Avoiding the Intrinsic Unfairness of the Trolley Problem

Tobias Holstein
Mälardalen University
Västerås, Sweden
tobias.holstein@mdh.se

Gordana Dodig-Crnkovic
Chalmers University of Technology
Gothenburg, Sweden
gordana.dodig-crnkovic@chalmers.se

ABSTRACT

As an envisaged future of transportation, self-driving cars are being discussed from various perspectives, including social, economical, engineering, computer science, design, and ethical aspects. On the one hand, self-driving cars present new engineering problems that are being gradually successfully solved. On the other hand, social and ethical problems have up to now being presented in the form of an idealized unsolvable decision-making problem, the so-called "trolley problem", which is built on the assumptions that are neither technically nor ethically justifiable. The intrinsic unfairness of the trolley problem comes from the assumption that lives of different people have different values.

In this paper, techno-social arguments are used to show the infeasibility of the trolley problem when addressing the ethics of self-driving cars. We argue that different components can contribute to an "unfair" behaviour and features, which requires ethical analysis on multiple levels and stages of the development process. Instead of an idealized and intrinsically unfair thought experiment, we present real-life techno-social challenges relevant for the domain of software fairness in the context of self-driving cars.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

KEYWORDS

Unfairness, Self-Driving Cars, Autonomous Cars, Trolley Problem, Decision Making, Ethics, Social Aspects, Software Engineering, Challenges

ACM Reference Format:
Tobias Holstein and Gordana Dodig-Crnkovic. 2018. Avoiding the Intrinsic Unfairness of the Trolley Problem. In *FairWare'18: FairWare'18:IEEE/ACM International Workshop on Software Fairness , May 29, 2018, Gothenburg, Sweden.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3194770.3194772

1 INTRODUCTION

Increasingly, prototypical self-driving vehicles are participating in public traffic [46, 56, 62] and are planned to be sold starting in 2020 [51, 55]. Public awareness and media coverage contribute to a manifold of discussions about self-driving vehicles. This is currently amplified through recent accidents with autonomous vehicles [15, 53].

Software is playing a key role in modern vehicles and in self-driving vehicles. Software in cars is growing by a factor of 10 every 5 to 7 years, and in some sense car manufacturers are becoming software companies. These novelties ask for a change on how the software is engineered and produced and for a disruptive renovation of the electrical and software architecture of the car, as testified by the effort of Volvo Cars [45].

Moreover, self-driving vehicles will be connected with other vehicles, with the manufacturer cloud for software upgrades, with Intelligent Transport Systems (ITS), Smart Cities, and Internet of Things (IoT). Self-driving vehicles will combine data from inside the vehicle with external data coming from the environment (other vehicles, the road, signs, and the cloud). In such a scenario, different applications will be possible: smart traffic control, better platooning coordination, and enhanced safety in general. However, the basic assumption is that future self-driving connected cars must be socially sustainable. Until now, ethical aspects of self-driving cars have been addressed in form of a thought experiment, so called "trolley problem" described in [20] and [63], and discussed in number of articles in IEEE [4, 6, 26], ACM [21, 34, 37], Scientific American [12, 30, 35], Science [8, 29], other high-profile journals [10, 25, 27], conference workshops [5, 48] and various other sources [1, 3, 40, 50]. Here is the general scenario being discussed:

A self-driving vehicle drives on a street with a high speed. In front of the vehicle a group of people suddenly blocks the street. The vehicle is too fast to stop before it reaches the group. If the vehicle does not react immediately, the whole group will be killed. The car could however evade the group by entering the pedestrian way and consequently kill a previously not involved pedestrian. The following variations have been proposed: (A) Replacing the pedestrian with a concrete wall, which in consequence will kill the passenger of the self-driving car; (B) Varying the personas of people in the group, the single pedestrian or the passenger. The use of personas allows including an emotional perspective [7], e.g., stating that the single pedestrian is a child, a relative, a very old or a very sick human, or a brutal dictator, who killed thousands of people, etc.

Even though the scenarios are similar, the responses of humans, when asked how they would decide, differ [8]. The problem is that the question asked has limited number of possible answers, which are all ethically questionable and perceived as bad or wrong. Therefore, a typical approach to this problem is to analyze the scenarios by following ethical theories, such as utilitarianism, other forms of consequentialism or deontological ethics [36]. For example, utilitarianism would aim to minimize casualties, even if it means to kill the passenger, by following the principle: the moral action is the

FairWare'18, May 29, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5746-3/18/05…$15.00
https://doi.org/10.1145/3194770.3194772

one that maximizes utility (or in this case minimizes the damage). Depending on the ethics framework, different arguments can be used to justify the decision.

Applying ethical theories to analyze a given dilemma and possible answers can presently only be done by humans. How would self-driving cars solve such dilemmas? There are numerous publications that suggest to implement moral principles into algorithms of self-driving cars [13, 14, 26]. We find that this does not solve the problem, but it reassures that the solution is calculated based on a given set of rules or other mechanisms, translating the problem to engineering, where it is implemented.

It is worth to notice that the real-world engineering problem is substantially different from the hypothetical trolley problem. While an ethical dilemma is an idealized constructed state that has no good solution, an engineering problem is always by construction such that it can differentiate between better and worse solutions. A decision making process that has to be implemented in a self-driving car can be summarized as follows. It starts with an awareness of the environment: Detecting obstacles, such as a group of humans, animals or buildings, and also the current context/situation of the car using external systems (GPS, maps, street signs, etc.) or locally available information (speed, direction, etc.). Various sensors have to be used to collect all required information. Gaining detailed information about obstacles would be a necessary step before a decision can be made that maximizes utility and/or minimizes damage. A computer program calculates solutions and chooses the solution with the optimal outcome. The self-driving car executes the calculated action and the process repeats.

The process itself can be used to identify concrete ethical challenges within the decision making by considering the current state of the art of technology and its development. In a concrete car both the parts of this complex system and the way in which it is created have a critical impact on the decision-making. This includes for instance the quality of sensors, code, and testing. We also see ethical challenges in design decisions, such as whether a certain technology is used because of its lower price, even though the quality of information for the decision making would be substantially increased if more expensive technology (such as sensors) would be used.

Besides the self-driving vehicle itself, it is also important to address yet another complex system: self-driving vehicles participating in public traffic among cars with human drivers. It also has to be taken into account that self-driving cars are highly connected with the infrastructure and with other self-driving cars. Therefore, it is important to investigate how self-driving vehicles are actually built, how ethical challenges are addressed in their design, production, and use and how certain decisions are justified. Discussing this before self-driving vehicles are officially introduced into the market, allows taking part in the setting and definition of ethical ground rules. McBride states that "Issues concerning safety, ethical decision making and the setting of boundaries cannot be addressed without transparency" [37]. We think that transparency is necessary but not sufficient, and it is important to start further investigations and discussions.

Identifying relevant ethical challenges that should currently be addressed is an important step before ethical aspects of self-driving cars can actually be meaningfully introduced from the point of view of societal and individual stakeholders including designers and producers [32]. It is important to focus not on abstract thought experiments but on concrete conditions that influence the behavior and properties of self-driving cars as being developed through different stages lading to deployment and inclusion in traffic. In this process evaluating software in terms of fairness can play a crucial role in the iterative development and deployment of self-driving cars.

The paper is structured as follows. A problem statement is described in 3. A short introduction to self-driving cars and their current state of the art is provided in Section 4, with the emphasis on the description of the decision making principles given in Section 4.1 and the role of software in Section 4.2. We explain the intrinsic unfairness introduced by the trolley problem discussion in 5 and point towards challenges for software fairness in 6. Conclusions and final remarks are presented together with recommendations in Section 8.

2 RELATED WORK

Most related works are disregarding the details on what sensors, components, and algorithms can actually do. It is an assumption, a speculative idea that self-driving cars will have access to private, medical, financial, and other records of every human on earth and that everyone can everywhere be tracked, recognized and valued.

Investigating scenarios based on the trolley problem by using surveys and experiments in various forms is probably interesting for the field of human behaviour and psychology. To determine whom people choose to kill will provide a deep insight into the human mind.

3 PROBLEM STATEMENT

However, the trolley-problem leads into the wrong direction. The research around self-driving vehicles should focus not on whom to kill, but on how not to kill at all, i.e. crash/incident avoidance. Researching solutions for the decision making based on ethical research, such as disregarding that all humans are equal, can only lead to bad, maybe even unethical, solutions. E.g., a differentiation based on age, social status, and other data is, at least in Germany, not allowed [18].

Solving technical problems towards self-driving cars that are safe, fair, ethically justified and integrated into our everyday life is the overall aim. It is important to discuss the current state of the art and to point out challenges for software fairness for current and future development of self-driving cars.

4 SELF-DRIVING CARS BASICS

The term "autonomous" could be ambiguous to some readers. It can be used to describe certain autonomous features or functions, such as advanced driver assistance systems, that for instance assist the driver in keeping the lane or adjust to the speed of vehicles ahead. Those systems are designed to assist, but the driver is always responsible and has to intervene if critical situation occur.

We use the term "self-driving" cars to avoid wrong interpretations of the terms "fully autonomous" or "driverless". Self-driving cars refer to cars that may operate self-driving without human help

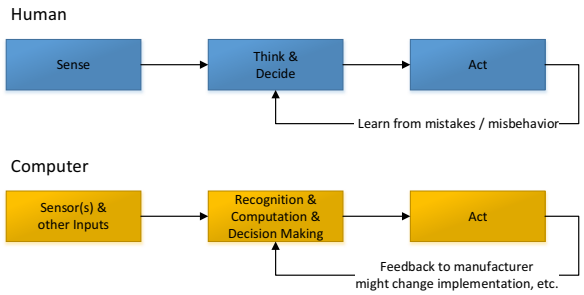

Figure 1: Comparison of human and computers sense, think and act process (cf. [24]) which we extended by adding a feedback loop.

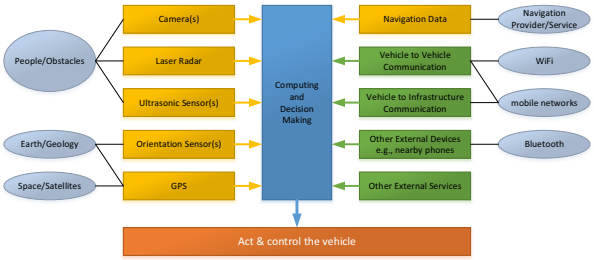

Figure 2: Abstract representation of decision making in autonomous vehicles composed from various sources (cf. [17, 23, 43, 54, 57, 59, 61])

or even without a presence of human being. This means that the unoccupied car can drive from place A to B to pick up someone. This is the highest level of autonomy for cars and corresponds to the last level of five, as defined by the Society of Automotive Engineers [49] and United States National Highway Traffic Safety Administration (NHTSA), who, since September 2016, adopted SAE's classification with level 0 (no automation), level 1 (driver assistance), level 2 (partial automation), level 3 (conditional automation), level 4 (high automation), and level 5 (full automation) [41, p.9].

A concrete example is the self-driving Waymo car [60], former known as the Google car [28], a fully autonomous and self-driving vehicle.

4.1 Decision Making in Self-Driving Cars

Developing self-driving cars that act without a driver means to replace a human, who today is performing the complex tasks of driving, with a computer system executing the same tasks. Figure 1 shows both variants and allows a comparison.

There is an important difference in the feedback loop. While humans continuously learn, for example from their mistakes or misbehaviour, automotive software might be confined to slow updates. Approaches with self-adaptive software, such as machine learning approaches, which learns and reacts immediately, aim to overcome this constraint. Extraordinary road signs for example, which are new to the self-driving car's software, present a risk as they can pass unnoticed/uninterpreted, while they could be understood by a human through context/interpretation. Also unexpected and dangerous situations, like an attack or threat near or even against the vehicle might not be correctly interpreted by a self-driving car compared to a human.

Depending on the technology and the amount of sensors, the type and quality of information that is gathered differs. This extremely complex process might be difficult to imagine and in order to give an idea of what self-driving cars "see" we refer to the visualization shown in [58] presented by Waymo [59]. It is based on the data gathered by multiple sensors installed in the self-driving car, including a laser radar (LIDAR) mounted on the top of the vehicle. Algorithms detect patterns in the data and calculate positions and sizes of objects, which then can be used by other components for decision making.

4.2 Complexity of Decision Making and the Role of Software

The amount and type of sensors used to detect objects around the vehicle and its surrounding environment differs among car manufacturers as well as in research [16, 23, 61]. The diagram in Figure 2 gives an abstract overview by categorizing different types of sensors mentioned in literature. This allows discussing the types of information used and how they relate to each other.

Most of the functionality in the automotive domain is based on software [9]. Software is written by software engineers and at least for important components extensively tested to ensure their correct functioning. In self-driving cars software relies on different disciplines, such as computer vision, machine learning, and parallel computing, but also on various external services. It is a complex process to calculate a decision, and it is also difficult to test those against all possible real world scenarios [57].

One of the problems is that all calculations are based on an abstraction of the real world. This abstraction is an approximate representation of a real world situation and thus the decision making will create decisions for an imperfect world. This is a twofold problem, because the more information is available the better the decisions might be, but at the same time more interpretation and filtering might have to be used to get the data that actually is useful for the decision making.

Engineers have to decide what kind of data to use, how reliable or trustworthy the data is and how to balance the different sources of information in their algorithms. Also different sensors have their specific limitations and to overcome those, a combination of multiple sensors might be used. The overall problem is usually referred to as sensor fusion [31]. This problem is acerbated in the case of connected vehicles since data will come not only from the sensors of the car, but also from other vehicles, street infrastructure, etc. In this case other factors should be taken into account since it is not possible to have a perfect knowledge about the devices that are used to sense information and about their status.

Imagine heavy weather conditions, the navigation reports a street ahead, the radar is reporting a clear street, but the visual

camera reports an obstacle straight ahead. How will this "equation" be solved and what will be the result? The wrong decision might lead to an accident, when important information of some sensors is disregarded and other sensors do not detect the obstacle or hazard in front of the vehicle [53]. Car manufacturers are constantly improving and testing the recognition capabilities of their systems [54]. It is a multi-factor optimization task, which aims to find an optimal solution under consideration of costs, quality, and potential risk factors.

As a measure of reliability, some manufacturers are thinking to count miles covered without any accident, however this might be infeasible since a vehicle should cover around 11 billion of miles to demonstrate with 95% of confidence and 80% power that autonomous vehicle failure rate is lower than the human driver failure rate [33]. Moreover, this calculation only holds if the software within the car does not change over time. Nowadays, manufacturers are increasingly interested in continuous integration and deployment techniques that promise to update the software even after the vehicle has been sold and is on the street, like a common smartphone. However, every change of code might require restarting the miles counter.

5 INTRINSIC UNFAIRNESS OF THE TROLLEY PROBLEM

Discussions of the trolley problem in the context of self-driving cars include the assumption that self-driving car can make a decision that will lead to a specific outcome, such as that someone will survive or will be killed. In the simplest case, a comparison between a different number of people is assumed, as for example in [8, 19, 52].

Further scenarios consider differentiating based on an individuals age, profession, gender or social rank [1, 8]. This has already been declared to be unethical by the German ethics commission for autonomous driving, which defined that all human lives are equal worth [18]. Therefore, the decision making in cars is not allowed to consider attributes that go into personal details. If it were to be allowed, a privacy and data protection problem would be the next challenge. A car would require access to all humans personal data, including medical records, police records, and so on. This would be an implementation of George Orwell's dystopian novel "Nineteen Eighty-Four" [44] in the context of self-driving cars. Considering approaches based on MIT's moral machine [1] that require personal data as input is therefore misleading.

Current sensor technologies can detect obstacles of different sizes and types depending on technology and distance to the object. This means the quality of detecting objects differs, and labeling a certain non-moving object as a human or a display dummy is difficult. Also, sensor "measurements aren't always detailed enough to distinguish one object from another" [38]. Therefore, it might not distinguish two groups of people based on the actual number of people but based on the volume of space that is occupied by them. Also, sensors are currently not able to count the number of passengers inside another car, which requires the other cars to report the number of people inside the car to all surrounding cars for instance via *Vehicle2X (V2X)*. The same problem exists for buildings or areas that sensors might not be able to cover correctly, such as coffee places inside the city that are surrounded by windshields made of wood or glass. Having a mixed environment of self-driving cars and cars or locations with or without technologies like V2X is another problem. When sensor technology and/or infrastructure is not as advanced or does not exist, the decision making cannot consider it.

After objects/obstacles have been detected, the self-driving car can determine the free, i.e. unoccupied, space around the car, which can be used to calculate emergency maneuvers [43]. Furthermore, it can be used to determine the current maximum speed of the car that would still allow the car to safely stop in the free area in case of an emergency [43].

The relevant trolley problem scenario in the context of the state of the art technology is thus more likely to be: hitting an obstacle that is correctly identified versus hitting another object that is unknown or incorrectly identified. It is unrealistic to assume that the self-driving car will have information about whether a human dies or not in a critical situation.

Taking the trolley problem as a basis for discussions of the ethics related to self-driving cars is neglecting the way technology works and simultaneously obfuscating greater ethical challenges, which should be considered if ethical values such as fairness are taken into account, as we describe in more detail in [32].

6 CHALLENGES FOR SOFTWARE FAIRNESS

Self-driving cars will be integrated incrementally. People will adapt to self-driving cars and self-driving cars have to be adapted to people. It is an iterative process where lawmakers, car manufacturers, and society play a crucial role in finding the correct behavioural rules for self-driving cars.

In this process, it will be important to continuously evaluate self-driving cars. Therefore, we point out a set of motivational challenges in the following sections that might be especially interesting in regard to software fairness.

6.1 Sensors

Sensor data is analyzed by algorithms. Recognition can be based on neural networks, i.e. machine learning, using recorded data from real vehicles or simulated environments to train the behaviour of the car [23, 61]. "By analyzing photos of pedestrians, for example, a neural network can learn to identify a pedestrian" [39].

When neural networks are trained by analyzing photos of pedestrians, are those photos subject to fairness? Can the set of photos be biased? Maybe because of the region the photos were taken in, by the type of clothes people wear, or by peoples behaviour/postures, e.g., in the USA compared to Saudi Arabia. If the set of photos is representative of one region, does it mean that people from outside the region become less likely to be detected correctly?

6.2 External Positioning Systems

Connected or interconnected systems, that report the position of obstacles to the car, are likely to become introduced. In Germany, there is already a project called "Schutzranzen" ("protective backpack") [11] that uses active transponders in backpacks to send position data to a cloud which distributes the data to nearby cars. Car drivers can use an App to be warned if pupils with protective

backpacks are near or in close range to the car. This is supposed to increase safety for pupils.

Let's take this example to a large scale and assume that all existing mobile devices become active transponders interconnected with self-driving cars. Can the position data be considered as an input for the decision making of self-driving cars? If yes, some phones might have better GPS, positioning sensors or a faster internet connection. Is that contrary to the principle that "all humans are equally worth"? What about people who don't have a mobile phone with them, or more likely have an empty battery?

6.3 External Services

External services providing up-to-date information to self-driving cars, such as map data, position, traffic information and so on can also be subject of concern when it comes to fairness.

Can external services change the behaviour of the car in some way? E.g., would it be possible for a map service to redirect or guide the car through a certain region that has more shops or advertisements than other regions? In the context of smart cities, i.e. green cities, the car might also be redirected due to traffic control systems that try to optimize the traffic flow throughout the city. How much control will the passenger of the self-driving car have, and how can we test whether the route of the car is biased in some way?

7 DISCUSSION

Environments and people change over time and with the introduction of self-driving cars, we will surely see people adapting to them. Therefore the introduction of self-driving cars becomes an iterative process that requires to constantly evaluate the quality of decision making both in the supporting socio-technical system and in the self-driving cars.

In a car, every component can introduce problems in terms of software fairness and therefore lead to an unfair behaviour. The overall complexity of self-driving cars being a system of systems that is highly interconnected will provide great challenges to test and to establish fairness. Therefore, it will be important for self-driving car manufacturers to test their software in regard to fairness and discrimination. This means to test on component level, on system level and also on system of systems level. Transparency will be necessary to allow external researchers and lawmakers to test self-driving cars and related services, which is an issue regarding intellectual property right from the point of view of the car, component or service manufacturer. Standards will have to incrementally adapt to the development and integration progress, learning from experiences and taking the current state of technology and society into consideration. Including tools that allow automatic discrimination tests, like Themis [22] that generates efficient test suites to measure discrimination, into the software development cycle is therefore a logical and promising improvement in software engineering. It is also a possibility to continuously check on neural networks and similar machine learning techniques, to make sure that their output is not biased when it is not supposed to be.

8 CONCLUSIONS AND FINAL REMARKS

Self-driving vehicles have been recognized as the future of transportation systems and will be successively introduced into the transport systems globally [2, 42, 47]. It is now the right time to start an investigation into the manifold of ethical challenges surrounding self-driving and connected vehicles [18]. As this new technology is being tested and gradually allowed on the roads under controlled conditions, the focus should be on the practical technological solutions and their social consequences, rather than on idealized unsolvable problems such as the much discussed trolley problem. The conclusion from idealized problem discussions is that it has no general solution under all circumstances. Moreover, we pointed out in this article that trolley problem is constructed under wrong assumptions. They include both belief in perfect predictability of complex systems involving vehicles and humans, and expectation that cars can and should make a difference between different people.

It is the right time to discuss the relationship between what is technically possible and what is ethically justifiable for self-driving cars. Even if this might limit the possibilities, it will set the necessary ground for further developments.

REFERENCES

[1] Moral Machine. http://moralmachine.mit.edu, 2016.

[2] Ethics commission on automated driving presents report: First guidelines in the world for self-driving computers. Technical report, Federal Ministry of Transport and Digital Infrastructure, 2017.

[3] J. Achenbach. Driverless cars are colliding with the creepy Trolley Problem. https://www.washingtonpost.com/news/innovations/wp/2015/12/29/will-self-driving-cars-ever-solve-the-famous-and-creepy-trolley-problem/, December 2015.

[4] E. Ackerman. People Want Driverless Cars with Utilitarian Ethics, Unless They're a Passenger. https://spectrum.ieee.org/cars-that-think/transportation/self-driving/people-want-driverless-cars-with-utilitarian-ethics-unless-theyre-a-passenger, June 2016.

[5] H. S. Alavi, F. Bahrami, H. Verma, and D. Lalanne. Is driverless car another weiserian mistake? In *Proceedings of the 2017 ACM Conference Companion Publication on Designing Interactive Systems*, DIS '17 Companion, pages 249–253, New York, NY, USA, 2017. ACM.

[6] S. Applin. Autonomous vehicle ethics: Stock or custom? *IEEE Consumer Electronics Magazine*, 6(3):108–110, July 2017.

[7] A. Bleske-Rechek, L. Nelson, J. P. Baker, M. Remiker, and S. J. Brandt. Evolution and the trolley problem: People save five over one unless the one is young, genetically related, or a romantic partner. 4:115–127, 01 2010.

[8] J.-F. Bonnefon, A. Shariff, and I. Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016.

[9] M. Broy, I. H. Kruger, A. Pretschner, and C. Salzmann. Engineering Automotive Software. *Proceedings of the IEEE*, 95(2):356–373, feb 2007.

[10] I. Coca-Vila. Self-driving cars in dilemmatic situations: An approach based on the theory of justification in criminal law. *Criminal Law and Philosophy*, Jan 2017.

[11] Coodriver GmbH. Schutzranzen App - anonymous, safe, commercial-free - makes it easier to see children near roads, September 2017. https://www.schutzranzen.com/files/2115/0478/2378/PM_Schutzranzen_IAA_2017_english.pdf.

[12] K. Deamer. What the First Driverless Car Fatality Means for Self-Driving Tech. https://www.scientificamerican.com/article/what-the-first-driverless-car-fatality-means-for-self-driving-tech/, 2016.

[13] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster. *Ethical Choice in Unforeseen Circumstances*, pages 433–445. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[14] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77(Supplement C):1 – 14, 2016.

[15] D. Dolgov. Google self-driving car project - monthly report - september 2016 - on the road. Technical report, Google, 2016.

[16] J. Dunbar and J. E. Gilbert. The human element in autonomous vehicles. In D. Harris, editor, *Engineering Psychology and Cognitive Ergonomics: Cognition and Design*, pages 339–362, Cham, 2017. Springer International Publishing.

[17] S. I. Earth Imaging Journal (EIJ): Remote Sensing, Satellite Images. Lidar boosts brain power for self-driving cars, 2012.

[18] Ethics Commission. Automated and connected driving. Technical report, Federal Ministry of Transport and Digital Infrastructure, 2017.

[19] A. K. Faulhaber, A. Dittmer, F. Blind, M. A. Wächter, S. Timm, L. R. Sütfeld, A. Stephan, G. Pipa, and P. König. Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles. *Science and Engineering Ethics*, Jan 2018.

[20] P. Foot. The problem of abortion and the doctrine of double effect. *Oxford Review*, 5, 1967.

[21] A.-K. Frison, P. Wintersberger, and A. Riener. First person trolley problem: Evaluation of drivers' ethical decisions in a driving simulator. In *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '16 Adjunct, pages 117–122, New York, NY, USA, 2016. ACM.

[22] S. Galhotra, Y. Brun, and A. Meliou. Fairness testing: Testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, ESEC/FSE 2017, pages 498–510, New York, NY, USA, 2017. ACM.

[23] General Motors. GM Self-Driving Safety Report, January 2018. https://www.gm.com/content/dam/gm/en_us/english/selfdriving/gmsafetyreport.pdf.

[24] G. Ghisio. Challenges for the Automotive Platform of the Future, 2016.

[25] N. J. Goodall. Vehicle automation and the duty to act. In *Proceedings of the 21st world congress on intelligent transport systems*, pages 7–11, 2014.

[26] N. J. Goodall. Can you program ethics into a self-driving car? *IEEE Spectrum*, 53(6):28–58, June 2016.

[27] B. Goodman and S. Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *ArXiv e-prints*, June 2016.

[28] Google. Google self-driving car project, 2016.

[29] J. D. Greene. Our driverless dilemma. *Science*, 352(6293):1514–1515, 2016.

[30] L. Greenemeier. Driverless Cars Will Face Moral Dilemmas. https://www.scientificamerican.com/article/driverless-cars-will-face-moral-dilemmas/, 2016.

[31] F. Gustafsson. Automotive safety systems. *IEEE Signal Processing Magazine*, 26(4):32–47, July 2009.

[32] T. Holstein, G. Dodig-Crnkovic, and P. Pelliccione. Ethical and Social Aspects of Self-Driving Cars. *ArXiv e-prints*, Feb. 2018.

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

[34] K. Kirkpatrick. The moral challenges of driverless cars. *Commun. ACM*, 58(8):19–20, July 2015.

[35] S. Kuchinskas. Crash Course: Training the Brain of a Driverless Car. https://www.scientificamerican.com/article/autonomous-driverless-car-brain/, 2013.

[36] B. MacKinnon. *Ethics: Theory and Contemporary Issues, Concise Edition*. Cengage Learning, 2012.

[37] N. McBride. The ethics of driverless cars. *SIGCAS Comput. Soc.*, 45(3):179–184, Jan. 2016.

[38] C. Metz. Former Apple Engineers Working on New Eyes for Driverless Cars, September 2017. https://www.nytimes.com/2017/09/20/technology/former-apple-engineers-driverless-cars.html.

[39] C. Metz. What Virtual Reality Can Teach a Driverless Car, October 2017. https://www.nytimes.com/2017/10/29/business/virtual-reality-driverless-cars.html.

[40] C. Mooney. Save the driver or save the crowd? Scientists wonder how driverless cars will 'choose'. https://www.washingtonpost.com/news/energy-environment/wp/2016/06/23/save-the-driver-or-save-the-crowd-scientists-wonder-how-driverless-cars-will-choose/, 2016.

[41] National Highway Traffic Safety Administration (NHTSA). Federal automated vehicles policy - accelerating the next revolution in roadway safety. Technical report, U.S. Department of Transportation, 2016.

[42] N. H. T. S. A. (NHTSA). "dot/nhtsa policy statement concerning automated vehicles" 2016 update to "preliminary statement of policy concerning automated vehicles". Technical report, National Highway Traffic Safety Administration (NHTSA).

[43] P. Oliver, B. Jan, and K. Sören. Automated driving on public roads: Experiences in real traffic , 2015.

[44] G. Orwell. *1984*. Secker & Warburg, 1949.

[45] P. Pelliccione, E. Knauss, R. Heldal, S. M. Ågren, P. Mallozzi, A. Alminger, and D. Borgentun. Automotive architecture framework: The experience of volvo cars. *Journal of Systems Architecture*, 77(Supplement C):83 – 100, 2017.

[46] M. Persson and S. Elfström. Volvo Car Group's first self-driving Autopilot cars test on public roads around Gothenburg, 2014.

[47] S. Pillath. Briefing: Automated vehicles in the EU. *European Parliamentary Research Service (EPRS)*, (January):12, 2016.

[48] A. Riener, M. P. Jeon, I. Alvarez, B. Pfleging, A. Mirnig, M. Tscheligi, and L. Chuang. 1st workshop on ethically inspired user interfaces for automated driving. In *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '16 Adjunct, pages 217–220, New York, NY, USA, 2016. ACM.

[49] SAE. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. *Global Ground Vehicle Standards*, (J3016):30, 2016.

[50] A. Shashkevich. Stanford professors discuss ethics involving driverless cars. https://news.stanford.edu/2017/05/22/stanford-scholars-researchers-discuss-key-ethical-questions-self-driving-cars-present/, may 2017.

[51] J. D. Stoll. Gm executive credits silicon valley for accelerating development of self-driving cars, 2016.

[52] L. R. Sütfeld, R. Gast, P. König, and G. Pipa. Using virtual reality to assess ethical decisions in road traffic scenarios: Applicability of value-of-life-based models and influences of time pressure. *Frontiers in Behavioral Neuroscience*, 11:122, 2017.

[53] Tesla. A tragic loss | tesla deutschland, 2016.

[54] Tesla. Upgrading Autopilot: Seeing the World in Radar | Tesla Deutschland, 2016.

[55] Toyota. New toyota test vehicle paves the way for commercialization of automated highway driving technologies | toyota global newsroom, 2015.

[56] D. WAKABAYASHI. WaymoâĂŹs Autonomous Cars Cut Out Human Drivers in Road Tests, November 2017. https://www.nytimes.com/2017/11/07/technology/waymo-autonomous-cars.html.

[57] M. M. Waldrop. Autonomous vehicles: No drivers required. *Nature*, 518:20–3, 2015.

[58] Waymo. Waymo - Navigating city streets, December 2016. https://youtu.be/fbWeKhAPMig?t=10.

[59] Waymo. Technology - Waymo, 2017. https://waymo.com/tech/.

[60] Waymo. Waymo, September 2017. https://waymo.com.

[61] Waymo. Waymo Self-Driving Safety Report, 2017. https://storage.googleapis.com/sdc-prod/v1/safety-report/waymo-safety-report-2017.pdf.

[62] Waymo. Waymo - On the Road, February 2018. https://waymo.com/ontheroad/.

[63] P. Wintersberger, A. K. Prison, A. Riener, and S. Hasirlioglu. The experience of ethics: Evaluation of self harm risks in automated vehicles. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 385–391, June 2017.