TECHNISCHE UNIVERSITÄT MÜNCHEN
Peter Löscher-Stiftungslehrstuhl für Wirtschaftsethik
TUM School of Management

Experimental and Normative Ethics:

The Case of Autonomous Cars

Jan Lennart Gogoll, M.Sc.

Vollständiger Abdruck der von der Fakultät für Wirtschaftswissenschaften der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Wirtschaftswissenschaften (Dr. rer. pol.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Michael Kurschilgen
Prüfer der Dissertation:	1. Prof. Dr. Christoph Lütge
	2. Prof. Dr. Claudia Peus

Die Dissertation wurde am 13.08.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Wirtschaftswissenschaften am 15.12.2018 angenommen.

Thesis advisor: Prof. Dr. Christoph Lütge Jan Lennart Gogoll, M.Sc.

Experimental and Normative Ethics: The Case of Autonomous Cars

Abstract

The introduction of autonomous cars is moving at a rapid pace. Almost all car manufacturers, as well as big technology companies, are working on this technology and its introduction might be just around the corner. Apart from technological questions and questions of liability, ethical issues play a major role in the academic discourse but also in the public domain. This dissertation provides two published articles that tackle different, yet related, ethical issues associated with the introduction of the self-driving car. First, using the traditional form of normative argument this dissertation discusses the question of who should decide about the "ethics setting" of an autonomous vehicle and if the famous trolley problem is an adequate tool to examine the problem of the "ethics of crashing". In the first publication, I argue that the trolley problem lacks three important characteristics that severely limit its application to the case of autonomous cars in a dynamic traffic scenario. Further, I argue that a mandatory ethics setting should be introduced to all autonomous cars because this would be in peoples' best interest. The second publication tries to shed some light on possible empirical reservations people have towards delegating a moral task to an algorithm. This study investigates whether the reluctance to delegate to a machine agent instead of a human agent can be explained by a misperception or a lack of trust. We are able to eliminate these explanations and find a general aversion of delegating a moral task to a machine agent in an economic lab experiment. Empirical research is important because every technology is built on acceptance. Given the strong moral case that can be made in favor of autonomous driving, e.g. the avoidance of accidents, it is vital to tackle problems regarding the acceptance of the technology before it is introduced. The introduction of this dissertation as well as the first two chapters give an overview about the debate of ethical issues regarding autonomous cars and makes the claim that normative considerations have to go hand in hand with empirical research.

Contents

In Memoriam Gorgi (†2011) and Ulla (†2017) Beiering

”Man darf zum Beispiel bei dem geldsammelnden Bankier nach dem Zweck seiner rastlosen Tätigkeit nicht fragen: sie ist unvernünftig. Die Tätigen rollen, wie der Stein rollt, gemäss der Dummheit der Mechanik. – Alle Menschen zerfallen, wie zu allen Zeiten so auch jetzt noch, in Sklaven und Freie; denn wer von seinem Tage nicht zwei Drittel für sich hat, ist ein Sklave, er sei übrigens wer er wolle: Staatsmann, Kaufmann, Beamter, Gelehrter.”

For example, one must not inquire of the money-gathering banker what the purpose for his restless activity is: it is irrational. Active people roll like a stone, conforming to the stupidity of mechanics. Today as always, men fall into two groups: slaves and free men. Whoever does not have two-thirds of his day for himself, is a slave, whatever he may be: a statesman, a businessman, an official, or a scholar.

– Friedrich Nietzsche
Menschliches, Allzumenschliches, 1. Hauptsück

Acknowledgments

I would like to express my gratitude to my advisor, Christoph Lütge for his open-mindedness, which gave me the chance to engage in this topic. I would also like to thank Prof. Claudia Peus for taking the task of reviewer. A special thank you goes to Dr. Matthias Uhl, who has not only been one of my co-authors but also a spiritual guide in his role as Mr. Miyagi, introducing me to experimental economics. I would like to thank my former colleagues, Johanna Jauernig and Julian Müller – the latter also for being a great partner in writing one of this thesis' articles. I would like to very specially thank Wally Eichler for all the efforts she has put in and for her ability to anticipate and solve problems before they occur.

Thanks to the DFG and the excellence initiative as well as the Friedrich-Naumann-Stiftung for their support during my thesis. However, since these grants are not manna falling from the sky but paid for with very real taxpayers' money, I'd like to express my grattitude to the German taxpayers. I sincerely hope that this dissertation and the research it contains are not a waste of their money.

Last but not least, I want to thank my family and friends. A special shout out goes to my parents, Frank and Jutta, as well as to my brother Jörn. Their support has always far exceeded mere kinship-based altruism.

I believe in horses, automobiles are a passing phenomenon.

Wilhelm II, German Emperor in 1905

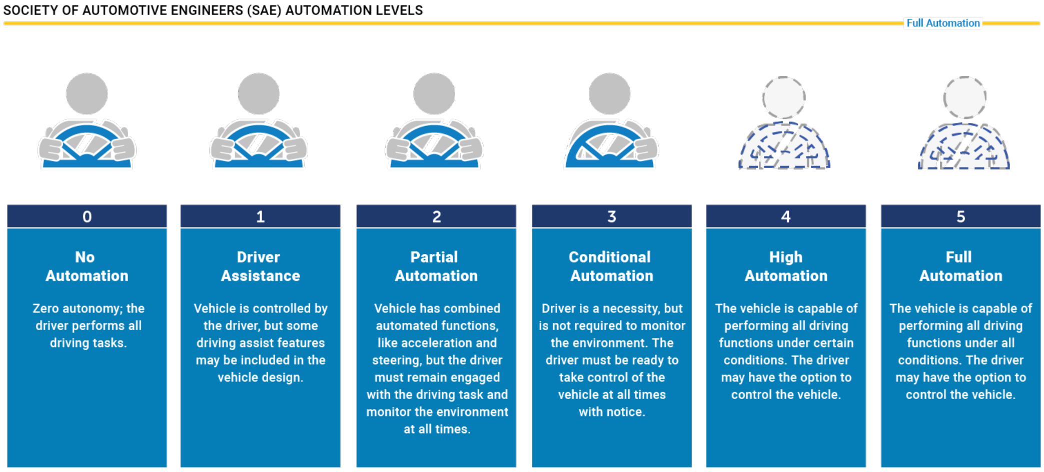
We are still at the 'horseless carriage' stage of this technology, describing these technologies as what they are not, rather than wrestling with what they truly are.

Peter Singer

0

Introduction: Autonomous Cars

Recent development in technology has seen many interesting and fascinating products emerge. The digital revolution has started, yet, we are merely in the very early stage of the changes that will conquer many areas of our life. Keywords like *big data, machine learning* and *cryptocurrency* are dominating the news and even everyday conversations. While all of these developments have great potential regarding future use cases, a few advancements in technology are already at a point of common agreement that the technology will be the future: One – and perhaps the biggest – of these advancements is the concept of the self-driving vehicle or *autonomous car*. All of the major car manufacturers are currently working on its introduction. Additionally, since the technology of an autonomous vehicle relies as heavily on software as it does on the hardware, big players from other industries are also pushing into the market. In fact, software companies are at the forefront of raising awareness of this development as can be seen by the fact that the term "google car" has sometimes been used as an umbrella term for the entire space. Being highly complex systems operating in one of the most dangerous surroundings – traffic – the last years have seen many soft- and hardware manufacturers form joint ventures to tackle the development and win the race to be the first to enter the market with a working and safe product. For instance, BMW and Intel have joined forces with MobileEye and are predicting to have the first autonomous cars on European roads as early as 2020. Japan also wants autonomous cars on the roads in time for the 2020 Olympics (White, 2018). At this time such a prediction does not sound to be far-fetched when one acknowledges the fact that in recent years and even decades, cars have been increasingly equipped with Driver Assistance Systems which, although far from turning modern cars into self-driving cars, have slowly taken over many aspects of the driving experience. The most prominent example might be the anti-lock braking system, which has become an integral part of basically every car on the street and is in place to avoid locking up the wheels when braking on slippery surfaces to maintain control over the vehicle. Essentially, this system takes over control in certain well-defined circumstances. The National Highway Traffic Safety Administration (NHTSA) defines five different levels of autonomous driving ranging from zero (see figure 1), which means no automation at all up to the stage 5 indicating that no human input is necessary or even possible (NHTSA, 2013). It is important to note that most political actors have acknowledged the importance of driverless traffic. This dissertation, for instance, has been partly performed within the Munich Center for Technology in Society (MCTS) lab "Automation & Society: The Case of Highly Automated Driving" as part of the Excellence Initiative by the German Research Foundation (DFG), which shows at least indirect support of governmental actors in this field. The German Federal Ministry of Transport and Digital In-



SOCIETY OF AUTOMOTIVE ENGINEERS (SAE) AUTOMATION LEVELS

0	1	2	3	4	5
No Automation	**Driver Assistance**	**Partial Automation**	**Conditional Automation**	**High Automation**	**Full Automation**
Zero autonomy; the driver performs all driving tasks.	Vehicle is controlled by the driver, but some driving assist features may be included in the vehicle design.	Vehicle has combined automated functions, like acceleration and steering, but the driver must remain engaged with the driving task and monitor the environment at all times.	Driver is a necessity, but is not required to monitor the environment. The driver must be ready to take control of the vehicle at all times with notice.	The vehicle is capable of performing all driving functions under certain conditions. The driver may have the option to control the vehicle.	The vehicle is capable of performing all driving functions under all conditions. The driver may have the option to control the vehicle.

Figure 1: Levels of automation, source: NHTSA (2018)

frastructure has also released an "Ethics Code for Automated and Connected Driving" in 2016 in order to help building a useful framework for future developments (Luetge, 2017, BMVI, 2016). Modern cars offer features like the Lane Keeping Assist System which controls the horizontal movement of a car to keep it in its lane if the driver is distracted. Since the development of the technology moves at a rapid pace, this dissertation is exclusively concerned with the fully automated car, that is a stage 5 completely autonomous vehicle. Basically, such a vehicle does not need to have a steering wheel or a throttle and might resemble a train cabin rather than a traditional car. Figure 2 provides an overview of the complexity and ability of level-5 autonomous cars.

As with every new technology, there is a myriad of ethical issues that arise with its introduction. Especially, if this technology is disruptive and in the domain of traffic – an institution which is responsible for the death of an estimated 1.25 million people globally in the year 2013 alone according to the World Health Organisation (World Health Organisation, 2015). While the technology has a huge potential to significantly reduce this terrifying num-

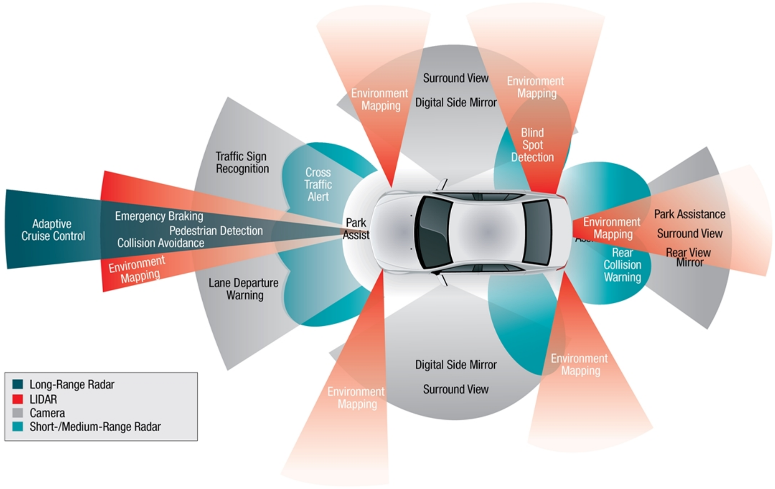


Figure 2: Sensors and electric equipment of autonomous cars, source: Narnakaje (2017)

ber, there will be situations in which the technology does not function as intended an might cause severe harm. In fact, the first fatal crash with a pedestrian has already happened in 2018 (Wakabayashi, 2018). Given the fact that autonomous cars might very well be on the streets in a matter of years, ethical issues need to be addressed as soon as possible. This dissertation is based on work that started in 2014 within the project "Automation & Society: The Case of Highly Automated Driving", which was meant to bring empirical evidence to illuminate several issues regarding autonomous cars in general. Ethics, however, also needs a normative base from which empirical evidence can be judged from and made fruitful. This dissertation is concerned with integrating empirical findings and normative arguments. Normative ethics regarding autonomous cars broadly deals with questions like: *Should we allow or even foster the development of autonomous cars? Is this ethically justifiable?*; – and while many of these questions eventually rely on empirical facts about the world to be true (or false), the method is not empirical. I will argue that empirical research in ethics is necessary and indeed helpful, especially for an endavour like the introduction of a technology that has the potential to severely impact peoples' lives. When I talk about empirical research in ethics, I am referring to sciences like economics, psychology and even experimental philosophy. These approaches look for descriptive rather than normative answers: *How do people actually think about giving up control to a machine? Is there an aversion to the use of algorithms when it comes to morally relevant decisions?* This distinction – between normative and empirical research – serves as the basis for this dissertation. Therefore, this dissertation will proceed as follows: Chapter 1 on page 7 will provide the reader with an overview of ethical issues regarding autonomous cars. Ethical questions regarding autonomous cars will be discussed here, for instance: Should we embrace this technology because of ethical reasons and, if so, how should we judge or prescribe behavior of self-driving cars in particular situations like trolley cases? In chapter 2 on page 24, I will make the case that when it comes to applied ethics normative reasoning alone is insufficient. Ignoring empirical evidence, especially when it comes to brand new technologies, makes the research program fragmentary at best. More likely, however, any research program that remains uninformed about attitudes and possible fears of people towards the new technology will have a negative effect on social welfare. To this effect I will discuss the research about trolley cases from an empirical standpoint, that is, firstly, how peoples' opinions about what is the right way to act may vary across different settings and, secondly, argue that there is a need to address a potential aversion against machine use in the moral domain, which is, essentially, a question of *acceptance*.

In this dissertation, I will use the terms "autonomous car", "self-driving car" and "driverless car" interchangeably. I will attempt to answer two ethical issues regarding autonomous cars using different methodological approaches – the traditional form of normative philosophical argument, the use of game theory and finally, the method of experimental economics in the form of controlled and incentivized laboratory experiment.

The publications on which this dissertation is built will be shortly discussed in chapter 3 on page 42. Here, the reader will find an extended abstract of both publications. The reason for this is that while I have the permission to print the articles in this dissertation, I can only do so in their published form including the format of the respective journal (Science and Engineering Ethics and Journal of Behavioral and Experimental Economics). This fact would make it unbearable for the reader to include them in the main body of the text. Instead, they will be provided in the Appendix.

I have arrived at the conviction that the neglect by economists to discuss seriously what is really the crucial problem of our time is due to a certain timidity about soiling their hands by going from purely scientific questions into value questions.

Friedrich August von Hayek

1

Autonomous Cars: Ethical Issues

In a modern, pluralistic and democratic society justifying and evaluating institutions is of the utmost importance. Especially, when it comes to questions about the design of a vital institution such as traffic in which everyone partakes. Traffic with all its rules and regulations is indeed a fascinating phenomenon. There is a widespread acceptance of the institution, yet, it is responsible for the death of about 3,500 people each year in Germany alone. Additionally, about 400,000 people are injured in Germany (Statistisches Bundesamt, 2017). When there is disagreement it usually revolves around minor changes in the law that regulates traffic such as decreasing the speed limit on certain roads and the like. The institution itself is hardly – if ever – challenged. *Prima facie*, this seems odd. What other institution can one think of that has an equal death toll and is not seen in a highly critical light or at least spoken of in terms of being a "necessary evil"? It is only when we realize that everybody has a stake in traffic that we can understand why the institution of traffic enjoys high approval ratings. It seems that the benefits far outweigh the costs – at least from a societal standpoint. Cost-benefit-analysis and compromises are, however, in the domain of politics rather than ethics. Still, it does give the ethical analysis a benchmark as to what cannot possibly be justifiable. In the case of autonomous cars this means that if it were the case that driverless cars are worse in every relevant aspect compared to their human-driven counterparts, there would be no need for an ethical discussion. In fact, if driverless cars would lead to more or more severe accidents, had a higher negative impact on the environment, were significantly more expensive and wasted valuable human time in the process, the research program and this dissertation along with it would lose significance. However, following Gao et al. (2014), who predict that the introduction of autonomous cars will lead to a decrease in traffic accidents by up to 90%, the technology has great potential for society as a whole and determining its ethical implications seems a worthwhile endeavor. Blanco et al. (2016) second this prediction by stating that even though there is yet not enough data available to make any predictions with certainty, "current data [do] suggest that self-driving cars may have low rates of more-severe crashes [...] when compared to national rates or to rates from naturalistic data sets." Roughly two-thirds of accidents can be attributed to human error, including drunk driving, speeding and distractions (Fagnant and Kockelman, 2015). Sparrow and Howard (2017) even briefly elaborate whether human driving should still be legal in case of a successful introduction of autonomous cars, calling humans "the moral equivalent of drunk robots" (Sparrow and Howard, 2017). While lower crash rates and hence traffic deaths seem like a *pro tanto* good reason to foster the development of autonomous cars from an ethical viewpoint, there exists a myriad of ethical points that have been raised in favor of driverless cars but also few that address potential obstacles that need to be taken into consideration. In general, one can divide the existing literature into two main categories: First, research of the "ethics of crashing", that deals with the reaction of an autonomous vehicle during a critical situation and its consequences. This research often talks about the (in)famous trolley case and its implications for programmable cars. This approach has created vast amounts of literature and is probably the dominating topic of ethical discussion with regards to autonomous cars. The second line of research is lower in volume and deals with ethical issues that are not concerned with explicit life-or-death scenarios but usually provide a more global approach. Examples of this research are questions about liability, environmental issues and social welfare improving features like the avoidance of traffic jams and the possible restructuring of inner-city areas.

I will first provide an introduction to the second category and outline some interesting possible developments due to the technology of driverless cars. Afterward, this dissertation will deal with the "ethics of crashing" and the trolley problem in more detail. The reason for this is twofold: 1.) This topic has gotten a lot of attention in academia – especially in the field of moral philosophy, but also within the social sciences – as well as in popular media outlets. One could say that it dominates the moral debate regarding autonomous cars. 2.) The first publication on which this dissertation is built also deals with the ethics of crashing, therefore an extended introduction seems appropriate.

1.1 Broader ethical issues

As mentioned above, when one judges the possibility of new technologies from an ethical standpoint, a good starting point would be to look for Pareto improvements or win/win-situations. As soon as a technology has the ability to create a benefit for all or to reduce a risk associated with the status quo technology, there is at least a *pro tanto* good reason to foster its development. Indeed, many arguments have been made in this direction. Here, I will describe the improvements that – in my view – provide the greatest benefits to society and the individual. Alas, as with every technology there are risks and downsides that come along with it. The case of autonomous vehicles is no exception. This subsection will end with a brief discussion about the risks and dangers regarding autonomous vehicle technology.

First, however, I will talk about the potential benefits of the technology. It is often argued that the introduction of the driverless car has several ecological benefits. Autonomous cars may be able to lower fossil fuel consumption significantly (Mccarthy, 2017, Torbert and Herrschaft, 2013). The reasons are manifold: Self-driving cars could operate in such a way that it optimizes the speed and breaking on energy-saving grounds, getting rid of the "all too human" behavior of exaggerating acceleration followed by heavy breaking for instance between traffic lights and crossroads in inner city areas (Alexander-Kearns et al., 2016). A fleet of autonomous cars could even, due to the predictable driving patterns, be coordinated to avoid busy roads or, if necessary, to adjust speeds in such a way that traffic jams will be prevented. This does not only foster energy saving due to the avoidance of stop-and-go driving but also increases commuters spare time – a point I will address later. Additionally, fuel can be saved by complete design overhauls of the car. Even though the additional sensors will initially increase the weight of a self-driving car, eventually manufacturers could redesign the cars and get rid of the crumple zone, airbags or other safety devices effectively decreasing the weight and hence consumption. Some companies like Tesla and GM try to link two trends in car development: autonomous driving and electrical driving. While efforts around the world to decrease reliance on fossil fuels are taking off, the usage of electric cars still lacks behind considerably. One reason often cited is the low range of electric cars and ergo the constant need to recharge the battery.

Because the infrastructure of charging sites is still underdeveloped, many people rather opt for a traditional engine or a hybrid alternative. Yet, autonomous cars could foster the acceptance of electric vehicles. An obvious reason is that self-driving cars could take the recharging process out of its owner's hands by looking for possibilities to recharge whenever the car is not needed and autonomously return when called to duty. This will help reduce carbon emissions in general and those that are produced by older Diesel engines in particular. According to Caiazzo et al. (2013) air pollution is responsible for the deaths of around 200.000 people in the United States with vehicle emissions as the biggest contributor. Reducing the number of this indirect danger would certainly be an improvement. It does not end there, however. Business models like car sharing are gaining popularity as can be seen by services like Uber pool [*]. As the fleet of autonomous cars grows, car sharing will become more convenient as well as cheaper compared to actually owning one's own car (see figure 1.1). Simulations for the cities of Berlin, Austin and Lisbon indicate that one shared autonomous car could replace up to ten conventional cars that are individually owned and used (cf., OECD (2015), Fagnant and Kockelman (2016) and Bischoff and Maciejewski (2016)). All else being equal car sharing also, obviously, reduces energy consumption – regardless of whether a car relies on fossil fuels or is powered by electricity.

Second, the introduction of autonomous cars will foster participation in traffic and reduce the amount of wasted time during commutes. A level-5 autonomous car will per definition not require any involvement of a human "driver". Therefore, there is no more need to exclude certain groups of people from using a car and constrain their mobility. These groups include people with no license, but also those that have been excluded from partaking in traffic for understandable reasons, e.g. children, people with disabilities like blindness or other restrictions

[*] Basically, Uber pool is a taxi service that allows customers who need to go to the same area to share the ride and pay a lower price. Today, this service is active in major cities.

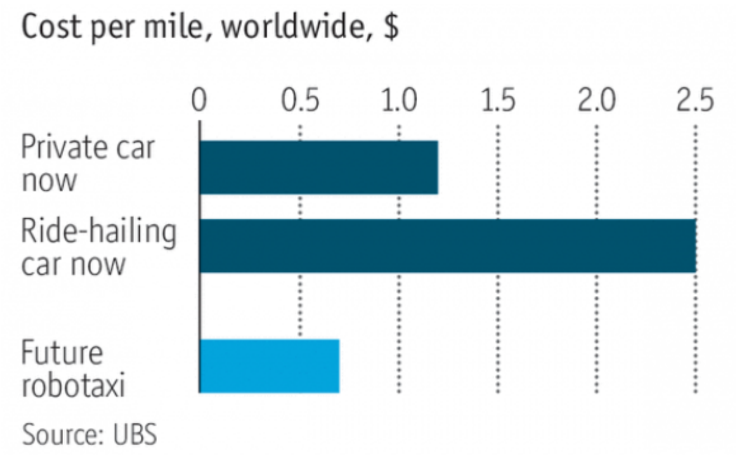
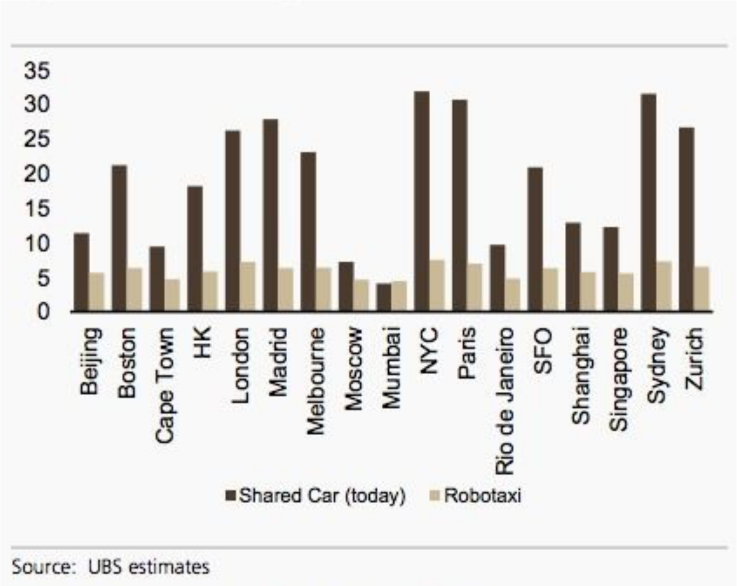


Figure 1.1: Cost per mile in $

like mental disabilities (Howard, 2013). It is hard to think of another technology that has the ability to enrich handicapped people's lives in such a profound way. Apart from the impaired the other group that will benefit from the introduction of autonomous cars clearly are commuters. According to the Statistical Office of Germany, about 68% of those fully employed use their personal car to commute between their home and their place of work. Roughly 30% of those spend more than 30 minutes on the road (Statistisches Bundesamt, 2017). Commuting is linked to stress and other illnesses due to the amount of time spent driving, blocked roads and an overall increase of the time that is spent away from home. Autonomous cars offer great potential to better this situation. Not only does it eliminate the task of driving which – especially in heavy traffic – can be mentally taxing on the driver, they also allow former drivers to use the time they spend commuting more effectively. Whether this time is spent for recreational purposes like reading or watching tv or used for work, e.g. answering emails etc. before the arrival at the office. It is not far-fetched to believe that autonomous cars may offer its passengers the possibility for physical training, for instance by offering spinning bikes that can be used during the commute. This also applies to people living in rural areas who want to make



Source: UBS estimates

Figure 1.2: Cost of shared cars vs. robotaxis in different cities

use of the services urban areas can offer, e.g. cultural events. According to Swiss bank UBS the total costs of commuting will decrease with so-called 'robotaxis', thereby saving resources (see figure 1.2).

Finally, there are benefits associated with the introduction of autonomous cars that will take longer to materialize, yet, will have a meaningful impact on everyday life. These benefits include the possibility to completely redesign urban areas since today's cities are planned in such a way as to enable traffic which includes the appropriation of sufficient parking space etc. Autonomous cars, however, do not need to park in urban areas but can be sent outside and called again when needed. Cities will be able to use the freed space that had been used for street parking to either increase traffic flow by increasing the number of lanes or even dedicate the space for new building projects. The demand for urban living has gone up in recent years, leading to an increase in housing and rent prices. Autonomous cars could, therefore, help to free valuable properties in order to construct living space as parking lots or parking garages which by design take up a lot of space will no longer be needed.

Autonomous cars, as any new technology, may also pose new ethical questions and challenges. The most debated are mainly discussed in the next subsection – the ethics of crashing. However, there are some issues that belong to the category of broader ethical issues, which I will present here. First, autonomous cars will receive and gather a lot of data. In fact, the software of driverless cars uses machine learning to increase its knowledge about the world and its surroundings, gathers feedback and implements features on its own (Kuderer et al., 2015). Additionally, cars will also have the need to communicate with each other and even with a central authority if one considers traffic planning in order to avoid traffic jams. This raises privacy concerns – especially in European countries that are traditionally more strict with the privacy of data. It is not hard to imagine that an autonomous car, that has access to large amounts of personal data, can choose a route that explicitly exposes its user to things she adores for commercial purposes, e.g. always driving past her favorite fast food chain (Lin, 2014). Some argue that the potential benefits of the new technology would dwarf any privacy concerns that are associated with a constant GPS/cellular connection of their cars (Lee, 2013). The analogy of a cell phone is often used to dissolve any concerns, claiming that we already live in a world of no privacy due to our dependence on our mobile devices. Yet, there is an important difference between the two cases. While I can easily leave my phone behind or turn it off whenever I want to travel without being potentially tracked (by advertisers or the government), it will probably be impossible to turn off the internet connection of an autonomous car because the technology will depend on constant information via these networks. Regulators have started to weigh in on this issue. In California, for instance, lawmakers already demand that car manufacturers "provide a written disclosure to the purchaser of an autonomous vehicle that describes what information is collected by the autonomous technology equipped on the vehicle" (cf., CAL. VEH. CODE §38750-b). Certainly, more regulations will be put in place in the future. But whether this might be an actual solution is doubtful. Citizens will have to trust in the regulator's ability to enforce these laws and in companies to comply. Past data breach scandals, e.g. Facebook in 2018 regarding the Cambridge Analytica case, have shed further doubt on the security of personal data. Boeglin (2015) argues to "only allow autonomous vehicles to infringe on user freedom and privacy to the extent that (1) reductions in freedom and privacy lead to equivalent reductions in liability for the users of self-driving cars; and (2) the social costs incurred by forfeiting these values will be outweighed by administrative efficiencies or other identifiable social benefits." While this seems like a reasonable approach, the difficult part still lies ahead of us. To figure out the exact balance of this cost to benefit relationship will demand public discourse and gives room to future scientific exploration. This debate mirrors the traditional trade-offs between freedom and security. In the light of recent events, in which vehicles have been used to execute terrorist attacks, e.g. Nice, Berlin and London, authorities might find it useful to create the ability to remotely control any vehicle when they see fit. In fact, plans to introduce such a remote control feature are already on the table (Council of the European Union, 2013). The ethical challenge here is obvious because even if the people were to trust their governments with this power, it would be naive to expect that this system would be safe from malicious hacking attempts. As a matter of fact, there has not been a single network-based complex technology that has not been successfully hacked. Even time-tested systems like those employed by credit card companies have shown severe vulnerabilities in the past (cf., Daily Mail (2012)). This well-intended proposal may even backfire, allowing the cyber-hijacking of cars by thieves or, even worse, terrorists (Lin, 2016). A hacking attack could even be performed with the goal to cripple the traffic of an entire city or country as means of cyber warfare, causing negative economic effects and bringing public life to a halt.

Second, the introduction of autonomous cars will have severe consequences for a lot of industries and organizations. The (car-)insurance industry has evolved for the sole reason that there are, in fact, car accidents that cause damage or injuries for which compensation is demanded. If it turns out to be true that self-driving vehicles will hardly ever crash, there might be no need to ensure a car any longer or at least not at the level of premiums that we experience today. Auditor KPMG estimates that the market for car insurance-related products will decline by 60% within the next 25 years (KPMG, 2015). Another group of people that will be at a disadvantage are people whose jobs rely on driving, e.g. taxi drivers, truck drivers and maybe even people in packet delivery service jobs. Every revolutionary technology that increased production has seen some of these negative effects. The introduction of personal computing reduced the need for secretary services, email-technology decreased the numbers of physical letters that need to be sent and the introduction of automated production machinery has eliminated many blue-collar jobs.

1.2 Ethics of Crashing

Certainly, the most discussed ethical issue when it comes to autonomous cars is their potential reaction to so-called "dilemma situations" – that is a situation in which whatever decision will be made, negative consequences will follow. The first publication on which this dissertation is built deals with this prominent and highly debated issue (see, chapter A on page 46). Two thoughts are vital to understanding if one wants to grasp the appeal of the discussion. First, the trolley case, a decade old thought experiment on which this discussion is based, illustrates nicely the differences between ethical schools and has therefore enabled debate and research along the lines of dilemmas. Second, this thought experiment has now – or so it seems – reached a point in which it might turn into a realistic scenario. Humans, who are involved in accidents, make split second decisions. Due to the short time frame, these decisions are usually not motivated by ethical considerations but by pure instinct. Autonomous cars, however, will have the possibility to actually behave according to preset ethical standards and can probably follow through on their programming in a fraction of a second. *Prima facie* this is a strong case for an ethical analysis of (1) *how* a car should react, and (2) *who* should decide how a car should react. The discussion so far has been mostly focused on the former.

Before we move into the discussion of how to program an autonomous car, I will first give a brief overview of the trolley case and its application to the case of highly automated driving. The trolley case was first introduced by Phillipa Foot in the late 1960s (Foot, 1967) and has since created a vast amount of literature thus sometimes dubbed "trolleyology". The basic formulation of this well-known dilemma is as follows: "To make the parallel as close as possible it may rather be supposed that he is the driver of a runaway tram which he can only steer from one narrow track on to another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed. [...] [T]he exchange is supposed to be one man's life for the lives of five." (ibid.) This original formulation has been modified multiple times to investigate what intuitions might be at work when we judge the trolley case. The most famous variation is the so-called "fat man" scenario in which it is necessary to push a heavy man off a bridge to stop the trolley, rather than switching a lever to steer the train. In this dissertation, I am not too much concerned with all the different variations and Edmonds (2013) provides a thorough overview of the discussion. The two mainly discussed scenarios – classic and the *fat man* – offer interesting insights (cf., (Greene, 2014)). Additionally, I will introduce a third scenario later because it offers some insights into the case of autonomous cars. In the two former cases, however, most people have the intuition to pull the lever and save the five workmen in the classic case, the opposite is true when it comes to pushing the *fat man* down the bridge to stop the trolley. Even though the consequences are essentially the same (save 5, sacrifice 1) most people evaluate the two cases differently. In fact, when asked about the two cases people overwhelmingly opt to save the five workmen in the first case (about 90 %), while only 10% of people agreed to push the fat man in the second scenario (Hauser, 2006) (cf., Sokol (2006) who conducted an online survey which comes to similar, yet, slightly less clear results of around 75% agreement and objection, respectively). The curious task is then to explain why this difference in intuitive evaluation occurs. The most prominent answer is usually given by the "doctrine of double effect" – a concept that goes back to Thomas Aquinas. This principle is often invoked to explain the permissibility of an action that causes a serious harm as long as it is merely a side effect of promoting some good end (McIntyre, 2014). Following Mangan (1949) one can identify four conditions that need to be fulfilled for an action that causes a good as well as a bad effect to be justified by this principle: First, that the action in itself must be good. Second, that the good effect and not the evil effect is actually intended and, third, that the good effect is not produced by means of the bad effect. Finally, that there is a proportionately important reason for permitting the bad effect. The "doctrine of double effect" has the potential to explain why we evaluate the two cases differently. While in the standard case we merely sacrifice the one person as a side effect, in fact, we wish the other track was empty, we do need the fat man as a *means* to achieve our goal of stopping the runaway trolley. In the *fat man* case, we produce the good effect by the means of the bad effect, therefore violating the third condition. In order for the transition to trolley problems and autonomous cars to be complete, there is one more variation that should be taken into account. Thomson (2008) introduces a third option to the standard case in which the bystander can pull the lever. Now the bystander has three options: (1) do nothing and 5 workmen die, (2) pull the lever and redirect the trolley to hit a single workman (standard case) and (3) redirect the trolley onto himself or herself, thereby sacrificing his or her own life in order to save the five. Thomson argues that we cannot possibly justify (2) if we are not also willing to do (3), saying: "Since he wouldn't himself pay the cost of his good deed if he could pay it, there is no way in which he can decently regard himself as entitled to make someone else pay it" (ibid.). The discussion about trolley cases is an ongoing debate and I will shed some more light on it in chapter 2, where I will talk about empirical findings in different cases and their (possible) implications.

Given the scenarios described above, one can immediately see how the trolley case might translate into a dilemma of traffic in general and of the autonomous vehicle in particular. First, the dilemma already incorporates a traffic scenario. Sure, a trolley is constricted in its movements by the tracks that it is on but this is merely a minor detail. Second, a bystander makes a decision about who is to live and who is to die (pull the lever, push the fat man etc.). Analogously, in the case of autonomous cars, an algorithm will decide what action to take. Certainly, there are some aspects in which trolley cases might not lend themselves for ethical analysis regarding the case of autonomous cars. In fact, the first publication on which this dissertation is built argues that trolley problems lack three important features that would be necessary to adequately address this issue. I will discuss these objections briefly at the end. Since road vehicles are not constricted by tracks but can move rather freely within two dimensions, the trolley problem, if applied to road vehicles, must be altered a bit. These alterations are often called the "tunnel problem" or the "bridge case" (cf., Millar (2014), Goodall (2014)). The first uses a scenario in which an autonomous vehicle is about to enter a tunnel when suddenly an obstacle (usually a playing child) is standing on the right lane in front of the car. Oncoming traffic is approaching on the other lane. This leaves the car with three choices: crash into the obstacle, swerve into oncoming traffic or drive itself against the wall. Obviously, all choices lead to unwanted consequences. The bridge case is similar but the third option is for the car to drive itself off the bridge. Depending on the purpose, these scenarios can be modified. It can be a school bus that is in front of the autonomous car that suddenly brakes or some other form of obstacle. Oncoming traffic might be a motorcycle or a van. The specifics are merely details. What is vital to understand is that whatever action is chosen a bad outcome will unfold. The interesting part, as mentioned above, is that when a human driver finds herself in a situation like this, she would simply react. Maybe she will just break and hit the object in front, maybe she will try to swerve in either direction. Whatever choice she makes, it would not be a morally relevant choice for she merely reacts in a split second and no deliberation takes place. The main idea and also what makes it so interesting to philosophers is that with the introduction of autonomous cars we could potentially now make informed judgments about what to do. Given enough data about all participating parties in this dilemma, the car could potentially minimize human deaths, maximizing life years (saving younger rather than older people, healthy over sick etc.) or minimize total damage (including equipment). Granted, situations like *tunnel* and *bridge* are highly unlikely – maybe a dilemma situation like this occurs only once every ten years (or never). With all the lives being saved by autonomous cars one might think why would this ethical discussion even matter at all? But this argument misses the point. Even if a situation like this happens only once in a lifetime the autonomous car will act according to its algorithm by default. This could be to always break as hard as possible and hope for the best. In any case, an action will be programmed into the car – it cannot just dissolve into thin air to avoid the dilemma. Any society who claims to care about the welfare of its members should discuss how such a situation should be dealt with. In other words: We cannot have no preprogrammed behavior in a car, therefore the discussion about what behavior this should be matters.

Who to target?

In scenarios like *tunnel* or *bridge*, there are limited options which each carry unwanted consequences. Nevertheless, a decision must be made. Thought experiments are often used to elicit peoples' intuitions while changing the parameters of the scenario. In the case of autonomous cars, this is usually the type of person who occupies a certain role in the dilemma. The question then becomes: Given the dilemmatic situation what choice would be the least objectable? Assume that one has the choice to break and crash into a school bus, sverwing into oncoming traffic or to sacrifice oneself. A strategy like maximizing life seems *prima facie* to be a good starting point. If there are ten school kids at risk and the car in oncoming traffic carries a family, then someone who is alone in her car should have her car opt for the self-sacrificing option. The problem here lies in the trading of lives. Consider a similar case: A man enters a hospital because he has broken his leg, which – though painful – is definitely a non-life-threatening injury. At the same time there are five people in the hospital who have been waiting for an organ donation but have not received any and now their time is running out. The doctor could kill the patient with the broken leg and harvest his organs in order to save the five people who had been on the waiting list to receive an organ (Thomson, 1985). Yet, this action seems to be unjustifiable. To always choose to save the greatest number of lives does not hold up to scrutiny if it is only used in a one-shot scenario. One could easily modify the scenario and replace the school bus with a bus of convicted murderers. In fact, the German constitution (other countries have similar laws in place. The 14th Amendment to the Constitution of the United States of America guarantees equal rights and due process) recognizes the value and dignity of every person regardless of any specific attributes like age, gender or religion. Moving away from life-and-death situations, targeting in such a way as to minimize damage might also lead to perverse results. At first glance, it seems obvious that a car, provided with the data, should orchestrate an unavoidable crash with the sole aim to minimize harm *ceteris paribus*. But this would mean to crash into the motorcyclist wearing a helmet, rather than into the one who did not care for his safety and thus is not wearing protection. The helmed cyclist does have a higher chance of survival (lower severity of injury) after all. Another example would be to crash into the SUV with higher safety standards than into a cheap tiny car. This sets perverted incentives because people who actually cared for their safety (by wearing a helmet and spending more on a safer car) will be at a disadvantage regarding their safety. If people were to believe that the probability to be involved in a *tunnel-like* scenario is sufficiently high, a race to the bottom would be initiated. If people rush to buy unsafe cars and engage in reckless behavior only to not be targeted would lead into a paradox situation. If we only look at a dilemma situation as a single event, there will be no chance of an agreement. Everybody has their own ethical approach and rightly so: Reasonable disagreement regarding ethical judgments always exists in a free society. When people have an equal right to not be harmed, but the situation is such that somebody has to be hurt, a random draw might be an alternative. Chance does not discriminate in a morally relevant way. While some would certainly welcome this solution – especially economists who value procedural justice the highest – it seems contra-intuitive to most people to leave a life-and-death decision to pure chance. Whatever scenarios we construct there will be no single right answer – at least as long as we treat such trolley-like scenarios as an isolated one-shot game. If there is no right answer to how to solve dilemma situations, yet, they need to be addressed nonetheless, the question then becomes: *Who* should decide on the ethics-setting of autonomous vehicles? The first publication on which this dissertation is built deals with this very question, arguing in favor of a mandatory ethic setting that has to be implemented in every vehicle, rather than a personal setting which could be determined by the individual owner of the car (cf., Gogoll and Müller (2017) in Appendix A). If we think about traffic as an institution that provides great benefits while – at the same time – exposes everyone to a small risk of getting injured or killed, then we understand that the proper way to argue about dilemma situations is a systemic question. Traffic is an "equitable social system of risk-taking" (Hansson, 2003) that we accept because it offers enormous upsides, yet, some people pay the ultimate price. The question of risk-taking is thus not a static one: Every time we step foot outside our house, drive in our cars or sit on a bus there is a tiny chance of being involved in an accident. We do not explicitly make this choice every day, but we make it nonetheless. The static trolley situation, however, models only a one-shot scenario. Who has to die in this specific case? Under these peculiar circumstances? But the reality is more complex than this. In Gogoll and Müller (2017) I make the case that the trolley scenario is, in fact, not a good fit for modeling possible dilemma situations regarding autonomous cars. Trolley cases lack three important characteristics that are essential to understanding traffic as a system of risk-taking: Strategic interaction, iteration and the fact that everybody can be a subject as well as an object of targeting.

Fanatics may suppose, that dominion is founded on grace, and that saints alone inherit the earth; but the civil magistrate very justly puts these sublime theorists on the same footing as common robbers, and teaches them by the severest discipline, that a rule, which, in speculation, may seem the most advantageous to society, may yet be found, in practice, totally pernicious and destructive.

David Hume

2

Why Ethics has an Empirical Dimension

It seems odd that one has to justify why empirical evidence should play a role in moral philosophy. It is even more bewildering that the subfield of ethics, that evaluates, steers and deals with the actions of real people in often times complex situations should turn a blind eye to empirical analysis. Philosophical and ethical inquiry have, after all, been intertwined for the most part of scientific history and no distinction had been made between philosophy and the (natural) sciences. In fact, Isaac Newton's theory of gravity is published in a book with the name "Philosophiæ Naturalis Principia Mathematica" – literally referring to Mathematical Principles of a *Natural Philosophy*. Due to the specialization of science this former joint project of understanding has split up into different 'sciences' that usually have a clear and well-defined domain and only rarely interact. When the natural sciences split from philosophy followed by the social sciences sometime later, philosophy mainly became an armchair science, relying on thought experiments, logic and normative argument. In recent decades, however, there has been a movement to the opposite. Empirical sciences have started to tackle questions that belong in the moral domain and provided an understanding of people's behavior. Specifically, the research done in fields like experimental or behavioral economics and moral psychology have had major influence even on classical philosophical theorists. One example certainly is John Rawls who, in his *opus magnum* "Theory of Justice", builds part of his main argument on insights from economics and psychology (c.f., Rawls (1971) and his reliance on risk aversion in order to obtain his 'difference principle'). Recently, a different trend in philosophy – experimental philosophy – has gained momentum as well. Experimental philosophers try to elicit so-called *folk intuitions* by using the method of a vignette study. They use a (thought-) experimental approach describing hypothetical scenarios and altering parts of it to get a treatment design. Interesting research has emerged from this line of research, e.g. Knobe (2003).

As noted above, ethics, especially in philosophy, is usually seen as a normative endeavor. It is concerned with the question of how people *ought* to act rather than how people actually behave. The main reason why philosophers have mostly stayed clear of using empirical findings is the well-known argument by David Hume that one cannot derive an *ought* from an *is*. This argument is known as the 'naturalistic fallacy' (Hume, 1978). This means that one cannot infer a normative claim from a descriptive empirical observation. Normative claims are the conclusion of normative arguments and therefore have to rely on normative premises. But in order to reach a normative conclusion, one has to inject a normative claim in the premises of an argument because in logic an argument is only valid if the conclusion follows from the premises. To start an inquiry with a certain claim already built in, however, is rejected by most modern scientists. There are ways to overcome this fallacy (cf., (Jauernig, 2017) for an overview). One way to get empirics back into ethics is provided via Immanuel Kant who stated that "ought implies can" (Kant, 2003). The question of whether someone is able to act according to a normative prescription is essentially an empirical one. Homann (2015) identifies a three-step process of how ethical convictions can manifest into moral action:

1 Justification/good reasons

2 The will to act morally (motivation to do so)

3 Moral action

Two transitions are necessary to get from step 1 to 2 and from 2 to 3, respectively. After the justification people need to have an insight into the good reasons in order to produce the motivation to form the will to act morally, which in turn then motivates the person to act accordingly, that is to perform the moral action. In this model, Homann (2015) argues, any violation of morality is attributed to either a lack of insight or a bad or at least weak will to perform the morally prescribed task. Given the vast amount of immoral actions, it appears that the constant plea for more morality does not seem to be very efficient. Therefore, Homann proposes a different approach. He argues that morality is enabled as well as constrained by certain empirical and contextual conditions that – at least sometimes – make acting in a moral way hard or even almost impossible. The empirical (social) sciences, Homann claims, can help unravel these conditions and shed some light on structures that prohibit moral conduct in an effort to overcome unsatisfying structures in society.

Let us see what is meant by empirical conditions and why people might find it hard to act according to what they themselves believe to be the right way. A normative claim that demands the impossible cannot be valid. Imagine for instance the normative command that you ought to walk on water. For all but one human in history this is an impossible task and whether the one person actually could do it, is highly debatable. Granted, this is an extremely obvious case, one that does not need scientific research to be uncovered. But the 'can' in Kant's statement is not necessarily restricted to the logical or physical possibility. Rather, there is a myriad of factors in each situation that influence human decision making. We live in a complex world and before we make a decision a lot of information needs to be processed. To make the 'right' decision could sometimes lead to what economists call prohibitively high search costs which describes the act of collecting data to the point of running out of time and resources when the costs start to outweigh the additional (or marginal) benefits. But even if we assume that we could collect all the relevant data, there is the problem of accurately processing them to be a useful tool in the decision-making process. We are, after all, natural creatures who have been subject to evolutionary sorting processes for over 2 million years. Most of this time has been spent in small ancestral bands and relatively plain surroundings. Consider decision theory, which is also a normative approach. It offers tools to make rational decisions. Mathematics tells us how to deal with exponential growth, for example. Yet, our minds are incapable to intuitively grasp the concept of non-linear growth. This is certainly a limitation with regards to our 'can', even though it is not impossible for us to understand it – mathematics is a product of the human mind after all. However, according to an early study of Wagenaar and Sagaria (1975), neither mathematical sophistication nor the experience of subjects could eliminate the fact that they underestimated exponential growth. Coming back to ethical considerations: People often times have strong normative moral beliefs, yet, they are unable to follow through with the appropriate action. More often than not do people experience a substantial gap between their stated intent and their performed action. One reason for this is that people regularly overestimate the robustness of their moral conduct. To simply say that people are either not really committed to their moral convictions or just lack the willpower to follow through with them is a simplification that does not allow to address the problems that might arise from this gap. It is also not the case that most people are evil and deliberately ignore their convictions. Rather, we find ourselves in a world of incentives and structures that often make it hard to act according to our ethical beliefs. Empirical research plays an important role in identifying structural characteristics of situations that make it harder for people to behave the way they intended due to unfavorable incentives. Altering these structures by changing the incentives is a task that economists, who are interested in ethical questions, are especially fond of.

Another point in which ethics relies on empirical data is the following: In the first chapter I talked about the ethical issues regarding autonomous cars. Even though some of these ethical challenges relied on particular states of the empirical world, e.g. whether autonomous cars *actually* decrease traffic deaths, they were normative in nature. Whenever a specific action is normatively prescribed references to the empirical world are a necessity. After all, they address humans who act under empirical constraints and ignoring these constraints would make the entire normative claim useless. Consider the following argument:

P1 Autonomous cars will reduce deaths in traffic

P2 It is better to save more lives than fewer lives, all other things being equal

---

C Autonomous cars are an improvement of the status quo and their introduction should be fostered

Whether P1 holds or not can only be decided by using observational data. P2 introduces a normative claim and enables the following normative conclusion C. The argument is *valid* because the conclusion follows from its premises. To determine whether the argument is *true*, however, empirical research is indispensable.

2.1 Trolleys reloaded: Empirics to the rescue?

Since the introduction of autonomous cars is a relevant topic in many scientific fields and the discussion of trolley problems is no longer an exclusively philosophical endeavor, a lot of empirical research has emerged. I talked about the different issues associated with implementing a decision rule for these cases. If we are unable to agree on the ethically correct way to solve these dilemma situations, collecting folk intuitions may shed some light on why this is the case. Furthermore, it is interesting to see what factors may influence peoples' decisions about trolley cases. Ethical theorizing often starts with intuitions about a particular case and what principles might underlie these intuitions. Kahane (2013) argues that empirical research "exposes psychological presuppositions implicit in armchair ethical theorizing" and therefore "if our moral intuitions are reliable, then psychological evidence should play a surprisingly significant role in the justification of moral principles." (ibid.). There have been some attempts to illuminate why and when people disagree on the proper course of action an autonomous car should take. These studies are not normative but rather descriptive, trying to answer the question: How do people think they want to react to the situation given a set of particular circumstances?* To uncover patterns and dispositions that can explain reactions to different scenarios might indeed increase our understanding of the psychological mechanisms at work. The studies that have been published thus far use the methods of vignette studies or surveys, but also advanced methods from neuro-psychology. The classical trolley problem has been the topic of empirical research for a couple of years. Greene et al. (2001) use functional magnetic resonance imaging (fMRI) to see how subjects' brains react to the impersonal case (pulling the lever) and the personal case (pushing the fat man down a bridge). While the personal case activates brain regions that are

*I phrased it this way because what people answer in these studies might not be their true assessment of the situation. A problem I will talk about at the end of this chapter.

responsible for emotional reactions, the impersonal cases activate the regions that are associated with controlled and logical reasoning (see also Greene (2014)). Hence, Greene et al. (2004) put forward a dual process theory of decision making: The first is based on emotions and more in line with deontological reasoning, while the other is a reasoning-based system that is often in line with utilitarian ethics. Ciaramelli et al. (2007) study the effects of certain types of brain damage and the subsequent reactions to trolley cases, finding that patients who suffered from lesions in the ventromedial prefrontal cortex, an area of the brain associated with emotions, are more likely to opt for the maximizing lives strategy in personal cases (*fat man scenario*) as opposed to a healthy control group. Experimental philosophers have used the trolley case in various vignette studies to shed light on seemingly irrelevant factors that influence peoples' decisions (Liao et al., 2012). Ahlenius and Tännsjö (2012), for instance, show that intuitions differ between cultures – testing western and Chinese subjects. Another important task of empirical research in ethics is to sensitize for situational factors and contextual cues since they can have a huge impact on moral reasoning. Bleske-Rechek et al. (2010) looked for an evolutionary explanation and manipulated the sex, age, genetic relatedness and potential reproductive opportunity of the one person tied to the track and found that subjects were "less likely to sacrifice one life for five lives if the one hypothetical life was young, a genetic relative, or a current mate" (ibid.). Framing effects also play an important role when people are to judge what path the trolley should take. Petrinovich and O'Neill (1996) show that the wording of the dilemma influences subjects' decision making. People are more likely to "save" five workmen than to "kill" the single workman on the other track, even though the outcome is essentially the same. Manipulating the degree of personal involvement also alters decision making, the more personal the dilemma is framed, the fewer utilitarian choices are made (Tassy et al., 2013).

In experimental philosophy, researchers rely on paper-and-pencil surveys to elicit peoples' opinions about how to decide in trolley scenarios (cf. Huebner and Hauser (2011), Hauser (2006), Di Nucci (2013)). They are interested in folk intuitions and especially whether there is a structural difference between these folk intuitions and expert's intuitions. In general these findings, too, find a strong tendency for utilitarian answers. However, when they added the third option in the trolley scenario – self-sacrifice (see chapter 1) – interesting findings occur. In Huebner and Hauser (2011) 38% of subjects opted for self-sacrifice and Faulhaber et al. (2017) even find more than 50% of subjects are willing to self-sacrifice in the autonomous car case. Di Nucci (2013) found that if subjects are confronted with the trilemma first (that is: sacrifice either one workman or yourself to save the five workmen, the trolley is heading towards the five workmen by default.) they will subsequently change their assessment of the standard trolley case with more than 60% saying that it is not right to pull the lever to save the five. Intuitions, so it seems, fluctuate and are easily altered by situational factors like *order effects* and *salience of wording*. I will discuss the shortcomings of this line of research at the end of this section.

The applied trolley case of autonomous driving has also been subject to research. Generally, a subject is put in the hypothetical situation of a bridge or tunnel case (see, chapter 1). The subject is then asked to decide how she would prefer the autonomous car to react in each case. Finally, the case is altered each time regarding the nature of the other vehicles or the specific people involved (see, 2.1). Brandom (2016) states that this approach "would be surveying the public on who they would most like to see hit by a car, and then instructing cars that it's less of a problem to hit those people" which he argues would be "horrific". Yet, it is just a fact of life that we sometimes face dilemmatic situations that force us to make such horrific trade-offs. First responders with limited resources face problems like this all the time. To elicit folk intuitions about trolley cases, therefore, might be a useful approach after all. The MIT Media Lab has launched a website which allows the public to voice their opinion about what they feel the appropriate decision in each case is (see, http://moralmachine.mit.edu). This project is based on Bonnefon et al. (2016) who show in three conducted studies that laypersons express a general tendency for utilitarian solutions to trolley scenarios. Note that these findings should not be interpreted as a normative claim, saying that we, in fact, *should* have utilitarian ethic settings for autonomous cars. The argument of Bonnefon et al. (2016) is rather that on a political level people might be willing to accept a law that solves dilemma situations with a utilitarian decision principle – the question of acceptance plays an important role regarding the introduction of autonomous cars and will be discussed in subsection 2.2. An important finding of Bonnefon et al. (2016) is that the preference of the utilitarian rationale only holds up as long as the subjects themselves are not the ones who will be targeted. This may inform normative considerations about whether the trolley case is the appropriate scenario that we should look at when considering autonomous cars in the system of traffic. As is argued in the first publication of this dissertation the trolley case lacks the important feature that one can be the object as well as the subject of targeting (Gogoll and Müller, 2017).

Another approach that has been taken is the use of virtual reality to simulate trolley cases for autonomous cars. In classic trolley scenarios Navarrete et al. (2012) and Skulmowski et al. (2014) show that when people are put in virtual reality scenarios, their answers to the dilemma are in line with the survey data from vignette studies. Sütfeld et al. (2017) used simulated road traffic scenarios in which participants controlled a virtual car. Subjects had to choose which of two given obstacles they would sacrifice to save the other. They used random samples from a variety of inanimate objects, animals and humans and looked for consistent patterns in peoples' assessment. Additionally, they manipulated the time subjects were given to make the decision. Suter and Hertwig (2011) found that the more time is available to deliberate on the decision, the more likely it is that a given subject arrives at a utilitarian standpoint.

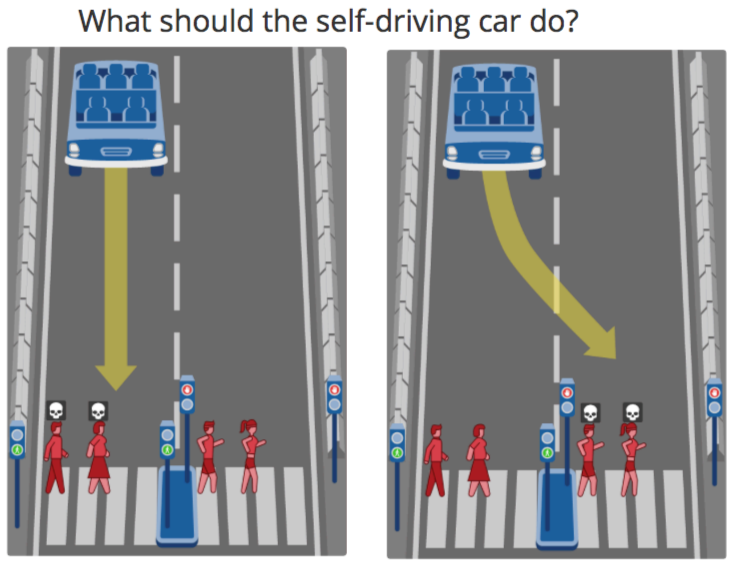


Figure 2.1: Example screenshot of "the moral machine"-project, source: http://moralmachine.mit.edu

Faulhaber et al. (2017) support these claims, finding that a vast majority opts for a utilitarian decision in trolley cases. Scientists who use VR technology to research subjects' reactions to trolley dilemma situations often claim that the more realistic approach has several benefits over a paper-and-pencil questionnaire, claiming that the latter often ignore contextual and situational cues.

There are, however, some methodological issues one needs to take into consideration when interpreting the results of the studies above. Granted, these studies do not make any normative claims, but to be a useful tool in determining what behavior people have towards trolley dilemmas it is crucial that we can elicit peoples' *true* preferences on how the autonomous vehicle should react. Some of these studies present data that seem implausible. Empirical research in the social sciences deals with human action, which is complicated since human beings have all kinds of reasons for their actions – some of which they are aware of and others that are unconscious. Well established phenomena like the *social desirability bias* or the *experimenter demand*

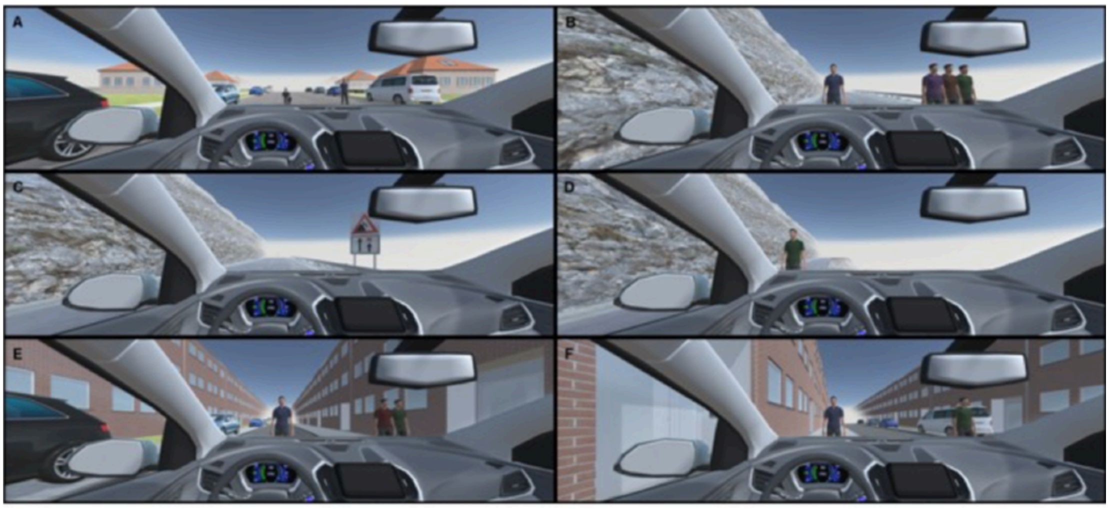


Figure 2.2: Illustration of the virtual reality design used by Faulhaber et al. (2017)

effect cast some doubt on the validity of these findings. While this criticism does not so much apply to the fMRI studies – after all, it is very hard if not impossible to consciously alter which part of the brain makes a decision –, it does certainly apply to questionnaires. This is especially problematic if the questions refer to hypothetical situations. Take, for instance, the findings on self-sacrifice: Huebner and Hauser (2011) report that 38% of subjects chose the option to sacrifice themselves in the hypothetical situation of a trilemma trolley case. Faulhaber et al. (2017) even find that roughly half of the subjects opted for self-sacrifice if they can only save one other person. The number goes up to around 60% when the lives of six people can be saved. These findings are as astonishing as they are problematic. It seems obvious that these findings are completely in conflict with common experiences about human nature. Yet, Faulhaber et al. (2017) simply claim that "the results of this module, therefore indicate that people are still acting in favor of the quantitative greater good even when their own life is at stake." If the data contradicts what some people might call "human nature" one would expect that the methods used in eliciting the data might be questioned and the findings not just simply be taken at face value. Huebner and Hauser (2011) do just that by acknowledging that "altruistic self-sacrifice is rare, supererogatory, and not to be expected of any rational agent." After all, why would armies give out medals of honor and how could these medals improve one's social status if it were the case that every other person would behave the same way anyway? It seems more likely that subjects chose the option of self-sacrifice because they felt that this would be the socially desired answer. Since the subjects – obviously – did not actually have to bear the consequences of their actions this seems plausible. Economists refer to that phenomenon as "cheap talk": Subjects thus were able to keep their self-image of being an outstanding moral person, yet, they achieved this goal at no costs. What does this mean for empirical studies of the trolley dilemma? One solution would be to implement some sort of incentives to the situation. It is impossible to implement the actual trolley dilemma in an experiment since this would entail life or death decision about actual human beings. But a milder version could pave the way to a more accurate assessment of the situation. Gold et al. (2014) implement the trolley situation using a game show paradigm. Subjects had to answer questions to a general knowledge quiz and earned money for every correctly answered question. As soon as six players were above a certain threshold the game stopped and the other players were told that five players were about to be knocked out of the game, losing all their earnings. However, there is a button that could be activated and would randomly select another single player who is above the threshold and destroy his or her earnings instead of the other five. In this setting, the decision actually produces consequences for the players involved in the game. Further research is necessary, yet, the prediction seems plausible that the possibility of self-sacrifice (the option to destroy one's own earnings while saving all six of the other players) will probably not be chosen by half of the subjects – even with somewhat low stakes, as were used in the game show scenario.

Another issue with these studies is that subjects may perceive them to be highly artificial up to a point where it influences their intuitions about the situation. See figure 2.2 for instance. In the upper right screenshot, one can see a total of five people just standing in the middle of the street. A single person on the left lane and a group of three on the right. Subjects could interpret the situation differently. Some may realize that this is an abstract situation and their decision is actually based on the numbers of lives to be saved as was intended by the researchers. Yet, we cannot know this with certainty. In fact, it seems plausible that other reasons might influence their decision. For instance, subjects may have the feeling that it is irresponsible to walk on a street next to each other and that this reckless behavior and the risk associated with it makes them less "worthy" of being saved. In other words: Many factors influence our moral intuitions. Statements like "people prefer utilitarian cars" should, therefore, be taken with a grain of salt.

2.2 Acceptance, Trust and Aversion

The experimental economist Alvin Roth once said that experimental research in economics serves three purposes: First, the testing and modifying of formal economic theories which he calls "Speaking to Theorists", second, inquiries that may have a direct impact on the policy-making process, which he calls "Whispering in the Ears of Princess", and, finally, to collect data on interesting phenomena and institutions, in the hope of detecting unanticipated regularities. An approach he calls "Searching for Facts" (Roth, 1986). The second publication on which this dissertation is based falls in the last two categories. It was designed to elicit peoples' beliefs about the use of automation in the moral domain with the aim to find out about potential reservations. This could then be used to influence decision making on a political level and could help to address potential problems people have with delegating moral decisions to machine agents before the introduction of the technology (Whispering in the Ears of Princess). Secondly, we also looked for regularities in peoples' evaluation of these delegation decisions and what reasons people might or might not have that could serve as an explanation for this (Searching for facts).

First, however, I will give a brief overview of the research regarding the acceptance of autonomous cars as well as a short introduction in the literature that has evolved around the topic of human-machine-interaction and why these issues are so important when it comes to autonomous driving. As in politics, every new technology must rely on a broad rate of acceptance. Autonomous cars might be able to improve on many things as I have outlined in chapter 1. Yet, if people are not willing to use the technology, it will probably not succeed and, as a consequence, not deliver any of the potential advantages. This is hardly news: Long before we acquired the technical skills to produce autonomous cars, Van Der Laan et al. (1997) pointed out that "it is unproductive to invest effort in designing and building an intelligent co-driver if the system is never switched on, or even disabled." Given the strong moral case for autonomous vehicles from chapter 1, it might not just be unproductive but also unethical to ignore obstacles to the introduction of self-driving cars. In some sense, we can view the empirical research regarding acceptance and trust as the second step of the ethical endeavor. While the topic of chapter 1 was about potential moral reasons that may endorse the technology of autonomous driving, this chapter tries to shed some light on the question of implementation. The empirical research on trolley cases in the previous section serves as a good example of this. Trolley cases like *tunnel* or *bridge* are unlikely to occur if they occur all. Yet, the research in this field is plenty. The reason for this is clear: Trolley scenarios might be unimportant in the sense that they will actually happen, nevertheless, they have a huge influence on acceptance. If people are uncomfortable with delegating moral decisions to self-driving vehicles, then this could have a great impact on their willingness to buy and use the technology. Whenever a new technology emerges one cannot possibly foresee every relevant aspect like use cases, disruptive potential and acceptance of it. The same problem applies to automated driving. In the same way that automobiles were more than just horseless carriages, autonomous cars might prove to be more than just driverless cars. An interesting thing to consider is the *Collingridge Dilemma*. Collingridge (1980) states that with many new technologies we cannot possibly foresee the entire impact of a technology and the possible problems associated with it. This makes regulation hard since one does not want to ban it completely because of the promising features of the technology (see, chapter 1). Yet, as soon as the technology has arrived at the market and is a success, regulators, again, only have very limited power because of the then widespread use of the technology.

Given the uncertainty of the future, there are, however, possibilities to shed some light on the transformations associated with the technology. A first step is to look at analogies. Automated vehicles in a broad sense have been around for decades. Commercial planes use autopilot systems and so do some rail-based forms of transportation. Autonomous robots have been introduced to factories and warehouses in order to cut costs and increase the speed of production or delivery. Yet, there is one crucial difference between these technologies and autonomous cars – the scope of application. Obviously, there are far fewer planes than cars, which in turn carry more people on a single trip. This allows for the luxury to employ human pilots who, even though they hardly do anything to control the plane, provide a sense of security to people who would fear an autonomous system that acts on its own without any supervision. Autonomous trains are limited by the tracks they run on. Especially, when one considers underground trains one must notice that many factors that make the development of autonomous cars so difficult are absent. Underground trains simply do not have to deal with upcoming traffic, because they run according to timetables. Additionally, these trains usually do not face the problem of pedestrians unwittingly crossing their paths. Therefore, the scope of things that could go wrong is highly limited when compared to the case of autonomous cars. Autonomous production robots are, by design, also limited to their specific tasks in a well-defined and safe area with hardly any unpredictable issues to occur. Automation technology in other areas may serve as a starting point for the evaluation of self-driving cars, but one needs to realize that due to the fact that autonomous cars will be able to drive everywhere, for instance in highly populated urban areas, the dangers and, in turn, the perceived view of the technology might drastically differ from the examples above. In the literature, one can find several different models of acceptance regarding new technologies (cf. Johnsen et al. (2017)). For the purpose of this argument it is sufficient to realize that the intention to use a technology (cf., Venkatesh and Davis (2000)) and the intention to purchase a product based on said new technology (cf., Arndt (2011)) as well as subjects' willingness to pay serve as decent proxies for acceptance. To be able to form an impression of peoples' attitudes towards autonomous driving a number of surveys have been conducted. Johnsen et al. (2017) offer a condensed overview of the findings. They list eight categories that have been surveyed – among them "Safety" and "Trust and Control" which are of importance for the purpose of this dissertation. Concerns about the safety of autonomous cars is a salient issue and thus it was included in all of the reviewed studies. There are two sides one has to take into consideration when discussing self-driving cars. First, as described in chapter one the issue of crash avoidance is marketed as a prominent feature. With so many accidents being caused by human error, the advantages of an algorithm that never drinks or is tired is convincing to most people (Piao et al., 2016). On the other hand, there is the issue of system failure. System failure is subject to uncertainty. While we have gotten used to human errors, the fear of the – in a way – random possibility of system failure is perceived as worrisome (Kyriakidis et al., 2015, Schoettle and Sivak, 2014). Trust is a very complicated concept. Most studies interpret it as "the attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" (Lee et al., 2015). According to the findings there is still a majority of people who have the belief that they are better drivers than their automated counterparts (Schoettle and Sivak, 2014). Trust could, therefore, act as a barrier to the introduction of autonomous driving. Lee and Moray (1992, 1994) argue that estimates of the trustworthiness of the automated aid are relative to estimates of their own ability. Yet, trust attitudes can easily change due to more data and familiarization with the new technology.

The feeling of control is highly aligned with the notion of trust. If a person were to trust an automated system completely, the fear of losing control would be nonexistent. As the case of commercial airplanes suggest, however, people prefer having a human supervisor to be able to intervene, even if that often leads to worse outcomes (Carr, 2015). Research in automation thus far has dealt with problems like the "automation bias" and the *misuse* or *disuse* of automated aids. This line of research, however, focuses on human-computer teams, e.g. a pilot and the autopilot of the plane. In the case of both parties contributing to an outcome *misuse* would be due to over-relying on the automated aid if it is, in fact, not able to perform the task appropriately, whereas *disuse* would be described by ignoring the automated aid, if it is, in fact, able to perform the task better than the human agent. Consider contemporary cars: a driver can choose to use automated assistance such as lane-keeping systems or the cruise-control, yet, he or she does not have to. Instead, they can opt in or out of these technologies according to their personal perception of the trustworthiness of the device. On the other hand, with autonomous cars, at least with the level-5 form of completely autonomous vehicles, there is no longer a joint effort between a human supervisor and the automated agent. In fact, a means of control and supervision like a steering wheel might be completely absent in future generations of these cars. The decision of whether an automated aid (in this case an autonomous vehicle) is used is therefore made before the task begins. In a survey by Piao et al. (2016) asked respondents about the choice between a conventional bus and an automated bus with and without a human staff on board. 37% opted for the conventional and 38% for the automated bus with a human supervisor. This gives further reason to believe that – as of now – people are ill at ease with fully autonomous vehicles. Further research into the cause of this aversion is necessary. One reason for this might be that people see the delegation of driving autonomously to a machine as a delegation of a moral task. After all, participating in traffic might very well have bad effects on third parties as well as on oneself. Whereas in the case of a human supervisor there is still a human who is ultimately responsible for the outcome, in the case of fully autonomous vehicles the moral decision needs to be transferred to an algorithm that relies on soft- and hardware. The second publication on which this dissertation is built examined this very distinction. In Gogoll and Uhl (2018) we conducted a laboratory experiment which analyzes the delegation of a moral task to either a human or a machine. To isolate a potential aversion against machine use in the moral domain we controlled for the perception of the ability of either agent as well as for trust towards a human and a machine, respectively. We conclude that neither of these two prominent factors can serve as an explanation for subjects' preferences to 1) choose a human over a machine and 2) the punishment of subjects by their peers if they had chosen to delegate to a machine. In the article, we identify an aversion *per se* against machine use in the moral domain. Further research in this area is necessary. The next step of inquiry would be to see if subjects feel that responsibility for the outcome cannot be delegated to a machine but to a human and whether this may also be part of the aversion. Whatever the reason, it seems obvious that the question of acceptance should be researched and tackled before the introduction of autonomous cars will be initiated.

In the case of all things which have several parts and in which the totality is not, as it were, a mere heap, but the whole is something beside the parts, there is a cause; for even in bodies contact is the cause of unity in some cases, and in others viscosity or some other such quality.

Aristotle

3

Two Articles – Extended Abstracts

3.1 Autonomous Cars - In Favor of A Mandatory Ethics Setting

The recent progress in the development of autonomous cars has seen ethical questions come to the forefront. In particular, life and death decisions regarding the behavior of self-driving cars in trolley dilemma situations are attracting widespread interest in the recent debate. Assuming the tremendous computational power of computers in autonomous cars as well as the probable access to a lot of data regarding traffic participants, one can easily imagine the capabilities of a self-driving car in a dilemma situation. Frankly, all kinds of decisions about the behavior in dilemma situation could be implemented. Yet, the most important thing to realize is that some sort of behavior *has to be* implemented. This fact leads to the question of who should decide how a car reacts in a certain dilemma situation. In this essay we want to ask whether we should implement a mandatory ethics setting (MES) for the whole of society or, whether every driver should have the choice to select his own personal ethics setting (PES). While the consensus view seems to be that people would not be willing to use an automated car that might sacrifice themselves in a dilemma situation, we will defend the somewhat contra-intuitive claim that this would be nevertheless in their best interest. There are some arguments as to why a PES should be considered: One of the essential answers of modern political philosophy to the problem of reasonable moral disagreement is to partition the moral decision space. Instead of searching for a binding rule, modern societies often leave it to the individual to decide. Furthermore, leaving the decision to the individual doesn't only have the virtue that – at least in a circumscribed space – the individual can live according to her own normative ideals and understanding of the good. It also has the virtue that leaving the decision to each individual also pays equal respect to each of the members of society. Despite these reasons we argue in favor of a MES because, simply put, a PES regime would most likely result in a prisoner's dilemma. The incentives for the individual will crowd out moral PES and drive people to choose a selfish PES. The result of this situation is that everybody (the moral as well as the selfish agents) is worse off compared to a mandatory rule that is enforced by a third party. Since informal sanctions in anonymous large societies do not possess the force needed to prevent the individual to choose a selfish PES, we advocate for a mandatory rule that aims at minimizing overall harm. State regulation seems to be the most obvious as well as practical way to achieve that. Additionally, we argue that the discussion about trolley-cases is essentially misguided because the well-known thought experiment lacks attributes that are vital for the decision in a dynamic system such as traffic. The three main factors that are important, yet, the trolley problem lacks, are: strategic interaction, iteration and the fact that we could be subjects and objects of targeting. Therefore, our analysis trades the static model of trolley cases for an interaction analysis – a more appropriate way of thinking about the ethical questions that arise from the ethics setting of automated cars is in terms of game theory.

3.2 Rage Against the Machine: Automation in the Moral Domain

The introduction of ever more capable autonomous systems is moving at a rapid pace. The technological progress will enable us to completely delegate to machines processes that were once a prerogative for humans. Progress in fields like autonomous driving promises huge benefits on both economical and ethical scales. Yet, there is limited research that investigates the utilization of machines to perform tasks that are in the moral domain and little attention has been paid to possible empirical reservations that might influence the acceptance of the new technology. This is of the utmost importance, since any form of public reservation regarding the introduction of new technology could impede the implementation of a technology that could be beneficial overall. This study explores whether subjects are willing to delegate tasks that affect third parties to machines as well as how this decision is evaluated by an impartial observer. We examined two possible factors that might coin attitudes regarding machine use—perceived utility of and trust in the automated device. The experiment consisted of three parts: (1) the delegation and execution of a task that affected a third party, (2) a perception guess, and (3) a trust game. The aim of part 1 was to elicit attitudes toward machine use in the moral domain from the perspectives of actors and observers. Part 2 was designed to test whether a given divergence in judgments towards humans and machines could stem from systematically different perceptions of the errors committed by humans versus machines. Part 3 was designed to test whether different levels of trust in humans and machines could account for diverging judgments. We found that subjects preferred to delegate a task that affects a third party to a human and that delegators were rewarded less for delegating a task that affects a third party to a machine than for delegating it to a human. In the next two steps, we investigated whether the aversion to machine use in the moral domain identified is based on a lower "perceived utility" of the machine or on a general lack of trust in machines. Neither perceived utility nor trust, however, can account for this pattern. Machine errors are not perceived significantly different from human errors and the level of trust toward machines and toward humans does not differ significantly in the experiment. Alternative explanations that we test in a post-experimental survey also do not find support. We may thus observe an aversion *per se* against machine use in the moral domain. From our findings, it seems that most people rather intuitively dislike machine use in the moral domain—an intuition which turns out being hard to rationalize. We identified this aversion per se by experimentally equalizing humans' and the algorithm's performance. In practice, however, algorithms will usually not be simulating human moral behavior but will be programmed to implement a specific normative rationale. This attachment to rules may induce a dismissal of their decisional inflexibility in people. Such an instrumental aversion would then come in addition to the non-instrumental aversion that we identified. In the case of self-learning algorithms, we might observe an additional instrumental aversion to decisional opacity. The non-instrumental aversion identified suggests that the emphasis on the superior performance of automated cars, which is currently the main argument for automation in traffic, may not be sufficient or even decisive in convincing the general public. It might be as important to address the perceived moral problems that are necessarily associated with the introduction of automated vehicles.

A

Publications

A.1 Autonomous Cars: In Favor of a Mandatory Ethics Setting

First publication details:

Title	Autonomous Cars: In Favor of a Mandatory Ethics Setting
Authors	Gogoll, Jan & Müller, Julian
Journal	Science and Engineering Ethics, Springer
Impact Factor	2.229
Submitted	9th of May 2016
Accepted	27th of June 2016
Citations	20 (as of Juli 2018) according to scholar.google.com

Sci Eng Ethics
DOI 10.1007/s11948-016-9806-x

ORIGINAL PAPER

Autonomous Cars: In Favor of a Mandatory Ethics Setting

Jan Gogoll[1] · **Julian F. Müller**[2]

Received: 9 March 2016 / Accepted: 27 June 2016

Abstract The recent progress in the development of autonomous cars has seen ethical questions come to the forefront. In particular, life and death decisions regarding the behavior of self-driving cars in trolley dilemma situations are attracting widespread interest in the recent debate. In this essay we want to ask whether we *should implement a mandatory ethics setting (MES) for the whole of society or, whether every driver should have the choice to select his own personal ethics setting (PES).* While the consensus view seems to be that people would not be willing to use an automated car that might sacrifice themselves in a dilemma situation, we will defend the somewhat contra-intuitive claim that this would be nevertheless in their best interest. The reason is, simply put, that a PES regime would most likely result in a prisoner's dilemma.

Keywords Autonomous driving · Automation · Ethics · Morality · Dilemma

Introduction

The introduction of autonomous cars[1] as well as the development of ever more capable driver assistance systems are moving at a high pace. Big companies like BMW, Mercedes, Ford, GM, Toyota, Nissan, Volvo, Audi and, most prominently, Google are currently working on projects that aim to get humans away from the

[1] Henceforth, we will use the terms autonomous car, robot car and self-driving car interchangeably.

✉ Jan Gogoll
jan.gogoll@tum.de

[1] Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany

[2] University of Hamburg, Von-Melle-Park 6, 20146 Hamburg, Germany

Published online: 14 July 2016

steering wheel. Tesla has even gone so far as to release an update that enables their cars to drive on autopilot (McHugh 2015).

From an ethical perspective, the introduction of autonomous cars promises huge progress: Car accidents resulted in the deaths of roughly 32,000 people in the year 2013 in the U.S. alone.[2] The WHO estimates about 1.2 million traffic deaths worldwide each year (WHO 2011). According to a study by the ENO Center of Transportation, about 93 % of the 5.5 million crashes in the U.S. have been attributed to human error as the primary cause of the crash. This statistic includes all reported crashes—most of them without serious consequences for the people involved. Yet, out of these 93 % of human attributed crashes, more than a third is caused by intoxication (mainly alcohol, but also illegal drugs), speeding (30 %), distracted drivers (20 %), and other human errors due to external factors such as weather conditions or personal shortcomings e.g. lack of proper driving skills (Fagnant and Kockelman 2013). Most experts agree that the introduction of self-driving cars will lower the overall number of traffic accidents and traffic deaths. Based on the evidence currently available, it seems fair to suggest that the number of traffic-related deaths will go down significantly as more and more self-driving cars are introduced into the market. Some believe that autonomous cars will decrease traffic accidents by 90 % (Gao et al. 2014). A study by the Virginia Transportation Research Institute compared crash rates of cars in autonomous mode to manually steered cars, accounting for different levels of severity. The study states that "current data suggest that self-driving cars may have low rates of more-severe crashes […] when compared to national rates or to rates from naturalistic data sets, but there is currently too much uncertainty in self-driving rates to draw this conclusion with strong confidence" (Blanco et al. 2016). Nevertheless, given the fact that the prominent "Google car" has—as of this writing—managed to drive autonomously for over 1.7 million miles of testing with just 11 minor incidents (in which the Google car has never been the cause of the incident), an improvement in safety seems to be a fair assumption.

Although on the one hand, there is—from a normative standpoint[3]—pro tanto good reason to welcome the introduction of autonomous cars, there is no doubt that automated driving also poses new ethical challenges. Self-driving cars—if introduced—will crash eventually and will kill or seriously hurt someone in the process. There has never been a technology that has not failed at one point, and self-driving cars will be no exception. Autonomous cars are highly dependent on software and sensors, which are prone to fail eventually. Yet, even if we assume that a malfunction of the system does not occur, unlucky circumstances might lead to the following situation.

[2] This is according to the data of the Insurance Institute for Highway Safety. Note that the traffic-related death rate per 100,000 inhabitants is lower for first world countries due to safer (newer) technology, functioning regulation and enforcement of traffic laws.

[3] We emphasize the normative point here, since there might be other perspectives from which the introduction of autonomous cars seems to pose a problem. People who enjoy having a steering wheel in their hand might fear, for instance, that autonomous cars will prove so much safer than regular cars that Elon Musk's prediction comes true and the government might outlaw non-autonomous cars (Hof 2015).

Imagine you are sitting in your autonomous car going at a steady pace entering a tunnel. In front of you is a school bus with children on board going at the same pace as you are. In the left lane there is a single car with two passengers overtaking you. For some reason the bus in front of you brakes and your car cannot brake to avoid crashing into the bus. There are three different strategies your car can follow: First, brake and crash into the bus, which will result in the loss of lives on the bus. Second, steer into the passing car on your left—pushing it into the wall, saving your life but killing the other car's two passengers. Third, it can steer itself (and you) into the right hand sidewall of the tunnel, sacrificing you but sparing all other participants' lives.[4]

In a world without autonomous cars, the *tunnel case* is a philosophically interesting problem, which is usually discussed in the literature under the rubric of 'trolley problems', but not an ethically relevant "real world" issue. The reason for this is mainly that the driver behind the bus needs to make a split second decision based on very limited information. In such a situation, there is simply no time to form a—what philosophers sometimes call—deliberate judgment and, thus, there are thin grounds for assigning responsibility. In a world with autonomous cars, the case is different. Here, an agent—for instance the driver of a particular car or a regulative agency—essentially needs to tell the car beforehand what it should do in such a case. Or to put it differently: an agent must decide for a specific *ethics setting*. From a normative perspective, this raises an immediate question, namely: What is the right ethics setting? In this essay, however, we want to deal with another—although related—normative question: *Should we collectively mandate a specific ethics setting for the whole of society, or should every driver have the choice to select his own ethics setting?* Let us look at both options a little more closely. First, a society could agree on one ethical rule that is mandatory for every car under its jurisdiction. For this to be a sensible approach, one would have to show that there is a rule, for instance, that could be agreed on ex ante by all members of society. Secondly, there is the option to let each individual choose his own ethical setting privately for his own car. In theory, she could determine how her car should behave in a scenario like the *tunnel case* by setting her car to value her life above all (or not) as well as set a threshold of possible lives being saved at which she would be willing to sacrifice herself. In this essay, we will defend a mandatory ethics setting for all cars. More specifically, we will claim that a mandatory ethics setting should be in the best interest of all members of society.

We will unfold our argument in three sections. In the first section, we will talk briefly about the current prospects of automatic driving. Furthermore, we will give a quick overview of the existing literature that deals with normative questions and provide some context to the question this article attempts to solve. In the second section, we will motivate and discuss the arguments that point in the direction of a *personal ethics setting*. In the third part, we will argue that there is compelling reason to accept a *mandatory ethics setting*, since implementing a PES regime would most likely result in a prisoner's dilemma, i.e. a socially inferior outcome.

[4] This scenario is based on Millar's tunnel problem (Millar 2014a). Marcus (2012) and Goodall (2013) give a similar scenario called the "bridge scenario".

Autonomous Cars and Ethics

Introduction: Autonomous Cars

The idea of autonomous cars goes back to General Motor's vision for the future of transportation at the 1939 New York World's Fair (Becker et al. 2014). Although the idea of driverless cars has never disappeared completely from the world of imagination, in recent years it has experienced an unprecedented uptake. The reason the idea of the driverless car has gained traction again is twofold. First, considerable advancements in technology have led to a situation in which driverless cars are essentially within our reach. The second reason is that big automobile manufacturing companies such as BMW, Mercedes, Ford, GM and Toyota as well as leading tech companies such as Google and Apple (Harris 2015) back the idea of autonomous driving. Recently, the first autonomous pods were introduced to public roads in the Netherlands (Murgia 2015) and the Japanese government will launch an experiment with an unmanned taxi service as early as 2016 (Hongo 2015).

When it comes to automated vehicles, it is important to emphasize that there is a continuum of vehicle automation. The US National Highway Traffic Safety Administration (NHTSA), for instance, distinguishes five levels of vehicle automation. They mainly differentiate between cars that "do not have any of their control systems automated" (level 0), from cars in which the human driver is still mainly in control (level 1–2) and cars that are fully automated such that a human driver can cede full control to the car, whenever she chooses (level 3–4) (NHTSA 2013). The current state of automation does not allow the driver to cede full control, but "automobile manufacturers and technology companies are working towards adding more and more autonomous functions to newly manufactured vehicles" (Marshall and Niles 2014). In general, experts "emphasize incremental automation over full automation, contrast research platforms with production vehicles […]." (Smith 2014) By now it seems evident that different players in the market for autonomous vehicles will rely on different strategies when it comes to introducing automated driving. While automobile manufactures especially favor a gradual, "evolutionary development path of stepwise improvements from advanced driver assistance systems" (Meyer et al. 2015) to fully automated driving, tech companies like Google favor a revolutionary, disruptive approach (Davies 2015). Although, it is not certain at the moment when—or on which route—autonomous cars will conquer the streets, it seems more likely than not that they will succeed in the end. The members of the Institute of Electrical and Electronics Engineers (IEEE), for instance, predict that self-driving cars "will account for up to 75 percent of cars on the road by the year 2040." (IEEE 2012)

Ethical Issues Regarding Autonomous Cars

Since autonomous cars are a relatively new technology and its development is fostered mainly by automotive companies and engineers, much of the current debate revolves around the question of liability. Although other ethical challenges have been introduced to the debate, they remain of minor impact. Many favorable ethical arguments for the introduction of the autonomous car have been made on environmental grounds. Autonomous cars could reduce fuel usage and pollution by strictly following hypermiling strategies, and provide the possibility to position themselves closely behind other cars, since self-driving cars react faster and need not have the same safety margin as humans (Spieser et al. 2014; Torbert and Herrschaft 2013; Silberg et al. 2012; Schrank et al. 2011; Coelingh and Solyom 2012). Interestingly enough, car manufacturers might be in a position to build lighter cars, as it may be the case that additional safety features from the crumple zone to the air bags are no longer needed, additionally reducing fuel consumption. Other arguments focus on economic benefits such as an increase in spare time, the lower frequency of congestions and the possibility to install shared-car business models. Due to the reasons above, a Morgan Stanley report forecasts about $507 billion in productivity gains in the US alone (Shanker et al. 2013). Societal arguments focus mainly on the ability of the impaired to gain independence and the possibility to redesign roads and parking opportunities in urban areas, since autonomous cars need less space to operate (Silberg et al. 2012). On the other hand, difficulties arise because the environmental advantages could be nullified by a higher total number of car users (e.g. children and the impaired). Additionally, some raise privacy concerns due to the need of autonomous cars to communicate constantly for the network to work efficiently (Lin 2014a). Mladenovic and McPherson (2015) raise the question of how to engineer social justice into traffic control, especially concerning the dimensions of safety sustainability, and privacy.

There is a rapidly increasing literature on the ethical issues surrounding self-driving cars, which focuses on the potential net benefit of lives saved and the issue of liability if an autonomous car does crash. These two topics are intertwined for a reason. If autonomous cars actually reduce the number of fatalities, this seems to be a reason to foster their development and incentivize companies and research facilities to invest heavily in the new technology. At this point, the question of liability comes into play: If an autonomous car causes a crash, it itself cannot be held morally accountable for the outcome since it is not a moral agent. If lawmakers were to shift the responsibility towards the developers, it will create a financial barrier for the companies due to the high-anticipated costs usually associated with a lawsuit. Hevelke and Nida-Rümelin (2014) provide a detailed analysis of the ethical issues related to the attribution of responsibility to either the manufacturers or the driver, proposing a tax or a mandatory insurance to cope with any damages that any autonomous car might cause.

This article, however, is an attempt to address a special moral issue that is discussed under the rubric of the trolley problem. First introduced by Philippa Foot, an Oxford-based philosopher, then taken up by American philosopher Judith Jarvis Thompson, the trolley problem has generated a vast amount of literature—sometimes referred to as "trolleyology" (Robinson 2014).[5] With the introduction of

[5] In this paper we will not discuss the trolley problem in detail. Readers who are not familiar with the thought experiment are referred to Foot (1967), Thomson (1976) and Thomson (1985). For a complete overview see Robinson (2014).

autonomous cars, the nature of the trolley problem changes dramatically. So far it has been used as a thought experiment to elicit people's intuitions and to strengthen or weaken an underlying moral concept like utilitarianism or deontic ethics. In the case of driverless cars, the issue gets a very practical relevance as Lin (2013) observes when he writes that "programmers will [still] need to instruct an automated car on how to act for the entire range of foreseeable scenarios, as well as lay down guiding principles for unforeseen scenarios." When we think of a human driver who suddenly finds herself in a scenario like *tunnel*, we do not expect her to follow a certain moral guiding principle and we certainly do not blame her afterwards if we find that her choice does not line up with our own intuitions or convictions. Instead, we would rather understand the nature of this dilemma and, given the short reaction time, would argue that she had no choice but to act out of pure instinct. In short, we would refrain from assigning moral responsibility and ergo moral blame. With autonomous cars, the case is quite different: Firstly, a computer is not deluded by mere instincts and is not pumped up with adrenalin when it finds itself in a moral dilemma. Secondly, a computer capable of controlling a vehicle autonomously in everyday traffic situations can be expected to take huge amounts of information (e.g. number of possible victims) into consideration, even if the time horizon for a decision is limited. Thirdly, it has to have some kind of default reaction if there is no specific order on how to react in a case like the *tunnel case*. Assuming that the default setting would be to brake and go straight ahead, this would already be a morally relevant decision made by the developer of the underlying algorithm. In any case, the automatic system will act and the consequences cannot be considered accidental because they are determined beforehand. As with the original trolley problem, there are different moral arguments that propose divergent strategies as to what conduct should be considered morally preferable in this scenario. This line of thought can be described with the umbrella term of "ethics of crashing", which tries to shed light on which decision is morally justified given the dilemma-like characteristic of trolley situations.

The central ethical issue with regard to trolley problems simply put is then: How should an autonomous car react in a trolley situation? Much of the current debate revolves around the question whether there is good moral reason to have the autonomous car react according to deontological or utilitarian considerations. While the first requires the ethical decision to be made according to a set of rules that must be adhered to under any circumstances, the latter seeks to maximize utility with every decision made, that is, it places the consequences of a morally relevant act in the foreground. Goodall (2013) notes that these "rational approaches" are appealing to engineers and software developers since machines are, by nature, destined to follow a specific set of rules (deontology) or maximize preset functions for optimization (utilitarianism). Others stress the importance of a virtue ethics approach, which is fostered by professional engineering organizations and therefore influence the decision-making of engineers (Kumfer and Burgess 2015).

However, if one takes into consideration the broader spectrum of machine ethics, one finds additional approaches evaluating the possibility of *Kantian machines* (Powers 2006), empathy based machines called *Smithian machines* (Powers 2013) and *descriptive ethics* based machines, which mimic the entire spectrum of actual human ethical opinions of society using some mechanism of randomization (Goodall 2014). In a sense then, the autonomous vehicle version of the trolley problem just reproduces the debate—and thus the disagreement—of the original trolley problem. The question from a normative perspective then becomes: how should we proceed given widespread normative disagreement about the appropriate ethics setting of autonomous cars?

In philosophy, such disagreements are ubiquitous. Since ethics—and in particular political philosophy—is faced with such normative stand-offs on a regular basis, philosophy has developed certain tools to approach those disagreements. The most common approach in liberal society is to partition the moral decision space and thus give individuals the freedom to act according to their own normative standards. In the next section, we will discuss this approach to facing disagreement. Although, at first, it seems very attractive, we will argue that such an arrangement would be to the detriment of everybody in the case of autonomous cars.

Personal Ethics Setting (PES)

When it comes to ethical problems, modern societies usually face pervasive disagreement. While it might be the case that reasonable people might be able to agree on very general rules of justice and the distribution of rights, political philosophers are usually much more skeptical when it comes to questions of applied ethics. Gerald Gaus (2005) writes: "although we may be able to obtain knowledge of abstract principles of right, particular judgments and specific issues involve conflicting principles, and [thus] it is exceedingly difficult to provide answers to these questions that have any claim to being clear and definitive." As Rawls (1993) has pointed out in his seminal work "a plurality of reasonable, yet incompatible, comprehensive doctrines is the normal result of the exercise of human reason within the framework of the free institutions of a constitutional democratic regime." In short, Rawls—and many others believe—that the institutions of modern democracies, which are based on toleration and acknowledgment of what economists call bounded rationality, and what Rawls dubbed the burdens of judgment, will inevitably produce a plethora of different beliefs and moral stances.

One of the essential answers of modern political philosophy to the problem of reasonable moral disagreement is to partition the moral decision space. Instead of searching for a binding rule, modern societies often leave it to the individual to decide. Furthermore, leaving the decision to the individual doesn't only have the virtue that—at least in a circumscribed space—the individual can live according to her own normative ideals and understanding of the good. It also has the virtue that leaving the decision to each individual also pays equal respect to each of the members of society. Jason Millar gives the following example: "In medical ethics, there is general agreement that it is impermissible to impose answers to deeply personal moral questions upon the [patient]. When faced with a diagnosis of cancer, for example, it is up to the patient to decide whether or not to undergo chemotherapy." (Millar 2014b) A personal ethics setting reflects the value of autonomy and is in that sense sensitive to the moral views of the members of society. In such a world, an old couple might decide that they have lived a fulfilled life and thus are willing to sacrifice themselves in a *tunnel case* scenario. On the other hand, a family father might decide that even if he drives his car alone to work that his car should never be allowed to sacrifice him. Even if it is his life against a family or a school bus. At least prima facie, devolving the ethical decisions space seems to be the appropriate solution; a solution that is in accordance with the values of a liberal society. Sandberg and Bradshaw (2013) argue along these lines proposing that an autonomous car should have different ethics settings consistent with several ethical theories to allow each individual owner to decide what ethics setting her car should have. In this case, a self-driving vehicle would be considered a "moral proxy" as opposed to a "moral agent" or a "moral patient" (see Millar 2014a). A recent web poll by robohub.org supports this result. The poll asked who should determine how an automated car responds in ethical dilemma situations such as the trolley problem. Most of the participants (44 %) thought that the passengers should decide, while 33 % thought that lawmakers should have the final say (Millar 2014b). In his short essay "Here is a terrible idea: Robot Cars With Adjustable Ethics Settings", Patrick Lin (2014b), however, takes a stance against—as the headline suggests—an adjustable ethics setting. The argument Lin presents in his short piece is mainly about manufacturer liability and does not directly confront the normative issue of whether a personal ethic setting would be justified or not. Nevertheless, Lin—en passant—mentions two interesting moral reasons against a PES that we want to consider here. The first reason is that a PES might allow options that seem morally troubling: For instances targeting black people over white people, poor people over rich ones, and gay people over straight. Lin undoubtedly touches an important point here. But there is an important counter-argument to this objection. Allowing for a PES does not mean that the PES itself allows for all conceivable trade-offs. Think about one of the central rights in modern liberal states: religious freedom. Modern states allow for a wide range of religious practices, but there are nonetheless certain practices that are ruled out. In Germany, for instance, shechita, a special Jewish tradition of slaughtering animals in a kosher fashion is banned because the practice stands in conflict with animal rights. A PES, thus, as every "moral free space" (Donaldson and Dunfee 1999) would have clearly defined limits. Presumably, modern societies could achieve a far-reaching overlapping consensus to prohibit deeply racist or sexist settings or even forbid the allocation of demographic data that such a targeting mechanism would require. Furthermore, it does not seem likely that any automotive company would indeed offer a vehicle that permitted discrimination against a certain minority in the case of an accident (see Millar 2014c).

The second objection that Lin mentions is, basically, that a PES would be too much of a burden for the individual. From a philosophical point of view, however, an argument along these lines would be puzzling. Who else, if not the citizens, should decide these moral conundrums? Lin points to two alternative agents: The car manufacturers and the government. Although at first glance, punting the responsibility to the manufacturers and the government seems to be a feasible option, a more careful analysis suggests that this is not a viable alternative. First, automobile manufacturers are faced with fierce international competition. This means that the individual manufacturers need to be responsive to the demand of customers. If customers want automated cars with a PES, manufacturers will have no other option than to produce robo-cars with a PES.[6] The other alternative is shifting the responsibility to government agencies. From a normative point of view though, the government should only pass laws that reflect the values, ideals and preferences of its citizens. Thus, a necessary condition to determine which regulations the government should pass is to elicit the values and preferences of the citizens. Again, we are back to the citizens as the primary moral authority. The crucial point we make is that, in any case, the citizen needs to make up her mind about these new ethical conundrums. Neither the government nor the automobile manufacturers have the moral authority to decide these questions, even if they had the opportunity to do so.

Mandatory Ethics Setting (MES)

In this section, we want to argue that despite the advantages of a PES, a mandatory ethics setting (MES) is actually in the best interest of society as a whole. Our argument will proceed in three steps. In "PES in an interaction analyses" section we will argue that implementing a PES would lead to a prisoner's dilemma. To be more specific, we argue that implementing a PES will lead to a situation that will crowd out the ethical PES and lead to a socially unwanted outcome. In "Why a mandatory rule is necessary" section, building on the result of the preceding section, we will argue that a MES is the only way to solve the prisoner's dilemma and that a MES would be in the interest of selfish as well as morally motivated agents. In particular, we will argue that a MES that minimizes the risk of people being harmed in traffic is in the considered interest of society. As a corollary, we will defend the somewhat contra-intuitive idea that automated cars—at least under some circumstances—should sacrifice their drivers in order to save a greater number of lives. In "Objections" section we will review a few objections against our approach.

PES in an Interaction Analyses

In the second part, we argued, that in liberal societies a common response to disagreement is partitioning the moral decision space. In applying this insight to the question of ethics settings, we developed and justified the idea of a PES. Although this idea seems intuitively appealing, implementing a PES will—or so we argue—most likely lead to a social state that is unappealing from a wide variety of views. In this section, we want to explain why implementing a PES leads to a prisoner's dilemma. However, before we go into medias res, we first want to comment on some methodological issues with regard to the application of trolley problems to the issue of autonomous cars.

[6] One could argue that the manufactures could come together and agree on industry standards. There are two things to say to this. First, industry-wide standards are pretty hard to achieve in a globalized world with important car manufacturers all over the globe. Second, it is especially difficult if the industry standards do not reflect the preferences of consumers.

Since the ethical questions of automated driving are often discussed with reference to trolley problems, we want to explain how our approach relates to the current debate. Trolley problems, as we discussed earlier, are philosophical thought experiments used to elicit moral intuitions. Collecting moral intuitions about certain cases, in turn, allows philosophers to infer underlying moral principles that, in part, explain our reactive moral attitudes. Thus, in applied ethics, we then use thought experiments as proxies for moral problems in the real world. Thought experiments in applied ethics are useful only insofar as they manage to abstract away distracting details, while retaining the important moral properties and variables of the initial problem X. If we fail to include an important variable of the initial problem in our thought experiment, then the elicited intuitions and the corresponding underlying moral principles will not teach us anything about how to regulate problem X. Creating a moral thought experiment is then essentially similar to what is called model building in the (social) sciences. In creating a model, it is important that we are able to identify the relevant variables at work in a certain situation. The tricky part in modeling, of course, is identifying the correct set of variables. If we miss important variables in modeling a problem, our explanations and predictions will suffer. If we are missing important moral variables in an ethical thought experiment, our moral judgments will be most likely inadequate. Basically, the question is then whether trolley cases adequately model the moral problems we are interested in when thinking about the ethics settings of automated cars.

We think that standard trolley problems miss three morally important aspects of the moral problem at hand and, thus, are inadequate, at least for the question we raise in this paper. The first two aspects missing in the trolley case are strategic interaction and iteration. In trolley problems, we are faced with a non-strategic dilemma situation. Our actions alone determine the result of the dilemma. If we pull the lever, the trolley will turn right; if we do nothing, the trolley will go straight ahead and will kill whoever is tied to the tracks. Furthermore, our decision is not dependent on the actions of other participants. This is very different in the case of ethics settings. Think about it this way: if you live in a society, in which everybody is known to have an altruistic ethics setting, you might consider having an altruistic ethics setting as well. On the other hand, if you know that everybody around you set their cars to protect themselves no matter what, you will most likely not be inclined to sacrifice yourself for the greater number in case of a crash. Closely related to that, trolley dilemma situations are essentially one-shot games. You make a decision and that is it. Your decision, importantly, does not take into account the response to your choice in the future. Again, this is different when it comes to the dilemma we are grappling with. As our last example suggested, the distribution of ethics settings might shift over time as a result of a myriad of individual strategic decisions.

The third aspect has to do with the decision situation of the trolley problem. In its standard form, the trolley problem puts the ethical inquirer in the position of the agent who needs to decide about life and death. However, when deliberating about an adequate ethics setting for an automated car, it is important to view the dilemma at hand from both perspectives, from the perspective of the subject and of the object. This is because every participant in traffic is equally concerned with the possibility of making the call in a dilemma situation, but also with being the target in such a situation. The agent, furthermore, can be singled out as a target or can be part of a group that is targeted, for instance, if he is sitting in a bus, is carpooling or in a group of pedestrians. The bottom line is that our fate in a trolley-like situation is not only determined by the ethics setting of our own car, but by all other road users and their ethics settings, respectively.

Since so many relevant moral aspects for the correct choice of ethics settings do not come into the picture in the classical trolley choice problem, we think it is not well suited to generate adequate intuitions and answers for the problem at hand. The argument so far suggests we look for a choice situation that models:

(a) strategic interaction
(b) iteration
(c) the fact that we could be subjects and objects of targeting.

A more appropriate way of thinking about the ethical questions that arise from the ethics setting of automated cars, we maintain, is in terms of game theory. Game theory is essentially about strategic interactions. Modeling the strategic interaction between drivers who can choose their ethics setting will give a new and important insight into the ethical question at hand.

PES: Crowding Out of Morality

We want to start here with a very simple game theoretic model. Imagine a social world, in which autonomous cars have the capability to communicate with each other and the relevant infrastructure about a wide range of potentially morally relevant issues, such as the number of persons within a car. Further, imagine for sake of simplicity that there are only two types of agents: moral agents and selfish agents. In general, moral agents are disposed to act altruistically as long as most of their fellows do so as well. Thus, their attitude towards moral behavior is conditioned upon a certain degree of overall reciprocity. Applied to traffic dilemma cases like the *tunnel case,* moral agents are disposed to sacrifice themselves in at least some situations. Moral agents in our story are then disposed to minimize harm. Note that the moral agents are not adhering to utilitarianism. Utilitarian agents would need to sacrifice themselves for the greater good regardless of whether other agents would do so or not. Selfish agents on the other hand, as one might expect, are solely interested in minimizing harm to themselves.

Now, it seems clear that in a population that is constituted solely by moral agents, every moral agent has good reasons to believe that every autonomous car on the road is programmed 'morally', which gives him sufficient reason to choose a moral PES as well. But consider now that a moral agent is put in a society in which he cannot be sure what the actual distribution of moral and selfish agents is. In this circumstance, even a moral agent might think to herself: Well, I am not disposed to sacrifice myself for people I don't know and who might well not do the same for me. I want to be moral, but I do not want to be a sucker. A standard way to model such a case is the well-known prisoner's dilemma.

		Player 1	
		cooperate	defect
Player 2	cooperate	3,3	0,4
	defect	4,0	1,1

Fig. 1 The Prisoner's Dilemma

In this situation, two players have the choice between cooperation and defection. Both recognize that they could maximize the social good by choosing to cooperate. Yet, each of the players has the opportunity to get a higher payoff if she defects, while the other player cooperates. Anticipating this line of thought, each player will choose to defect in order to not be exploited, thus leading to the socially unwanted outcome of (1,1) in the lower right quadrant. Obviously, the prisoner's dilemma depicted in Fig. 1 is a great simplification of any social situation that might occur, since in actual scenarios, countless variables and uncertainties enter the equation. The complexity of the dilemma also grows with an increasing number of players and possible strategy options. However, following Brennan and Buchanan (1985), we believe that the prisoner's dilemma does "contain most of the elements in its structure required for an understanding of the central problems of social order, those of reconciling the behavior of separately motivated persons so as to generate patterns of outcomes that are tolerable to all participants". How does the prisoner's dilemma then translate to our discussion of ethics settings? Let us first begin with a *strategic analysis* of the situation. The individual, let's call her Johanna, plays the game against all other people who participate in traffic. If every participant chooses the moral PES, traffic would be maximally safe for everyone.

This can be shown displaying the case of a society that consists only of three people. These people have to commute every day but, since they happen to have two sports cars, they cannot carpool together. Instead, they have to split up in parties of two and one. Before they leave, they decide how their autonomous cars should behave in case of a dilemma situation in which one car has to be sacrificed. To mix up the daily routine, they also decide to switch positions every time they leave, so that, ultimately, the probability of each person occupying any single spot (being alone in one car or being the (co-)driver in the other) is identical. If they decide on a selfish PES setting the expected value[7] of the situation would be:

$$E(PES) : 0.5 * 2 + 0.5 * 1 = 1.5$$

Since one position in the dilemma is at an advantage and it is equally likely that either car occupies this position, each car would have a chance of 50 % to survive

[7] In this case, the expected value equals the expected number of deaths.

the dilemma. This means that the expected value of a dilemma in a PES world is 1.5 deaths.

Setting the car according to a (mandatory) MES setting, which is programmed to always spare the car that has two passengers, however, leads to the following expected value:

$$E(MES) : 0 * 2 + 1 * 1 = 1$$

Since $1 < 1.5$ the social outcome of a PES is worse compared to the MES world. From the standpoint of each individual, the expected value to die in a dilemma is therefore:

$$E_{MES}(I) : \frac{1}{3} * 1 + \frac{2}{3} * 0 = \frac{1}{3}$$

Being randomly distributed to the two cars, each of the three people in this society survives in two out of three cases, because the two-person car is never the one that has to sacrifice.

Contrariwise, if the three decide on a selfish PES for each car the expected value of a dilemma would be:

$$E_{PES}(I) : \frac{1}{3} * 0.5 + \frac{2}{3} * 0.5 = \frac{1}{2}$$

Obviously, this is a deterioration compared to the former scenario since the expected value to die for each individual is 50 % higher than before.

Coming back to Johanna, the problem that arises is that even if Johanna believes that everyone else set their PES setting to minimize harm, she still has a strong reason to set her car's ethics setting privately to value her life above all. If everyone chooses a moral PES, Johanna can maximize her personal safety by choosing the selfish PES unilaterally. In dilemma situations, Johanna's car would then be the only one who would save its driver no matter what, while all other cars in traffic would sacrifice their driver given it minimizes total harm. Instead of cooperating and reaping the overall higher social benefit (in this case a lower probability of being harmed or killed), Johanna then could defect and gain additional security by avoiding those cases in which a strategy to minimize harm would mean self-sacrifice on her part, thus increasing the probability of not being harmed at the expense of other road users. At the same time, choosing the selfish PES is not only the best strategy for Johanna in a world populated by (mostly) moral agents, but also in a social world that is inhabited by mostly selfish agents. In game theoretic terms, defecting is, thus, the optimal choice regardless of what the others do.

Up to this point, we have analyzed the strategic decision that Johanna, and, thus, every agent, faces in the traffic game. However, as we explicated earlier, choosing PES has an important temporal aspect. The decision by another agent—let us name him Matt—for or against a moral PES in t_2 will, at least in part, depend on the PES Johanna and others have chosen. If Matt, who is generally inclined to choose a moral PES, is convinced that most of society has chosen a moral ethics setting, there is a good chance that he will choose a moral PES as well. There are many people, of course, who would follow a general rule even if the individual incentive to deviate is high and the chance of being sanctioned is low. Nevertheless, if there is a sufficient number of people who will not choose a moral PES, the moral equilibrium will not be stable. There is a strong incentive for each individual to defect from the minimizing harm strategy. Therefore, even if Matt accepts the minimizing strategy to be morally superior to a selfish PES, defecting will increase his safety. Yet, if such a defection is possible, there is no reason to believe that only Matt would take this opportunity. If a sufficient number of people realize that this strategy maximizes their utility, the benefits of the minimizing harm strategy to society will eventually evaporate. This phenomenon can be observed in many circumstances. In such situations, theory as well as experiments show that conditional-cooperators—moral agents in our case—will usually become crowded out rather quickly. To conclude, the first result is that a PES, even in a population of mostly moral agents, will lead to a prisoner's dilemma. To put it differently, the result is that there is good reason to believe that morality will become crowded out in a world where people can choose their own ethics setting.

One might think that, since morality becomes crowded out, at least the selfish agents end up with what they want. Readers familiar with the prisoner's dilemma know that this is not the case. The unintended result of letting everybody choose their personal ethics setting is also not in the interest of selfish agents. Again, selfish agents are defined as agents aiming to minimize harm to themselves and their friends and family. As becomes evident from our small game theoretic exercise above, if everybody tries to minimize the expected harm to him or herself, the expected likelihood of everyone becoming harmed actually rises. This game theoretic exercise is easily confirmed. Think about a world in which everybody is moral and, thus, is ready to sacrifice themselves for a greater number of people. Evidently, in such a world, fewer people in total will be killed. Therefore, by this logic, a world in which nobody is ready to sacrifice themselves for the greater number, the number of actual traffic casualties is necessarily higher. This leads to our second, and maybe unexpected, result that selfish as well as moral agents have a strong reason against implementing PES.

Why a Mandatory Rule is Necessary

So far, we have argued that moral agents as well as selfish agents prefer a social world—albeit for different reasons—in which the risk of serious injury in traffic is minimized. It is important to emphasize here that the result of our discussion is derived from a contractarian thought experiment. We arrived at the answer by asking what would be in the interest of a diverse set of individuals (moral and selfish ones). We have further argued that to achieve such a world, every participant in 'traffic' needs to have a moral PES, i.e. a PES that would allow the car to sacrifice its driver for the greater number. Unfortunately, as we have shown, due to the logic of the iterated prisoner's dilemma, the moral PES would eventually be crowded out.

Given that moral as well as selfish agents are interested in establishing a social world in which everybody uses a moral PES, the question becomes how to solve the generalized prisoner's dilemma that prevents our agents to achieve the socially preferred result? In general, there are two types of solutions to collective action problems. The first kind of solution involves the introduction and sanctioning of informal rules. Nobel Prize laureate Elinor Ostrom has shown that under certain conditions, a group of people can overcome collective action problems such as the prisoner's dilemma (Ostrom 2005: 258–270). There are, however, certain conditions for overcoming collective action problems. In general, solving collective action problems by informal rules works best in relatively small groups, since effective monitoring as well as informal punishing of rule violation must be comparatively cheap. The bigger the group, the more expensive monitoring and punishing becomes. In the social dilemma 'traffic' however, monitoring and sanctioning is very complicated. There is no way to know about the ethics settings of the other cars participating in traffic. In general, in anonymous large-scale societies, informal sanctioning mechanisms do not work.

This leaves us with the classical solution to collective action problems: governmental intervention. The only way to achieve the moral equilibrium is state regulation. In particular, the government would need to prescribe a mandatory ethics setting (MES) for automated cars. The easiest way to implement a MES that maximizes traffic safety would be to introduce a new industry standard for automated cars that binds manufactures directly. The normative content of the MES, that we arrived at through a contractarian thought experiment, can easily be summarized in one maxim: *Minimize the harm for all people affected!*[8]

If applied ethics wants to generate useful solutions to real world ethical problems, it is important that the solutions suggested not stray away too far from the normative beliefs held by the people affected by the normative proposal. While in traditional ethics, we are usually not concerned with the normative beliefs that people actually hold, applied ethics has to be concerned with popular sentiment. The reason for this is simply that any proposal not properly reflecting the values of the affected people will certainly not be picked up by lawmakers or by the people affected, respectively. Thus, what we need here is a 'sanity check'. Regarding trolley situations with autonomous cars, there is already some empirical evidence that corroborates the results of our philosophical thought experiment. Three studies performed by Bonnefon et al. (2015) show that subjects being presented vignettes of dilemma situations involving self-driving cars are generally comfortable with utilitarian autonomous cars, "programmed to minimize an accident's death toll" (ibid.). What Bonnefon et al. call the "utilitarian autonomous vehicle" is completely in line with our notion of minimizing harm in trolley situations.[9]

[8] Unfortunately, we cannot debate the various ways in which such a maxim could be implemented. Although this maxim, on the face of it, seems quite simple, the implementation will surely raise many morally relevant follow-up questions. For instance, how should we weight lives? Should one person count equally regardless of, say, their age? Furthermore, who should count as 'all people affected'—should this include just motorized participants in traffic or should this also include pedestrians?

[9] However, note that our approach is contractarian by nature.

Our proposal of a MES that minimizes harm for all affected is further vindicated by a recent experimental study that tests a new version of Thomson's trolley dilemma. In this version, the initial dilemma becomes a trilemma. In the example of Bryce Huebner and Marc Hauser, an agent named Jesse has the option to sacrifice himself or another person for the benefit of a small group of strangers. Alternatively, he can also do nothing, which results in the death of the aforementioned group. Huebner and Hauser (2011) found that when confronted with the trilemma, "the largest number of participants (43 %) judged that Jesse should flip the switch to the right (killing the lone stranger) and a surprisingly large proportion of participants (38.3 %) judged that Jesse should engage in an act of altruistic self-sacrifice to save the five people on the main track." Adding up the numbers, this means that 81.3 % of the people in this study preferred a solution to the trilemma that minimizes the harm for all affected. The limited evidence available then seems to corroborate our proposal.[10] Before we conclude our argument though, we want to discuss a few objections.

Objections

In this essay we presented a contractarian argument for a *mandatory ethics setting*. In this final part of the essay, we want to discuss whether our argument holds under scrutiny. Let us then turn to the first objection. Firstly, one might ask, whether our proposed mandatory ethics setting is not biased against people who are usually or exclusively single drivers, since single drivers would be targeted over vehicles with more than one passenger in any case.

This question implicitly attacks one of the fundamental premises of our model. Our model rests upon the concept of the average participant in traffic. This participant spends an equal amount of time as a single driver, in groups of two, three, four and so forth. While our mathematical example has shown that the average participant of traffic has an increase in safety with a MES, it is not so clear as to what the benefit to single drivers is. On the contrary, the calculations suggest that people who always drive alone might incur a loss in safety relative to a PES world. We define a marked individualist driver as someone who (almost) always drives his car alone. To assess this objection, we first need a better understanding of its importance. There are a few things to note. First off, even somebody who rarely drives with other people will benefit from a MES under many circumstances. The maxim 'minimize harm for all affected' applies not only to single vehicles, but, more generally, to traffic. Therefore, even if a marked individualist is usually alone in his car when he participates in traffic, he will nevertheless be treated as part of a group by the AI of an autonomous car under many circumstances. To highlight just a few cases: (a) Think about the following dilemma. An automated truck can decide whether it sacrifices its driver or collides with the oncoming traffic, which would save the truck driver but put the lives of

[10] It should be noted, though, that the data just weekly confirms our argument. The reason is that there is a difference between what an individual deems as the right course of conduct and whether she wants that particular course of action to become a law that is applied to everyone.

the car drivers in danger. (b) An individualist car driver might sometimes use public transportation and, thus, be counted as part of a group by the AI of an automated car. (c) A third case that would make him part of a group from the vantage point of an AI, is him taking a stroll on a somewhat populated boardwalk. In all these cases, even an individualist would gain from a MES. Furthermore, even an individualist might strongly care about her family and friends and, thus, would prefer if his loved ones were as secure as possible in traffic. Taking these arguments together, we think that the idealization of the average participant at work in our model can be defended.[11]

Let us now turn to a second objection. Our model defined a moral agent as an agent that is ready to act altruistically as long as others do so as well. Our moral agent is then a conditional cooperator. One might object that morality consists of more than reciprocal altruism. This is certainly true. However, within the limits of an essay, it is not possible to discuss various strains of ethical theory in detail. Furthermore, it is important to note that ethical theories such as deontic ethics and utilitarianism are themselves abstractions. Real world agents usually do not judge a case on purely deontic or utilitarian grounds. Instead, real world actors usually rely on rather eclectic normative standards in evaluating certain actions or regulations. At the same time, altruism as well as reciprocity are core ideals of our everyday morality. While there is much ethical disagreement, it is reasonable to assume that the absolute majority of real world reasoners would judge someone moral who is ready to sacrifice her life for a greater number of strangers. Considering both points, we think our modelling of moral agents is sufficiently justified.

Furthermore, there is a third and very plausible objection. A liberal might be not impressed by the advantages of a MES. He might hold that the government is nevertheless not justified to restrict the choices of reasonable people. Millar for instance argues that owners of autonomous cars “ought to be morally responsible” for their car’s ethics setting and that any interference to their choice by either the companies or the government would be paternalistic (Millar 2014a). The question then becomes, under which circumstances liberals in general accept infringements on choice sets. One reason that liberals in general accept for coercing individuals and limiting their choices is the prevention of negative externalities. This explains why liberals in general might be in favor of granting a life or death decision to a cancer patient, as in the Millar example, but are nevertheless in favor of prohibiting drunk driving.[12] The reason why liberals are in favor of granting autonomy in the first case, but not in the latter is because drunk driving does not only endanger the

[11] An interesting question that arises from this line of argument would be whether a MES would incentivize people to car-share to minimize their risk of being targeted. The answer to that depends on many variables, for instance, to what degree people value time alone. From an ecological perspective, an incentive to carpool would surely not be a bad thing. Furthermore, more carpooling or the use of public transportation would mean less traffic, and less traffic might decrease the possibility of accidents. On the other hand, people could choose to pay people to accompany them in their cars to increase their safety. While this is not impossible, it seems highly unlikely to play a role.

[12] For that reason we are also highly skeptical of Millar’s suggestion to apply ethical norms from medicine and bioethics to the case of autonomous cars.

life of the driver, but also imposes risks on others. Acts that are justified because they limit unwanted externalities are therefore never paternalistic. If that is the case, then liberals should be in favor of a MES since self-prioritizing PES unilaterally impose additional risks on others.[13]

A valid fourth objection would be that the proposed moral MES would simply not be moral enough from the viewpoint of at least some agents. Take, for instance, the elderly couple Ann and Joe. They might feel that they have already had a great life and enjoyed much good fortune during their fifty years of marriage. It is then intelligible if Ann and Joe preferred to sacrifice themselves in a dilemma situation rather than killing, say, a young driver or a single mother. The MES setting we proposed, however, would make it impossible for them to act on their altruistic judgment. We are not sure how many people there are that really have such high-end altruistic preferences. At the same time, we do not think, in principle, this objection poses much of a problem to our approach. There seems to be prima facie no reason why our proposed MES should not allow for an 'altruistic add-on'. There are neither game theoretic nor any moral reasons that speak against the option to allow people to confirm to moral standards that go beyond the MES. Furthermore, there also seem to be no important technical problems to allow for such an altruistic add-on.

Conclusion

The question of how an autonomous vehicle should behave in trolley-like situations has caused much debate over the last 2 years. Debates about the autonomous vehicle version of the trolley problem have largely reproduced the moral disagreement of the original trolley problem. In this article, we presented two ways of dealing with moral disagreement about trolley dilemmas. We argue that the default option in liberal societies to deal with moral disagreement is to partition the moral decision space in order to enable each individual to live according to her own normative ideals and understanding of the good and thus to respect individual autonomy (within limits). Applied to the case of autonomous cars this would peak in favor of a personal ethics setting (PES). However, allowing for a PES, we argued, will likely lead to a situation that has the structure of a prisoner's dilemma. The incentives for the individual will crowd out moral PES and drive people to choose a selfish PES. The result of this situation, so we argued, is that everybody (the moral as well as the selfish agents) is worse off compared to a mandatory rule that is enforced by a third party. While the consensus view seems to be that people would not be willing to use an automated car that might sacrifice themselves in a dilemma situation, we argued that such a MES is in the considered interest of everybody. Since informal sanctions in anonymous large societies do not possess the force needed to prevent the individual to choose a selfish PES, we advocate for a mandatory rule that aims at minimizing overall harm. State regulation seems to be

[13] We want to express our gratitude towards two anonymous reviewers who brought this point to our attention.

the most obvious as well as practical way to achieve that. Furthermore, we made the case that the classic trolley problem is conceptually inadequate for discussing the case of ethics settings. The reason for this is that the trolley problem fails to model three important structural aspects of the traffic dilemma discussed: strategic interaction, iteration as well as the varying position an individual might occupy.

References

Becker, J., Colas, M. A., Nordbruch, S., Fausten, M. (2014). Bosch's vision and roadmap toward fully autonomous driving. In G. Meyer & S. Beiker (Eds.), *Road vehicle automation* (pp. 49–59). Lecture Notes in Mobility. Springer International Publishing.

Blanco, M., Atwood, J., Russell, S., Trimble, T., McClafferty, J., & Perez, M. (2016). Automated vehicle crash rate comparison using naturalistic data. Virginia Tech Transportation Institute.

Bonnefon, J., Shariff, A., Rahwan, I. (2015). Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars? CoRR, arXiv:1510.03346. Accessed February 2016.

Brennan, G., & Buchanan, J. (1985). *The reason of rules: Constitutional political economy*. Indianapolis: Cambridge University Press.

Coelingh, E., & Solyom, S. (2012). All aboard the robotic train. *Ieee Spectrum* 49.

Davies, A. (2015). Google's Plan to eliminate human driving in 5 years. *Wired*. http://www.wired.com/2015/05/google-wants-eliminate-human-driving-5-years/. Accessed February 2016.

Donaldson, T., & Dunfee, T. (1999). *The ties that bind: A social contract approach to business ethics*. Boston: Harvard Business Press.

Fagnant, D., & Kockelman, K. (2013). Preparing a nation for autonomous vehicles. *Eno Center of Transportation*. https://www.enotrans.org/wp-content/uploads/2015/09/AV-paper.pdf. Accessed February 2016.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect in virtues and vices. *Oxford Review, 5*, 5–15.

Gao, P., Hensley, R., & Zielke, A. (2014). A road map to the future for the auto industry. McKinsey Quarterly, Oct.

Gaus, G. (2005). Should philosophers 'apply ethics'? *Think, 3*(09), 63–68.

Goodall, N. (2013). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board, 2424*, 58–65.

Goodall, N. (2014). Machine ethics and automated vehicles. In *Road vehicle automation* (pp. 93–102). Springer International Publishing.

Harris, M. (2015). Documents confirm Apple is building self-driving car. *The Guardian*. http://www.theguardian.com/technology/2015/aug/14/apple-self-driving-car-project-titan-sooner-than-expected. Accessed February 2016.

Hevelke, A., & Nida-Rümelin, J. (2014). Responsibility for crashes of autonomous vehicles: An ethical analysis. *Science and Engineering Ethics, 21*(3), 619–630.

Hof, R. (2015). Tesla's Elon Musk thinks cars you can actually drive will be outlawed eventually. *Forbes*. http://www.forbes.com/sites/roberthof/2015/03/17/elon-musk-eventually-cars\-you-can-actually-drive-may-be-outlawed/. Accessed February 2016.

Hongo, J. (2015). RoboCab: Driverless taxi experiment to start in Japan. *Wall Street Journal*. http://blogs.wsj.com/japanrealtime/2015/10/01/robocab-driverless-taxi-experiment-to-start-in-japan/. Accessed February 2016.

Huebner, B., & Hauser, M. D. (2011). Moral judgments about altruistic self-sacrifice: When philosophical and folk intuitions clash. *Philosophical Psychology, 24*(1), 73–94.

IEEE. (2012). This lane is my lane, that lane is your lane. http://www.ieee.org/about/news/2012/5september_2_2012.html. Accessed February 2016.

Kumfer, W., & Burgess, R. (2015). Investigation into the role of rational ethics in crashes of automated vehicles. *Transportation Research Record: Journal of the Transportation Research Board, 2489*, 130–136.

Lin, P. (2013). The ethics of autonomous cars, The Atlantic. http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/. Accessed February 2016.

Lin, P. (2014a). What if your autonomous car keeps routing you past Krispy Kreme? *The Atlantic*. http://www.theatlantic.com/technology/archive/2014/01/what-if-your-autonomous-car-keeps-routing-you-past-krispy-kreme/283221/. Accessed February 2016.

Lin, P. (2014b). Here's a terrible idea: Robot cars with adjustable ethics settings. *WIRED*. http://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/. Accessed February 2016.

Marcus, G. (2012). Moral machines. *The New Yorker Blogs*. http://www.newyorker.com/online/blogs/newsdesk/2012/11/google-driverless-car-morality.html. Accessed February 2016.

Marshall, S., & Niles, J. (2014). Synergies between vehicle automation, telematics connectivity, and electric propulsion. In G. Meyer & S. Beiker (Eds.), *Road vehicle automation*. Berlin: Springer.

Mchuch, M. (2015). Tesla's cars now drive themselves, Kinda. *WIRED*. http://www.wired.com/2015/10/tesla-self-driving-over-air-update-live/. Accessed February 2016.

Meyer, G., Dokic, J., & Müller, B. (2015). Elements of a European roadmap on smart systems for automated driving. In Gereon Meyer und Sven Beiker (Eds.), *Road vehicle automation 2* (pp. 153–159). Springer International Publishing.

Millar, J. (2014a). Proxy prudence: Rethinking models of responsibility for semi-autonomous robots. Available at SSRN 2442273.

Millar, J. (2014b). An ethical dilemma: When robot cars must kill, who should pick the victim? *Robohub.org*. http://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim. Accessed February 2016.

Millar, J. (2014c). You should have a say in your robot car's code of ethics. *WIRED*. http://www.wired.com/2014/09/set-the-ethics-robot-car. Accessed April 2016.

Mladenovic, M. N., & McPherson, T. (2015). Engineering social justice into traffic control for self-driving vehicles? *Science and Engineering Ethics*. doi:10.1007/s11948-015-9690-9.

Murgia, M. (2015). First driverless pods to travel public roads arrive in the Netherlands. *The Telegraph*. http://www.telegraph.co.uk/technology/news/11879182/First-driverless-pods-to-travel-public-roads-arrive-in-the-Netherlands.html. Accessed February 2016.

National Highway Traffic Safety Administration. (2013). Preliminary statement of policy concerning automated vehicles. http://www.nhtsa.gov/staticfiles/rulemaking/pdf/Automated_Vehicles_Policy.pdf. Accessed February 2016.

Ostrom, E. (2005). *Understanding institutional diversity* (pp. 258–270). Princeton: Princeton University Press.

Powers, T. M. (2006). Prospects for a Kantian machine. *IEEE Intelligent Systems, 21*(4), 46–51.

Powers, T. M. (2013). Prospects for a Smithian machine. In: *Proceedings of the international association for computing and philosophy*, College Park, MD.

Rawls, J. (1993). *Political liberalism*. New York: Columbia University Press.

Robinson, J. (2014). Would you kill the fat man? *Teaching Philosophy, 37*, 449–451.

Sandberg, A., & Bradshaw, H. G. (2013). Autonomous vehicles, moral agency and moral proxyhood. In *Beyond AI conference proceedings*. Springer.

Schrank, D., Lomax, T., Eisele, B. (2011). TTIs 2011 urban mobility report, Texas Transportation Institute. http://d2dtl5nnlpfr0r.cloudfront.net/tti.tamu.edu/documents/mobility-report-2011-wappx.pdf. Accessed February 2016.

Shanker, R., Jonas, A., Devitt, S., Humphrey, A., Flannery, S., Greene, W., et al. (2013). Autonomous cars: Self-driving the new auto industry paradigm. Morgan Stanley Blue Paper, November.

Silberg, G., Wallace, R., Matuszak, G., Plessers, J., Brower, C., & Subramanian, D. (2012). *Self-driving cars: The next revolution* (pp. 10–15). KPMG and Center for Automotive Research.

Smith, B. W. (2014). A legal perspective on three misconceptions in vehicle automation. In G. Meyer & S. Beiker (Eds.), *Road vehicle automation* (pp. 85–91). Berlin: Springer International Publishing.

Spieser, K., Ballantyne, K., Treleaven, R., Frazzoli, E., Morton, D., & Pavone, M. (2014). Toward a systematic approach to the design and evaluation of automated mobility-on-demand systems: A case study in Singapore. In G. Meyer & S. Beiker (Eds.), *Road vehicle automation*. Berlin: Springer.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist, 59*(2), 204–217.

Thomson, J. J. (1985). Double effect, triple effect and the trolley problem: Squaring the circle in looping cases. *Yale Law Journal, 94*(6), 1395–1415.

Torbert, R., & Herrschaft, B. (2013). Driving Miss Hazy: Will driverless cars decrease fossil fuel consumption? Rocky Mountain Institute. http://blog.rmi.org/blog_2013_01_25_Driving_Miss_Hazy_Driverless_Cars. Accessed February 2016.

World Health Organisation. (2011). Road traffic deaths: Data by country. WHO. http://apps.who.int/gho/data/node.main.A997. Accessed February 2016.

This Agreement between Technical University Munich -- Jan Gogoll ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4276140058014
License date	Jan 25, 2018
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Science and Engineering Ethics
Licensed Content Title	Autonomous Cars: In Favor of a Mandatory Ethics Setting
Licensed Content Author	Jan Gogoll, Julian F. Müller
Licensed Content Date	Jan 1, 2016
Licensed Content Volume	23
Licensed Content Issue	3
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	yes
Title	Autonomous Cars: Ethical Analysis and Experiments
Instructor name	Lütge
Institution name	Technical University Muncih
Expected presentation date	Dec 2018
Portions	Full article
Requestor Location	Technical University Munich Arcisstraße 33 Munich, 80333 Germany Attn: Technical University Munich
Billing Type	Invoice
Billing Address	Technical University Munich Arcisstraße 33 Munich, Germany 80333 Attn: Technical University Munich
Total	0.00 EUR
Terms and Conditions	

A.2 Rage Against the Machine: Automation In the Moral Domain

Second publication details:

Title	Rage Against the Machine: Automation In the Moral Domain
Authors	Gogoll, Jan & Uhl, Matthias
Journal	Journal of Behavioral and Experimental Economics
Impact Factor	0.809
Submitted	14th of March 2017
Accepted	8th of April 2018
Citations	2 (as of March 2018)

Note: Additional material (Instructions, screenshots and survey questions of the experiment) is provided at the end. This material is also available on the publisher's homepage.

Journal of Behavioral and Experimental Economics 74 (2018) 97–103

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

Rage against the machine: Automation in the moral domain[☆]

Jan Gogoll[*], Matthias Uhl

ZD.B Junior Research Group "Ethics of Digitization", TUM School of Governance, TU Munich, Richard-Wagner-Straße 1, Munich 80333, Germany

ABSTRACT

The introduction of ever more capable autonomous systems is moving at a rapid pace. The technological progress will enable us to completely delegate to machines processes that were once a prerogative for humans. Progress in fields like autonomous driving promises huge benefits on both economical and ethical scales. Yet, there is little research that investigates the utilization of machines to perform tasks that are in the moral domain. This study explores whether subjects are willing to delegate tasks that affect third parties to machines as well as how this decision is evaluated by an impartial observer. We examined two possible factors that might coin attitudes regarding machine use—perceived utility of and trust in the automated device. We found that people are hesitant to delegate to a machine and that observers judge such delegations in relatively critical light. Neither perceived utility nor trust, however, can account for this pattern. Alternative explanations that we test in a post-experimental survey also do not find support. We may thus observe an aversion *per se* against machine use in the moral domain.

> *"I know I have made some very poor decisions recently, but I can give you my complete assurance that my work will be back to normal. I have still got the greatest enthusiasm and confidence in the mission. And I want to help you."*
> – HAL9000 (2001: A Space Odyssey)

1. Introduction

Due to the constant progress of automation over the past decades, we find ourselves ever and anon in a situation in which we have the possibility of employing an automated companion to help us take some work off our shoulders. In a perfect collaboration scenario the human operator delegates part of the work to her automated aid while she keeps an eye on its performance and takes back control whenever she sees fit. Yet, as technology progresses we (will) find ourselves in situations in which this dichotomy of work and supervision might crumble—even up to a point where human supervision during a task is neither needed nor wanted. The planned introduction of a technology that need and will not be monitored by human operators during its performance therefore poses new ethical challenges and questions. In the absence of a human operator who serves as an ultimately responsible moral agent, we have to address questions of responsibility and liability (Hevelke and Nida-Rümelin, 2014). Recently, the case of autonomous cars is gaining substantial interest.

Almost all car manufacturing firms have fostered the development of automated devices. While traditional car companies follow a step by step approach of adding pieces of automation to their latest models, such as "Active Lane Keeping Assist" systems, Google and Tesla are taking a disruptive approach that aims directly at the creation of a completely autonomous vehicle. The economic opportunities of autonomous driving are great. A Morgan Stanley report estimates a productivity gain of about $500 billion annually for the U.S. alone (Shanker et al., 2013). But there is also a moral case that can be made: Since most traffic accidents are due to human error (drunk driving, speeding, distraction, insufficient abilities) some estimate that the introduction of autonomous cars will decrease the number of traffic accidents by as much as 90% (Gao et al., 2014).

While a small literature on the moral case of autonomous driving exists, it mainly focuses on utilitarian benefits of the technology (Fagnant and Kockelmann, 2015) or deals with ethical decision-making in dilemma situations (see, e.g., Goodall, 2014; Gogoll and Müller, 2017). Little attention has been paid to possible empirical reservations that might influence the acceptance of the new technology. The delegation of a task that could carry severe consequences for a third party to an unmonitored machine might invoke popular resistance to the technology in cases of malfunction. This is of the utmost importance, since

[☆] The authors thank Johanna Jauernig and Julian Müller for helpful comments. Ro'i Zultan has opened our eyes to the appropriate title. Friedrich Gehring did an excellent job in programming. This work was partly performed within the Munich Center for Technology in Society (MCTS) lab "Automation & Society: The Case of Highly Automated Driving" as part of the Excellence Initiative by the German Research Foundation (DFG).

[*] Corresponding author.

E-mail addresses: jan.gogoll@tum.de (J. Gogoll), m.uhl@tum.de (M. Uhl).

https://doi.org/10.1016/j.socec.2018.04.003
Received 14 March 2017; Received in revised form 26 March 2018; Accepted 8 April 2018
Available online 13 April 2018
2214-8043/

any form of public reservation regarding the introduction of new technology could impede the implementation of a technology that could be beneficial overall.

The relationship between human operators and automated devices has generated vast amounts of literature. The primary focus has been set on understanding this relationship. To our knowledge, however, the question of whether the delegation of tasks that affect a third party to an automated device is being welcomed or condemned has not received any attention. This may largely be due to the fact that the usual role of a human operator is to supervise and control an automated device that carries out a specific task. A typical example is the duties of the pilot of an airplane, which is, essentially, capable of flying on its own. The primary role of a human operator is therefore to supervise and—if need be—to intervene in case of automation failure or unforeseen scenarios that are not in the domain of the automated device. Consequently, a large part of the literature has investigated what factors influence the usage of an automated device.

Dzindolet et al. (2001) have created a framework of automation use indicating a variety of parameters that can be used to predict the use of automation in a human-computer "team". There is evidence that people who can opt in for automation use sometimes fear a loss of control when delegating to an automated device (Muir and Moray, 1996; Ray et al., 2008). This study, however, investigates the attitudes toward delegating tasks that affect a third party to a machine rather than to a human being, as opposed to the more general question of under which circumstances people are willing to relinquish control. The latter would refer to people's general propensity to delegate, as is also the case when people take a bus or taxi instead of driving themselves. To abstract from this issue, we wittingly forced subjects to give up control by delegating to either a machine agent or a human agent, thus keeping a loss of control constant between groups.

First, our study elicits attitudes toward machine use in the moral domain from the perspective of actors and observers: Do subjects prefer to delegate a task that affects a third party to a machine or a human? To what extent do subjects get blamed or praised for their delegation decision? Specifically, our first two hypotheses are as follows:

Hypothesis 1. People's delegation of a task that affects a third party to a human or to a machine is not balanced.

Hypothesis 2. Delegators are rewarded differently for delegating a task that affects a third party to a machine than for delegating it to a human.

In a second step, we investigate potential reasons for any negative or positive preference concerning machine use in the moral domain. Specifically, we test two factors that recur throughout the literature.

A major factor that could influence the decision of a subject to delegate to a machine is the "perceived utility" of an automated aid, which is defined as a comparison between the perceived reliability of an automated device and manual control (Dzindolet et al., 2002). If a subject judges that her ability exceeds that of an automated device, she usually does not allocate a task to the machine (Lee and Moray, 1994). This judgment might also be due to self-serving biases that see people overestimating their own abilities (Svenson, 1981) or their contribution to a joint task (Ross and Sicoly, 1979). Additionally, the perceived abilities of an automated aid might be influenced by a higher or lower salience of errors an automated device commits. There are controversial findings in cognitive psychology as to whether a violation of expectation (expectancy-incongruent information) is more readily remembered than decisions that are in line with prior anticipation (expectancy-congruent information) (Stangor and McMillan, 1992; Stangor and Ruble, 1989). While people initially tend to have high expectations of the performance of automation, humans may be judged according to a "errare humanum est" standard—decreasing the salience of an observed mistake made by a human delegatee due to a priced-in expectancy of errors. In Dzindolet et al. (2002) subjects chose to be paid according to their performance rather than that of their automated aids. This was even the case when they were informed that the automated device was far superior, stating salient errors of the automated device they perceived earlier to justify their decision. This is astonishing since an important factor in the decision to employ an automated device lies in the goal-oriented nature (Lee and See, 2004) of the task. Prima facie, a subject should be more likely to use automation if she rates the device's ability to successfully perform the delegated task positively (Davis, 1989), i.e., if the machine is seen as a reliable entity. We isolate the potential effect of machine-error salience by forcing subjects to relinquish control, thus abstracting from a self-serving bias. Our third hypothesis is as follows:

Hypothesis 3. Machine errors are perceived differently to human errors.

Another important factor that is known to influence the decision to delegate to an automated device is trust. The concept of trust has attracted a lot of attention regarding its influence on automation. While some researchers have seen trust between human agents and machines to be closely related to the traditional concept of trust between humans, others stress important differences regarding trust relationships between humans and machines (de Visser et al., 2012). Trust is a notoriously broad term but one characteristic that is commonly shared by most authors is a state of vulnerability the trustor has to be in. That is, a trust relationship requires the trustor's willingness to set herself in a vulnerable position by delegating responsibility to the trustee (Rousseau et al., 1998). Obviously, if the outcome of a delegation is completely determined and the process fully transparent, there is no need to incorporate trust. In this study, we use a simple trust game to isolate the mere aspect of trust, since it requires no capabilities on the trustee's side about which the trustor might have biased beliefs. The trust game only requires the trustee to reciprocate. It thus abstracts from the aspect of perceived utility discussed above, which is closely related to the specific task at hand. Finally, our fourth hypothesis is as follows:

Hypothesis 4. The level of trust toward machines and toward humans is different.

2. Experiment design

The experiment consisted of three parts: (1) the delegation and execution of a task that affected a third party, (2) a perception guess, and (3) a trust game. The aim of part 1 was to elicit attitudes toward machine use in the moral domain from the perspectives of actors and observers (Hypotheses 1 and 2). Part 2 was designed to test whether a given divergence in judgments towards humans and machines could stem from systematically different perceptions of the errors committed by humans versus machines (Hypothesis 3). Part 3 was designed to test whether different levels of trust in humans and machines could account for diverging judgments (Hypothesis 4).

Subjects received instructions for the experiment on screen. They were informed at the beginning that the experiment consisted of three independent parts and that they could earn money in each of these three parts. In the end of the experiment, one of these parts was selected at random and subjects were paid according to their respective payoff in this part. Prior to the experiment there were two preparatory sessions which provided us with the necessary data to calibrate machine performances and were also used to create perception tasks. We will first explain the three parts of the experiment and then provide some details on the preparatory sessions.

2.1. Part 1: Task affecting third party

Part 1 of the experiment consisted of the delegation of a calculation task to either another human or to a machine and in the subsequent solving of the task by human task-solvers and the machine. The benevolent effort of the other human or the preprogrammed actions of the machine then determined the payoff of a third party.

For part 1 of the experiment, half of the subjects were randomly assigned the role of actors, and the other half, observers. One observer was randomly assigned to each actor. Each actor played all roles consecutively. First, actors as delegators had to delegate the calculation task either to another human or to a machine. Second, actors as task-solvers had to perform the task themselves. Third, actors as third parties were the recipient of the payoff created by the benevolent effort of another human task-solver or by the performance of a machine. The fact that the successful or unsuccessful performance of the task determined a third party's payoff made its solving, and—more importantly—its prior delegation, morally relevant.

Actors were first informed that they were randomly assigned to two other subjects in the lab, say X and Y. They were told that their own payoff would depend on the decision of Y and that their own decision would determine the payoff of X. The calculation task in part 1 of the experiment was then explained to actors. For the task each subject was confronted with a block of ten calculation exercises, each consisting of seven digits, lined up on the screen. The sum of the seven digits had to be entered in an input field. Finally, one line was selected at random. If the respective exercise was solved correctly, the third party received 70 ECU. Otherwise, the third party received nothing.

Before the actors made their delegation decision, we wanted them to form an impression about the relative capability of human task-solvers and the machine. Because we were interested in a potential systematic misperception of human and machine errors, we did not simply provide subjects with statistics on actual performances. Instead, all subjects were visually presented with past performances of 24 subjects from a preparatory session. They were also shown the corresponding performance of a preprogrammed algorithm (see section 2.4 for details).

Relative performances of humans and machine in the task were visualized on a split screen. The caption "human" and "machine" was shown on the respective half of the screen. In total, subjects were shown 240 (24 subjects solved 10 lines each) past solutions of humans and the corresponding machine performances. If a single exercise was solved correctly by the human subject or by the algorithm respectively, it appeared in white. Otherwise, it appeared in red.[1] Exercises solved by human and machine appeared alternately and one by one. Each exercise appeared for only 0.5 s making it extremely difficult to simply count the number of red lines. The side of the screen on which the performance of the machine was presented was randomized across subjects. In fact, subjects in the earlier preparatory sessions, and consequently also the tailored algorithm, solved about 20% of the lines incorrectly.

Once delegators had formed an impression of the performance of humans and the machine, they made their delegation decision. Note that every actor solved the calculation task in the task-solver's role for her recipient. Each actor did this without knowing whether her delegator had actually delegated the task to her or to a machine. This was done to prevent a general tendency to delegate to the machine to spare fellow subjects the work. The performance of a task-solver was only relevant for her recipient if the task-solver's delegator decided to delegate to her and not to a machine. Observers solved the calculation task as well, without any consequence for another subject, in order to give them an impression of the task.

Each actor was rewarded or punished for her delegation decision by her assigned observer. An observer could reduce the actor's initial endowment of 30 ECU by any integer amount, down to a minimum of zero or increase it up to a maximum of 60 ECU without any influence on her own payoff. The observer could, of course, also leave the actor's endowment unaltered. Reward and punishment choices were elicited via the strategy method (Selten, 1967). This means that each observer made her reward or punishment choice conditional on the delegation decision, as well as its outcome. Thus, judgment was contingent upon whether the delegator had delegated to a human task-solver or to a machine *and* upon whether the randomly drawn exercise was solved correctly or not. An observer thus gave his full evaluation profile behind the veil.

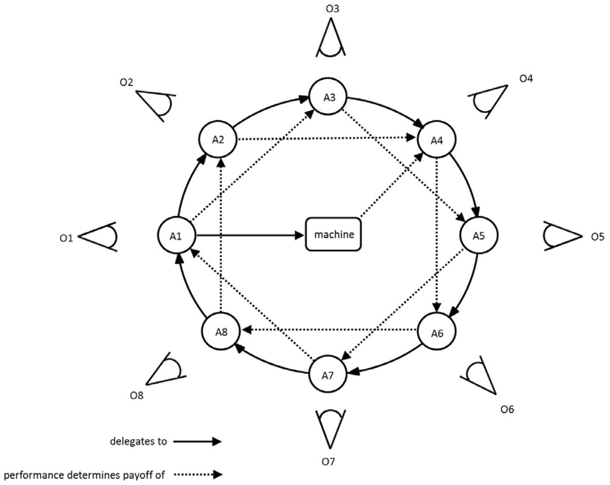


Fig. 1. Matching for delegation decision.

For the first round, actors thus received their altered endowment, ranging from 0 to 60 ECU plus 70 ECU, if their task-solver had calculated the randomly drawn line correctly. Observers received a flat payment of 100 ECU for the first round.[2]

The dependencies between subjects and the matching procedure for part 1 of the experiment are illustrated in Fig. 1. Here, actors are denoted by the letter A, while observers are denoted by the letter O. Consider the case of A1. A1 delegates the calculation task to A2 or to a machine (solid arrows). A2's or the machine's performance in the calculation task then determines the payoff of A4 (dotted arrows).[3] In this constellation, A1 is the delegator, A2 is the task-solver, and A4 is the recipient. A1, however, is also a task-solver, because A8 delegates to him or to a machine. Finally, A1 is a recipient. His payoff depends on the calculation performance of A7, if A6 has decided to delegate the calculation task to A7. Otherwise, it depends on the machine's performance. As can be seen, the design made sure that there were no direct interdependencies between any actors in the experiment. Potential feelings of reciprocity were thus excluded.[4] Subjects were explicitly informed about this feature of the design. O1 rewarded or punished the delegation decision of A1.

In the example above, the task-solving performance of A2 only determined A4's payoff if A1 had actually delegated the decision to A2. Otherwise, the performance of the machine was relevant. In either case, one of the ten solved exercises was selected at random and the recipient

[1] Subjects were told the following: "To evaluate the performance of a person and a machine, you will subsequently see a comparison of the performance of a past run. One line is shown per column, each respectively calculated either by a person or a machine. If calculated correctly, the line will be displayed in white. If calculated incorrectly the line will be displayed in red."

[2] This equalized the observer's own payoff with the payoff of an as-yet unrewarded or unpunished actor who had received the 70 ECU from the randomly-drawn exercise solved successfully by the task-solver on whom he depended. This is the case because he was additionally equipped with an initial endowment of 30 ECU. Thus, we established a conservative measure of reward and punishment, since any alteration of actors' endowment by a generally inequality-averse observer would require good reasons.

[3] For reasons of visual clarity, delegation and payoff dependency between actors and machine in Fig. 1 are shown for A1 and A4 only, by way of an example.

[4] See Greiner and Levati (2005) regarding the issue of indirect reciprocity in small groups.

Table 1
Payoffs for accuracy of guess.

Deviation(%)	Payoff
≤ 20	70 ECU
≤ 40	40 ECU
≤ 60	20 ECU
> 60	0 ECU

received her earnings if it was solved correctly. If A1 delegated to the machine, the performance of the tailored algorithm determined the payoff to A4.

2.2. Part 2: Perception guess

For part 2, the role differentiation of subjects was abolished. Subjects were informed that they would soon be confronted with yet another visualization of actual previous performances of the known calculation task by humans and a machine.[5] Their task was then to guess the number of errors of either the humans or the machine as accurately as possible. When seeing the visualization, they did not yet know whether they would later be asked to guess the errors of the humans or of the machine. All subjects were shown the data of the 24 subjects (240 lines) from a second preparatory session (i.e., different data than used for visualization in part 1) and the performance of the tailored algorithm. As in part 1, the relative performance of humans and the machine was presented on a split screen. The side of the screen on which the performance of the machine was presented was randomized across subjects. Exercises solved correctly again appeared in white, while those solved incorrectly appeared in red. In order to prevent subjects from counting, each exercise was only shown for 0.3 s. The interval was even shorter than in part 1, because subjects were already used to this kind of visualization.

After the actual past performances were shown, subjects had to state how many of the 240 exercises shown had been solved incorrectly, i.e., how many errors had been made. Subjects' payoff for part 2 depended on the accuracy of their guess. Payoffs were calculated according to Table 1.

Half of subjects were randomly asked to guess humans' performance, while the other half was asked to guess the machine's performance. It was ensured that an equal number of actors and observers from part 1 of the experiment were distributed between both of these treatments. Furthermore, subjects who delegated to a human and those who delegated to a machine were also divided equally between the two treatments.

2.3. Part 3: Trust game

In part 3, subjects were randomly assigned to one of two treatments. In the Human Treatment, subjects played a standard trust game. Trustors were endowed with 50 ECU. They could transfer 0, 10, 20, 30, 40 or 50 ECU to the trustee. The sent amount was tripled and credited to the trustee. The trustee could then reciprocate any integer amount she wished. The Machine Treatment was identical to the Human Treatment except for the fact that the reciprocation decision was made by a machine agent on behalf of the trustee who had no chance to intervene. Before subjects were informed about the treatment to which they were assigned, the setup of both treatments was carefully explained to them.

In the Human Treatment, before subjects learned their role, they made their choice for the trust game via the strategy vector method (Selten, 1967) and submitted their full strategy profile for both roles. If a subject was ultimately assigned the role of a trustee, her reciprocation decision conditional on the amount actually transferred by the trustor was returned.

Because the trustee had no voice in the Machine Treatment, subjects only took a decision for the case of ending up in the role of the trustor.[6] The reciprocation decision of the machine was determined according to the actually previously submitted strategy profiles of the subjects in the preparatory sessions. The algorithm was programmed such that it picked one of all 48 reciprocation profiles submitted in the preparatory sessions at random, and applied the respective conditional choice to a trustor's actually chosen transfer.

Before subjects made their choices, they were given an impression of the reciprocation choices of humans and the machine on a split screen.[7] For each subject, the choices of the machine were shown on either the left or right side of the screen at random. For this purpose, subjects were shown the actual reciprocation profiles of all 48 subjects from both preparatory sessions. These choices were contrasted with the reciprocation profiles of the machine algorithm. Each profile consisted of five choices, i.e., the returned amount for each possible transfer. The five choices of a human and the machine profile selected at random appeared alternately and one by one in blocks. Each choice was shown for only 0.7 s.[8]

As in part 2, random assignment to the Human and Machine Treatment was contingent upon the subjects' role and delegation decisions from part 1. Thus, they were assigned in equal proportions to both treatments.

2.4. Preparatory sessions

The preparatory sessions were necessary for two reasons. First, they were needed in order to produce actual data from human task-solvers, which would later be presented to subjects in the experiment. Second, they were needed in order to tailor the machine's task-solving performance and decisions in the trust game to the performance and decisions of the humans. Keeping the de-facto performance of humans and machine constant allowed us to test for a potential systematic misperception of relative performances.

In the first part of the preparatory sessions, subjects processed the same calculation task as in the experiment. Also, each subject solved the task not for herself but for another subject with whom she was randomly matched. This receiver was paid according to the task-solvers performance. For this purpose, one of the ten exercises was selected at random and if this exercise was solved correctly, the receiver was given 70 ECU. Otherwise, she received nothing. It was ensured that no pair of subjects solved tasks for each other. So, the mechanism of the matching was the same as described in 2.1 in order to eliminate any potential feelings of reciprocity. Twenty-four subjects took part in each calibration session for the calculation task. Thus, 24 blocks of ten exercises were solved in each session.

[5] Subjects were told the following: "Now you will see the performance of humans and machines again. Please be aware of the fact that the data has been collected in a different past run than the performance that you saw in the first part."

[6] Subjects were told the following: "Every participant is able to send money to the participant assigned to him. This participant cannot decide how much money he wants to send back. This decision is made by a machine. You and the participant assigned to you decide simultaneously, but only one decision is going to be implemented. (...) The amount you transfer will be subtracted from your initial endowment. Subsequently, it will be tripled and send to the participant assigned to you. (...) Afterwards, the machine, deciding for the participant assigned to you, determines the amount of ECU that is returned to you. The participant assigned to you cannot influence the machine's decision. (...). As mentioned above, the participant assigned to you is also able to transfer money. The procedure is the same as already outlined above, meaning the returned amount is determined by your machine agent."

[7] Subjects were told the following: "To be able to form your personal expectations about how much will be sent back, you will be shown the return transfer of participants from an earlier session. To form an expectation about the return transfers of machines, you will see the decisions of the machine agent next to those of human participants."

[8] A choice was indicated by the returned amount for each possible transfer, e.g., "transfer: 30 → return:45.″

The algorithm of the machine that solved the calculation task was programmed in such a way that it resembled the error distribution of the human subjects exactly.[9] So, for instance, if few subjects tended to make many errors, while many subjects made few errors, this was resembled by the algorithm: It made many mistakes in few of the 24 blocks and solved many blocks with few mistakes. The clustering of errors was important to equalize error distribution and account for risk preferences.[10]

Recall that past data on calculation performance was presented twice in the experiment, once before the delegation decision in part 1 and once before the perception guess in part 2. Therefore, two preparatory sessions were performed. We used the data from the first session for part 1 of the experiment, and the data from the second session for part 2.[11]

In the second part of the preparatory sessions, subjects were randomly rematched to new pairs and played a trust game with the same parameters as in the experiment. Using the strategy vector method, each subject gave a reciprocation profile for the case of ending up in the role of a trustee. A random draw then assigned the roles and payoffs were determined according to their own decision for that role and the decisions of their match. The collected 48 reciprocation profiles constituted the pool of data from which the machine agent in part 3 of the experiment randomly picked one and applied it to a trustor's chosen transfer.

Finally, one of the two parts of the preparatory sessions was selected at random and subjects were paid according to their payoff in this part.

3. Experiment results

The experiment took place in a major German university in September 2015. It was programmed in z-Tree (Fischbacher, 2007), subjects were recruited via ORSEE (Greiner et al., 2003). A total of 264 subjects participated in twelve sessions. Subjects received a show-up fee of € 4.00 and could earn additional money in the experiment. A session lasted about 45 min. and the average payment was € 10.38 (*sd*= € 3.45). Task-solvers solved on average 8.58 (*sd* = 2.34) of the ten exercises correctly. The conversion rate was 10 ECU = € 1.00.

First, we checked whether subjects preferred delegating a task that affects a third party to a human over delegating it to a machine. Overall, 132 subjects made a delegation decision. Ninety-seven of these subjects (73.48%) delegated to a human, 35 of them (26.52%) delegated to a machine. The fraction of subjects deciding to delegate to a machine is therefore significantly lower than half ($p < .001$, according to an Exact Binomial Test). This confirms our first hypothesis.

Result 1. Subjects preferred to delegate a task that affects a third party to a human than to a machine.

We now turn to analyze the observers' evaluation of a delegation to a machine as compared to a human. Remember that each observer evaluated both cases—the delegation to a human and to a machine—in a random order. Furthermore, he made choices contingent upon whether the respective task-solver had successfully solved the task or made an error. This means that each observer provided four choices.

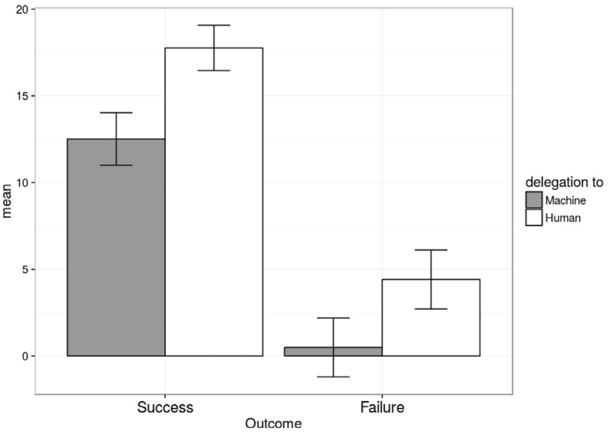


Fig. 2. Observers' rewarding of delegation to machines and to humans.

Observers' levels of rewarding delegators are illustrated in Fig. 2. If the respective task-solver was successful, observers rewarded delegations to a machine with an average of 12.52 ECU (*sd*= 17.38 ECU), while they rewarded delegations to humans with an average of 17.77 ECU (*sd*= 15.01 ECU). If the respective task-solver made an error, observers rewarded delegations to a machine with an average of 0.49 ECU (*sd*= 19.53 ECU), while they rewarded delegations to humans with an average of 4.42 ECU (*sd*= 19.53 ECU). Delegators to machines are thus evaluated significantly worse than delegators to humans regardless of whether the outcomes are successful or unsuccessful ($p < .001$ and $p = .002$, respectively, according to two-sided Wilcoxon signed-rank tests). This confirms our second hypothesis.

Result 2. Delegators were rewarded less for delegating a task that affects a third party to a machine than for delegating it to a human.

In the next two steps, we investigated whether the aversion to machine use in the moral domain identified is based on a lower "perceived utility" of the machine or on a general lack of trust in machines.

First, we compare the number of machine errors guessed by subjects to the number of human errors they guessed. Note that the number of actual errors, i.e., red-colored exercises, presented to subjects was the same for humans and the machine. Specifically, 50 of the 240 exercises shown to subjects on each side of the screen, which was split between human and machine, were shown in red. Subjects who were incentivized to guess the number of machine errors made an average guess of 58.11 (*sd* = 24.88), while those who were incentivized to guess the number of human errors made an average guess of 59.84 (*sd* = 24.43). This difference in guesses is insignificant ($p = .632$ according to a two-sided Mann–Whitney U-test). We thus reject our third hypothesis.

Result 3. Machine errors are not perceived significantly different from human errors.

Second, we tested whether the amount in the trust game transferred by trustors to a machine agent was lower than that sent to a human trustee. Those who were randomly matched with a machine agent sent an average amount of 30.83 ECU (*sd* = 16.67 ECU), while those matched with a human trustee sent an average of 33.64 ECU (*sd*= 15.78 ECU). The difference is insignificant ($p = .170$ according to a two-sided Mann-Whitney U-test).

One might suspect that this insignificance is only an aggregate phenomenon: It may result from the leveling of diverging levels of trust

[9] The algorithm was programmed such that it could not solve exercises in which the sum of numbers was higher than 34. The algorithm was fed with calculation data which led it to reproduce the historical error distribution from the calibration session precisely. Assume the first task-solver in a calibration session had made one mistake, while the second had made three mistakes, and so on. The algorithm was thus fed with an initial block of ten exercises in which one exercise added up to more than 34 and with a second block of ten exercises in which three exercises added up to more than 34. One of the 24 blocks resembling the performance of the task-solvers from the calibration study was randomly drawn to be decisive. The machine then actually calculated this block of ten exercises. Due to its inability to calculate the exercises adding up to more than 34 correctly, it made the same number of errors as the respective human task-solver.

[10] If the machine would have taken the average error rate of all 24 humans and "applied" it to each block, it would have caused a more uniform distribution of errors over the 24 blocks than the humans. In this case, a risk-averse subject might have preferred to delegate the task to a machine, because she feared a particularly weak fellow human subject more than she appreciated a particularly strong fellow human subject.

[11] The average number of lines solved correctly was 8.00 (*sd* = 2.10) in the first session and 7.92 (*sd* = 1.93) in the second session. They were thus very close to each other.

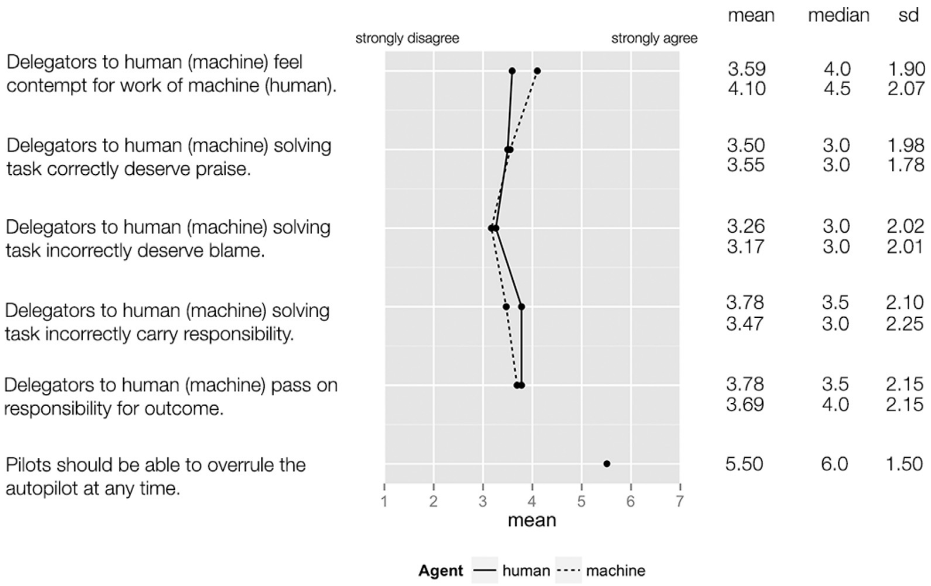


Fig. 3. Results of post-experimental survey.
NOTE: Upper rows of numbers to the right of the graph represent means, medians and standard deviations (sd) for human agent, lower rows represent measures for machine agent.

toward humans and machines between subjects who made different delegation decisions. In particular, one might expect that delegators to humans are generally more skeptical toward machines and express a lower level of trust. Subjects who delegated to a human task-solver in the first part, however, did not, on average, transfer any less to a machine than to a human (31.63 ECU (*sd* = 15.46) vs. 32.92 ECU (*sd* = 16.37), *p* = .627 according to two-sided Mann–Whitney U-tests). Therefore, our fourth hypothesis is also rejected.

Result 4. The level of trust toward machines and toward humans does not differ significantly.

4. Post-experimental survey

Because both potential explanations for the very clear relative aversion to machine use in the moral domain could not be supported in the experiment, we conducted a survey study in February 2018. In this survey, we recruited 78 new participants via ORSEE (Greiner, 2004) and confronted them with a concise description of part 1 of the experiment. They were then asked to indicate their agreement to several statements on a 7-point Likert scale.

We confronted subjects with five pairs of statements that were identical except for the words “human task-solver” and “machine”. The order in which statements were presented was randomized within each pair. Specifically, we investigated the following alternative explanations for the observed aversion to a delegation to a machine. First, people might feel that a delegator who delegated the task to a machine holds a human task-solver’s benevolent effort in contempt (first pair of statements). Second, people might be biased against delegators to machines when attributing praise and blame for a resulting outcome (second and third pair). Third, delegators might successfully pass on the responsibility for negative outcomes to another human but not to a machine (fourth and fifth pair). The eleventh statement was not directly related to part 1 but represented a general remark on automation in a morally relevant domain that we included as a control.

Fig. 3 illustrates the average agreement to the five pairs of statements testing the alternative explanations for an aversion to machine use in the moral domain and to the eleventh statement that served as a control.

The figure suggests that the agreements for each pair of statements concerning delegators to a human and to a machine are quite similar to each other. In fact, none of the differences between the delegators to humans and machines is significant (*p* > .100 according to two-sided Wilcoxon signed-rank tests) except for the first pair. People do indeed more readily agree with the idea that delegators to machines hold a human’s work in contempt than that delegators to humans hold the machine’s work in contempt (*p* = .031). The agreement with the former statement, however, is still close to neutrality. The effect is thus likely to be driven by a perceived implausibility of the idea that a machine’s effort can be held in contempt.

The only statement where participants’ answers tend to clearly deviate from neutrality is the eleventh statement, where people firmly express the opinion that a human pilot should always be able to overrule the decision of an autopilot. While we can thus also identify an aversion to automation in the moral domain in the post-experimental survey, none of the alternative explanations tested in this survey can be supported.

5. Discussion

In this study, we compared the frequency of delegation decisions of a task that affects a third party to machines and humans and elicited their respective evaluation by impartial observers. It should be stressed again that the question we posed here was about people’s preference relation over a machine agent and a human agent, and not about delegating versus performing the task oneself. Consequently, subjects had to delegate in either case and could thus not be blamed merely for shifting responsibility.

We found that subjects express an aversion to delegating tasks that fall into the moral domain to machines rather than humans. First, this manifests in the relatively small fraction of delegators that mandate a machine rather than a human. Second, it is clear that observers evaluate the decision to delegate to a machine in the moral domain less favorably than if the delegation is to a human. Interestingly, machine use is viewed more critically, irrespective of whether the delegation ultimately caused positive or negative consequences for the person affected.

The experiment tested two potential explanations for an aversion to machine use in the moral domain: an oversensitivity to machine errors, and a lack of trust in machines. Both explanations could be ruled out in our experiment. Subjects did not perceive machine errors more saliently than human errors, as subjects' incentivized guessing of failure rates demonstrates. The phenomenon identified, therefore, seems to be an aversion to delegating tasks that affect a third party to machines *per se* as opposed to an instrumentally justified attempt to minimize the risk of failure for those affected. Analogously, the level of trust expressed by subjects toward a machine agent was very similar to the trust level expressed toward a human. Thus, we were unable to identify a general distrust in machines in a self-regarding trust game. This latter finding indicates that the unconditional aversion to machine use seems to be rather specific to the delegation of tasks that affect a third party.

Finally, we used a post-experimental survey with fresh subjects to investigate three alternative explanations for the observed phenomenon: the feeling that delegators to machines hold humans' effort in contempt, a bias against delegators to machines when attributing praise and blame, and a difference in a delegator's ability to pass on responsibility to another human and to a machine. Neither of these potential alternative explanations, however, could account for the aversion at hand.

From our findings, it seems that most people rather intuitively dislike machine use in the moral domain—an intuition which turns out being hard to rationalize. We identified this aversion per se by experimentally equalizing humans' and the algorithm's performance. In practice, however, algorithms will usually not be simulating human moral behavior but will be programmed to implement a specific normative rationale. This attachment to rules may induce a dismissal of their decisional inflexibility in people. Such an instrumental aversion would then come in addition to the non-instrumental aversion that we identified. In the case of self-learning algorithms, we might observe an additional instrumental aversion to decisional opacity.

Our results underline the importance of an open discussion of machine use in the moral domain. The case of automated driving certainly qualifies as such a domain, since errors of the machine may cause substantial externalities to third parties. The non-instrumental aversion identified suggests that the emphasis on the superior performance of automated cars, which is currently the main argument for automation in traffic, may not be sufficient or even decisive in convincing the general public. It might be as important to address the perceived moral problems that are necessarily associated with the introduction of automated vehicles.

Against this background, Chris Urmson, head of Google's self-driving car project, might be mistaken in downplaying the role of moral considerations in the context of automated driving by calling them "a fun problem for philosophers to think about" (McFarland, 2015). As this empirical study suggests, concerns regarding the involvement of machines in the moral domain are not only an issue for armchair philosophers but may reflect a larger societal phenomenon, viz. a folk aversion (see also Kohl et al. (2018)). So far, the industry seems to mainly be occupied with engineering issues and has, due to a déformation professionnelle, predominantly neglected or downplayed the possibility of public resistance to the new technology. It may, however, be well-advised to take moral concerns against automated driving seriously, since citizens' resistance may slow down the automation process substantially. This, however, would mean to preserve a status quo that involves an avoidably high number of traffic deaths, injuries and damages.

Research that investigates how the feeling of unease can be addressed prophylactically (Feldhütter et al., 2016) is just emerging. Enabling people to experience, and thus better understand, the technology in order to dissipate reservations and fears may pave the way for a trouble-free introduction of autonomous driving. A deeper investigation of the causes of people's aversion to the use of automated cars in the moral domain seems to us a promising venue for future research.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:http://dx.doi.org/10.1016/j.socec.2018.04.003.

References

Davis, F.D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 13 (3), 319–340.

de Visser, E.J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., Parasuraman, R., 2012. The world is not enough: trust in cognitive agents. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 56, 263–267. http://dx.doi.org/10.1177/1071181312561062.

Dzindolet, M.T., Beck, H.P., Pierce, L.G., Dawe, L.A., 2001. A Framework of Automation Use. Army Research Laboratory, Aberdeen Proving Ground.

Dzindolet, M.T., Pierce, L.G., Beck, H.P., Dawe, L.A., 2002. The perceived utility of human and automated aids in a visual detection task. Human Factors: The Journal of the Human Factors and Ergonomics Society 44 (1), 79–94.

Fagnant, D.J., Kockelman, K., 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. Transportation Research Part A 77, 167–181.

Feldhütter, A., Gold, C., Hüger, A., Bengler, K., 2016. Trust in automation as a matter of media and experience of automated vehicles. Proceedings of the Human Factors and Ergonomics Society Sixtieth Annual Meeting.

Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10 (2), 171–178.

Gao, P., Hensley, R., Zielke, A., 2014. A road map to the future for the auto industry. McKinsey Quarterly (4), 42–53.

Gogoll, J., Müller, J.F., 2017. Autonomous cars: in favor of a mandatory ethics setting. Science and Engineering Ethics 23 (3), 681–700.

Goodall, N., 2014. Ethical decision making during automated vehicle crashes. Transportation Research Record: Journal of the Transportation Research Board 24, 58–65.

Greiner, B., Levati, M.V., 2005. Indirect reciprocity in cyclical networks: an experimental study. Journal of Economic Psychology 26 (5), 711–731.

Greiner, B., 2004. The online recruitment system orsee 2.0-a guide for the organization of experiments in economics. University of Cologne, Working paper series in economics 10 (23), 63–104.

Hevelke, A., Nida-Rümelin, J., 2014. Responsibility for crashes of autonomous vehicles: an ethical analysis. Science and Engineering Ethics 21 (3), 619–630.

Kohl, C., Knigge, M., Baader, G., Böhm, M., Krcmar, H., 2018. Anticipating acceptance of emerging technologies using twitter: the case of self-driving cars. Journal of Business Economics. Forthcoming.

Lee, J.D., Moray, N., 1994. Trust, self-confidence, and operators' adaptation to automation. International Journal of Human-Computer Studies 40, 153–184.

Lee, J.D., See, K.A., 2004. Trust in automation: designing for appropriate reliance. Human Factors 46 (1).

McFarland, M., 2015. Google's chief of self-driving cars downplays dilemma of ethics and accidents. The Washington Post. https://www.washingtonpost.com/news/innovations/wp/2015/12/01/googles-leader-on-self-driving-cars-downplays-the-trolley-problem/.

Muir, B., Moray, N., 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics 39 (3), 429–460. http://dx.doi.org/10.1080/00140139608964474.

Ray, C., Mondada, F., Siegwart, R., 2008. What do people expect from robots? Proceedings of the IEEE/RSJ International Conference Intelligent Robots and Systems. pp. 3816–3821.

Ross, M., Sicoly, F., 1979. Egocentric biases in availability and attribution. Journal of Personality and Social Psychology 37 (3), 322–336.

Rousseau, D., Sitkin, S., Burt, R., Camerer, C., 1998. Not so different after all: a cross-discipline view of trust. Academy of Management Review 23, 393–404.

Selten, R., 1967. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In: Sauermann, H. (Ed.), Beiträge zur experimentellen Wirtschaftsforschung. JCB Mohr, Tübingen.

Shanker, R., Jonas, A., Devitt, S., Huberty, K., Flannery, S., Greene, W., Swinburne, B., Locraft, G., Wood, A., Weiss, K., Moore, J., Schenker, A., Jain, P., Ying, Y., Kakiuchi, S., Hoshino, R., Humphrey, A., 2013. Autonomous cars: self-driving the new auto industry paradigm. Morgan Stanley & Co. LLC Morgan Stanley Blue Paper.

Stangor, C., McMillan, D., 1992. Memory for expectancy-congruent and expectancy-incongruent information: a review of the social and social developmental literatures. Psychological Bulletin 111 (1), 42–61.

Stangor, C., Ruble, D.N., 1989. Strength of expectancies and memory for social information: what we remember depends on how much we know. Journal of Experimental Social Psychology 25 (1), 18–35.

Svenson, O., 1981. Are we all less risky and more skillful than our fellow drivers? Acta Psychologica 47 (2), 143–148.

Appendix 1: Instructions (translated from German)

Explanations

During the experiment subjects were assigned to different roles. The role is indicated at the top of each slide. Generally, slides are in chronological order. Although, the different roles acted simultaneously, the slides of each role are depicted consecutively to improve readability.

Some slides (e.g., waiting screens) and features (e.g., buttons to click in order to advance) were eliminated as long as they did not serve any significant purpose.

Part 1 - Delegation and Blaming

Role: Everyone

Welcome!

You will now receive a participation letter.

RANDOM DRAW

The letter *X* has been assigned to you.

Role: Everyone

The experiment consists of three parts. You can earn money in each of these parts but only your earnings in one of these parts will actually be paid out to you. The relevant part is chosen at random. The three parts are independent and do not build on each other.

Your identity remains unknown and will not be revealed during the experiment.

Role: Delegator (1)

During the first part of the experiment, two participants out of this room will be assigned to you at random.

RANDOM DRAW

Participants *Y* and *Z* have been assigned to you.

Your personal payment depends on the decision of *Z*.

The payment of participant *Y* depends on your decision.

Role: Delegator (2)

You will be presented with 10 lines. Each line consists of 7 fields. Each field displays a number between 1 and 9. All fields of any given line must be summed up. Subsequently, the result has to be entered into the accompanying input field and confirmed by clicking on the “OK” button. Below you can see an example of such a line. Please sum up the numbers of this line probationary and enter the result into the accompanying input field and confirm your input by clicking “OK”.

TEST LINE TO SOLVE

Role: Delegator (3)

The payment of participant *Y* depends on your decision regarding the calculation task. You can either delegate the calculation of the lines to another person (different from the participants assigned to you) or to a machine. The person is someone different than the assigned participants *Y* and *Z*. Afterwards, your delegation decision will be evaluated by a person that is unknown to you. This person has the option to add or subtract money from your endowment, depending on the consequences of your decision.

Role: Delegator (4)

You receive an endowment of 30 ECU (€ 3) that can be increased during the first part. Your personal payment also depends on the decision of the participant assigned to you. Out of the ten lines to be solved, one is picked at random. Hereby every single line has the same probability to be selected. If participant *Z* calculated the result correctly, you will receive 70 ECU (€ 7) additionally leading to a total payment of ECU 100 (€ 10.00). If the result of this line has been calculated incorrectly, you will not receive any additional payment to your original endowment of ECU 30 (€ 3).

The procedure is the same for all participants, meaning that participant *Y* will be rewarded according to your delegation and the performance of your agent.

Role: Delegator (5)

To evaluate the performance of a person and a machine, you will subsequently see a comparison of the performance of a past run. One line is shown per column, each respectively calculated either by a person or a machine. If calculated correctly, the line will be displayed in white. If calculated incorrectly the line will be displayed in red.

Role: Delegator (6)

PERCEPTION TASK IS SHOWN

Role: Delegator (7)

Your decision affects participant *Y*. You have to delegate the calculation of the lines to another person or to a machine. The result of this delegation is relevant for participant *Y*, who has been assigned to you.

CHOOSE MACHINE *CHOOSE HUMAN* (*RANDOMIZED ORDER*)

Role: Delegator (8)

You now have to solve the addition tasks for the case that you have been chosen to calculate the lines of another participant. Please consider that your results are relevant to another participant that is unknown to you. This is only relevant if you have been selected instead of a machine.

Role: Observer (1)

During the first part of the experiment, one participant out of this room will be assigned to you randomly. This participant is Y. For the first part of the experiment you will receive a fixed payment of 100 ECU (€10). You will be presented with 10 lines. Each line consists of 7 fields. Each field displays a number between 1 and 9. All fields of any given line must be summed up. Subsequently, the result has to be entered into the accompanying input field and confirmed by clicking on the “OK” button. Below you can see an example of such a line. Please sum up the numbers of this line probationary and enter the result into the accompanying input field and confirm your input by clicking “OK”.

TEST LINE TO SOLVE

Role: Observer (2)

The participant assigned to you now has the opportunity to delegate this task to another person or to a machine. Therefore, his decision becomes relevant to a person that is unknown to you. The payment of participant *Y* depends on a delegation decision as well. This decision is made by an unknown participant who also delegates the calculation to another person or to a machine. Thus, the participant assigned to you is responsible for the delegation, relevant for a third person that is unknown to you. Additionally, the payment of participant *Y* depends on a participant that is unknown to you, who is in exactly the same situation as you are in.

Role: Observer (3)

The payment of participant *Y* depends on the decision of another participant, who is assigned to your participant. Out of the ten lines to be solved, one is picked at random. Each single line has the same probability to be selected. If the participant, who is deciding for participant *Y*, delegated the calculation to another person or to a machine, solving it correctly, your participant receives 70 ECU (€ 7). If the result is wrong, participant *Y* receives no additional payment. The procedure is the same for all participants, meaning that another participant is rewarded according to the delegation decision of the participant assigned to your participant and the performance of the agent who was selected at random.

Role: Observer (4)

To evaluate the performance of a person and a machine, you will subsequently see a comparison of the performance of a past run. One line is shown per column, each respectively calculated either by a person or a machine. If calculated correctly, the line will be displayed in white. If calculated incorrectly the line will be displayed in red.

Role: Observer (5)

PERCEPTION TASK IS SHOWN

Role: Everyone

Please calculate the lines underneath. Also consider, that the result of each line has to be entered in the input field on the right hand side. Please confirm your input by clicking the "OK" button. You have 105 seconds to solve the whole task.

Role: Observer (6)

In what follows you have the opportunity to judge the delegation decision of participant *Y*. Note that participant Y's delegation decision impacts the earnings of another participant that is unknown to you. You have the opportunity to add or subtract money from the endowment of participant *Y*. This is only possible in integer numbers, while the maximum increase or reduction is 30 ECU.

The participant assigned to you delegates to	Add	Subtract
Machine—a correctly solved line is drawn (unknown participant receives 70 ECU)	*INPUT FIELD*	*INPUT FIELD
Machine—an incorrectly solved line is drawn (unknown participant receives 0 ECU)	*INPUT FIELD*	*INPUT FIELD
Human—a correctly solved line is drawn (unknown participant receives 70 ECU)	*INPUT FIELD*	*INPUT FIELD
Human—an incorrectly solved line is drawn (unknown participant receives 0 ECU)	*INPUT FIELD*	*INPUT FIELD

ORDER IN WHICH CHOICE CONCERNING MACHINE AND HUMAN IS PRESENTED IS RANDOMIZED

Part 2 - Perceived Ability

Role: Everyone

Now you will see the performance of humans and machines again. Please be aware of the fact that the data has been collected in a different past run than the performance that you saw in the first part. You will subsequently see a comparison of the performance of a past run. One line is shown per column, each respectively calculated either by a person or a machine. If calculated correctly, the line will be displayed in white. If calculated incorrectly the line will be displayed in red.

Afterwards, you will be questioned about the error rate and rewarded depending on the accuracy of your guess.

Role: Perceiver of machine (1)

Please estimate the performance of the machine. Give your estimation on how many lines out of 240 have been solved incorrectly by the machine. The following table displays the payments per deviation of your guess.

Deviation	Payoff
$\leq$ 20 %	70 ECU
$\leq$ 40 %	40 ECU
$\leq$ 60 %	20 ECU
$>$ 60 %	0 ECU

Role: Perceiver of machine (2)

Please enter your estimated number of errors made by the machine in the input field below.

INPUT FIELD

Role: Perceiver of human (1)

Please estimate the performance of the machine. Give your estimation on how many lines out of 240 have been solved incorrectly by the human. The following table displays the payments per deviation of your guess.

Deviation	Payoff
$\leq$ 20 %	70 ECU
$\leq$ 40 %	40 ECU
$\leq$ 60 %	20 ECU
$>$ 60 %	0 ECU

Role: Perceiver of human (2)

Please enter your estimated number of errors made by the person in the input field below.

INPUT FIELD

Part 3 - Trust

Role: Everyone

In the third part, you will be assigned to another participant. You form a pair with him.

Role: Everyone (2)

You and the participant assigned to you decide simultaneously, but only one decision is going to be implemented. You will start with an endowment of 50 ECU for the case that your decision is going to be implemented. Now you have the opportunity to send money to the participant assigned to you. You can transfer a maximum of 50 ECU and a minimum of 0 ECU. Amounts of 10, 20, 30 or 40 ECU are also possible. The amount you transfer will be subtracted from your initial endowment. Subsequently, it will be tripled and send to the participant assigned to you. As an example a transferred amount of 30 ECU results in a credit of 90 ECU on your participant's account.

The decision of how much will be returned to you is either made by your partner or by a machine that is acting on behalf of your partner. In the latter case your partner has no say in the decision of how much is sent back.

Role: Everyone (3)

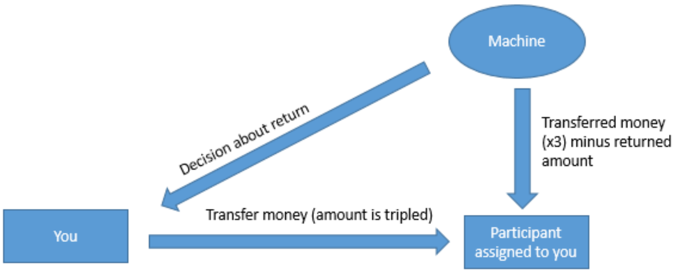
To be able to form your personal expectations about how much will be sent back, you will be shown the return transfer of participants from an earlier session. To form an expectation about the return transfers of machines, you will see the decisions of the machine agent next to those of human participants.

PERCEPTION TASK IS SHOWN

Role: Matched with machine (1)

Every participant is able to send money to the participant assigned to him. This participant cannot decide how much money he wants to send back. This decision is made by a machine.

Role: Matched with machine (2)


Machine
Decision about return
Transferred money (x3) minus returned amount
You
Transfer money (amount is tripled)
Participant assigned to you

You and the participant assigned to you decide simultaneously, but only one decision is going to be implemented. You will start with an endowment of 50 ECU for the case your decision is implemented. Now you have the opportunity to send money to the participant assigned to you. You can transfer a maximum of 50 ECU and a minimum of 0 ECU. Amounts of 10, 20, 30 or 40 ECU are also possible. The amount you transfer will be subtracted from your initial endowment. Subsequently, it will be tripled and send to the participant assigned to you. As an example, a transferred amount of 30 ECU results in a credit of 90 ECU on your participant's account.

Afterwards, the machine, deciding for the participant assigned to you, determines the amount of ECU that is returned to you. The participant assigned to you cannot influence the machine's decision. The returned amount can vary between 0 ECU (minimum) and 150 ECU (maximum).

As mentioned above, the participant assigned to you is also able to transfer money. The procedure is the same as already outlined above, meaning the returned amount is determined by your machine agent.

One of the two transactions (your transaction or the one of the participant assigned to you) is selected at random and classified as a payment for the third part. If the decision of the participant assigned to you is implemented, you will receive the tripled amount that has been transferred to you subtracted by the amount returned by the machine acting on your behalf.

For the case your decision is going to be implemented, please put the amount you want to transfer into the following input field.

INPUT FIELD

Role: Matched with human (1)

You and the participant assigned to you decide simultaneously, but only one decision is going to be implemented. You will start with an endowment of 50 ECU for the case that your decision is going to be implemented. Now you have the opportunity to send money to the participant assigned to you. You can transfer a maximum of 50 ECU and a minimum of 0 ECU. Amounts of 10, 20, 30 or 40 ECU are also possible. The amount you transfer will be subtracted from your initial endowment. Subsequently, it will be tripled and send to the participant assigned to you. As an example a transferred amount of 30 ECU results in a credit of 90 ECU on your participant's account.

He can now decide what amount is returned to you. The returned amount can vary between 0 ECU (minimum) and 150 ECU (maximum).

As mentioned above, the participant assigned to you is also able to transfer money. The procedure is the same as already outlined above. One of the two transactions (your transaction or the one of the participant assigned to you) is selected at random and classified as a payment for the third part. If the decision of the participant assigned to you is implemented, you will receive the tripled amount that has been transferred to you subtracted by the amount returned by you.

For the case that your decision is going to be implemented, please put the amount you want to transfer into the following input field.

INPUT FIELD

Role: Matched with human (2)

Simultaneously with your decision, the participant assigned to you made his decision about the amount of money to be transferred for the case that his decision is going to be implemented.

Please enter the amount of money you would like to return if your the participant assigned to you selected the shown amount in the table below. If you do not want to transfer any money, insert the number "0". Please note that every amount transferred to you by the participant assigned to you will be tripled.

If the participant assigned to me transferred 10 ECU (you receive 30 ECU), I will transfer: *INPUT FIELD*
If the participant assigned to me transferred 20 ECU (you receive 60 ECU), I will transfer: *INPUT FIELD*
If the participant assigned to me transferred 30 ECU (you receive 90 ECU), I will transfer: *INPUT FIELD*
If the participant assigned to me transferred 40 ECU (you receive 120 ECU), I will transfer: *INPUT FIELD*
If the participant assigned transferred 50 ECU (you receive 150 ECU), I will transfer: *INPUT FIELD*

Payment (Example)

Role: Everyone (1)

Thank you. The decision to be implemented will now be picked at random.

RANDOM DRAW

The decision of the participant assigned to you has been implemented. He transferred 10 ECU. 30 ECU are credited to your account. You returned 15 ECU.

You receive 15 ECU.

Role: Everyone (2)

Your guess for the second part of the experiment is now evaluated.

The machine made 50 errors and you estimated 42 errors.

You receive 70 ECU.

Role: Everyone (3)

Now the part of the experiment relevant for your payment (part 1, part 2 or part 3) will be picked at random. All three parts are equally likely to be chosen as relevant for payment.

RANDOM DRAW

Part *x* is relevant for your payment.

You receive *y* ECU for this part. This equals *€ $0.10 \cdot y$*. Additionally, you receive a show-up fee of € 4.00.

In total, you receive *€ $0.10 \cdot y$ + € 4.00*.

Appendix 2: Post-Experimental Survey (translated from German)

Description of the Situation

Recently, a scientific experiment was conducted. In this experiment, deciders had to delegate the solving of ten calculation exercises. The task could either be delegated to a human task-solver or to a machine. The deciders did not have the possibility to solve the task themselves.

The payoff of a passive third party depended on the correct solving of the task by the machine or the human task-solver. For each correctly solved exercise, a green ball was put into an urn, while for each incorrectly solved exercise a red ball was put into it. Then, one ball was randomly drawn and the passive third party received a payment of €7.00, if a green ball was drawn and nothing, if a red ball was drawn.

The deciders who had to delegate the task to a human task-solver or a machine knew the following:

a. The machine solves the same fraction of exercises incorrectly as the human task-solvers. It is thus neither superior nor inferior in its capabilities.

b. All human task-solvers had to solve the task. But only the performance of those task-solvers was considered whose decider had determined that their solution should be relevant for the third party's payment. The performance of the task-solvers whose decider had chosen the machine was thus irrelevant. The task-solvers, however, did not find out whether their performance was ultimately relevant. Therefore, by delegation to a machine, nobody was spared from working.

THE ORDER WITHIN EACH PAIR OF STATEMENTS WAS RANDOMIZED

1. People who delegate the task to another human hold the work of the machine in contempt.

   7-POINT LIKERT SCALE RANGING FROM 1 (STRONGLY DISAGREE) to 7 (STRONGLY AGREE)*

2. People who delegate the task to a machine hold the work of other people in contempt.

3. People who delegate the task to another human who then commits an error are to blame for this outcome.

4. People who delegate the task to a machine that then commits an error are to blame for this outcome.

5. People who delegate the task to another human who then solves it correctly deserve praise for this outcome.
6. People who delegate the task to a machine that then solves it correctly deserve praise for this result.
7. People who delegate the task to another human pass on their responsibility for the outcome.
8. People who delegate the task to a machine pass on their responsibility for the outcome.
9. People who delegate the task to another human who then commits an error carry the responsibility for this outcome.
10. People who delegate the task to a machine that then commits an error carry the responsibility for this outcome.
11. Pilots should be able to overrule the decisions of the autopilot at any time.

Appendix 3: Additional Illustrations

Figure 4: Visualization of Calculation Performance from Preparatory Sessions—Example 1

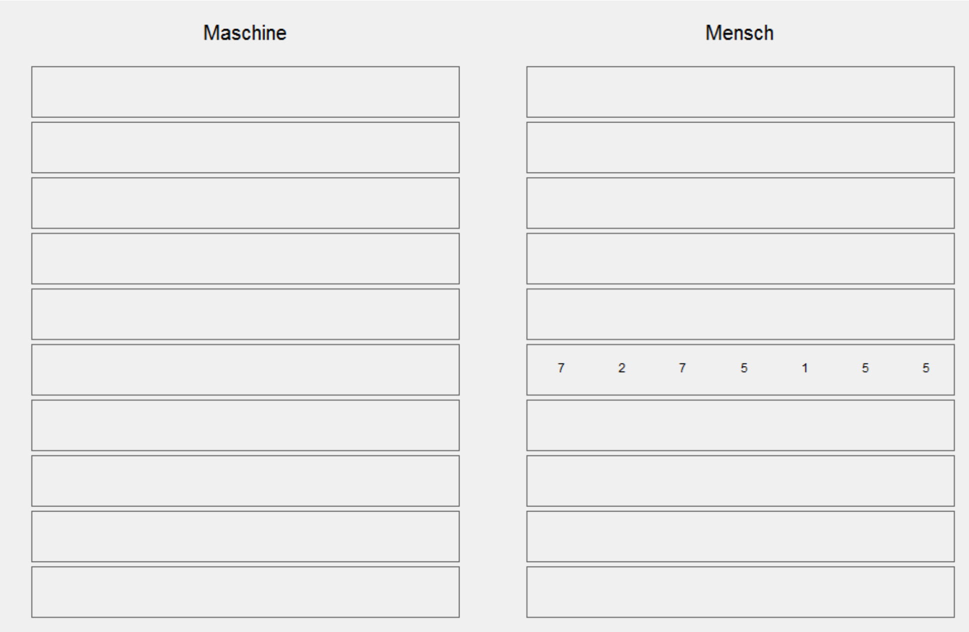


Figure 5: Visualization of Calculation Performance from Preparatory Sessions—Example 2

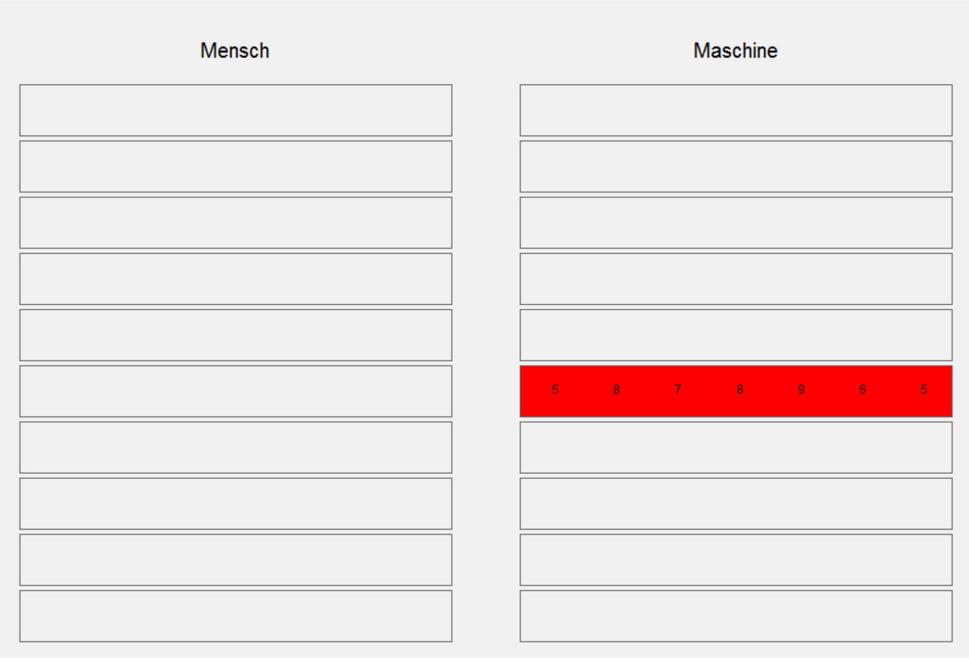
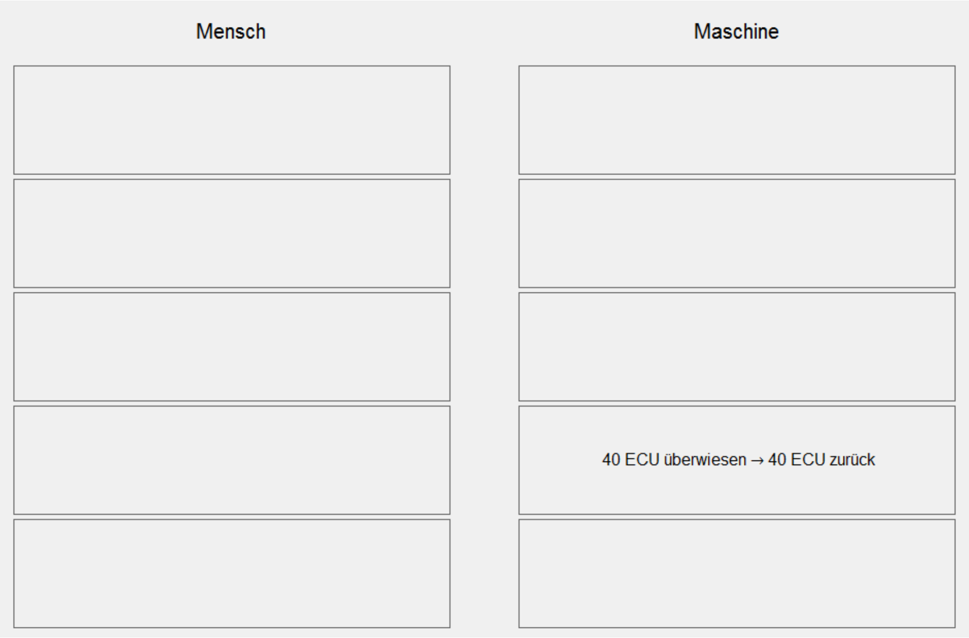
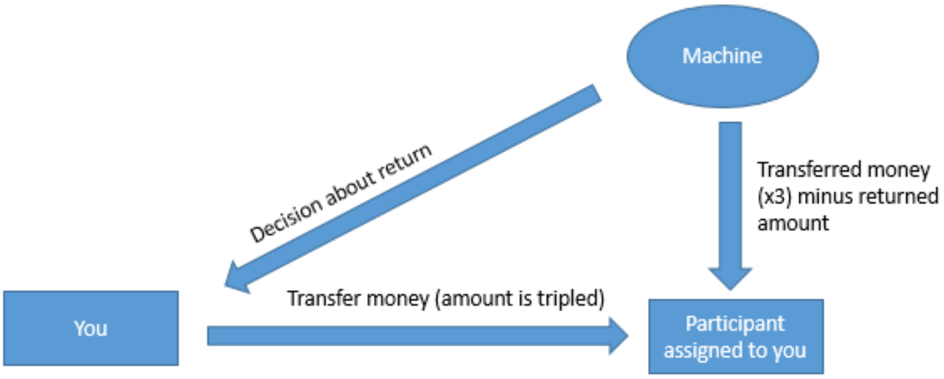


Figure 6: Visualization of Reciprocity Choices from Preparatory Sessions



Mensch	Maschine
	40 ECU überwiesen → 40 ECU zurück

Figure 7: Visualization of Trust Game with Machine


Machine
Decision about return
Transferred money (x3) minus returned amount
You
Transfer money (amount is tripled)
Participant assigned to you

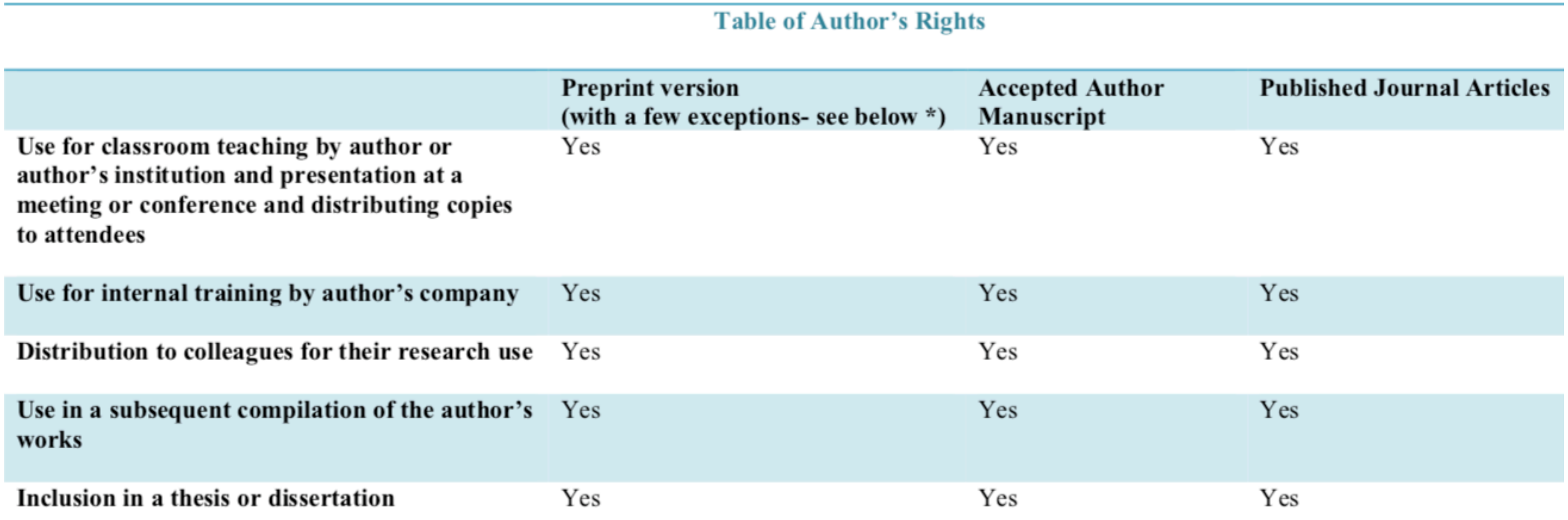


Table of Author's Rights

	Preprint version (with a few exceptions- see below *)	Accepted Author Manuscript	Published Journal Articles
Use for classroom teaching by author or author's institution and presentation at a meeting or conference and distributing copies to attendees	Yes	Yes	Yes
Use for internal training by author's company	Yes	Yes	Yes
Distribution to colleagues for their research use	Yes	Yes	Yes
Use in a subsequent compilation of the author's works	Yes	Yes	Yes
Inclusion in a thesis or dissertation	Yes	Yes	Yes

Figure A.1: Rights of authors. Source: `https://www.elsevier.com/_data/assets/pdf_file/0007/55654/AuthorUserRights.pdf`

B

Contributions to Publications

Publication 1: Autonomous Cars: In Favor of a Mandatory Ethic Setting

Published in *Journal of Science and Engineering Ethics*, 2017, 23(3):681–700 with Julian Müller.

- I was responsible for the development of the research question
- I was responsible for the literature review
- I created the game-theroretic model
- As corresponding author I was responsible for the submission process and the coordination of the revise and resubmit process
- The writing of the article was a joint effort by the team of authors
- The revision was a joint effort by the team of authors

Jan Gogoll (lead author)

Julian Müller (Co-author)

Publication 2: Rage against the machine: Automation in the moral domain

Published in *Journal of Behavioral and Experimental Economics*, 2018, 74:97–103 with Matthias Uhl.

- I was responsible for the development of the research question
- I conducted all the laboratory sessions
- I conducted the post-experimental survey
- The design of the experiment was developed by the team of authors
- I was responsible for writing the article
- As corresponding author I was responsible for the submission process and the coordination of the revise and resubmit process
- The revision was a joint effort by the team of authors

Jan Gogoll (lead author)

Matthias Uhl (Co-author)

References

Ahlenius, H. and Tännsjö, T. (2012), 'Chinese and westerners respond differently to the trolley dilemmas', *Journal of Cognition and Culture* 12(3-4), 195–201.

Alexander-Kearns, M., Peterson, M. and Cassady, A. (2016), 'The impact of vehicle automation on carbon emissions', *Center for American Progress* .

Arndt, S. (2011), Untersuchung zur Überprüfung der Akzeptanzvorhersage, *in* 'Evaluierung der Akzeptanz von Fahrerassistenzsystemen', Springer, pp. 173–207.

Bischoff, J. and Maciejewski, M. (2016), 'Autonomous taxicabs in berlin–a spatiotemporal analysis of service performance', *Transportation Research Procedia* 19, 176–186.

Blanco, M., Atwood, J., Russell, S., Trimble, T., McClafferty, J. and Perez, M. (2016), Automated vehicle crash rate comparison using naturalistic data, Technical report, Virginia Tech Transportation Institute.

Bleske-Rechek, A., Nelson, L. A., Baker, J. P., Remiker, M. W. and Brandt, S. J. (2010), 'Evolution and the trolley problem: People save five over one unless the one is young, genetically related, or a romantic partner.', *Journal of Social, Evolutionary, and Cultural Psychology* 4(3), 115.

BMVI (2016), 'Ethics commission – automated and connected driving'. https://www.bmvi.de/SharedDocs/EN/Documents/G/ethic-commission-report.pdf?__blob=publicationFile.

Boeglin, J. (2015), 'The costs of self-driving cars: reconciling freedom and privacy with tort liability in autonomous vehicle regulation', *Yale JL & Tech.* 17, 171.

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

Brandom, R. (2016), 'Who will you decide to kill with your self-driving car? let's find out!', *The Verge* .

Caiazzo, F., Ashok, A., Waitz, I. A., Yim, S. H. and Barrett, S. R. (2013), 'Air pollution and early deaths in the united states. part i: Quantifying the impact of major sectors in 2005', *Atmospheric Environment* 79, 198–208.

Carr, N. (2015), *The glass cage: Where automation is taking us*, Random House.

Ciaramelli, E., Muccioli, M., Làdavas, E. and di Pellegrino, G. (2007), 'Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex', *Social cognitive and affective neuroscience* 2(2), 84–92.

Collingridge, D. (1980), 'The social control of technology (london: Frances pinter, 1980)', *CollingridgeThe Social Control of Technology* .

Council of the European Union (2013), 'Law enforcement technology services (enlets) 2014 - 2020 - work programme'. http://www.statewatch.org/news/2014/jan/eu-enlets-wp-2014-2020.pdf.

Daily Mail (2012), '1.5million account numbers hacked after visa and mastercard card data theft', *Daily Mail* . http://www.dailymail.co.uk/news/article-2123854/1-5million-account-numbers-hacked-Visa-Mastercard-card-data-theft.html.

Di Nucci, E. (2013), 'Self-sacrifice and the trolley problem', *Philosophical Psychology* 26(5), 662–672.

Edmonds, D. (2013), *Would you kill the fat man: The trolley problem and what your answer tells us about right and wrong*, Princeton University Press.

Fagnant, D. J. and Kockelman, K. (2015), 'Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations', *Transportation Research Part A: Policy and Practice* 77, 167–181.

Fagnant, D. J. and Kockelman, K. M. (2016), 'Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in austin, texas', *Transportation* pp. 1–16.

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütfeld, L. R., Stephan, A., Pipa, G. and König, P. (2017), 'Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles', *Science and Engineering Ethics* pp. 1–20.

Foot, P. (1967), 'The problem of abortion and the doctrine of double effect', *Oxford Review* 5.

Gao, P., Hensley, R. and Zielke, A. (2014), 'A road map to the future for the auto industry', *McKinsey Quarterly, Oct* .

Gogoll, J. and Müller, J. F. (2017), 'Autonomous cars: in favor of a mandatory ethics setting', *Science and engineering ethics* 23(3), 681–700.

Gogoll, J. and Uhl, M. (2018), 'Rage against the machine: Automation in the moral domain', *Journal of Behavioral and Experimental Economics* 74, 97–103.

Gold, N., Pulford, B. D. and Colman, A. M. (2014), 'The outlandish, the realistic, and the real: contextual manipulation and agent role effects in trolley problems', *Frontiers in psychology* 5, 35.

Goodall, N. (2014), 'Ethical decision making during automated vehicle crashes', *Transportation Research Record: Journal of the Transportation Research Board* (2424), 58–65.

Greene, J. (2014), *Moral tribes: Emotion, reason, and the gap between us and them*, Penguin.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M. and Cohen, J. D. (2004), 'The neural bases of cognitive conflict and control in moral judgment', *Neuron* 44(2), 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. and Cohen, J. D. (2001), 'An fmri investigation of emotional engagement in moral judgment', *Science* 293(5537), 2105–2108.

Hansson, S. O. (2003), 'Ethical criteria of risk acceptance', *Erkenntnis* 59(3), 291–309.

Hauser, M. (2006), *Moral minds: How nature designed our universal sense of right and wrong.*, Ecco/HarperCollins Publishers.

Homann, K. (2015), 'Das Können des moralischen Sollens: Die ökonomische Problematik', *Ethica, Jg* 23, 243–259.

Howard, D. (2013), 'Robots on the road: The moral imperative of the driverless car', *University of Notre Dame.* . http://donhoward-blog.nd.edu/2013/11/07/robots-on-the-road-the-moral-imperative-of-the-driverless-car/#.Wt8I6i9rT64.

Huebner, B. and Hauser, M. D. (2011), 'Moral judgments about altruistic self-sacrifice: When philosophical and folk intuitions clash', *Philosophical Psychology* 24(1), 73–94.

Hume, D. (1978), 'A treatise of human nature 1739', *London: John Noon* .

Jauernig, J. (2017), Using Experiments in Ethics, Dissertation, Technische Universität München.

Johnsen, A., Kraetsch, C., Možina, K., Rey, A., Patten, C. and Takman, J. (2017), 'D2. 1 literature review on the acceptance and road safety, ethical, legal, social and economic implications of automated vehicles', *Brave project* .

Kahane, G. (2013), 'The armchair and the trolley: an argument for experimental ethics', *Philosophical studies* 162(2), 421–445.

Kant, I. (2003), *Kritik der praktischen Vernunft*, Felix Meiner Verlag.

Knobe, J. (2003), 'Intentional action and side effects in ordinary language', *Analysis* 63(279), 190–194.

KPMG (2015), 'Marketplace of change: Automobile insurance in the era of autonomous vehicles – whitepaper'. https://assets.kpmg.com/content/dam/kpmg/pdf/2016/06/id-market-place-of-change-automobile-insurance-in-the-era-of-autonomous-vehicles.pdf.

Kuderer, M., Gulati, S. and Burgard, W. (2015), Learning driving styles for autonomous vehicles from demonstration, *in* 'Robotics and Automation (ICRA), 2015 IEEE International Conference on', IEEE, pp. 2641–2646.

Kyriakidis, M., Happee, R. and de Winter, J. C. (2015), 'Public opinion on automated driving: Results of an international questionnaire among 5000 respondents', *Transportation research part F: traffic psychology and behaviour* 32, 127–140.

Lee, J. D. and Moray, N. (1992), 'Trust, control strategies and allocation of function in human-machine systems', *Ergonomics* 35(10), 1243–1270.

Lee, J. D. and Moray, N. (1994), 'Trust, self-confidence, and operators' adaptation to automation', *International journal of human-computer studies* 40(1), 153–184.

Lee, J.-G., Kim, K. J., Lee, S. and Shin, D.-H. (2015), 'Can autonomous vehicles be safe and trustworthy? effects of appearance and autonomy of unmanned driving systems', *International Journal of Human-Computer Interaction* 31(10), 682–691.

Lee, T. (2013), 'Self-driving cars are a privacy nightmare. and it's totally worth it.', *The Washington Post* . https://www.washingtonpost.com/news/wonk/wp/2013/05/21/self-driving-cars-are-a-privacy-nightmare-and-its-totally-worth-it/?noredirect=on&utm_term=.5c84861e109b.

Liao, S. M., Wiegmann, A., Alexander, J. and Vong, G. (2012), 'Putting the trolley in order: Experimental philosophy and the loop case', *Philosophical Psychology* 25(5), 661–671.

Lin, P. (2014), 'What if your autonomous car keeps routing you past krispy kreme?', *The Atlantic* . https://www.theatlantic.com/technology/archive/2014/01/what-if-your-autonomous-car-keeps-routing-you-past-krispy-kreme/283221/.

Lin, P. (2016), Why ethics matters for autonomous cars, *in* 'Autonomous Driving', Springer, pp. 69–85.

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

Mangan, J. T. (1949), 'An historical analysis of the principle of double effect', *Theological Studies* 10(1), 41–61.

Mccarthy, J. F. (2017), Sustainability of Self-Driving Mobility: An Analysis of Carbon Emissions Between Autonomous Vehicles and Conventional Modes of Transportation, PhD thesis. http://nrs.harvard.edu/urn-3:HUL.InstRepos:33813411.

McIntyre, A. (2014), Doctrine of double effect, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', winter 2014 edn, Metaphysics Research Lab, Stanford University.

Millar, J. (2014), 'Proxy prudence: Rethinking models of responsibility for semi-autonomous robots', *We Robot 2014 Proceedings* . http://robots.law.miami.edu/2014/wp-content/uploads/2013/06/Proxy-Prudence-Rethinking-Models__-of-Responsibility-for-Semi-autonomous-Robots-Millar.pdf.

Narnakaje, S. (2017), Ti's smart sensors ideal for automated driving applications, Technical report, Texas Instruments.

Navarrete, C. D., McDonald, M. M., Mott, M. L. and Asher, B. (2012), 'Virtual morality: Emotion and action in a simulated three-dimensional "trolley problem".', *Emotion* 12(2), 364.

NHTSA (2013), 'Preliminary statement of policy concerning automated vehicles'. http://www.nhtsa.gov/staticfiles/rulemaking/pdf/Automated_Vehicles_Policy/.

NHTSA (2018), 'Automated vehicles for safety'. https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety.

OECD (2015), 'Urban mobility system upgrade. how shared self-driving cars could change city traffic'.

Petrinovich, L. and O'Neill, P. (1996), 'Influence of wording and framing effects on moral intuitions', *Ethology and Sociobiology* 17(3), 145–171.

Piao, J., McDonald, M., Hounsell, N., Graindorge, M., Graindorge, T. and Malhene, N. (2016), 'Public views towards implementation of automated vehicles in urban areas', *Transportation Research Procedia* 14, 2168–2177.

Rawls, J. (1971), 'A theory of justice (cambridge', *Mass.: Harvard University* .

Roth, A. (1986), 'Laboratory experimentation in economics', *Economics and Philosophy* 2, 245–273.

Schoettle, B. and Sivak, M. (2014), 'A survey of public opinion about autonomous and self-driving vehicles in the us, the uk, and australia'.

Skulmowski, A., Bunge, A., Kaspar, K. and Pipa, G. (2014), 'Forced-choice decision-making in modified trolley dilemma situations: a virtual reality and eye tracking study', *Frontiers in behavioral neuroscience* 8, 426.

Sokol, D. (2006), 'What if...', *BBC* . `http://news.bbc.co.uk/2/hi/uk_news/magazine/4954856.stm`.

Sparrow, R. and 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.

Statistisches Bundesamt (2017), 'Unfälle und verunglückte im straßenverkehr', *Destatis* . `https://www.destatis.de/DE/ZahlenFakten/Wirtschaftsbereiche/TransportVerkehr/Verkehrsunfaelle/Tabellen/UnfaelleVerunglueckte.html`.

Suter, R. S. and Hertwig, R. (2011), 'Time and moral judgment', *Cognition* 119(3), 454–458.

Sütfeld, L. R., Gast, R., König, P. and Pipa, G. (2017), '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.

Tassy, S., Oullier, O., Mancini, J. and Wicker, B. (2013), 'Discrepancies between judgment and choice of action in moral dilemmas', *Frontiers in psychology* 4, 250.

Thomson, J. J. (1985), 'The trolley problem', *The Yale Law Journal* 94(6), 1395–1415.

Thomson, J. J. (2008), 'Turning the trolley', *Philosophy & Public Affairs* 36(4), 359–374.

Torbert, R. and Herrschaft, B. (2013), 'Driving miss hazy: Will driverless cars decrease fossil fuel consumption?', *Rocky Mountain Institute* . `http://blog.rmi.org/blog_2013_01_25_Driving_Miss_Hazy_Driverless_Cars`.

Van Der Laan, J. D., Heino, A. and De Waard, D. (1997), 'A simple procedure for the assessment of acceptance of advanced transport telematics', *Transportation Research Part C: Emerging Technologies* 5(1), 1–10.

Venkatesh, V. and Davis, F. D. (2000), 'A theoretical extension of the technology acceptance model: Four longitudinal field studies', *Management science* 46(2), 186–204.

Wagenaar, W. A. and Sagaria, S. D. (1975), 'Misperception of exponential growth', *Perception & Psychophysics* 18(6), 416–422.

Wakabayashi, D. (2018), 'Self-driving uber car kills pedestrian in arizona, where robots roam', *The New York Times* . `https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html`.

White, S. (2018), 'Japan looks to launch driverless car system in tokyo by 2020', *Business Insider UK* . `http://uk.businessinsider.com/r-japan-looks-to-launch-driverless-car-system-in-tokyo-by-2020-2018-6`.

World Health Organisation (2015), 'Number of road traffic deaths', *Global Health Observatory (GHO) data* . http://www.who.int/gho/road_safety/mortality/number_text/en/.