Ethical issues in focus by the autonomous vehicles industry

Martinho, Andreia; Herber, Nils; Kroesen, Maarten; Chorus, Caspar

DOI
10.1080/01441647.2020.1862355

Publication date
2021

Document Version
Final published version

Published in
Transport Reviews

Transport Reviews

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ttrv20

Ethical issues in focus by the autonomous vehicles industry

Andreia Martinho , Nils Herber , Maarten Kroesen & Caspar Chorus

To cite this article: Andreia Martinho , Nils Herber , Maarten Kroesen & Caspar Chorus (2021): Ethical issues in focus by the autonomous vehicles industry, Transport Reviews, DOI: 10.1080/01441647.2020.1862355

To link to this article: https://doi.org/10.1080/01441647.2020.1862355

Published online: 07 Jan 2021.

TRANSPORT REVIEWS
https://doi.org/10.1080/01441647.2020.1862355

Ethical issues in focus by the autonomous vehicles industry

Andreia Martinho, Nils Herber, Maarten Kroesen and Caspar Chorus

Engineering Systems & Services, Delft University of Technology, Delft, The Netherlands

ABSTRACT

The onset of autonomous driving has provided fertile ground for discussions about ethics in recent years. These discussions are heavily documented in the scientific literature and have mainly revolved around extreme traffic situations depicted as moral dilemmas, i.e. situations in which the autonomous vehicle (AV) is required to make a difficult moral choice. Quite surprisingly, little is known about the ethical issues in focus by the AV industry. General claims have been made about the struggles of companies regarding the ethical issues of AVs but these lack proper substantiation. As private companies are highly influential on the development and acceptance of AV technologies, a meaningful debate about the ethics of AVs should take into account the ethical issues prioritised by industry. In order to assess the awareness and engagement of industry on the ethics of AVs, we inspected the narratives in the official business and technical reports of companies with an AV testing permit in California. The findings of our literature and industry review suggest that: (i) given the plethora of ethical issues addressed in the reports, autonomous driving companies seem to be aware of and engaged in the ethics of autonomous driving technology; (ii) scientific literature and industry reports prioritise safety and cybersecurity; (iii) scientific and industry communities agree that AVs will not eliminate the risk of accidents; (iv) scientific literature on AV technology ethics is dominated by discussions about the trolley problem; (v) moral dilemmas resembling trolley cases are not addressed in industry reports but there are nuanced allusions that unravel underlying concerns about these extreme traffic situations; (vi) autonomous driving companies have different approaches with respect to the authority of remote operators; and (vii) companies seem invested in a lowest liability risk design strategy relying on rules and regulations, expedite investigations, and crash/collision avoidance algorithms.

ARTICLE HISTORY
Received 4 January 2020
Accepted 25 November 2020

KEYWORDS
Autonomous vehicles; ethics; moral dilemma; trolley; industry; companies

Introduction

The onset of autonomous driving has provided fertile ground for discussions about ethics in recent years. In addition to the ongoing debates regarding ethical issues particular to automated driving systems-equipped vehicles, the disruptive yet mundane nature of this

CONTACT Andreia Martinho a.m.martinho@tudelft.nl Engineering Systems & Services, Delft University of Technology, Jaffalaan 5 Room A3060, 2628 BX, Delft, The Netherlands

technology dictated its archetypal role in broader conversations about ethics and Artificial Intelligence (Etzioni & Etzioni, 2017; Hengstler, Enkel, & Duelli, 2016; Lin, Abney, & Jenkins, 2017). The vehicle at the centre of these discussions is a machine designed to provide conveyance on public streets, roads, and highways with conditional, high, or full driving automation (SAE On-Road Automated Vehicle Standards Committee, 2014). Such a machine is commonly known as Autonomous Vehicle (AV)[1] and is expected to yield a multitude of social benefits to traffic safety, mobility, and accessibility (Anderson et al., 2014; Fleetwood, 2017; Milakis, Van Arem, & Van Wee, 2017).

The ethics discussions associated with AVs which are documented in the scientific literature have mainly revolved around extreme traffic situations depicted as moral dilemmas, i.e. situations in which the AV is required to make a difficult moral choice between actions in traffic which will result in different combinations of lives saved and sacrificed (Awad et al., 2018; Bonnefon, Shariff, & Rahwan, 2016; Goodall, 2016; Lundgren, 2020). Scholars have debated exhaustively the relevance of the AV moral dilemma (Himmelreich, 2018; Keeling, 2020; Wolkenstein, 2018), the merits of using different ethical frameworks, such as Deontology, Utilitarianism, or Rawlsianism, as control algorithms for AVs (Bergmann et al., 2018; Huang, Greene, & Bazerman, 2019; Keeling, 2017; Leben, 2017; Lin, 2016), and the moral preferences and societal expectations about the ethics to be encoded in AVs (Awad et al., 2018; Bonnefon et al., 2016). Other ethical issues have been addressed in the scientific literature, such as ethical design, accountability, human meaningful control, sustainability, and privacy (Aydemir & Dalpiaz, 2018; Collingwood, 2017; Endsley, 2017; Hevelke & Nida-Rümelin, 2015; Santoni de Sio & Van den Hoven, 2018; Taiebat, Brown, Safford, Qu, & Xu, 2018).

Quite surprisingly, little is known about the ethical issues in focus by the AV industry. General claims have been made about the struggles of companies over such issues but these lack proper substantiation (Awad et al., 2018; Fagnant & Kockelman, 2015; Kirkpatrick, 2015). As private companies are highly influential on the development and acceptance of AV technologies (Van den Hoven, Vermaas, & Van de Poel, 2015), their stance on ethics should be taken into account for the purposes of a meaningful debate about the ethics of AVs.

In order to assess the ethics awareness and engagement of industry, we inspected the narratives in official business and technical reports of companies operating in the AV field. In this research, we focused on the companies with an AV testing permit in California, where there was an early adoption of comprehensive regulations governing the testing of AVs (Favarò, Eurich, & Nader, 2018; Soriano, Dougherty, Soublet, & Triepke, 2014). We believe that the analysis of reports from a wide range of technology and manufacturing companies in the forefront of AV technology allows us to draw important insights about ethics within the AV industry.

We first provide an overview of the ethics narratives both in the scientific literature and industry reports. At this point, it should be noted that the main aim of this paper is not to present an exhaustive review of the scholarly literature concerning ethical issues surrounding the development and deployment of AVs. Rather, our aim is to explore how the discussion of ethical issues in industry reports and its counterpart in the academic literature relate and compare to one another.

For reasons of brevity, we focus on the matters of *safety and cybersecurity*, *accountability*, and *human oversight, control, auditing of AVs* as presented in the scientific literature

thus raising critical yet practical questions for which we will look for answers in the industry narratives. These three issues, which we selected as our focus points, have generated a particularly rich debate in both streams of literature, and are often discussed in relation to one another. We expect that, by providing empirical insights from industry, we can make a contribution for a richer, less speculative, and more meaningful debate on the ethics of AVs.

Methodology

The ambiguous nature of ethics makes systematisation challenging. Here, we attempt to alleviate ambiguity by building our research around a list of ethical issues compiled from 22 major guidelines of AI ethics (Hagendorff, 2020). We use this list of AI ethical issues to guide us in identifying the ethics within the scientific and industry narratives.

Interestingly, Hagendorff is reluctant about the effectiveness of AI ethics guidelines. He argues that these sorts of guidelines, traditionally based on a deontological approach to ethics which relies on fixing a set of principles and maxims, should be augmented with a virtue ethics oriented approach aiming at addressing values, attitudes, and behavioural dispositions that would ultimately help professionals refraining from unethical actions (Hagendorff, 2020). We acknowledge the limitations of these deontology-based ethics guidelines in promoting a robust ethics culture within organisations. And moreover, it is noted that, because we are using a list of ethical issues based on deontological guidelines as a guidance tool in this research, our results will necessarily reflect such top-down deontological approach, thus leaving out other potential relevant ethical approaches and principles related, for instance, to informed consent and risk acceptance (Menon & Alexander, 2020).

While acknowledging these limitations, we believe this list is adequate for our research given that it includes a comprehensive and state-of-the-art compilation of ethical issues in the field of AI ethics.

The original list featuring 22 ethical issues in published guidelines about AI (Hagendorff, 2020) was adjusted for this research. We removed one ethical issue ("field-specific deliberations") as well as all "AI" references as we focus on the AV as a particular AI-powered technology. The final list of 21 ethical issues can be found in Figure 1.

We first reviewed the scientific literature, with the aim of outlining the AV ethics debates, by identifying the ethical issues prioritised by the scientific community and the main empirical findings.

For this purpose, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework (Liberati et al., 2009). Using the combination of keywords "Autonomous Vehicles" AND "Ethics"; "Autonomous Vehicles" AND "Moral"; "Self driving" AND "Ethics"; "Self driving" AND "Moral"; "Driverless" AND "Ethics"; and "Driverless" AND "Moral" in Scopus and Google Scholar (in the latter only records within the 2015–2020 timeframe in the first five pages of the database were considered), 715 records were identified. Five additional records were identified through other sources, such as citation chaining. Upon initial screening of the 720 records, 324 duplicates were removed, which meant that 396 records were assessed for eligibility. Only published scientific documents with available full text written in English language and addressing the moral or ethics dimensions of automated driving systems were considered eligible

for this study. Therefore, based on this eligibility criteria, 158 records were excluded (23 records were not published scientific documents; 47 records did not have full text available; 3 records were not written in English language; and 85 records were not about the moral or ethics dimensions of automated driving systems) thus leaving a total of 238 records for further review and analysis.

It is recalled that, as far as this literature review is concerned, our aim is quite modest as we intend solely to provide an outline of the AV ethics debates in the scientific literature. We started by reading and analysing the articles in order to identify the ethical issues prioritised by the scientific community. Thereafter, we divided the articles into theoretical (195) or empirical (43), depending on the type of research employed. And finally we reflected on the theoretical propositions and main empirical findings related to the ethical issues which are the focus points of this research.

Following the review of the scientific literature, we proceeded to the document review of AV business and technical reports, relying on the list of ethical issues mentioned above, to identify the relevant issues within the industry narratives. For the selection of companies, we used the record of 66 companies with an AV testing permit both with and without a driver in California made available by the Department of Motor Vehicles as of June 2020.

The technical and business reports from the past five years were requested from the companies and also screened through standard online searches on their websites. For reasons of reproducibility of this research we only considered reports that could be downloaded and saved as portable document format (pdf) files thus excluding articles, blog entries, or other materials made available by the companies online but which carried the risk of not being accessible in the future. In total, we used 86 documents from 29 companies.

These reports are curated documents that serve the purpose of communicating corporate information to investors, consumers, and regulatory agencies. For that reason, such reports may depict augmented or abbreviated accounts of the range of actions taken by AV companies with respect to ethics. These documents are, nevertheless, important pieces of information to learn the industry's formal stance on the complex ethical issues associated with AVs.

Initially, we proceeded with the reading of the documents and selection of statements that signalled ethical considerations in the context of AVs. Although contextual analysis is crucial for this investigation, we acknowledge the limitations of the manual approach. Therefore, on a second occasion, we relied on linguistic-based text data analytics in order to assess the validity of our initial results. We started by creating lexicons, i.e. groups of search keywords organised to investigate a concept (Schuelke-Leech, Jordan, & Barry, 2019), associated with each ethical issue. Subsequently, we applied a text mining algorithm using the previously created lexicons as regular expressions in order to locate the keywords associated with each ethical issue in the 86 documents. The output generated by this algorithm is a report stating the number of occurrences of the keywords in the lexicons associated with each ethical issue in each one of the documents issued by the AV companies. And lastly, we compared the results of the text mining algorithm and the manual approach and made the necessary adjustments with reference to the contexts of the narratives. For the quantitative analysis of the results, we did a standard descriptive statistical analysis of the ethical issues found in the AV industry reports[2].

Overview of the ethics of AV technology in scientific literature

The amount of attention that AV technology ethics has received in recent years is quite new to the field of Transportation. Traditionally, the ethics debates in this field have revolved around less sensational issues, such as cost–benefit analysis of transport projects or fairness in pricing (Van Wee, 2011). The advent of autonomous driving is a remarkable scientific and engineering achievement that has given rise to novel and controversial ethical issues.

Our review showed quite clearly that the scientific literature on AV ethics is dominated by considerations about *safety and cybersecurity* concerning the programming of extreme traffic situations. This controversial issue is commonly known as the *trolley problem* in reference to a thought experiment popularised by Philippa Foot in 1967 in which an agent needs to make a difficult choice of allowing a runway trolley to proceed its course and kill five track workers or divert the trolley from its course killing only one worker (Foot, 1967; Thomson, 1984). There are many variations and extensions to this thought experiment but its core can be defined as a moral choice between actions in traffic which will result in different combinations of lives saved and sacrificed. Because extreme traffic situations need to be programmed in advance, AV technology seemed to bring this textbook thought experiment to life thus capturing the attention of scholars and the media.

We found references to the trolley problem in more than half of the 238 reviewed articles. Most of these articles are theoretical pieces of research, often written as argumentative or normative essays, about different perspectives and dimensions of the AV moral dilemma. While these debates are certainly very rich, we found this stream of the literature to be quite fragmented. For instance, there is still little consensus about the relevance of the trolley problem in the context of AVs (Bonnefon, Shariff, & Rahwan, 2019; Cunneen et al., 2020; Himmelreich, 2018; Keeling, 2020; Wolkenstein, 2018).

The empirical findings reported in the literature are also quite controversial, as they reveal potential challenges in adapting societal expectations to moral decision-making driving algorithms (Kallioinen et al., 2019). The AV social dilemma, i.e. a conflict between individual and collective interest in the context of autonomous driving technology, illustrates such challenge. It has been reported that people approve and would like others to buy utilitarian AVs which sacrifice their passengers for the greater good, yet prefer to ride in AVs that protect their passengers at all costs thus disapproving utilitarian regulation of AVs (Bonnefon et al., 2016; Huang et al., 2019; Morita & Managi, 2020).

A substantial amount of research, namely the Moral Machine Experiment (MME) (Awad et al., 2018), has focused on collecting and analysing moral preferences and societal expectations about the ethics to be encoded in AVs. However, the methodological soundness and value of such investigations for the purpose of defining moral algorithms for AVs have been questioned (Bigman & Gray, 2020; Harris, 2020). The main contribution of the MME, regarding moral preferences in AV moral dilemmas, conflicts with current ethical guidelines, such as rule 9 of the German Ethics Code for Automated and Connected Driving, which prohibits distinctions based on personal features in the case of unavoidable accident situations (Harris, 2020; Kallioinen et al., 2019; Luetge, 2017). The proponents of the MME acknowledge that AV policy should not necessarily follow public expectations and preferences but they believe that such preferences should not be

completely dismissed. They argue that, given the strong preference for sparing children, it would be challenging to explain the rationale for not assigning a special status to children (Awad et al., 2018). Recently, it has been hypothesised that AV fatalities carry more weight because those are rare events and not so much due to intrinsic differences in public perception between AV and conventional vehicles fatalities (Huang, van Cranenburgh, & Chorus, 2020).

Several scholars have reported an overstatement of the *AV trolley problem* and called for the ethics community to focus on other ethical issues associated with AVs (Goodall, 2019; Himmelreich, 2018; Keeling, Evans, Thornton, Mecacci, & de Sio, 2019).

Other issues debated in the literature include ethical design, accountability, human meaningful control, sustainability, and privacy (Aydemir & Dalpiaz, 2018; Collingwood, 2017; Endsley, 2017; Hevelke & Nida-Rümelin, 2015; Santoni de Sio & Van den Hoven, 2018; Taiebat et al., 2018). Particularly, matters related to accountability and human meaningful control have received considerable attention in the literature recently and, along with safety and cybersecurity, will be further explored later in this research.

Overview of the ethics of AV technology in AV industry reports

In this research, we focus on the AV industry in California, a State that has been an early and strong proponent of this technology and hosts many R&D programs (Brown et al., 2018). As of June 2020, the California DMV had listed 66 permit holders for testing with a driver, 2 permit holders for driverless testing, and 0 permit holders for AV deployment.

For our analysis, we used a total of 86 documents issued by 29 companies[3] in the forefront of AV technology. Therefore, we consider their official reports as important pieces of information about the industry's formal stance on the complex ethical issues associated with AVs.

Overall, the AV reports reviewed in this research show a overwhelmingly positive tone about AV technology, which needs to be interpreted in light of such reports being written for a particular audience of investors, consumers, and regulatory agencies. Although lacking the rigour and depth of the narratives in the scientific literature, a plethora of ethical issues are referenced in the AV industry reports.

A quantitative synopsis of the ethics prioritised by companies with an AV testing permit in California, based on the number of companies that addressed each one of the 21 ethical issues in their reports, can be found in Figure 1. It shows that *safety and cybersecurity* is the ethical issue that was addressed by more companies, followed by *common good, sustainability, well-being*; *human oversight, control, auditing*; and *science-policy link*. In contrast, issues such as *hidden costs*, *fairness, non-discrimination, justice*, or *responsible research funding* were not addressed in these reports.

The relevance of particular ethical issues within industry narratives requires a balance between frequency and comprehensiveness. We recall that three ethical issues had previously been selected from the scientific literature to be further explored in this research (*safety and cybersecurity*, *accountability*, and *human oversight, control, auditing*). Safety and human oversight issues are frequently addressed by AV companies and, whereas accountability is addressed by a lower number of companies, it is comprehensively explored in the reports we analysed.

It should be noted that, even when we found quite comprehensive accounts on particular ethical issues, the narratives in the AV reports are consistently pragmatic and

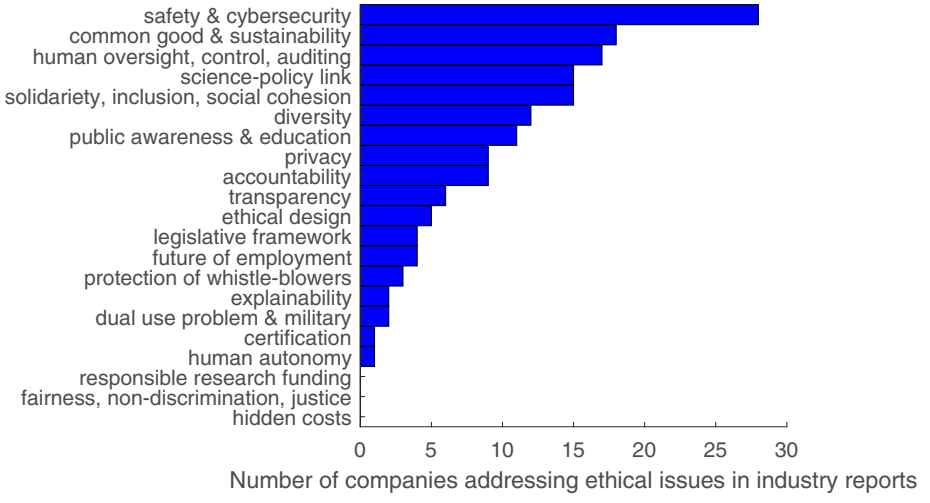

Figure 1. Ethical issues in AV industry.

oriented towards technical solutions. For instance, in its reports, Mercedes-Benz acknowledges the importance of data privacy while emphasising practical data-protection-friendly solutions that provide privacy by design in compliance with privacy laws (Daimler Sustainability Report 2018 and Reinventing Safety: A Joint Approach to Automated Driving Systems).

Safety and cybersecurity

Safety and security are both related to the integrity of systems but, whereas *safety* concerns the adequate functioning of a system, *security* is about the ability of a system to resist intentionally malicious actions. There are unsettled considerations about the acceptable safety and cybersecurity levels of AVs, both in mundane and extreme situations, in order to secure the well-being of users and other traffic agents (Kalra & Paddock, 2016; Parkinson, Ward, Wilson, & Miller, 2017; Sparrow & Howard, 2017).

Safety and cybersecurity of AVs in the scientific literature

Mundane and extreme traffic situations

Mundane traffic situations are the day-to-day interactions of traffic agents (pedestrians, cyclists, animals) that require some flexibility, such as crossroads, highway entrances, or crosswalks with limited visibility. These interactions are challenging for AVs not only because these systems lack human intuition and flexibility but also because of the large scale fleet programming that is needed (Borenstein, Herkert, & Miller, 2019; Himmelreich, 2018). Coordination over different technical approaches to these traffic situations is crucial to ensure safety but it is unclear how such technological coordination can be facilitated in the competitive AV market (Himmelreich, 2018).

Extreme traffic situations are unexpected occurrences in the traffic environment which entail danger for vehicle occupants and other traffic agents (e.g. the unexpected

appearance of an animal on a highway). Some of these situations are depicted in the scientific literature as moral dilemmas. As mentioned above, these difficult moral situations are highly explored in the scientific literature with reference to the *trolley problem* thought experiment (Foot, 1967; Thomson, 1984).

The underlying argument in the debates that take place in the scientific literature about the *AV trolley problem* can be outlined as follows: (i) *AVs ought to save lives*. (ii) *However, upon deployment of AVs, extreme traffic situations will not be completely avoided*. (iii) *Some of these extreme traffic situations will require AVs to make difficult moral decisions*. (iv) *Difficult moral decisions in traffic resemble the trolley problem*. (v) *The best option to assist AVs in managing the AV trolley problem is* x. (vi) *Option* x *is programmable*. (vii) *Therefore AVs should be programmed with* x. The disputes in the literature about the AV trolley problem are mainly related to premises (iv) regarding the relevance of the trolley problem in the AV context and (v) regarding the merits of different approaches to assist AVs in moral decision-making in extreme traffic situations.

Relevance of the trolley problem

The relevance of the trolley problem in the AV context concerns its value as a model to investigate a relevant AV ethical challenge. It has been acknowledged that using trolley cases as inputs for crash optimisation algorithms invites a myriad of criticism (Keeling, 2020; Nyholm & Smids, 2016). Scholars have argued that trolley cases are of limited usefulness for the ethics of AVs because such cases would not only be highly improbable occurrences, but also their assumptions are unrealistic (outcomes of the different moral decisions available to the agent are known rather than probabilistic), inconsistent (agent has control over a vehicle yet a collision is imminent and unavoidable), and limited with respect to design (trolley cases assume a top-down approach in which an agent makes a decision explicitly, thus failing to encompass different design approaches to decision-making) (Fried, 2012; Himmelreich, 2018).

It is accepted that trolley cases are dramatic, stylised, black-and-white situations that have little resemblance to real-life extreme traffic situations. However, it is also widely acknowledged that AVs will not eliminate crashes (Bagloee, Tavana, Asadi, & Oliver, 2016; Bonnefon et al., 2019; Favarò, Nader, Eurich, Tripp, & Varadaraju, 2017; Fleetwood, 2017). Therefore, weak trolley cases seem to be plausible. An example of a weak trolley case from the literature is an AV which is travelling across a two-lane bridge when a bus in the other lane swerves into its lane and the AV needs to decide either to brake, which would result in a collision with the bus, or to swerve into the other lane, thus hitting the side of the bridge (Goodall, 2014a). Such extreme traffic situations, entailing decisions about who is put at marginally more risk of being sacrificed, may be rare occurrences when AVs are deployed but they need to be addressed (Bonnefon et al., 2019).

Recently, it has been proposed that the relevance of the trolley cases in the AV context is associated with the prospect of development of novel ethical principles. These principles, formulated upon analyses of the moral intuitions that emerge in stylised cases, would ultimately guide the AV design process (Keeling, 2020; Wu, 2019).

Approaches to assist AVs in extreme traffic situations

Another debate in the scientific literature concerns the conflicting approaches that have been advanced by scholars to address extreme traffic situations. We recall premise (v)

above, *The best option to assist AVs in managing such extreme traffic situations is* x, to clarify that in the AV ethics literature x tends to be proposed within the realm of Machine Ethics.

Scholars have debated the merits of using ethical frameworks such as Rawlsianism, Deontology, or Utilitarianism as the control algorithms of AVs (Bergmann et al., 2018; Huang et al., 2019; Keeling, 2017; Leben, 2017; Lin, 2016; Thornton, Pan, Erlien, & Gerdes, 2017) or aggregating societal moral preferences (Awad et al., 2018; Etienne, 2020; Harris, 2020; Noothigattu et al., 2018) to encode ethics in AVs, thus assisting them in navigating extreme traffic situations that would require moral choices.

It was shown earlier that safety considerations are central both in scientific literature and industry reports. However, scholars have mostly debated the ethics in extreme traffic situations with reference to trolley cases. We further investigate the industry's approach to extreme traffic situations and raise two relevant questions: (i) Are extreme traffic situations resembling trolley cases addressed by industry? and (ii) What are the solutions proposed by industry to address extreme traffic situations?

Safety and cybersecurity in the AV industry reports

Safety and trust

Considering that the commercial success of AV technology depends greatly on the trust of consumers, it is hardly surprising that the industry narratives focus mainly on safety issues. In *A Matter of Trust Ford's Approach to Developing Self-Driving Vehicles*, it is stated that *for autonomous vehicles to be accepted by the public it needs to be established that they can be trusted* (Marakby, 2018) and in Intel's white paper *A Matter of Trust: How Smart Design Can Accelerate Automated Vehicle Adoption*, trust is also emphasised when it is stated that *before driverless AVs can be widely accepted, people must be willing to trust them with their lives and the lives of those they care about hence AVs must behave, react, and communicate in ways that make it easy for people to trust them – not only the passengers inside, but also pedestrians and the other drivers who encounter them on the road* (Weast, Yurdana, & Jordan, 2016).

Trust and business-related considerations may not be the only reasons for the prevalence of safety considerations in the AV reports. Autonomous driving is a complex and disruptive technology which is expected to have a major societal impact. Unlike other social and ethical issues, such as fairness or human autonomy in an AI-dominated society, safety challenges are prone to be solved by technical or engineering approaches (Hagendorff, 2020). Therefore, AV companies tend to prioritise these issues, for which technical solutions are presented.

In the reports reviewed in this research, we found extensive safety considerations both for mundane and extreme traffic situations. In order for AVs to successfully deal with mundane traffic situations, companies propose advanced sensing and AI-powered solutions. Mercedes-Benz and Bosch designed a Object and Event Detection and Response (OEDR) system for AVs which is based on sensors, actuators, and computing resources that is expected to assist the AV in handling these traffic situations - *Reinventing Safety: A Joint Approach to Automated Driving Systems* (Daimler, 2018), whereas Valeo proposes an AI-based approach, building on the thought that in order to negotiate complex traffic conditions where there are many unknowns, AVs need to learn the data – *Meet the Future 2016 Activity and Sustainable Development Report* (Valeo, 2016).

Extreme traffic situations: crashworthiness, collisions, and moral dilemmas

Safety considerations with respect to extreme traffic situations are also explored in the industry reports. Companies focus on the crashworthiness of AV technology, which is quite relevant for our investigation of the AV moral dilemma as, at its core, the trolley case – either in its weak or strong version – is a convoluted crash optimisation problem. By inspecting the industry reports regarding crashworthiness, we expect to clarify some of the critical elements of the AV moral dilemma.

The first element concerns the risk of crashing. Indeed the AV moral dilemma could be promptly dismissed on the account that autonomous driving will eliminate crashes. While companies express their vision of a future without accidents (*Advanced Driver Assistance Systems continues to evolve in order to realise autonomous driving and zero-accident smart vehicles, the essence of the fourth industrial revolution – Mando Sustainability Report* Mando, 2018), such ambition is mitigated by the plethora of statements on the inevitability of AV crashes and collisions which leave no room to entertain the thought of a complete elimination of accidents (*driving environments can be extremely complex and difficult and no automated driving system – regardless of how capable it may be – is likely to prevent crashes entirely – Automated Driving at Toyota: Vision, Strategy and Development* (Toyota, n.d.); *While our top priority is to avoid collisions, we recognise it is possible that we could be involved in a collision at some point – Delivering Safety: Nuro's Approach* Nuro, 2018).

Accepting that the future will not be crash and collision-free leads us to further considerations about the AV moral dilemma. Indeed crashes and collisions are a necessary condition for such extreme situations. In our document review, we did not find any reference to trolley cases as described in the scientific literature, i.e. situations that require the AV to make difficult moral choices (Bonnefon et al., 2016; Goodall, 2014b), but we identified nuanced allusions to this matter.

Companies acknowledge that AVs will face rare extreme traffic situations, often mentioned in the industry reports as *edge cases*, and emphasise simulation and validation methods used to test these scenarios (*we test and validate our self-driving vehicles in the wide variety of environmental conditions that the vehicle might face in its operational design domain – from driving scenarios the vehicle would face daily to the rare edge cases – General Motors Self-Driving Safety Report* (GM, 2018); *AI-powered autonomous vehicles must be able to respond properly to the incredibly diverse situations they could experience, such as emergency vehicles, pedestrians, animals, and a virtually infinite number of other obstacles – including scenarios that are too dangerous to test in the real world – Nvidia Self-Driving Safety Report* (Nvidia, 2018); *decision making is one of the most challenging tasks in the A.I. development of an autonomous vehicle … there are infinite edge cases that may be difficult or dangerous to reproduce in reality, such as illegal driving behaviours or sudden traffic accidents – The Autox Safety Factor* AutoX, 2018).

We found one statement that somewhat resembles the AV moral dilemma, with one important caveat regarding the nature of harms at stake. In the Nuro report *Delivering Safety: Nuro's Approach*, it is stated that, in the *unlikely case of a Nuro shuttle ever encountering an unavoidable collision scenario the driverless passengerless vehicle has the unique opportunity to prioritise the safety of humans, other road users, and occupied vehicles over its contents* (Nuro, 2018). Whereas we can not legitimately consider Nuro's account as an AV moral dilemma, we consider it as yet another indication that companies are

aware of convoluted situations akin to weak versions of moral dilemmas. We speculate that Nuro's slightly more transparent stance on this matter could be explained by the fact that it focuses on passengerless self-driving delivery technology (*Our custom vehicle is engineered to make delivery of everything more accessible – from groceries to pet food, prescription drugs to dry cleaning … with no driver or passengers to worry about, our vehicle can be built to keep what's outside even safer than what's inside … it's lighter, nimbler, and slower than a passenger car, and is equipped with state-of-the-art software and sensing capabilities that never get distracted* Nuro, 2018).

At the root of extreme or edge cases are often blind spots that prevent the AV from performing an accurate evaluation of the traffic context and having enough emergency braking time. The solutions advanced by companies to address this problem rely on radars and speed limitation when the visual field of the AV is obstructed. Although blind spot detection and assistance is considered a low level automation feature, the narratives we found about this issue and its implications for pedestrians' safety are yet another substantiation of the concerns of AV companies about extreme traffic situations (*in the case of pedestrians who are occluded from the vehicle … it should adjust the speed such that if a child would emerge from behind some object there would be no accident … even in a worst case scenario where the pedestrian emerges from behind some sensing obstruction (e.g. a parked car) even at that maximal speed - Intel Implementing the RSS Model on NHTSA Pre-Crash Scenarios*[4] (Mobileye, n.d.); *If the view is blocked Perception will flag that area as unknown … if an object is hard to see because of rain or fog or because it is hidden behind a truck the computer brain knows that and adjusts its decision-making and performance accordingly … this allows prudent decision-making and operation based upon both what the sensors "see" as well as what may be hidden from view* GM, 2018).

As a result of our review of AV reports, we conclude that moral dilemmas resembling trolley cases are not addressed in these reports in the terms described in scientific and media publications, but there are nuanced allusions that unravel underlying concerns about these extreme traffic situations. Regarding the solutions proposed by industry to address extreme traffic situations, we report for now that companies rely on radars and speed limitation to address the problem of blind spots which are often at the root of these traffic situations.

Human oversight, control, auditing of AVs

Human oversight, control, and auditing of autonomous systems imply the surveillance of the development and performance of the technology. It is expected that remote oversight of the performance of autonomous driving ensures trust and safety in this technology as human operators are able to take-over the vehicle. It has been reported, however, that several technical problems take place precisely during the transfer of control over the AV (Heikoop et al., 2019).

Human oversight, control, auditing of AVs in the scientific literature

A philosophical account of meaningful human control over automated systems has been proposed by Santoni de Sio and Van den Hoven to ground the design guidelines with

respect to human oversight, control, and auditing of AV technology (Fischer & Ravizza, 2000; Santoni de Sio & Van den Hoven, 2018). According to this account, AVs should meet tracking and tracing conditions to allow for a meaningful form of human control (Santoni de Sio & Van den Hoven, 2018). An AV should therefore be able to track the relevant human moral reasons in a sufficient number of occasions, thus adjusting its behaviour in accordance to the intentions of a human designer or operator – tracking condition – and its actions should be traceable to a proper moral understanding on the part of the humans who design and deploy the system – tracing condition (Santoni de Sio & Van den Hoven, 2018).

The tracking condition ensures that the AV complies with the intentions of a human operator. It should be noted, however, that humans are poor fallback systems. As more autonomy is added to a system and it becomes more reliable and robust, the situation awareness of human operators decreases and they are less likely to take over manual control (*automation conundrum*) (Endsley, 2017). Therefore, in critical situations, it could be that, by meeting the tracking condition, the AV is complying with an instruction issued by a low situation awareness operator.

The tracing condition requires the presence of at least one human agent that can understand the real capabilities of the system and bear the moral consequences of the actions of the system (Santoni de Sio & Van den Hoven, 2018). This condition is especially relevant to tackle the responsibility gaps, i.e. situations where it is unclear who should be responsible for an outcome (Matthias, 2004; Nyholm, 2018), that are expected to arise in the context of AV technology as a result of the fragmentation of the technology action (*many-hands* problem Van de Poel, Fahlquist, Doorn, Zwart, & Royakkers, 2012).

Meaningful human control has been heralded as the standard for AVs to meet the appropriate level of safety and accountability (Keeling et al., 2019). We will revisit this theory in the section below about the accountability of AVs. From this section, we raise one relevant question, regarding the tracking condition, to be investigated in the industry reports: according to the autonomous driving industry, which decision prevails in traffic, the decision of the AV or the decision of the human operator?

Human oversight, control, auditing in the AV industry reports

Remote and onsite human oversight of AV operations is addressed in the industry reports analysed in this research. We recall that few companies have a driverless testing permit in California, which means that in general companies rely heavily on onsite human oversight for the testing of AVs. *Mission Specialists are trained on the governing operational design domain, and are prepared to take manual control of the vehicle when presented with a scenario that is not included in the current operational design domain – Uber Advanced Technologies Group A Principled Approach To Safety* (Uber, 2018). In addition to onsite oversight, companies also rely on the remote control of AV operations. In a report issued by Zoox, it is stated that their *remote operations support centre will have operators available to remotely guide vehicles at any time, day or night, when a vehicle encounters an uncertain driving situation such as a traffic light outage or a road obstruction – Safety Innovation at Zoox: Setting the bar for safety in autonomous mobility* (Zoox, 2018).

Building on the notion of *tracking*, introduced above in the context of Meaningful Human Control, we report different approaches with respect to the authority of remote

operators. Companies such as Mercedes-Benz and Intel seem to prioritise the autonomy of the vehicle (*while automated driving vehicles take under consideration data received from an infrastructure, particularly data that can be strongly authenticated and validated, the vehicles ultimately maintain their own decision authority, not the infrastructure – Safety First of Automated Driving 2019* Wood et al., 2019) whereas other companies, such as AutoX, seem to prioritise the decisions made by remote operators (*operators at the remote support system can check the AI decision results and correct or overwrite them when unexpected errors occur* AutoX, 2018).

The statements that we identified in the AV reports regarding human oversight relate more to the first condition of the Human Meaningful Control theory, but we also report one statement which relates to the tracing condition with respect to the understanding of the system. AImotive states that *test operators face their own unique challenges. The debug screen of a complex autonomous system is incomprehensible to the untrained eye. These engineers and developers have a deep understanding of the code at work in our prototypes allowing them, at times, to predict when the system may fail. This allows our test crews to retake control of the vehicle preemptively, in a controlled manner – Ensuring Safe Self-Driving AImotive's Development Puts Safety First* (Csizmadia, 2018). The tracing condition in the Meaningful Human Control theory has another dimension, related to responsibility, which will be addressed below.

Accountability

Accountability issues associated with AV technology have received substantial attention in the scientific literature. We refer to accountability in broad terms, thus encompassing closely related concepts, such as responsibility and liability. It is clarified that accountability entails responsibility, but unlike the latter it requires explanations about actions and it cannot be shared (Mulgan, 2000); responsibility for an action traditionally requires at least a *control condition*, i.e. an agent is responsible if it is the agent of the action, and an *epistemic condition*, i.e. awareness or knowledge of the agent regarding the action (Coeckelbergh, 2020); and liability is legal or financial responsibility (Collingwood, 2017). These matters are challenging in the AV domain, mainly because of the fragmentation of the technology action, which can result in responsibility gaps.

Accountability in the scientific literature

Different approaches to AV responsibility have been proposed in the literature (Borenstein, Herkert, & Miller, 2017; Coeckelbergh, 2016; Misselhorn, 2015; Nyholm, 2018). The theory of Human Meaningful Control, which was introduced above, encompasses a tracing condition that requires the presence of at least one human agent who can bear the moral consequences of the actions of the AV (Santoni de Sio & Van den Hoven, 2018). It has been asserted that, in order for the tracing condition to be met in higher-order levels of automation, a transition of responsibility from the driver to designers or remote operators is required. At such levels of automation, how the AV is designed to execute its tasks is more important than how the human driver ought to execute its tasks (Heikoop et al., 2019).

An argument has been presented in the scientific literature particularly concerning liability, in which it is roughly stated that AVs have the potential to save lives but crushing

liability may discourage manufacturers from developing and deploying AVs, and as such this technology would not meet its potential to save lives (Hevelke & Nida-Rümelin, 2015; Marchant & Lindor, 2012). As legal scholars are working on extensions to criminal and civil law (Funkhouser, 2013; Gurney, 2013, 2015), it is questioned whether liability legal frameworks should be designed in such a way that would not impede, but rather promote, the development and improvement of AVs (Hevelke & Nida-Rümelin, 2015). This argument has been undermined by some scholars who claim that increased manufacturer liability will not be problematic, as AVs will be safer and will bring down the overall cost of litigation and insurance (Garza, 2011). Indeed, it has been reported that thus far governments have avoided strict measures in order to promote AV developments (Taeihagh & Lim, 2019).

Another issue that has been presented in the literature is the *liability dilemma of the AV manufacturer* which showcases the conflict between ethics and law when it comes to liability. When designing a crash collision algorithm, a manufacturer is assumed to face three options while balancing ethics and liability: (i) program an algorithm to swerve in a direction that would sacrifice fewer lives but would entail high liability due to compensatory and punitive damages for intentional conduct caused by targeting the sacrificed people; (ii) allow the AV to run its course which would entail a larger number of lives sacrificed but lower liability which would then be restricted to compensatory damages; and (iii) avoid a collision, which if successful, entails that no lives are sacrificed but if unsuccessful, entails the largest number of lives sacrificed, but in either case entails the lowest liability (Wu, 2019). It is therefore concluded that what is easier in a lawsuit may not be the more ethical solution (Wu, 2019). By featuring a trolley case in the background and making simplistic and general assumptions about the law, the AV liability dilemma suffers from the same sort of shortcomings that have been pointed out earlier about the AV moral dilemma. Despite its limitations, the liability dilemma of the AV manufacturer sheds light on the tension between ethics and liability, which should not be ignored.

The three issues we explored above regarding the transition of responsibility in higher order levels of automation, the liability and technology development argument, and the liability dilemma can be further investigated within the industry narratives by considering the design strategies with respect to accountability. From this section we raise the question: which accountability design strategy is being adopted by the AV industry?

Accountability in the AV industry reports: the case of the super-humanly fast runner

We found several statements in the industry reports that allow us to further reflect on the approach of the industry with respect to accountability. In general, AV companies seem invested in a lowest liability risk design strategy relying on rules and regulations, expedite investigations, and crash/collision avoidance algorithms.

Companies have stated the need for clear rules to be set in advance (*It is necessary to put legal frameworks in place in order to clarify where the responsibility lies in case of the occurrence of an accident after the realisation of fully automated driving – Honda Sustainability Report 2015* Honda, 2015), while also presenting technical solutions aimed at expediting investigations, such as the use of a "black-box" akin to a flight recorder designed to store accident data, or responsibility algorithms based on mathematical models (*With regard*

to liability issues a "black-box" is required that stores certain data necessary to sort out accident liability which can contribute towards allocating responsibility between human and machine when accidents happen – BMW Sustainable Value Report 2016 (BMW, 2016); *What will happen when a collision occurs? There will be an investigation, which could take months … our solution is to set clear rules for fault in advance, based on a mathematical model … the investigation can be very short and based on facts, and responsibility can be determined conclusively. This will bolster public confidence in AVs when such incidents inevitably occur and clarify liability risks for consumers and the automotive and insurance industries – Intel A Plan to Develop Autonomous Vehicles. And Prove it.* Shashua & Shalev-Shwartz, 2017).

It was mentioned earlier that companies acknowledge that AV technology will not eliminate accidents, yet some companies are invested in developing AVs which will never cause or be responsible for accidents (*By formally defining the parameters of the dangerous situation and proper response, we can say that responsibility is assigned to the party who did not comply with the proper response. Therefore, the Responsibility-Sensitive-Safety model guarantees that when applying it to any "driving policy" (the decision-making mechanism of the AV), the self-driving car will never initiate a dangerous situation and thus, it will never cause an accident* (Mobileye, n.d.); *Over time, though, Guardian capability will grow steadily as technology improves, with a goal of creating a vehicle never responsible for a crash regardless of errors made by a human driver* Toyota, n.d.). In one of the Intel reports (*Intel A Plan to Develop Autonomous Vehicles. And Prove it.*) it is stated that their Responsibility-Sensitive-Safety system will always brake in time to avoid a collision with a pedestrian unless the pedestrian is running super-humanly fast (Shashua & Shalev-Shwartz, 2017). By providing the super-humanly fast runner illustration, Intel is not only emphasising that their AV will not be responsible for a collision with a pedestrian, but it is also promoting trust in their technology.

Notwithstanding the positive accounts found in the industry reports regarding the development of minimally responsible AV technology, we found a statement in Nissan's Financial Information 2018 bracing the company for potential liability losses related to AVs. It is stated that *If the autonomous driving technology is developed and its use becomes quickly widespread in the future, the responsibility of automobile manufacturers might be brought into question in connection with the decline in drivers engaged in driving … If the recalls that the Group has implemented for the benefit of customers' safety become significant in volume and amount, the Group would not only incur significant additional expenses but also experience damage to its brand image, which could adversely affect its financial position and business performance* (Nissan, 2017).

Conclusion

Despite the wealth of discussions about the ethics of AVs, little is known about the awareness and engagement of the industry on this matter. In this research we have provided an overview of the narratives on the ethics of AVs as presented both in scientific literature and in industry reports issued by companies with an AV testing permit in California. Subsequently, we focused on *safety*, *accountability*, and *human oversight*, and we raised critical yet practical questions, for which we looked for answers in the industry narratives. A combination of contextual analysis and text mining techniques was employed to select statements signalling AV-related ethical considerations within the industry reports.

The overall conclusion that can be drawn from our analyses is that industry and academia look at the ethics of AV technology through rather different lenses. For example, while the scientific literature has been largely preoccupied with deep considerations of abstract moral dilemmas (trolley problem), industry reports adopt a much more pragmatic, technology-infused and perhaps overly optimistic narrative when discussing the potential of so-called edge cases where accidents cannot be avoided and loss of life and damage need to be minimised. While this discrepancy may perhaps not come as a surprise to many, it is disappointing to see that on matters that are of such great importance to the general public, science and industry seem to diverge so profoundly. While we certainly do not advise to try and establish some form of agreement between industry's views regarding the ethical issues surrounding AVs and those of academia (which would be a tall order anyway, given the wide variety of such views within industry and within academia), we do believe that it would be valuable to both sides of the aisle to inform one another of one's viewpoints.

More specifically, the findings in this research suggest that: (i) given the plethora of ethical issues addressed in the reports, autonomous driving companies seem to be aware of and engaged in the ethics of autonomous driving technology; (ii) scientific literature and industry reports prioritise safety and cybersecurity; (iii) scientific and industry communities agree that AVs will not eliminate the risk of accidents; (iv) scientific literature on AV technology ethics is dominated by discussions about the trolley problem; (v) moral dilemmas resembling trolley cases are not addressed in industry reports but there are nuanced allusions that unravel underlying concerns about these extreme traffic situations; (vi) autonomous driving companies have different approaches with respect to the authority of remote operators; and (vii) companies seem invested in a lowest liability risk design strategy relying on rules and regulations, expedite investigations, and crash/collision avoidance algorithms.

Clearly, our study has its limitations, which we would like to highlight once more at this point. Despite our efforts to alleviate ambiguity surrounding terms such as accountability, we could not successfully remove it entirely from this research. This is unavoidable as academic ethicists amongst themselves have rather diverging views on what a term like accountability means, and how it should be distinguished from related concepts such as responsibility. It should therefore not come as a surprise that this ambiguity at an abstract level may translate into different (implicit) meanings attached to the same word, in different industry reports. As a consequence, our analysis of these reports which uses a combination of "manual reading" and text mining, risks conflating different meanings attached to the same vocabulary. One promising way to alleviate or at least diminish this problem is to use techniques that are popular in the field of Anthropology, such as participant observation, in-depth interviews and focus groups. These techniques offer a potential window into how particular terminology is being used in the AV industry, as such providing a base for more carefully discussing how different industry actors differ from one another in terms of their approach to, e.g. accountability in the context of AVs and how industry as a whole differs from academia in this regard.

Such techniques could also help remedy a second limitation of our study, which is that we focused on curated reports that were made publicly available by industry actors for a particular audience. Although, as we argued above, we believe that such

documents hold important clues regarding the views of industry actors – e.g. in providing insight into how they like to be seen by others – there is clearly scope and need for more and other types of data collection here. For example, participant observation in which a scholar would be allowed to be embedded in an AV-company for a longer period of time, and to do a range of in-depth interviews with employees at various levels of the organisation, is likely to add significantly to our knowledge of industry's dealings with the ethical conundrums that surround the development and deployment of AVs. We trust that our study would provide a useful stepping stone for such follow up research.

Notes

1. Different nomenclatures are used for highly automated vehicles such as *autonomous vehicles*, *automated vehicles*, *self-driving cars*, or *driverless cars* (Gandia et al., 2019). Here we adopt *autonomous vehicles* when referring to automated driving systems-equipped vehicles (levels 3, 4, or 5 driving automation systems according to the Society of Automotive Engineers International Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles Committee, 2014) for reasons of consistency with the nomenclature favoured by the industry. In this context *autonomy* is associated with the ability of a vehicle to determine its operational environment, thus modulating its behaviour according to relevant norms, needs or constraints (Danks & London, 2017).
2. Information regarding the companies with a testing permit in California, the reports used in this study, and the lexicons is available in the dataset stored in the 4TU. Center for Research Data in doi:10.4121/13348535 (Martins Martinho Bessa, Chorus, Kroesen, & Herber, 2020).
3. Waymo LLC; Tesla Motors; Nissan; BMW; Ford; Valeo North America Inc.; AutoX Technologies Inc.; Nuro Inc.; Apple Inc.; TuSimple; Aurora Innovation; Toyota Research Institute; Intel Corp; TORC Robotics Inc.; EasyMile; Ridecell; Mercedes Benz; Bosch; GM Cruise LLC; Honda; Zoox Inc.; NVIDIA Corporation; Navya Inc.; Udelv; Pony.AI; Continental Automotive Systems; Mando America Corporation; Uber Advanced Technologies Group; and Almotive Inc.
4. RSS stands for Responsibility-Sensitive Safety and NHTSA stands for National Highway Traffic Safety Administration.

Acknowledgments

The authors acknowledge Nicolas Cointe for his assistance in the linguistic-based text data analytics, the three anonymous reviewers for providing helpful comments on earlier versions of the manuscript, and the European Research Council for financial support of this research (ERC Consolidator grant BEHAVE – 724431).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research has received funding from the European Research Council: Consolidator Grant BEHAVE (grant agreement 724431).

ORCID

Andreia Martinho http://orcid.org/0000-0003-2982-3476
Maarten Kroesen http://orcid.org/0000-0001-6623-9848
Caspar Chorus http://orcid.org/0000-0002-6380-4853

References

Anderson, J. M., Kalra, N., Stanley, K. D., Sorensen, P., Samaras, C., & Oluwatola, O. A. (2014). *Autonomous vehicle technology: A guide for policymakers*. Santa Monica, CA: Rand Corporation.

AutoX (2018). *The autox safety* factor (Tech. Rep.). Auto X. Retrieved from https://www.autox.ai/download/pdf/safety.pdf

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., … Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64.

Aydemir, F. B., & Dalpiaz, F. (2018). A roadmap for ethics-aware software engineering. In *2018 IEEE/ACM international workshop on software fairness (fairware), Gothenburg, Sweden* (pp. 15–21). IEEE.

Bagloee, S. A., Tavana, M., Asadi, M., & Oliver, T. (2016). Autonomous vehicles: Challenges, opportunities, and future implications for transportation policies. *Journal of Modern Transportation*, *24*(4), 284–303.

Bergmann, L. T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S., … Stephan, A. (2018). Autonomous vehicles require socio-political acceptance – An empirical and philosophical perspective on the problem of moral decision making. *Frontiers in Behavioral Neuroscience*, *12*, 31.

Bigman, Y. E., & Gray, K. (2020). Life and death decisions of autonomous vehicles. *Nature*, *579*(7797), E1–E2.

BMW (2016). *Sustainable value report* (Tech. Rep.). BMW. Retrieved from https://www.bmwgroup.com/content/dam/grpw/websites/bmwgroup_com/ir/down loads/en/2016/2016-BMW-Group-Sustainable-Value-Report.pdf

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2019). The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. *Proceedings of the IEEE*, *107*(3), 502–504.

Borenstein, J., Herkert, J., & Miller, K. (2017). Self-driving cars: Ethical responsibilities of design engineers. *IEEE Technology and Society Magazine*, *36*(2), 67–75.

Borenstein, J., Herkert, J. R., & Miller, K. W. (2019). Self-driving cars and engineering ethics: The need for a system level analysis. *Science and Engineering Ethics*, *25*(2), 383–398.

Brown, A., Rodriguez, G., Best, B., Hoang, K. T., Safford, H., Anderson, G., … D'Agostino, M. C. (2018). *Federal, state, and local governance of automated vehicles*. Davis, CA: Institute of Transportation Studies & Policy Institute for Energy, Environment, and the Economy.

Coeckelbergh, M. (2016). Responsibility and the moral phenomenology of using self-driving cars. *Applied Artificial Intelligence*, *30*(8), 748–757.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, *26*(4), 2051–2068. doi:10.1007/s11948-019-00146-8

Collingwood, L. (2017). Privacy implications and liability issues of autonomous vehicles. *Information & Communications Technology Law*, *26*(1), 32–45.

Csizmadia, T. (2018). *Ensuring safe self-driving Almotive's development puts safety first* (Tech. Rep.). Almotive. Retrieved from https://medium.com/aimotive/ensuring-safe-self-driving-ad28cc89fa3b?.

Cunneen, M., Mullins, M., Murphy, F., Shannon, D., Furxhi, I., & Ryan, C. (2020). Autonomous vehicles and avoiding the trolley (dilemma): vehicle perception, classification, and the challenges of framing decision ethics. *Cybernetics and Systems*, *51*(1), 59–80.

Daimler (2018). *Reinventing safety: A joint approach to automated driving systems* (Tech. Rep.). Mercedes-Benz and Bosch. Retrieved from https://www.daimler.com/documents/innovation/other/vssa-mercedes-benz-a nd-bosch.pdf

Danks, D., & London, A. J. (2017). Regulating autonomous systems: Beyond standards. *IEEE Intelligent Systems*, *32*(1), 88–91.
Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(1), 5–27.
Etienne, H. (2020, February). When AI ethics goes astray: A case study of autonomous vehicles. *Social Science Computer Review*. doi:10.1177/0894439320906508
Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, *21*(4), 403–418.
Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, *77*, 167–181.
Favarò, F., Eurich, S., & Nader, N. (2018). Autonomous vehicles' disengagements: Trends, triggers, and regulatory limitations. *Accident Analysis & Prevention*, *110*, 136–148.
Favarò, F. M., Nader, N., Eurich, S. O., Tripp, M., & Varadaraju, N. (2017). Examining accident reports involving autonomous vehicles in California. *PLoS One*, *12*(9), e0184952.
Fischer, J. M., & Ravizza, M. (2000). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.
Fleetwood, J. (2017). Public health, ethics, and autonomous vehicles. *American Journal of Public Health*, *107*(4), 532–537.
Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5–16. Retrieved from https://philpapers.org/archive/FOOTPO-2.pdf
Fried, B. H. (2012). What does matter? The case for killing the trolley problem (or letting it die). *The Philosophical Quarterly*, *62*(248), 505–529.
Funkhouser, K. (2013). Paving the road ahead: Autonomous vehicles, products liability, and the need for a new approach. *Utah Law Review*, *2013*, 437.
Gandia, R. M., Antonialli, F., Cavazza, B. H., Neto, A. M., Lima, D. A. D., Sugano, J. Y., … Zambalde, A. L. (2019). Autonomous vehicles: Scientometric and bibliometric review. *Transport Reviews*, *39*(1), 9–28.
Garza, A. P. (2011). Look ma, no hands: Wrinkles and wrecks in the age of autonomous vehicles. *New England Law Review*, *46*, 581.
GM (2018). *Self-driving safety report* (Tech. Rep.). General Motors. Retrieved from https://www.gm.com/content/dam/company/docs/us/en/gmcom/gmsafetyreport.pdf.
Goodall, N. (2019). More than trolleys. *Transfers*, *9*(2), 45–58. Retrieved from https://www.berghahnjournals.com/view/journals/transfers/9/2/trans090204.xml.
Goodall, N. J. (2014a). Ethical decision making during automated vehicle crashes. *Transportation Research Record*, *2424*(1), 58–65.
Goodall, N. J. (2014b). Machine ethics and automated vehicles. In G. Meyer & S. Beiker (Eds), *Road vehicle automation. Lecture notes in mobility* (pp. 93–102). Cham: Springer.
Goodall, N. J. (2016). Can you program ethics into a self-driving car? *IEEE Spectrum*, *53*(6), 28–58.
Gurney, J. K. (2013). Sue my car not me: Products liability and accidents involving autonomous vehicles. *University of Illinois Journal of Law, Technology & Policy*, *2013*(2), 247.
Gurney, J. K. (2015). Crashing into the unknown: An examination of crash-optimization algorithms through the two lanes of ethics and law. *Albany Law Review*, *79*, 183.
Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*, 99–120. doi:10.1007/s11023-020-09517-8
Harris, J. (2020). The immoral machine. *Cambridge Quarterly Of Healthcare Ethics*, *29*(1), 71–79.
Heikoop, D. D., Hagenzieker, M., Mecacci, G., Calvert, S., Santoni De Sio, F., & Arem, B. van. (2019). Human behaviour with automated driving systems: A quantitative framework for meaningful human control. *Theoretical Issues in Ergonomics Science*, *20*(6), 711–730.
Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust – The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, *105*, 105–120.
Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for crashes of autonomous vehicles: An ethical analysis. *Science and Engineering Ethics*, *21*(3), 619–630.

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

Honda (2015). *Honda sustainability report* (Tech. Rep.). Honda. Retrieved from https://global.honda/about/sustainability/report/pdf-download/2015.html

Huang, K., Greene, J. D., & Bazerman, M. (2019). Veil-of-ignorance reasoning favors the greater good. *Proceedings of the National Academy of Sciences*, *116*(48), 23989–23995.

Huang, B., van Cranenburgh, S., & Chorus, C. G. (2020). Death by automation. *European Journal of Transport and Infrastructure Research*, *20*(3), 71–86.

Kallioinen, N., Pershina, M., Zeiser, J., Nosrat Nezami, F., Stephan, A., Pipa, G., … König, P. (2019). Moral judgements on the actions of self-driving cars and human drivers in dilemma situations from different perspectives. *Frontiers in Psychology*, *10*, 2415.

Kalra, N., & Paddock, S. M. (2016). Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice*, *94*, 182–193.

Keeling, G. (2017). Against Leben's Rawlsian collision algorithm for autonomous vehicles. In *3rd conference on "Philosophy and Theory of Artificial Intelligence", Leeds, UK* (pp. 259–272). Cham: Springer.

Keeling, G. (2020). Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics*, *26*(1), 293–307.

Keeling, G., Evans, K., Thornton, S. M., Mecacci, G., & de Sio, F. S. (2019). Four perspectives on what matters for the ethics of automated vehicles. In G. Meyer & S. Beiker (Eds.), *Automated vehicles symposium* (pp. 49–60). Cham: Springer.

Kirkpatrick, K. (2015). The moral challenges of driverless cars. *Communications of the ACM*, *58*(8), 19–20.

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology*, *19*(2), 107–115.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., … Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Journal of Clinical Epidemiology*, *62* (10), e1–e34.

Lin, P. (2016). Why ethics matters for autonomous cars. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte* (pp. 69–85). Berlin, HD: Springer.

Lin, P., Abney, K., & Jenkins, R. (2017). *Robot ethics 2.0: From autonomous cars to artificial intelligence* (P. Lin, R. Jenkins, & K. Abney). Oxford: Oxford University Press.

Luetge, C. (2017). The German ethics code for automated and connected driving. *Philosophy & Technology*, *30*(4), 547–558.

Lundgren, B. (2020, April). Safety requirements vs. crashing ethically: What matters most for policies on autonomous vehicles. *AI & Society*, 1435–5655. doi:10.1007/s00146-020-00964-6

Mando (2018). *Mando sustainability report 2018* (Tech. Rep.). Mando Corporation. Retrieved from https://www.mando.com/eng/sustainability/sustain13.jsp

Marakby, S. (2018). *A matter of trust: Ford's approach to developing self-driving vehicles* (Tech. Rep.). Ford Autonomous Vehicles LLC. Retrieved from https://media.ford.com/content/dam/fordmedia/pdf/Ford_AV_LLC_FINAL_HR_2.pdf

Marchant, G. E., & Lindor, R. A. (2012). The coming collision between autonomous vehicles and the liability system. *Santa Clara Law Review*, *52*, 1321.

Martins Martinho Bessa, A., Caspar, C., Kroesen, M. (Maarten), & Van Herber, N. (2020). *Dataset ethical issues in focus by the autonomous vehicles industry 4TU* ResearchData [Dataset] doi:10.4121/13348535.v1

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

Menon, C., & Alexander, R. (2020). A safety-case approach to the ethics of autonomous vehicles. *Safety and Reliability, 39*(1) 33–58.

Milakis, D., Van Arem, B., & Van Wee, B. (2017). Policy and society related implications of automated driving: A review of literature and directions for future research. *Journal of Intelligent Transportation Systems*, *21*(4), 324–348.

Misselhorn, C. (2015). Collective agency and cooperation in natural and artificial systems. In C. Misselhorn (Ed.), *Collective agency and cooperation in natural and artificial systems* (pp. 3–24). Cham: Springer.

Mobileye (n.d.). *Implementing the RSS model on NHTSA pre-crash scenarios* (Tech. Rep.). Intel Corp. Retrieved from https://static.mobileye.com/website/corporate/rss/rss_on_nhtsa.pdf

Morita, T., & Managi, S. (2020). Autonomous vehicles: Willingness to pay and the social dilemma. *Transportation Research Part C: Emerging Technologies*, *119*, 102748.

Mulgan, R. (2000). 'Accountability': An ever-expanding concept? *Public Administration*, *78*(3), 555–573.

Nissan (2017). *Financial information as of March 31, 2017* (Tech. Rep.). Nissan. Retrieved from https://group.renault.com/wp-content/uploads/2017/07/yuho-nissan-fr2016.pdf

Noothigattu, R., Gaikwad, S. S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., … Procaccia, A. D. (2018). *A voting-based system for ethical decision making*. In *Thirty-second AAAI conference on Artificial Intelligence*, New Orleans, Louisiana.

Nuro (2018). *Delivering safety: Nuro's approach* (Tech. Rep.). Nuro. Retrieved from https://static1.squarespace.com/static/57bcb0e02994ca36c2ee746c/t/5b9a00848a922d8eaecf65a2/15368193586

Nvidia (2018). *Nvidia self-driving safety report* (Tech. Rep.). Nvidia. Retrieved from https://www.nvidia.com/content/dam/en-zz/Solutions/self-driving-cars/safety-report/auto-print-safety-report-pdf-v16.5%20(1).pdf

Nyholm, S. (2018). The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, *13*(7), e12507.

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice*, *19*(5), 1275–1289.

Parkinson, S., Ward, P., Wilson, K., & Miller, J. (2017). Cyber threats facing autonomous and connected vehicles: Future challenges. *IEEE Transactions on Intelligent Transportation Systems*, *18*(11), 2898–2915.

SAE On-Road Automated Vehicle Standards Committee. (2014). *Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems*. SAE Standard J3016 201401. Warrendale, PA: SAE International.

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*, 15.

Schuelke-Leech, B. A., Jordan, S. R., & Barry, B. (2019). Regulating autonomy: An assessment of policy language for highly automated vehicles. *Review of Policy Research*, *36*(4), 547–579.

Shashua, A., & Shalev-Shwartz, S. (2017). *A plan to develop autonomous vehicles. And prove it*. (Tech. Rep.). Intel. Retrieved from https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/10/autonomous-vehicle-safety-strategy.pdf

Soriano, B. C., Dougherty, S. L., Soublet, B. G., & Triepke, K. J. (2014). Autonomous vehicles: A perspective from the California Department of motor vehicles. In G. Meyer & S. Beiker (Eds.), *Road vehicle automation* (pp. 15–24). Cham: Springer.

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.

Taeihagh, A., & Lim, H. S. M. (2019). Governing autonomous vehicles: Emerging responses for safety, liability, privacy, cybersecurity, and industry risks. *Transport Reviews*, *39*(1), 103–128.

Taiebat, M., Brown, A. L., Safford, H. R., Qu, S., & Xu, M. (2018). A review on energy, environmental, and sustainability implications of connected and automated vehicles. *Environmental Science & Technology*, *52*(20), 11449–11465.

Thomson, J. J. (1984). The trolley problem. *Yale Law Journal*, *94*, 1395.

Thornton, S. M., Pan, S., Erlien, S. M., & Gerdes, J. C. (2017). Incorporating ethical considerations into automated vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, *18*(6), 1429–1439.

Toyota (n.d.). *Automated driving at toyota: Vision, strategy and development* (Tech. Rep.). Toyota Motor Corporation. Retrieved from https://autodocbox.com/Electric_Vehicle/65566194-Automated-driving-at-toyota-vision-strategy-and-development.html

Uber (2018). *A principled approach to safety* (Tech. Rep.). Uber. Retrieved from https://docs.huihoo.com/car/Uber-ATGSafety-Report-2018.pdf.

Valeo (2016). *Meet the future 2016 activity and sustainable development report* (Tech. Rep.). Valeo. Retrieved from https://www.valeo.com/wp-content/uploads/2017/05/VALEO_RA_2016_GB_MEL.pdf

Van den Hoven, J., Vermaas, P., & Van de Poel, I. (2015). *Handbook of ethics, values and technological design*. Dordrecht: Springer.

Van de Poel, I., Fahlquist, J. N., Doorn, N., Zwart, S., & Royakkers, L. (2012). The problem of many hands: Climate change as an example. *Science and Engineering Ethics*, *18*(1), 49–67.

Van Wee, B. (2011). *Transport and ethics: Ethics and the evaluation of transport policies and projects*. Cheltenham: Edward Elgar Publishing.

Weast, J., Yurdana, M., & Jordan, A. (2016). *A matter of trust: How smart design can accelerate automated vehicle adoption* (Tech. Rep.). Intel. Retrieved from https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/trust-autonomous-white-paper-secure.pdf

Wolkenstein, A. (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*(3), 163–173.

Wood, M., Robbel, P., Maass, M., Tebbens, R. D., Meijs, M., Harb, M., … Robinson, K. (2019). *Safety first for automated driving* (Tech. Rep.). Mercedes-Benz, Daimler, Intel Corp. Retrieved from https://www.daimler.com/documents/innovation/other/safety-first-for-automated-driving.pdf

Wu, S. S. (2019). Autonomous vehicles, trolley problems, and the law. *Ethics and Information Technology*, *22*(1), 1–13.

Zoox (2018). *Safety innovation at Zoox: Setting the bar for safety in autonomous mobility* (Tech. Rep.). Zoox. Retrieved from https://zoox.com/wp-content/uploads/2018/12/Safety_Report_12Dec2018.pdf