Sci Eng Ethics (2017) 23:681–700
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 / Published online: 14 July 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

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.