ethics. Secondly, the paper goes into the aspects where the case of self-driving cars has turned out to lie in between both edge cases. Thirdly, the points where no analogies could be found are discussed regarding their relevance for the issue of designing ethical crash algorithms.

On Trolleys, Humans, and Machines

The Controversy about Trolley Cases

The well-known trolley problem goes back to philosopher Philippa Foot who designed it as a thought experiment asking people to imagine the situation of a driver of a runaway trolley which heads rapidly towards five persons working on the track. The driver can reroute the tram on another track where one person is working. What should she do? Foot uses this scenario to elaborate on the doctrine of double effect, an ethical principle "insisting that it is one thing to steer towards someone foreseeing that you will kill him and another to aim at his death as part of your plan" (Foot 1978: 23). She contrasts different variants of trolley cases against one another to explore how people's judgments change in terms of whether an action is morally permissible. Although there are several versions of trolley cases, the basic structure is, however, fairly similar: refusing to act causes the death of five persons, while deciding to act causes the death of one person.[3] Judith Jarvis Thomson (1976, 1985) took up Foot's ideas and presented modified versions that have received much attention throughout research literature in different fields: "Let us begin by looking at a case that is in some ways like Mrs. Foot's story of the trolley driver. I will call her case Trolley Driver; let us now consider a case I will call Bystander at the Switch. In that case you have been strolling by the trolley track, and you can see the situation at a glance: The driver saw the five on the track ahead, he stamped on the brakes, the brakes failed, so he fainted. What to do? Well, here is the switch, which you can throw, thereby turning the trolley yourself. Of course you will kill one if you do." (Thomson 1985: 1397). Besides shifting the burden of decision from the trolley driver to a bystander, Thomson introduces another variation that includes the option to push a fat man onto the track in order to stop the train from hitting any of the workers. JafariNaimi (2018: 318) brings Thomson's intentions of presenting new scenarios to the point: "Through these examples, she put Foot's distinction of *killing versus letting die* under scrutiny and highlighted how other principles, such as being treated as a means or the infringement of the rights of the people involved, are relevant to these cases."

By means of her modifications, Thomson especially draws attention to which ethical approaches such as Kantianism or consequentialism can be used to solve questions of this kind. The same issue is central to the ethical decision problem behind crash algorithms. It is for this reason that, as part of the contemporary debate on driverless car ethics, diverse variants of trolley cases have continuously been proposed as cut-down illustrations of crash algorithm scenarios. Wolkenstein (2018: 169) writes: "In a way, the car's decision is a projected TD, and so it faces the very situation that the programmer has foreseen." Facing an imminent crash that would strike five people crossing the road, should the car swerve and hit a single pedestrian instead? Or should it sacrifice its passengers by e.g. colliding with a heavy obstacle? Although scenarios like these appear similar to a trolley case at first glance, an increasing number of articles take a critical view on alleged analogies. They recognise that trolley cases and collisions in real life differ in non-trivial, morally relevant ways (e.g. Etzioni and Etzioni 2017; Goodall 2016, 2017; Liu 2017; Nyholm and Smids 2016). This paper takes up the apparent confusion in the pertinent research literature. Using trolley cases as edge cases, a thorough investigation that focuses specifically on the decision situation in trolley cases is presented on the basis of existing literature. By means of juxtaposing distinct aspects of the two cases against each other, the paper eventually aims at scrutinising the potential of trolley cases to inform research on ethical crash algorithms, both theoretically and practically.

Are Autonomous Systems the Better Drivers?

A frequently cited argument in the debate on the potential merits of autonomous vehicles is that their introduction will help to eliminate fatalities due to human driver error. Autonomous vehicles do not fall asleep at the wheel, do not drink alcohol, and are not mentally distracted. They also do not intentionally break traffic rules by e.g. wilfully exceeding the speed limit (e.g. Coca-Vila 2018; Gogoll and Müller 2017). However, this is only one side of the coin. While experts emphasise the considerable increase in road safety through the banning of human driving, they at the same time continuously support the implementation of machine learning methods that enable the software systems to adopt driving algorithms based on the model of

[3] For a comprehensive review and systematisation of proposals made in trolley literature, see Bruers and Braeckman (2014).

human driving behaviour. This seems to be a paradox at some point: on the one hand, human driving should be replaced by autonomous systems because it causes too many accidents, but on the other hand, human drivers are considered to be good examples and suitable 'teachers'. An even stronger argument becomes apparent when we turn towards the apparent discrepancy between expectations for self-driving cars and their actual impact. Technological systems are designed to avoid accidents, but it is obvious that they can never be fully safe. Technological shortcomings as well as unprepared and possibly unpredictable situations are only some of the causes that might hamper autonomous vehicles from completely preventing accidents (e.g. Brändle and Grunwald 2019; Goodall 2014). Still their performance is commonly considered as 'better' than those of human drivers if they reduce the number of road fatalities. But, and this point is often neglected in public perception, technological failure may cause certain types of accidents that would not occur with human failure, for example if they provoke specifically careless behaviour of other road users. It is implausible to stipulate that these types of accidents should be tolerated.

Against this background, it seems that self-driving vehicles are not categorically 'better' than human drivers, but they open up a realm of problems that are different from those that are prevalent in everyday accidents. They create a new need for regulation and ethical consideration. As real-life scenarios only allow for intuitive and split-second decisions about which option to take, a human driver may well be excused though her action might be considered unethical (Bonnefon et al. 2015; Gerdes and Thornton 2016). In contrast, the case of driverless cars is much more complicated. Results from the empirical studies by Bonnefon et al. (2015, 2016) reveal that actions by a human driver and an autonomous vehicle are assessed differently in the public perception. Interestingly enough, those judgments still hold even though the same option to solve a moral dilemma is chosen. One of the main reasons for this observed distance in the moral and legal perception of chosen courses of action seems to be that people are likely to expect a higher moral standard from machines (e.g. Etzioni and Etzioni 2017; Wallach and Allen 2009). It appears that for their moral judgments about an action of a driverless car, people also take into account the contextual circumstances in which a specific decision about an action is taken, i.e. how, when, and by whom. In this paper, it is argued that moral decision-making in the face of imminent crashes is fundamentally different for autonomous cars and human drivers. However, a closer consideration of manual driving as an edge case of the investigation seems promising, since it may help to create a deeper understanding of people's moral attitude towards self-driving car dilemmas. And, more importantly, by identifying where exactly the two cases differ, it is disclosed where new regulation is needed.

Designing Crash Algorithms – A Human Task

As Brändle and Grunwald (2019) explain, dilemma situations are of central importance to the issue of autonomous driving not because they pose a practical problem that most urgently needs an answer, but because they raise a core ethical question connected to self-driving vehicles in particular, and artificial intelligence systems in general: Ought an autonomous system be allowed to ultimately decide on human lives? The answer to this question touches the core of machine ethics, a discipline at the interface between ethics, robotics, and AI. It deals with the question of whether autonomous systems are capable of autonomous moral agency, i.e. whether they are able to take, perform, and reflect upon moral decisions themselves (Bendel 2018). There is currently widespread agreement in machine ethics that artificial systems cannot be considered full ethical agents in a sense that is equivalent to human morality (e.g. Misselhorn 2018b; Moor 2006). The main argument behind this position is that central properties of the human mind such as conscious understanding and free will do not go together with representations in computational technologies, especially the deterministic character of machines, being dependent on clear definitions of states and transition rules (e.g. Wallach and Allen 2009). Some authors emphasise that attributing *machine* autonomy does not imply that machines possess *moral* autonomy (e.g. Domingos 2015; Lucas Jr. 2013). Etzioni and Etzioni (2017: 409) write:

> "Some believe that machines can attain full autonomy: for instance, weapon systems that choose their own targets without human intervention, and whose missions cannot be aborted. In fact, even these machines are limited to the missions set for them by a human, and they are only 'free' to choose their targets because a human programmed them that way. Their autonomy is limited."

Following this view, the behaviour of autonomous vehicles in dilemma cases is entirely determined by so-called crash algorithms the programming of which remains ultimately a human task. Weber and Zoglauer (2019) even argue that it is not the vehicles that find themselves in moral dilemmas, strictly speaking, but humans that have to decide upon possible reactions. JafariNaimi (2018: 303) observes a "need for *algorithmic morality*, broadly construed as a set of variables and action scripts, that could decide the fate of people involved in such scenarios in order to bring self-driving cars to the masses." Against this background, the decision situation in the context of autonomous driving that is analysed in this paper is a situation of human decision-making on the design of crash algorithms – and not machines making autonomous decisions as full ethical agents.

Contrasting Decision Situations

This section offers a detailed analysis of dilemmatic decision situations in manual driving, trolley cases, and autonomous driving. The analysis is guided by a conceptual framework introducing three major categories: the decision maker (*who* takes the decision?), the design of the decision (*how* is the decision taken?), and the consequences of the decision (*what* outcomes are caused by the decision?). Each category comprises distinct factors that help to identify whether and to what extent there are potential analogies and discrepancies between the three cases.

The Decision Maker: *Who* Takes the Decision?

The first category of aspects focuses on the question of who is in charge of taking the decision.

Type of Decision In both manual driving and trolley cases, decisions are taken by individuals. Even if driver assistance systems might be activated, the human driver of a manually driven car is able to take voluntary decisions which are spontaneous and mostly intuitive due to the time constraints faced in cases of impending danger. In trolley cases, the decision is left either to the driver (as in Foot's version) or the bystander (as in Thomson's version). Conversely, decisions in autonomous driving dilemmas are characterised by a different type of decision. A variety of stakeholders such as, e.g., lawyers, philosophers, politicians, and manufacturers are involved in the process of elaborating behavioural responses to dilemma situations in autonomous driving contexts (e.g. Awad et al. 2018; Lin 2016; Nyholm and Smids 2016). They are expected to take into account the perspectives of all affected parties such as future passengers, other groups of roads users as well as uninvolved persons (e.g. Gogoll and Müller 2017; Hübner and White 2018). Therefore, moral decisions in the context of crash algorithms could be described as being collective.

Decision Perspective[4] The human driver takes her decision from a first-person decision perspective. She finds herself both physically and mentally involved. Basically, this applies also to the driver version of trolley cases designed by Foot. In Thomson's variants, however, the bystander agent merely observes the situation that must be resolved. As there is no physical involvement, decisions are taken from a third-person perspective. As stated in the introduction, this paper adopts the angle of a mandatory ethics setting regarding the case of autonomous cars, which seems to resemble the bystander trolley cases.[5] Decisions about crash algorithms for solving moral dilemmas are not taken by passengers themselves. Instead, a group of delegated people making decisions on crash algorithms are urged to consider the interests of all stakeholders involved while they are both physically and temporally dissociated from the real-life situation where the implemented algorithms might come into effect. Consequently, the decision perspective in the case of autonomous cars can be categorised as a third-person perspective.

The Decision Design: *How* Is the Decision Taken?

The aspects of the second category shed light on the conceptual design of the decision situation in each of the three cases.

Point of Time A closer look at the point of time when the actual decision is taken shows that the human driver can only take a spontaneous action when she faces an imminent collision. Her way of decision-making is more of a responsive reaction to given circumstances and her decision materialises in the very same moment she has taken it. In contrast, decision-making in trolley cases as well as in autonomous driving is prospective or anticipatory (Nyholm and Smids 2016). The bystander in the trolley case has the possibility to anticipatorily choose the option that is most appropriate in her view. Driverless cars operate on algorithms that have been implemented long before their market launch. The processes of taking a decision, on the one hand, and materialising it in actual situations, on the other hand, are temporally separated. This means that decisions about how the vehicle will react in a given situation are already taken when the machine is programmed, i.e. ahead of time and under conditions free from time pressure (Lin 2013).

[4] The decision perspective describes the relation between the decision maker and the decision, i.e. the degree to which the decision maker is personally involved in the situation and its consequences.

[5] In contrast, a personalised ethics setting appears to be comparable to Foot's driver trolley cases. JafariNaimi (2018: 307) emphasises that the relational bonds that tie the decision maker to the situation and potential victims is a highly relevant factor in determining decision circumstances: "There is a difference practically, emotionally, and intellectually to being in charge of the trolley and knowing firsthand about the brakes, the tracks, the terrain, the number of the passengers, and other specifics of the situation as opposed to being a bystander who is making inferences about the situation from a distance."

Basis of Information The basis of information available to decision makers plays a vital role for their epistemic situation. A human driver usually does not have detailed information at her disposal which would potentially enable her to ponder on different options. Even though driver assistance systems might serve as additional sources, the provided information is likely not to be sufficient to capture the complete picture of an emerging crash situation. Therefore, the quantity of available information remains sparse and its quality depends on human capacities of sensual perception and mental information processing. On the contrary, the design of trolley cases is based on the presumption of complete and perfect information on the environment. Some research articles criticise this as being highly unrealistic (e.g. Goodall 2016). Instead, they propose considering the alternative of imperfect information for real-life application contexts, as is the case with driving crash scenarios. Crash algorithms make use of a broad quantity of detailed data, e.g. about environment characteristics, that is collected by sensors, cameras, and other technical aids (Lin 2013, 2016). Technological components of autonomous systems are able to process and evaluate these data much faster than humans. Nevertheless, the gathered data is naturally error-prone and incomplete. The case of driverless cars is therefore found to lie, on the one hand, in between the sparse and low-quality information the human driver has at her disposal, and, on the other hand, the complete and perfect information stipulated in trolley cases.

Choice Design Trolley cases have often been criticised for their choice design, which only allows for single choices (Goodall 2014; Nyholm and Smids 2016). The trolley decision maker is faced with a decision that directly leads to the death of either one worker or five. But moral choices in real-life situations are complex and consist of several intertwined layers of interaction between different parties involved (Gogoll and Müller 2017). The human driver has the possibility of adjusting her intuitive reaction to a certain extent in real time, e.g., she can still try to swerve when she notices that her intuitive choice to perform an emergency-braking manoeuvre is not enough to prevent a collision. Of course, her options are practically very limited for reasons of time pressure. As the programming of autonomous vehicles is free from such constraints, a broad range of complex, multilayer choices are possible at the time of programming. Additionally, various courses of action can be coordinated and selected according to their desired outcomes.

Degree of Restrictiveness The trolley decision maker can only choose between two predetermined options: change tracks or not in the driver versions and pull the switch or not in the bystander variants. Respective outcomes are explicitly settled in any case. There is no possibility of elaborating individualised and more sophisticated solutions. This decision concept, especially its degree of restrictiveness, has been subject to occasional criticism by many researchers. The human driver, on the contrary, can react to arising circumstances more flexibly. Her options are not necessarily separated from one another, but can be combined, e.g. braking and swerving at the same time. This is also possible for crash algorithms where the programmer has more flexibility to choose from a great variety of possible actions.[6]

[6] Autonomous vehicles which are based on the top-down approach are subject to the general constraint that choices are implemented beforehand and cannot be changed dynamically at run time. Since this aspect is, however, not a constraint of the decision situation per se, it is neglected at this point.

Relation to Real-Life Situations The human driver finds herself directly engaged in a highly individual and specific decision situation that is framed by a unique context. In contrast, trolley cases are characterised by entirely constructed set-ups, comprising abstract decisions that are disconnected from any specific and real context. It is not least because of this hypothetic character that Himmelreich (2018), among others, questions the evidential value of trolley cases for real-life situations. As far as autonomous driving is concerned, algorithmic decisions are programmed as responses to hypothetic but still realistic scenarios which are expected to materialise in real life. They are not as abstract as in trolley cases, but still not attached to specific, real occurrences at the time of decision-making. Therefore, crash algorithms can be said to lie in between the two edge cases in terms of their relation to real-life situations.

Identity of Affected Parties Accidents that involve manually driven cars are always individuated events and involve real and specific persons. In general, it is possible to unequivocally determine the identity of affected parties, whereas the constructed set-up of trolley cases treats the potential victims as impersonal entities (Liu 2017). Since the bystander can see and distinguish the workers on either track, some authors maintain their identities to be specific (e.g. Hevelke and Nida-Rümelin 2015). In fact, the versions of trolley cases as described by Foot and Thomson only speak of five men on the one track and one man on the other. So apart from the fact that all potential victims are male, it is only clear how many workers will be sacrificed in either case, but their personal identity remains unspecific. It is plausible to assume that the workers' identities remain faceless and interchangeable, presenting merely placeholders in a decision that is basically about counting up numbers of lives, not valuing individuals. As concerns the case of driverless cars, the identity of potential individual victims is said to be unspecific for algorithms do not refer to real-life situations at the time when they are implemented (e.g. Luetge 2017). This means that potentially every individual could be affected by crash algorithms in one way or another.

The Decision Outcome: *What* Results from the Decision?

By considering the consequences that result from decisions taken, the third category provides further evidence for relevant differences in moral decision situations.

Complexity of Responsibility Issues A growing number of research articles critically remark that trolley cases are characterised by a restrained view of agents and affected persons as impersonal entities (Liu 2017). As Nyholm and Smids (2016) point out, trolley cases fail to take the significance of issues related to moral and legal responsibility into account, such as special obligations e.g. of manufacturers towards passengers. However, matters like these are integral parts of any modern society. It can be plausibly assumed that choices of individuals are influenced by an inherent human sense of responsibility and related expectations about possible sanctions. But even if the human driver is legally responsible for her actions, she would not be acting unlawfully and therefore must not fear criminal consequences – provided that she did not willfully cause the crash (Contissa et al. 2017; Federal Ministry of Transport and Digital Infrastructure 2017: 7, Rule No. 8). Dilich et al. (2002: 246) provide a possible reason why the blameworthiness of the human driver is limited: "Once it is determined that a driver was confronted with a sudden emergency which demanded an extraordinary response, outside the normal experience of most drivers, the outcome of the accident is dictated more by the chance of the circumstances than by the performance abilities of the driver and his vehicle." However, moral and legal responsibility are among the most pivotal – and also controversial – issues concerning crash algorithms. It seems plausible to assume that decision makers in this case have a far-reaching responsibility for the outcomes that result from implemented algorithms. Lin (2013) writes: "[…] it matters to the issue of responsibility and ethics whether an act was premeditated (as in the case of programming a robot car) or reflexively without any deliberation (as may be the case with human drivers in sudden crashes." Self-driving cars cannot excuse themselves with regard to psychological stress factors which characterise the human driver's decision situation (e.g. Birnbacher and Birnbacher 2016; Trappl 2016). The German Ethics Commission specifies: "In the case of automated and connected driving systems, the accountability that was previously the sole preserve of the individual shifts from the motorist to the manufacturers and operators of the technological systems and to the bodies responsible for taking infrastructure, policy and legal decisions." (2017: 7, Rule No. 10) One of the most critical points is that directing an autonomous vehicle towards a specific person may be considered an act of intentionally targeting individuals and therefore a discriminatory practice (e.g. Millar 2017; Santoni de Sio 2017).

Certainty of Outcomes The human driver is aware that the outcomes of her actions are largely uncertain, and so she is likely to intuitively choose the option she considers will result in the least harm to all affected parties. On the contrary, various research papers critically remark that trolley cases make unrealistic assumptions regarding the certainty of outcomes (e.g. Goodall 2016). The trolley decision maker can be perfectly sure that her expectations about the consequences resulting from her choice will eventually come into effect. Anyhow, it is widely agreed that consequences of real-life crashes can never be completely certain. JafariNaimi (2018: 306 f.) writes: "Literal readings of the trolley experiments mask the deep sense of uncertainty that is characteristic of ethical situations by placing us outside the problematic situations that they envision, proffering a false sense of clarity about choices and outcomes." Dilich et al. (2002: 245) provide further empirical evidence by referring to a study conducted by Lechner and Malaterre (1991) who "concluded that above all, the result of an emergency situation is completely uncertain and that the behavior of the obstacle to be avoided, in particular in the case of an intersection, is a determining factor." This is also true for the case of autonomous cars (Lin 2013; Wolkenstein 2018). However, due to the wide range of available information on environmental parameters and the system's technological capacities, probabilistic assessment methods can be used to detect manoeuvres that have a higher change of causing less harm than others by means of generating probability distributions over all possible outcomes (Goodall 2017; Himmelreich 2018; Liu 2017; Nyholm and Smids 2016). Therefore, decisions on crash algorithms can be classified as decisions under risk which are found to lie in between the two extreme positions of perfect certainty and extensive uncertainty.

Scale of Effects The scale of effects is one of the most challenging matters in the context of crash algorithms. In general, the effects of a human driver's actions remain restricted to a unique and individuated situation.[7] However, they might very well have long-term effects both

[7] Given the assumption that she would learn from her experiences and tell other persons about it, she would potentially influence their decisions in similar scenarios – as well as her own decisions in future situations. But this will only affect a limited number of cases and concerned parties.

for the victim and the potentially traumatised driver. Trolley cases, in contrast, are hypothetical thought experiments and the extent to which they have an impact on moral judgments in real-life situations is questionable. It is for this reason that Himmelreich (2018) doubts their usefulness for the attempt to meaningfully inspire self-driving car ethics. Anyhow, the scale of decisions implemented in autonomous cars is much broader. With an increasing number of driverless cars on the road, the number of situations that are solved by identical crash algorithms will also grow. This way, there is an increase of ethically consistent decisions on action. However, the repeated application of those algorithms might lead to an accumulation of effects, e.g. when two identically programmed vehicles are involved in a collision (Nyholm and Smids 2016). Owing to these cumulative effects, the scale of the consequences resulting from a moral decision is unpredictable at the time of programming.

Discussion of Findings

In this section the results of the preceding analysis are discussed. Table 1 presents a concise summary of the findings.

Can the Edge Cases Keep Up?

Among the great variety of research and press articles dealing with moral issues in autonomous driving, there are hardly any that do not connect them with some version of trolley cases. Although this seems tempting at first glance, the analysis presented in this paper reveals that there are actually very few touch points between the respective dilemma decision situations. Apart from the findings that decisions in both cases are taken anticipatorily from a third-person perspective and the identity of affected parties is unspecific, the conceptual design of trolley cases accounts for major differences from the decision situation related to autonomous cars. But is this enough to draw the conclusion that the trolley debate is not useful here at all?

The presented investigation leaves trolley cases as theoretically-constructed scenarios built on assumptions that are far from moral decision-making in real-life cases. JafariNaimi (2018:

Table 1 Overview of identified analogies between situations of moral decision-making in the context of self-driving cars, trolley cases, and manual driving

Category	Subcategory	TC	MD	IB	NA
Decision maker	Type of decision				x
	Decision perspective	x			
Decision design	Point of time	x			
	Basis of information			x	
	Choice design		x		
	Degree of restrictiveness		x		
	Relation to real-life situations			x	
	Identity of affected parties	x			
Decision outcome	Complexity of responsibility issues				x
	Certainty of outcomes			x	
	Scale of effects				x
∑ *sum (number of analogies identified)*		*3*	*2*	*3*	*3*

TC = Trolley cases; MD = Manual driving; IB = In-between; NA = No analogy.

309) writes: “Yet, ethical situations are not snapshots frozen in time but uncertain and living movements. We are not engaged with them as outside judges but as ethical characters.” Bringing another related aspect to the discussion, Applin (2017) remarks that the option of self-destruction of passengers is missing in trolley cases, which is due to the fact that the decision maker herself is not directly involved in the decision scenario. In contrast, the human driver has a valid choice to sacrifice herself in order to save others. Considering the delegated manner of decision-making, this is only true for driverless cars in an indirect way: decisions about sacrificing future passengers are taken by the committee of decision makers, not future passengers themselves. This constellation raises moral and legal issues of its own. Besides, Keeling (2020) notices that the nature of the differences between trolley cases and driverless cars needs to be defined more precisely. He argues that the fact that trolley cases exclude considerations of something that is morally relevant, e.g. responsibility, does not necessarily prove that trolley cases are of no relevance to the question of how a self-driving car should react to an upcoming crash. In his view, it is not sufficient that the differences are morally *relevant*, they also have to be *categorical*. Some authors say that there is a difference of this kind in terms of the aspect of certainty of outcomes, for instance Nyholm and Smids (2016: 1286):

> “Reasoning about risks and uncertainty is categorically different from reasoning about known facts and certain outcomes. The key concepts used differ drastically in what inferences they warrant. And what we pick out using these concepts are things within different metaphysical categories, with different modal status (e.g. risks of harm, on one side, versus actual harms, on the other).”

Keeling’s response to this is interesting. He says that what the self-driving car does not ‘know’, i.e. is not certain about, is not only what will result from its actions, but also what the real-life situation that eventually occurs will be like:

> “In risky cases, it is necessary to consider different possible worlds when evaluating acts, as the moral value of an act depends on the value of the outcomes in different possible worlds. This difference is insufficient to render trolley cases irrelevant to the moral design problem. This is because trolley cases are relevant only to the AV’s utility function, in the sense that these cases are used to identify which properties are relevant to the evaluation of actions (e.g. harm, responsibility, fairness, and so on).” (2020: 300)

All in all, the presented analysis supports the serious skepticism that has been raised in previous literature about the attempt to draw too wide-ranging analogies to driverless car dilemmas (e.g. Gogoll and Müller 2017; Goodall 2016; Liu 2017; Nyholm and Smids 2016). It stands in line with the position that trolley cases are not able to provide the entire answer to the issue of how to program crash algorithms. But the outlined findings suggest that trolley cases are not completely irrelevant either. In fact, they help to uncover some of the focal points that the case of driverless cars faces. Etzioni and Etzioni (2017: 415) stress that they “can serve as a provocative dialogue starter”. Hübner and White (2018) exhibit that trolley cases can help to carve out a distinction between ‘involved’ and ‘uninvolved’ parties. This is a central matter of crash algorithms as it emphasises the need to consider the perspectives of different stakeholders in order to provide a reasonable way of dealing with moral and legal issues in autonomous driving. Yet, a holistic ethical discussion of crash algorithms must go beyond an examination of single trolley cases.

There is, however, another angle from which the trolley perspective might be of value for the application context of autonomous car dilemmas. The range of possible scenarios involving driverless cars is broad and corresponding moral decisions are naturally complex. Relevant literature (e.g. Etzioni and Etzioni 2017; Gerdes and Thornton 2016; Goodall 2014) has exposed traditional ethical approaches such as utilitarian or deontological principles to reach their limits in certain scenarios for two reasons: either they are notoriously general and not specific enough to adequately solve particular real-life cases. For example, Isaac Asimov's (1950) 'Three Laws of Robotics' are often cited as an example of deontological ethics leading to contradictory implications when rigidly followed. Or they are too specific to be transferred to similar cases, i.e. they might lead to morally desirable results in one situation but have questionable implications for another. This means that, if a top-down approach to implementing crash algorithms is adopted, coherence is a crucial quality for justifiable ethical principles. The academic trolley debate is concerned with deriving implications from contrasting different cases against one another rather than focusing on single trolley cases. When confronted with single trolley cases, people usually follow their moral intuitions to decide on one of the given options for action. It is usually not before they are asked to respond to more than one case that they start to question their own implicit assumptions and to revise their intuitions in light of changed situational set-ups. This process of continuously reflecting on moral judgments is reminiscent of the so-called 'reflective equilibrium', a well-known – but at the same time often criticised – methodological approach to normative ethics that aims at achieving coherent ethical principles. Having become prominent by John Rawls' *A Theory of Justice* (Rawls 2009 [1971]), it states that intuitive moral judgments, either on general moral principles or specific relevant cases, are revised back and forth until systematic principles are achieved which prove suitable for practice. In this equilibrium state, the derived judgments are considered stable, free of conflict with each other, and providing consistent practical guidance. Conceived in this sense as a touchstone to scrutinize moral intuitions for their coherence, the trolley problem might be of relevance for the issue of ethical self-driving car algorithms in particular as well as for ethics of artificial intelligence systems in general. To say it with the words of Goodall (2016: 814): "While the trolley problem is valuable in isolating people's intuitions about morally ambiguous crash decisions and stress testing ethical strategies, it represents a fairly narrow area of automated vehicle ethics and suffers from a perceived lack of realism."

Furthermore, the analysis reveals that there is also very little common ground between decision-making in manual and autonomous driving. Both cases share a multilayer choice design and flexibility of choices. As Gogoll and Müller (2017) maintain, decisions in real-life situations are characterised by interaction between agents. Simulations of possible options for action need to take this into account. Decision makers are required to mutually respond to actions of others, whether those are other autonomous vehicles, pedestrians, car drivers, cyclists, or uninvolved parties.[8] Apart from these two aspects, there seems to be hardly any basis for drawing analogies between the cases. Instead, evidence is provided for the observation that diverging contextual factors of decision-making situations in manual and autonomous driving crucially affect the moral assessment of dilemma behaviour of driverless cars.

[8] A deeper discussion of action coordination issues in autonomous driving is beyond the scope of this paper. At this point, it should simply be remarked that it might be important to keep interaction aspects in mind with regard to implementing practicable crash algorithms.

The epistemic situations of decision makers in the two cases differ profoundly. The human driver's action is more of a reaction to given circumstances than a carefully considered decision, not least because of the short response time. It does not intend to express a preference for certain moral values in the strict sense. It is for this reason that although the human driver will be held legally responsible for her actions, she is likely not to be blamed morally. Brändle and Grunwald (2019: 286) even state that accidents involving manually driven cars are rather tragic accidents than dilemma constellations. The very opposite seems to be true for self-driving cars. Decision makers in this context can take a deliberate and intentional choice that is explicitly based on moral values and judgments. Hence, their moral decisions cannot be justified by calling upon human instinct but must be ethically grounded in an in-depth consideration of alternative options. The fact that machines cannot make intuitive decisions – as human drivers can – makes an additional point here,[9] namely that the comparison of decision situations in manual and autonomous driving is directed towards one main conclusion: overall, moral decision-making is much more complex in the context of crash algorithms than in manual driving. What makes the difference between the two cases is not only the fact that drivers become passengers, but also a completely new decision situation calling for adequate regulation that must go beyond the state-of-the-art perspective on road accidents.

Where Crash Algorithms Lie Somewhere In Between

The presented analysis additionally discloses some aspects where autonomous driving corresponds to neither of the two edge cases, but rather lies somewhere in between. Firstly, autonomous driving can draw on a large quantity of very detailed information collected at run time and during test drives. However, this data is naturally incomplete, and its quality is not perfectly reliable. On the one hand, the quality might suffer from error-prone technology of sensors and cameras, e.g. due to weather or lighting conditions (e.g. Gogoll and Müller 2017). On the other hand, the assessment, interpretation, and processing of data to make them usable for the algorithms might, for their part, be faulty. In this way, even high quality data might lead to suboptimal decisions.

Secondly, decision-making in autonomous driving lies in between the two edge cases in terms of the certainty of outcomes. It uses approaches which rely on risk estimation and calculation models of probabilities. Even though it benefits from a broad range of provided technical support, calculated results always remain uncertain to some extent (Nyholm and Smids 2016).

Thirdly, the relation to real-life situations can be interpreted as a summary of what makes autonomous driving so different from the edge cases. Decisions must be made for hypothetic, imaginable scenarios that have not come into reality yet. This tension between what *might* happen and what *really* happens makes algorithmic decision-making so challenging. These identified 'in-betweens' have an important implication. Because autonomous driving does not totally differ from the two edge cases in these aspects, we might tend to misleadingly interpret them as full analogies. Based on these wrong assumptions, expectations might be built towards the moral quality of solutions to driverless car dilemmas that will hardly be met. For this reason, we

[9] Himmelreich describes this observation as the *challenge of specificity* (Himmelreich 2018: 679).

should be careful when we invoke analogies on which to build desired ideals for autonomous cars.

Where Crash Algorithms Reach Further

Ultimately, the analysis reveals three aspects where the case of self-driving cars differs considerably from the edge cases. They point out the areas of particular regulatory need.

Firstly, decision situations in driverless car dilemmas are characterised as collective decisions. Decision-making about how to solve moral dilemmas in autonomous driving is not an individual task, but a process of considering the needs and claims of various parties. Respective decisions are delegated and free from any form of direct personal involvement due to the separateness of the decision on the one hand and its materialisation on the other. Procedures of algorithmic decision-making which are implemented into machines are ready to come into effect whenever the vehicle encounters a dilemma situation. They might potentially affect everyone. It is for this reason that authorised decision makers need to respect the interests of all parties concerning economic, ecological, social, moral, and legal aspects (e.g. Hevelke and Nida-Rümelin 2015).

Secondly, responsibility issues are particularly complex in the context of driverless cars. While most research articles deal with the question of *who* is responsible for harm caused by autonomous vehicles, this paper is more concerned with *in what way* one could be held responsible, both morally and legally. Lin (2016: 75) emphasises that the implementation of crash algorithms bears a special responsibility:

> “But the programmer and OEM do not operate under the sanctuary of reasonable instincts; they make potentially life-and-death decisions under no truly urgent time-constraint and therefore incur the responsibility of making better decisions than human drivers reacting reflexively in surprise situations.“

The German Ethics Commission points out that the assessment of an action varies with the perspective it is taken from: “It is true that a human driver would be acting unlawfully if he killed a person in an emergency to save the lives of one or more other persons, but he would not necessarily be acting culpably.” (2017: 7, Rule No. 8) The blameworthiness of a person seems to depend on what she ought to have considered *before* taking her decision. The personal involvement of a human driver into the dilemma situation influences her feelings and exposes her to pressure to take a decision (Birnbacher and Birnbacher 2016). Dilich et al. (2002: 240) argue that besides physical perception and reaction, “emergency-inflicted mental disturbances resulting from intense arousal, violated expectations and the uncertainty of handling circumstances that the driver has rarely, if ever, encountered” have an important impact on human actions in emergency situations. Human drivers need to act *ex ante* in situations that are framed by specific epistemic circumstances. Although the action might turn out to be disastrous afterwards, “such legal judgements, made in retrospect and taking special circumstances into account, cannot readily be transformed into abstract/general ex ante appraisals and thus also not into corresponding programming activities.” (Federal Ministry of Transport and Digital Infrastructure 2017: 7, Rule No. 8) Birnbacher and Birnbacher (2016) add that the German criminal law is hardly applicable to the programming of automated vehicles for another reason. It builds upon a normative distinction between acts of actively causing serious injury or death *by* a change of direction of the vehicle in order to prevent the death of uninvolved persons which are in the trajectory of the vehicle, on the one hand, and acts of causing serious injury or death *without* a change of direction, on the other hand. Transferred to the case of autonomous cars, this would imply that the entity that approves of programming a car in a way that makes a self-driving car change its direction and subsequently hurt or even kill a person, acts intentionally and is therefore to be held criminally liable. This conclusion seems problematic and reveals that attaching moral relevance to the distinction between active and passive acts of harming hardly makes sense for algorithmic decision-making. While a human driver only acts actively when taking evasive actions in the event of unavoidable collision, keeping the trajectory is considered a passive act. But a driverless car does not have a default behaviour that needs to be actively outvoted; it does not have intentions of its own. The programmer of the car's control system causes damage to affected parties actively – although indirectly – irrespective of which option is chosen; she cannot help but take an active decision. Anyway, jurisdictions are urged to elaborate a sophisticated solution that meets the complexity of personal and collective responsibility issues saturating driverless car ethics (e.g. Coca-Vila 2018; Hevelke and Nida-Rümelin 2015; Liu 2017; Loh and Loh 2017; Santoni de Sio 2017).

Thirdly, the decision of a human driver can be roughly characterised as a single reaction resulting from a spontaneous, split-second decision and consequently relates only to a single occurrence. The decisions taken in trolley cases are closely tied to the unique contexts framed by the structure of the respective version. Trolley cases are always single cases; they require individual solutions that might not be transferred to other cases. But for driverless cars, a moral solution is needed that fits not only one specific context but multiple others, too. The programming of individual solutions to make decisions context-sensitive to all facets of every imaginable scenario poses a great, if not impossible, practical challenge. It is for this reason that categorising scenarios by means of distinct criteria seems reasonable and necessary at the same time. However, real dilemma situations are always single cases and cannot be fully standardised (Federal Ministry of Transport and Digital Infrastructure 2017: 17). This tension between the unique nature of dilemmas and the practical need to generalise adequate responses seems to be one of the most complex issues related to self-driving car ethics. Determined actions are not unique but recur several times. A moral decision in the context of an autonomous driving dilemma is therefore not individual but may be described as a kind of recurring behavioural pattern. This implies that decision-making on crash algorithms has a much wider scope than the individual, narrow context given in manual driving or trolley cases.[10] The complexity of this issue is particularly apparent in dilemma situations that involve more than one driverless car. It is plausible to assume that a whole fleet of cars will use the same crash algorithms. This means that all of them will react in the same way when facing a specific scenario and thereby might provoke cumulative effects of unpredictable severity.[11] Linked to this is the fact that the standardised algorithms will suffer from biases in producing specific outcomes, for example when two identically programmed autonomous vehicles are involved in an accident (Nyholm and Smids 2016). The resulting network of cumulative effects might additionally produce inefficient outcomes. Imagine, for example, that two driverless vehicles move towards each other and then both swerve and sacrifice their passengers by driving against a wall. This scenario shows that it is indispensable to coordinate

[10] This observation largely corresponds to Himmelreich's distinction between "small-scale" and "large-scale" problems (Himmelreich 2018: 678).

[11] If the programming of moral preferences and corresponding action processes is made a centralised policy, this even implies that *all* registered autonomous vehicles will choose the very same course of action in a given scenario.

reactions among self-driving vehicles instead of blindly following pre-implemented specifications (Nyholm and Smids 2018).

Conclusion

In a nutshell, the paper argues that there are only very few touch points between dilemma decision situations related to self-driving cars on the one hand and the used edge cases of manual driving and trolley cases on the other hand. This finding has important implications: neither the practical real-life perspective of manual driving nor the theoretical concepts behind trolley cases qualify as wide-ranging landmarks when discussing ethics for crash algorithms. Instead, the presented analysis substantiates a fundamental epistemic and conceptual distance between decision situations in the context of self-driving cars and the edge cases. The discussed aspects pose challenges to the introduction of autonomous cars that are characterised by a need for regulation taking into account the specific factors that frame moral decision-making in driverless car dilemmas.

How could the humanistic paradigm contribute to the implementation of ethical crash algorithms? As an extensive discussion of this question is beyond the scope of this paper, just some tentative thoughts are sketched at this point. The central humanist notion of dignity has been conceptualised in various ways throughout the history of the humanities, the one by Immanuel Kant being among the most influential. As outlined in his *Groundwork for the Metaphysics of Morals* (Kant 1998 [1785]), he specifies dignity as the intrinsic value of what has no price and can therefore not be equivalently compensated for. The dignity of a person is subject to an ethically inadmissible act of violation if the person concerned is instrumentalised, i.e. is used merely as a means to realise (unrelated) purposes and not at the same time treated as an end in itself.[12] Applied to the context of autonomous vehicles, this interpretation of dignity becomes particularly relevant in scenarios characterised by dilemma structures. Interestingly enough, Hevelke and Nida-Rümelin (2015) argue that an approach designed to minimise sacrifices needs no longer be ruled out in principle, even from a deontological perspective. Their main argument is that harm minimisation would be in the interest of each individual when each person is not merely used as a means to achieve the minimal number of sacrifices, but also embodies the purpose of the regulation. This would exactly be the case when the identity of the victims is unspecific (as examined in section 3.2) since the risk of each individual is equally reduced. Is this a plausible assumption? To be precise, that would only be true if the algorithms are completely unbiased in terms of personal information, i.e. they would never consider one's age, sex, health condition, etc. As soon as any kind of preferences regarding target objects is involved, unknown identities can no longer be plausibly assumed. Not least because of related legal problems, the German Ethics Commission (2017: 7, Rule No. 9) declares that "any distinction based on personal

[12] The boundaries between what characterises Kantian and humanistic algorithms are blurred. A Kantian approach could be part of a humanistic algorithm, but since the latter addresses the protection of human dignity as well as the promotion of (societal) well-being, a distinctively humanistic algorithm would be one that aims at reconciling these goals likewise.

features (age, gender, physical, or mental constitution) is strictly prohibited." In any case, divergent views on the issue of harm minimisation algorithms expose a need for clarification regarding different ways of interpreting the instrumentalisation problem and evaluation of corresponding actions. As has been outlined in the introduction, humanistic management researchers have gained considerable expertise on decision situations on different levels and therefore seem to be particularly suitable to stimulate the respective debate.

Apart from that, the discussed case of driverless car dilemmas might be interpreted as having broader implications for the way artificial intelligence technology in general is perceived and assessed from an ethical perspective. As mentioned in section 2.3, there is wide agreement among machine ethics researchers that autonomous systems lack the capacity of full moral agency commonly ascribed to humans. Machines do not possess genuinely human capacities like phenomenal consciousness, emotions, higher-order reasoning, intentionality, and freedom of the will (e.g. Misselhorn 2018a). Although autonomous systems might be implemented in ways enabling them to act according to ethical principles, certain values or observed moral intuitions, they are not able to recognise persons as bearers of intrinsic values and dignity in a humanistic sense. It is for this reason that algorithmic procedures are commonly perceived as impersonal and rigid. A humanistic management perspective might be able to open up ways of integrating 'human factors' into what appears inhumanly mechanised. This way, it acts as a vital contributor to paving the way for algorithmic decision-making in general towards a meaningful and thoughtful management practice. Transferring its expertise to the upcoming issue of algorithmic moral decision-making will not only enrich the pertinent debate with a fresh angle, but also encourage moral philosophers to take a position against compromising on human dignity.

Acknowledgements I thank two anonymous reviewers for critically reading the manuscript and suggesting substantial improvements. Open Access funding enabled and organized by Projekt DEAL.

Funding This paper is related to a dissertation project which I am conducting as part of the interdisciplinary, cooperative PhD program "Ethics, Culture, and Education for the 21st Century" (financially supported by Hanns-Seidel-Foundation e.V., granted by funds from the German Federal Ministry of Education and Research) and the BayWISS Joint Academic Partnership "Mobility and Transport" (granted by funds from the Bavarian Academic Forum (BayWISS)).

Data Availability Not applicable.

Compliance with Ethical Standards

Conflict of Interest Not applicable.

Code Availability Not applicable.

References

Amann, Wolfgang, Michael Pirson, Claus Dierksmeier, Ernst von Kimakowitz, and Heiko Spitzeck, eds. 2011. *Business Schools Under Fire. Humanistic Management Education as the Way Forward. Humanism in Business Series*. London/New York: Palgrave Macmillan Publishers.

Applin, Sally. 2017. Autonomous vehicle ethics: Stock or custom? *IEEE Consumer Electronics Magazine* 6: 108–110. https://doi.org/10.1109/MCE.2017.2684917.

Asimov, Isaac. 1950. *Runaround. I, Robot (the Isaac Asimov Collection ed.)*. New York: Doubleday.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563: 59–64. https://doi.org/10.1038/s41586-018-0637-6.

Bachmann, Claudius, Laura Sasse, and Andre Habisch. 2018. Applying the practical wisdom lenses in decision-making: An integrative approach to humanistic management. *Humanistic Management Journal* 2: 125–150. https://doi.org/10.1007/s41463-017-0026-6.

Bal, Matthijs. 2017. Dignity in the workplace. *New Theoretical Perspectives*. Humanism in Business Series. London/New York: Palgrave Macmillan Publishers.

Bendel, Oliver. 2018. Überlegungen zur Disziplin der Maschinenethik. *Aus Politik und Zeitgeschichte* 68: 34–38.

Birnbacher, Dieter, and Wolfgang Birnbacher. 2016. Automatisiertes Fahren. Ethische Fragen an der Schnittstelle von Technik und Gesellschaft. *Information Philosophie* 4: 8–15.

Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. 2015. Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars? *ArXiv Preprint ArXiv* 1510 (03346): 1–15.

Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352: 1573–1576. https://doi.org/10.1126/science.aaf2654.

Brändle, Claudia, and Armin Grunwald. 2019. Autonomes Fahren aus Sicht der Maschinenethik. In *Handbuch Maschinenethik*, ed. Oliver Bendel, 281–300. Wiesbaden: Springer VS.

Bruers, Stijn, and Johan Braeckman. 2014. A review and systematization of the trolley problem. *Philosophia* 42: 251–269. https://doi.org/10.1007/s11406-013-9507-5.

Clark, Charles M.A., Alexander Buoye, Timothy Keiningham, Jay Kandampully, Mark Rosenbaum, and Anuar Juraidini. 2019. Some foundational factors for promoting human flourishing. *Humanistic Management Journal* 4: 219–233. https://doi.org/10.1007/s4146-019-00064-8.

Coca-Vila, Ivó. 2018. Self-driving cars in dilemmatic situations: An approach based on the theory of justification in criminal law. *Criminal Law and Philosophy* 12: 59–82. https://doi.org/10.1007/s11572-017-9411-3.

Contissa, Giuseppe, Francesca Lagioia, and Giovanni Sartor. 2017. The ethical knob: Ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law* 25: 365–378. https://doi.org/10.1007/s10506-017-9211-z.

Dierksmeier, Claus, Wolfgang Amann, Ernst von Kimakowitz, Heiko Spitzeck, and Michael Pirson, eds. 2011. *Humanistic Ethics in the Age of Globality. Humanism in business series*. London/New York: Palgrave Macmillan Publishers.

Dierksmeier, Claus. 2016. What is 'humanistic' about humanistic management? *Humanistic Management Journal* 1: 9–32. https://doi.org/10.1007/s41463-016-0002-6.

Dierksmeier, Claus. 2018a. Qualitative freedom and cosmopolitan responsibility. *Humanistic Management Journal* 2: 109–123. https://doi.org/10.1007/s41463-017-0029-3.

Dierksmeier, Claus. 2018b. Just HODL? On the moral claims of Bitcoin and ripple users. *Humanistic Management Journal* 3: 127–131. https://doi.org/10.1007/s41463-018-0036-z.

Dilich, Michael A., Dror Kopernik, and John Goebelbecker. 2002. Evaluating driver response to a sudden emergency: Issues of expectancy, emotional arousal and uncertainty. *SAE Transactions* 111: 238–248.

Domingos, Pedro. 2015. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. New York: Basic Books.

Etzioni, Amitai, and Oren Etzioni. 2017. Incorporating ethics into artificial intelligence. *The Journal of Ethics* 21: 403–418. https://doi.org/10.1007/s10892-017-9252-2.

Federal Ministry of Transport and Digital Infrastructure. 2017. Ethics commission "Automated and Connected Driving": *Report (extract) June 2017*. URL https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission-automated-and-connected-driving.pdf?__blob=publicationFile. Accessed 23 June 2020.

Ferguson, Ronald J., Kaspar Schattke, and Michèle Paulin. 2019. Persuasions by corporate and activist NGO strategic website communications: Impacts on perceptions of sustainability messages and Greenwashing. *Humanistic Management Journal*. https://doi.org/10.1007/s41463-019-00072-8.

Fisk, Raymond, Angie Fuessel, Christopher Laszlo, Patrick Struebi, Alessandro Valera, and Carey Weiss. 2019. Systemic social innovation: Co-creating a future where humans and all life thrive. *Humanistic Management Journal* 4: 191–214. https://doi.org/10.1007/s41463-019-00056-8.

Foot, Philippa. 1978. The problem of abortion and the doctrine of the double effect. In *Virtues and vices and other essays in moral philosophy*, ed. Philippa Foot, 19–32. Berkeley and Los Angeles: University of California Press.

Gerdes, J. Christian, and Sarah M. Thornton. 2016. Implementable ethics for autonomous vehicles. In *Autonomous driving: Technical, legal and social aspects*, eds. Markus Maurer, J. Christian Gerdes, Barbara Lenz, and Hermann Winner, 87–102. Berlin, Heidelberg: Springer Vieweg.

Glauner, Friedrich. 2019. The myth of responsibility: On changing the purpose paradigm. *Humanistic Management Journal* 4: 5–32. https://doi.org/10.1007/s41463-018-0048-8.

Gogoll, Jan, and Julian F. Müller. 2017. Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics* 23: 681–700. https://doi.org/10.1007/s11948-016-9806-x.

Gohl, Christopher. 2018. Weltethos for business: Building shared ground for a better world. *Humanistic Management Journal* 3: 161–186. https://doi.org/10.1007/s41463-018-0049-7.

Gonstead, Mariana Hernandez-Crespo, and Rachana Chhin. 2019. God's participatory vision of a global symphony: Catholic business leaders integrating talents through dispute and shared decision system design. *Humanistic Management Journal*. https://doi.org/10.1007/s41463-019-00073-7.

Goodall, Noah J. 2014. Machine ethics and automated vehicles. In *Road vehicle automation: Lecture notes in mobility*, eds. Gereon Meyer and Sven Beiker, 93–102. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05990-7_9.

Goodall, Noah J. 2016. Away from trolley problems and toward risk management. *Applied Artificial Intelligence* 30: 810–821. https://doi.org/10.1080/08839514.2016.1229922.

Goodall, Noah J. 2017. From trolleys to risk: Models for ethical autonomous driving. *American Journal of Public Health* 107: 496. https://doi.org/10.2105/AJPH.2017.303672.

Hevelke, Alexander, and Julian Nida-Rümelin. 2015. Selbstfahrende Autos und Trolley-Probleme: Zum Aufrechnen von Menschenleben im Falle unausweichlicher Unfälle. *Jahrbuch für Wissenschaft und Ethik* 19: 5–24.

Himmelreich, Johannes. 2018. Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice* 21: 669–684. https://doi.org/10.1007/s10677-018-9896-4.

Hormann, Shana. 2018. Exploring resilience: In the face of trauma. *Humanistic Management Journal* 3: 91–104. https://doi.org/10.1007/s41463-018-0035-0.

Hübner, Dietmar, and Lucie White. 2018. Crash algorithms for autonomous cars: How the trolley problem can move us beyond harm minimisation. *Ethical Theory and Moral Practice* 21: 685–698. https://doi.org/10.1007/s10677-018-9910-x.

JafariNaimi, Nassim. 2018. Our bodies in the Trolley's path, or why self-driving cars must *not* be programmed to kill. *Science, Technology, & Human Values* 43: 302–323.

Kabadayi, Sertan, Linda Alkire (née Nasr), Garrett M. Broad, Reut Livne-Tarandach, David Wasieleski, and Ann Marie Puente. 2019. Humanistic Management of Social Innovation in service (SIS): An interdisciplinary framework. *Humanistic Management Journal* 4: 159–185. https://doi.org/10.1007/s41463-019-00063-9.

Kant, Immanuel. 1998 [1785]. *Groundwork For the Metaphysics of Morals (Oxford philosophical texts)*. New York: Oxford University Press.

Keeling, Geoff. 2020. Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics* 26: 293–307. https://doi.org/10.1007/s11948-019-00096-1.

Kostera, Monika, and Michael Pirson, eds. 2017. *Dignity and the Organization. Humanism in Business Series*. London/New York: Palgrave Macmillan Publishers.

Lechner, Daniel, and Gilles Malaterre. 1991. Emergency manuever experimentation using a driving simulator. *SAE Technical Paper No. 910016*. https://doi.org/10.4271/910016.

Leisinger, Klaus M. 2018. Using the world ethos body of thought as a compass for managers some thoughts on the practical application of a philosophical concept. *Humanistic Management Journal* 3: 147–159. https://doi.org/10.1007/s41463-018-0043-0.

Lepeley, Maria-Teresa, Ernst von Kimakowitz, and Roland Bardy, eds. 2016. *Human Centered Management in Executive Education. Global Imperatives, Innovation and New Directions. Humanism in Business Series*. London/New York: Palgrave Macmillan Publishers.

Lin, Patrick. 2013. The ethics of autonomous cars. *The Atlantic*. URL http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/. Accessed 11 Mar 2020.

Lin, Patrick. 2016. Why ethics matters for autonomous cars. In *Autonomous driving: Technical, legal and social aspects*, eds. Markus Maurer, J. Christian Gerdes, Barbara Lenz, and Hermann Winner, 69–85. Berlin, Heidelberg: Springer Vieweg.

Liu, Hin-Yan. 2017. Irresponsibilities, inequalities and injustice for autonomous vehicles. *Ethics and Information Technology* 19: 193–207. https://doi.org/10.1007/s10676-017-9436-2.

Loh, Wulf, and Janina Loh. 2017. Autonomy and responsibility in hybrid systems: The example of autonomous cars. In *Robot ethics 2.0: From autonomous cars to artificial intelligence*, eds. Patrick Lin, Ryan Jenkins, and Keith Abney, 35–50. Oxford: Oxford University Press.

Lucas Jr., George R. 2013. Engineering, ethics and industry: The moral challenges of lethal autonomy. In *Killing by remote control: The ethics of an unmanned military*, ed. Bradley J. Strawser, 211–228. New York: Oxford University Press.

Luetge, Christoph. 2017. The German ethics code for automated and connected driving. *Philosophy & Technology* 30: 547–558. https://doi.org/10.1007/s13347-017-0284-0.

Lupton, Nathaniel C., and Michael Pirson, eds. 2014. *Humanistic Perspectives on International Business and Management.* Humanism in Business Series. London/New York: Palgrave Macmillan Publishers.

Melé, Domenec. 2016. Understanding humanistic management. *Humanistic Management Journal* 1: 33–55. https://doi.org/10.1007/s41463-016-0011-5.

Millar, Jason. 2017. Ethics settings for autonomous vehicles. In *Robot ethics 2.0: From autonomous cars to artificial intelligence*, eds. Patrick Lin, Ryan Jenkins, and Keith Abney, 20–34. Oxford: Oxford University Press.

Misselhorn, Catrin. 2018a. Artificial morality. Concepts, issues and challenges. *Society* 55: 161–169. https://doi.org/10.1007/s12115-018-0229-y.

Misselhorn, Catrin. 2018b. *Grundfragen der Maschinenethik*. Stuttgart: Reclam.

Moor, James H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 21: 18–21.

Nyholm, Sven, and Jilles Smids. 2016. The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice* 19: 1275–1289. https://doi.org/10.1007/s10677-016-9745-2.

Nyholm, Sven, and Jilles Smids. 2018. Automated cars meet human drivers: Responsible human-robot coordination and the ethics of mixed traffic. *Ethics and Information Technology* 9: 1–10. https://doi.org/10.1007/s10676-018-9445-9.

Pirson, Michael. 2017. *Humanistic management: Protecting dignity and promoting well-being*. Cambridge: Cambridge University Press.

Pirson, Michael. 2018. Reclaiming our humanity – a cornerstone for better management. *Humanistic Management Journal* 2: 103–107. https://doi.org/10.1007/s41463-018-0032-3.

Pirson, Michael. 2019. Managing towards a world that works for all. *Humanistic Management Journal* 4: 1–4. https://doi.org/10.1007/s41463-019-00062-w.

Pirson, Michael, and Jonathan Keir. 2018. Humanistic management: A universalist perspective based on a world ethos. *Humanistic Management Journal* 3: 141–145. https://doi.org/10.1007/s41463-018-0052-z.

Rawls, John. 2009 [1971]. *A theory of justice*. Cambridge: Harvard University Press.

Santoni de Sio, Filippo. 2017. Killing by autonomous vehicles and the legal doctrine of necessity. *Ethical Theory and Moral Practice* 20: 411–429. https://doi.org/10.1007/s10677-017-9780-7.

Sasse-Werhahn, Laura. 2019. The practical wisdom behind the GRI. *Humanistic Management Journal* 4: 71–84. https://doi.org/10.1007/s41463-019-00054-w.

Seele, Peter. 2018. Let us not forget: Crypto means secret. Cryptocurrencies as enabler of unethical and illegal business and the question of regulation. *Humanistic Management Journal* 3: 133–139. https://doi.org/10.1007/s41463-018-0038-x.

Thomson, Judith Jarvis. 1976. Killing, letting die, and the trolley problem. *The Monist* 59: 204–217. https://doi.org/10.5840/monist197659224.

Thomson, Judith Jarvis. 1985. The trolley problem. *The Yale Law Journal* 94: 1395–1415. https://doi.org/10.2307/796133.

Trappl, Robert. 2016. Ethical Systems for Self-Driving Cars. *Applied Artificial Intelligence* 30: 745–747.

Von Kimakowitz, Ernst, Michael Pirson, Heiko Spitzeck, Claus Dierksmeier, and Wolfgang Amann, eds. 2011. *Humanistic Management in Practice.* Humanism in Business Series. London/New York: Palgrave Macmillan Publishers.

Wallach, Wendell, and Colin Allen. 2009. *Moral machines: Teaching robots right from wrong*. New York: Oxford University Press.

Weber, Karsten, and Thomas Zoglauer. 2019. Maschinenethik und Technikethik. In *Handbuch Maschinenethik*, ed. Oliver Bendel, 145–163. Wiesbaden: Springer VS.

Wolkenstein, Andreas. 2018. What has the trolley dilemma ever done for us (and what will it do in the future)? On some recent debates about the ethics of self-driving cars. *Ethics and Information Technology* 20: 163–173. https://doi.org/10.1007/s10676-018-9456-6.

Wu, Stephen S. 2020. Autonomous vehicles, trolley problems, and the law. *Ethics and Information Technology* 22: 1–13. https://doi.org/10.1007/s10676-019-09506-1.