Note: This is the penultimate draft of a paper published in *Ethics and Information Technology*. The version of record can be found here: http://doi.org/10.1007/s10676-022-09668-5

No Wheel but a Dial: Why and how passengers in self-driving cars should decide how their car drives

Johannes Himmelreich

Abstract

Much of the debate on the ethics of self-driving cars has revolved around trolley scenarios. This paper instead takes up the political or institutional question of who should decide how a self-driving car drives. Specifically, this paper is on the question of whether and why *passengers* should be able to control how their car drives. The paper reviews existing arguments — those for passenger ethics settings and for mandatory ethics settings respectively — and argues that they fail. Although the arguments are not successful, they serve as the basis to formulate desiderata that any approach to regulating the driving behavior of self-driving cars ought to fulfill. The paper then proposes one way of designing passenger ethics settings that meets these desiderata.

1 Introduction

Self-driving cars are not a hypothetical technology.[1] Even as entrepreneurial enthusiasm has receded, ethical alarmism subsided, and the sense of social urgency waned, it still seems reasonable to expect that self-driving cars will replace today's cars over the coming decades. They might be introduced at first only in certain regions, they might be restricted to components of the traffic system, they may have mandatory safety drivers (either in the car or remotely) or be subject to speed limits. But whatever the timing and manner of their introduction, ethical questions around the design, driving behavior and the social implications of self-driving cars arise now (Millar 2017; Milakis, Arem, and Wee 2017; Nyholm and Smids 2020; Keeling et al. 2019).

The issue that arguably has gotten the most attention is the question of how self-driving cars should behave in trolley scenarios (Nyholm 2018). But, of course, the ethics of self-driving cars is much broader; and even when it comes to the distribution of risks

[1] By "self-driving cars," "autonomous vehicles" or "automated vehicles" (AV) I understand individually-owned passenger vehicles with automation levels 4 or higher according to the SAE definition. I concentrate on cars owned by individuals, in contrast to corporate-owned cars.

and harms, attention should not be restricted to trolley-like scenarios (N. J. Goodall 2014; 2016; 2017; Mladenovic and McPherson 2016; Nyholm and Smids 2016; JafariNaimi 2017; Himmelreich 2018; Epting 2019; Dietrich and Weisswange 2019; Nunes 2019; Cunneen et al. 2020; Gogoll and Müller 2017, 694).[2] Specifically, at the center are now questions of justice, power and the normative assessment of institutions and policies tracking a similar development in the literature on the ethics of artificial intelligence (Crawford and Calo 2016; Mladenovic and McPherson 2016; Borenstein, Herkert, and Miller 2017; Rahwan 2018; Susskind 2018; Nunes 2019; Zimmermann, Di Rosa, and Kim 2020; Gabriel 2022). This paper is part of this development (cf. Himmelreich 2020; Rodríguez-Alcázar, Bermejo-Luque, and Molina-Pérez 2021; Brändle and Schmidt 2021).

This paper is on the political-institutional question: Who should decide how self-driving cars drive? Today, with largely non-automated cars, the decision about how a car drives is in the driver's hands. But self-driving cars, of course, need no driver. Why, if at all, should passengers be allowed to decide how their car drives? How, anyway, should we think about the decisions that need to be made when driving? In what way, if any, do these decisions have any relevance for political philosophy? These are questions that I aim to address with this paper.

In this paper, I argue that passengers should be able to decide how a self-driving car drives. I argue in favor of personal ethics settings (PES) and against mandatory ethics settings (MES).[3] Where today cars have a wheel, tomorrow they should have a dial.

Some have argued in favor of PES before (Millar 2014b; Bonnefon, Shariff, and Rahwan 2016; Contissa, Lagioia, and Sartor 2017; Awad et al. 2020; Soltanzadeh, Galliott, and Jevglevskaja 2020). Others have argued against PES (Lin 2014; Gogoll and Müller 2017; Dietrich and Weisswange 2019). Interestingly, the arguments often start with the same assumptions but come to opposite conclusions. I begin by reconstructing and reviewing in detail many of these arguments and argue that they

[2] For arguments in favor of the relevance of trolley scenarios, however, see Lin (2017), Keeling (2020) and Awad et al. (2020)

[3] The nomenclature is from Gogoll and Müller (2017). The distinction between PES and MES depends on whether a passenger can meaningfully control a vehicle's driving style and macro path planning. The expression "meaningful control" is central to the ethics of robotics.

fail to establish their respective conclusions.[4] Nevertheless, arguments from both sides get at something important, which I formulate as desiderata. In this sense, the review of the arguments is constructive. In light of these desiderata, I then put forward a proposal on how passengers should be able to make settings that affect driving style and macro path planning.

The paper has three parts. First, I review arguments *in favor* of PES and identify their shortcomings. Second, I discuss arguments *against* PES and formulate objections. Because such a review has not been done before, and because the literature on this topic may lack some cohesion as a result, in addition to providing us with desiderata, conducting this critical review itself is a main aim of this paper. Third, I outline my proposal, illustrate how it goes beyond existing proposals, and consider objections.

In addition to this being the first systematic and critical review of arguments on the question of who should decide how a self-driving car drives, this paper contributes to the existing literature in two ways. First, I propose a novel, two-dimensional parameter space for PES. Existing proposals suggest that passengers should be able to choose how egoistically or altruistically they want their car to drive (Contissa, Lagioia, and Sartor 2017). I propose a further mobility–safety dimension that can help thinking about the ethical relevance of decisions that need to be made on the road. Second, I develop a practicable proposal that includes a new way of limiting the extent of PES as well as a signaling functionality that can be used to indicate to outsiders the driving settings on which a car is operating.

This paper — as the literature overall — proceeds on the basic assumption that the behavior of self-driving cars can be governed by policies, constraints, or considerations such as "avoid passing cyclists unless there are at least 3 feet of lateral distance". A skeptic might reject this assumption and maintain instead that the behavior of self-driving cars can be manipulated only *implicitly* by changing the machine learning (ML) training data (cf. Basl and Behrends 2020). On this view, it is misguided to discuss policies, rules, or considerations because the relevant challenge is instead how desired behavioral outcomes can be attained though a change in the

[4] In addition to arguments that address PES directly, I also review related arguments that can be applied to the issue of PES (Millar 2014a; 2015; Bonnefon, Shariff, and Rahwan 2016).

training data. The existing literature seems to ignore the realities of the core technologies that drives self-driving cars — or so a skeptic may argue.[5]

But the skeptic is mistaken, for three reasons. First, her challenge about "realities of the core technologies" may turn out to be ill-informed about the realities of the core technologies. Self-driving cars use a mix of technologies. Some of these technologies, such as reinforcement learning, allow to explicitly represent policies to govern outcomes. Tesla's path planning, for example, explicitly considers measures of safety and comfort.[6] Second, the skeptic seems to conflate normative and technical issues. The normative issue, how cars *should* behave and who should decide how cars should behave — the topic of this paper —, is in some ways prior to the technical question of *how* to achieve a certain desired behavior via a training regime.[7] Finally, even *if* the current technologies were all like supervised learning (which focuses on implicit behavior manipulation via the selection of training data), these technologies might have to change. Perhaps the core technologies *should* be such that they do allow to govern the car's behavior explicitly.

With this fundamental methodological challenge out of the way, let us investigate arguments over how the behavior of self-driving cars should be governed.

Arguments in Favor of Passenger Ethics Settings

What speaks in favor of PES? The current literature contains broadly two arguments in favor of PES — which I call the autonomy argument and the social dilemma argument — but each falls short in some way, or so I argue.

1.1 Autonomy and the Moral Proxy Argument

A first argument for PES rests on the idea that PES are an expression of an individual's autonomy. The argument comes in two forms. I call one the *autonomy argument* and

[5] My discussion here is prompted by comments by a peer reviewer for a different journal.

[6] Tesla's cost function for path planning minimizes traversal time, collision risk, lateral acceleration, and lateral jerk — the latter as a measure of comfort (Tesla 2021). The behavior of Teslas is hence governed via deliberately designed properties of the cost function.

[7] Technical and normative issues are not independent: Technological choices constrain the ethics of a system. This is an important insight in the value-alignment literature (cf. Gabriel 2020), of which the debate on the ethics of self-driving cars can be seen as a part.

the other the *moral proxy argument* (Millar 2014a; 2015). Both rest on the same idea and both face a similar problem.

The autonomy argument is prominent within the emerging movement for the right to drive (Roy 2018; O'Connor 2019). This argument makes a direct inference from the value of autonomy to PES (Soltanzadeh, Galliott, and Jevglevskaja 2020). The idea is, very simply, that autonomy is valuable, that we should adopt a policy which promotes autonomy (all other things being equal) and that, because PES promote autonomy, we should adopt PES.

The moral proxy argument, by contrast, is an indirect inference and argues by analogy (Millar 2014a; 2014b). The analogy is that decisions that engineers make about how to build self-driving cars are similar to decisions that doctors and medical proxies make about what treatments to administer. Passengers have cars as moral proxies just as patients may have family members as proxy decision-makers. And just as your medical proxy should act with your interest in mind, so your car — your moral proxy — should drive in a way that furthers your interests. This is the moral proxy argument for self-driving cars.

The moral proxy argument teaches an important lesson to engineers in that it challenges them to conceive of their role as analogous to that of medical professionals.[8] Just as doctors must adhere to a professional ethics that binds them to the good of their patients, so engineers must be clear about their professional ethics and be bound to the good of their products' users. This part of the argument — about grounding the role-obligations of developers, individually or collectively — strikes me as laudable. But the part of the argument having to do with PES is less compelling.

Unlike what the moral proxy argument assumes, things are not analogous. Unlike in the medial context, individual decisions in traffic have significant negative external effects. Whether I should receive a certain type of pacemaker in the case of a cardiac emergency — a central example used by Millar (2014a) — leaves the health of others virtually unaffected. The decision of how to drive, by contrast, especially in hazardous

[8] Things are actually more complicated because it is not clear *whose* proxy the cars ought to be – there is thus a "moral proxy problem" (Thoma 2022). Depending on whether cars are proxies for individuals or aggregates (such as developers or regulators), they should make risky decisions very differently (ibid.).

scenarios and dilemma situations, is a decision about the health of others just as much as it is a decision about my own health. Although the overall message of the moral proxy argument might be correct, the argument falls short in its analysis of the situation — and in establishing the conclusion that we have even *pro tanto* reasons in favor of PES.

Put differently, PES promote the autonomy of some to the detriment of others. Outsiders — pedestrians, cyclists and other vulnerable traffic participants — do not get a say in how PES are used all the while PES might be used in a way that sets back these outsiders' autonomy. Driving takes place in a public space and hence must heed to (shared) public concerns. Often, for that reason, decisions of public concern are made collectively and behavior in such public spaces is restricted. That is, we use our *collective* autonomy to legislate or govern in public spaces where issues of shared public concern are at stake (cf. Rodríguez-Alcázar, Bermejo-Luque, and Molina-Pérez 2021).

Proponents of the autonomy and moral proxy argument recognize this challenge. They argue that "there are some obvious limits to the kinds of ethics settings we should allow in our robot cars" and that PES should therefore cover only strict subset of the overall parameter space (Millar 2014b; 2017; Etzioni and Etzioni 2017; Soltanzadeh, Galliott, and Jevglevskaja 2020, n. 5).[9]

Nevertheless, even with limits, the problem of external effects remains. Having PES that are merely bounded, arguably, might not go far enough. The range of choices even within bounded PES still needs justification and may set back the autonomy of others. If a passenger decides to let their car drive somewhat more aggressively, then this will likely put cyclists who share the road at a greater risk of accidents. In this sense, even a small change in driving style could violate rights or encroach on the autonomy of others. Similarly, even small infractions of speed limits still need to be justified. Here the collective autonomy is encroached upon insofar as even moderate speeding violates democratically established norms. As Nyholm and Smids (2020) argue that, in domains as crucial traffic, which affects bodily safety, individuals are not allowed to decide on the limits of legal authority and "self-apply the law" in this way. Pointing

[9] What these limits should be and what considerations should guide our delineation of limits is often not clear. But see Contissa et al. (2017, 374) and Etzioni and Etzioni (2017).

out how each infraction — driving a bit more aggressively, speeding moderately — that doing so exercises or furthers the passengers' autonomy just will not do.

Traffic is replete with situations that have external effects in which decisions affect others. Many individuals' interests and their autonomy hence conflicts with the interests and the autonomy of others. Traffic has all the features of the kind of situations that we subject to collective decision making for this reason. The best way to further individual autonomy in traffic might be to decide *collectively* about what driving is acceptable. So, instead of an argument in favor, we seem to have an argument against PES.[10]

Nevertheless, the argument gets at something important. Considerations of autonomy should play a central role in the normative assessment of different policy proposals, at least as an injunction guarding against paternalistic intervention, but moreover as an expression of a commitment to liberties. A desideratum for *respect for autonomy* hence puts a thumb on the scale as one dimension and desideratum under which policy proposals generally ought to be assessed. The central challenge is to think of a justifiable mechanism that appropriately balances the autonomy of passengers and outsiders.

1.2 Social Dilemma Argument

The moral proxy argument for PES is broadly deontological. A different argument in favor of PES that is more consequentialist in spirit is what I call the *social dilemma argument*.[11] The social dilemma argument rests chiefly on the claim that PES are necessary in order to enable or further the adoption of self-driving cars (Bonnefon,

[10] Of course, there could be a collective decision in favor of PES; but this is not how PES are usually defended.

[11] I take the name for this argument from the title of a paper by Bonnefon, Shariff and Rahwan (2016), who present the empirical finding that motivates the argument that I present here (The main idea in the argument is also called the "ethical opt-out problem" (Bonnefon, Shariff, and Rahwan 2020)). However — to avoid misattribution — the argument I present here is not theirs. The argument is hinted at by Contissa, Lagioia, and Sartor (2017, 367) who write that "[i]f an impartial (utilitarian) ethical setting is made compulsory for, and rigidly implemented into, all AVs, many people may refuse to use AVs, even though AVs may have significant advantages, in particular with regard to safety, over human-driven vehicles." Bonnefon, Shariff, and Rahwan (2020, 110), however, advance a similar argument. They write: "[I]f people are not satisfied with the ethical principles that guide moral algorithms, they will simply *opt out* of using these algorithms, thus nullifying all their expected benefits."

Shariff, and Rahwan 2016; 2020; Shariff, Bonnefon, and Rahwan 2017; Contissa, Lagioia, and Sartor 2017, 367; Awad et al. 2020). The idea behind this claim is that if people are unable to set parameters of their car's driving style, then they will not, or will be unlikely to, buy or use self-driving cars. Or in other words: An "Ethical Knob may improve users' acceptance of AVs" (Contissa, Lagioia, and Sartor 2017, 377).[12] Without PES, self-driving cars would be a technology with high promise but low uptake. But because policymakers and manufacturers ought to do what is necessary to reap the benefits that self-driving cars promise, they ought to regulate or design for PES.[13]

This argument should resonate well with anyone who is familiar with non-cooperative game theory. Suppose everyone has a choice between two kinds of cars. The first kind of car maximizes everyone's welfare. Call this a *utilitarian car*. When the welfare interests of outsiders and passengers conflict, a utilitarian car always acts so as to harm its passengers if the aggregate harm to outsiders is greater. This might often be the case assuming that the number of outsiders tends to be greater than the number of passengers and the potential harms to outsiders tend to be greater than the potential harms to passengers. The second kind of car, by contrast, strictly prioritizes the welfare of its passengers. Call this an *egoistic car*. When the welfare interests of outsiders and passengers conflict, the egoistic car will always act as to harm outsiders.[14] This choice situation for consumers seems to resemble the prisoners' dilemma game (PD).[15]

A crucial moment in this argument is a strategic consideration. The argument anticipates the behavior of consumers — will they buy or use self-driving cars? — in

[12] Similarly, Ryan (2020) writes: "Very few people would buy [a self-driving car] if they prioritised the lives of others over the vehicle's driver and passengers."

[13] The social dilemma argument is motivated by an empirical finding: Although a majority of people agree that a driving style that maximizes overall welfare or health in a population is the preferable driving style from a moral point of view, many people would not actually want to use or buy a vehicle that drives in this way (Bonnefon, Shariff, and Rahwan 2016; Gill 2021). This is the social dilemma.

[14] What I describe is only an extreme version of an egoistic car. In fact, as has been argued, there could be a continuum (Contissa, Lagioia, and Sartor 2017).

[15] A prisoners dilemma is a two-person symmetric game with two pure strategies, "cooperate" and "defect", in which the payoffs of the four different outcomes satisfy the condition $T > R > P > S$, that is, *t*emptation to defect against a cooperator has a strictly greater payoff than *r*eward of mutual cooperation, *p*unishment for mutual defection, and the so-called *s*ucker payoff for cooperating with a defector.

a way that assumes them to be rational, self-interested, and highly concerned with safety. Incorporating strategic considerations is a theoretical virtue in discussions of practical political philosophy. I moreover agree with the conclusion in favor of PES. But the social dilemma argument has two problems and fails to establish its conclusion.

The first problem is that the argument presents false alternatives. The argument assumes that consumers have a choice either between a self-driving car with utilitarian MES on the one hand, or a self-driving car with PES on the other hand (or a car that they drive themselves). But the options are not so clear cut. There are more than these two options because policymakers can use a wide range of tools to influence consumer choices. Policymakers can use tax-breaks, grants, subsidies, advertising, education or outright bans to only name a few (cf. Cohen and Cavoli 2019). In this way, policymakers could change incentives and encourage the use of self-driving cars with utilitarian MES. After all, insofar as the social dilemma is just an instance of a prisoners' dilemma — essentially a collective action problem —, then it is precisely the kind of conflict that policymakers face all the time.[16] Market failures and free-riding problems result from the same conflict between individual rationality (incentives) and collective rationality (welfare) that we see in the social dilemma. The alternatives are thus not only a choice between self-driving cars with MES and PES (or manually driven cars), but also how policymakers should resolve this dilemma by changing incentives with the tools at their disposal.

The second problem of the social dilemma argument is that the argument might fail to establish its conclusion on its own terms. Consumers might in fact decide to use self-driving cars, even with the two alternatives that the argument assumes or when the choice is between self-driving cars with MES or manually operated cars.

First, consumers' decisions may not be driven exclusively by *self-interested* safety considerations. In fact, even when they find themselves in scenarios where they are passengers, the majority of participants in some surveys do *not* give a strict priority to the safety of passengers. Around 30% of participants even express altruistic preferences. They say the car should protect pedestrians even in scenarios where they imagine themselves to be passengers (Gill 2021, 669).

[16] This is acknowledged by some (Bonnefon, Shariff, and Rahwan 2016).

Second, norms and attitudes about safety might change. Expectations about AV safety have been very high (Liu, Yang, and Xu 2019),[17] in part due to algorithm aversion and the better-than-average effect (Shariff, Bonnefon, and Rahwan 2021). Yet, there are reasons to expect these attitudes to change: Up to a quarter of respondents in the US express an interest in purchasing a self-driving car (Gill 2021, 669).[18] Experiencing a ride in a self-driving car seems to positively shift attitudes towards this technology (Xu et al. 2018). Even skeptical authors concede that the relevant "norms can change quickly" (Shariff, Bonnefon, and Rahwan 2021, 9). Thus, the diffusion of self-driving cars might follow the usual S-curve pattern: It starts with some keen early adopters before the majority signs on — and some hold-outs remain in a long tail (Liljamo, Liimatainen, and Pöllänen 2018).

Third, several considerations beyond overall safety play a role when consumers decide which car to buy. Some people are "image shoppers" who are "concerned with what your vehicle says about you" (KBB Editors 2022). In other words: A self-driving car can have *expressive value*. In particular, it can be a means of virtue signaling (Shariff, Bonnefon, and Rahwan 2017, 695), and it can be a status symbol. Other people might be *convenience* shoppers. The most wanted features in cars currently are things like wireless charging pads, sunroofs, and parking sensors (AutoPacific 2022). Similarly, perceived usefulness seems to be one driver of intentions to use self-driving cars (Choi and Ji 2015). Again other people might care about *operating costs* of their cars.[19] Insofar as self-driving cars with MES are safer or more consistent than cars with PES, the insurance premium for MES self-driving cars might be lower than that for PES or manually driven cars.[20] Finally, some people have *medical conditions* that restrict their ability to travel in manually operated cars. Consider the young, some elderly, or anyone with certain motor or visual conditions — a significant market segment

[17] Respondents in China would find it "tolerable" if self-driving cars are four to five times as safe as human drivers and "acceptable" if the cars were safer by one to two orders of magnitude (Liu, Yang, and Xu 2019).

[18] For context: These are data from US participants. US participants can be expected to have relatively unfavorable attitudes towards AVs compared to India or China. A study in 2014 found that only 14% to 22% of respondents in the UK and US respectively hold very positive attitudes towards automated vehicles compared to 46% and 50% in India and China (Schoettle and Sivak 2014).

[19] The Kelley Blue Book calls these "value shoppers" (KBB Editors 2022).

[20] This is not a crucial assumption: Even if the nominal insurance costs might be higher, especially in the short term, they could be decreased by policy to make self-driving cars attractive (Ravid 2014).

(Harper et al. 2016). For this market segment, a self-driving car might be preferable since it is the only option to use individualized transportation, even if that car has MES.

In short, contrary to what the social dilemma argument suggests, self-driving cars, even if they have utilitarian MES, will likely find a significant market (esp. when the alternative is a manually operated car). The social dilemma argument assumes that consumers buy cars almost exclusively based on considerations of self-interested safety. But attitudes about safety might change and the expressive value of certain cars, convenience, operating costs, or a diversity of abilities to use manually driven cars are factors that may help self-driving cars find a market even if they are equipped with a utilitarian MES.

Even if safety is what consumers cared about most in self-driving cars, the social dilemma argument still makes another problematic assumption. The social dilemma argument assumes that a utilitarian self-driving car with MES would be less safe than the alternative with PES. But "safety" is a complex concept and hard to measure.[21] Specifically, safety is constituted by environmental factors and thus not an intrinsic property of a self-driving car (Fraade-Blanar et al. 2018). Compared to a manually driven car, a utilitarian car with MES would on average be safer even for its passengers, given that most fatal accidents are due to human error — the US Department of Transportation estimates the number to be between 90 and 94% (NHTSA 2017; 1995). Finally, if consumers are also pedestrians or cyclists and face *other* peoples' egoistic self-driving cars, it might also be safer for them to prefer utilitarian cars with MES for everyone: because they might be a driver today but a pedestrian tomorrow.

Although the social dilemma argument fails, it again gets at something important, which can be put in the form of a desideratum that I call *safety despite strategy*. This desideratum has it that benefits associated with self-driving cars should be attained even if agents choose strategically. More specifically, a policy should be designed such that under reasonable expectations subjects have self-interested reasons for abiding by the policy.

[21] Moreover, it would likely take decades to be able to have sufficient exposure to measure (as opposed to simulate or estimate) the safety of self-driving cars (Kalra and Paddock 2016).

2 Arguments Against Passenger Ethics Settings

Arguments against PES are legion (cf. Lin 2014). I concentrate on what I call the best interest to society argument (Gogoll and Müller 2017).[22] This argument contends that MES are necessary and sufficient to attain an outcome that is in the best interest of society.[23]

The best interest argument assumes that traffic situations present a prisoners' dilemma game.[24] In a PD, cooperation between players is unlikely to emerge and, if it were to emerge, it would not be stable.[25] Traffic situations, in virtue of being like a PD, can therefore only be solved through mandatory regulation. In other words, "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" (Gogoll and Müller 2017, 695). In short, since traffic situations are like a PD, and since a PD requires mandatory regulation to achieve the socially best outcome, MES are required for, and will achieve, the socially best outcome in the context of self-driving cars.

A first thing to note is that, despite arriving at the opposite conclusion, the best interest argument is surprisingly similar to the social dilemma argument. Both arguments turn on strategic considerations. Whereas the social dilemma argument applies strategic considerations to the decision of whether or not to buy or use a self-driving car, the best interest argument applies strategic considerations to decisions of how to behave in hazardous traffic situations.

But the best interest argument has two problems.

First, the assumption that traffic is a PD is not quite correct. What kind of game traffic is, is to some extent a matter of policy. Policymakers can set incentives, thereby change

[22] I concentrate on this argument because it is a recent and the best developed one.

[23] By "best interest of society" the authors mean that traffic injuries and fatalities are minimized in a given population.

[24] This differs from the social dilemma argument which assumed that *purchasing* decisions are a PD instead of *traffic* being a PD.

[25] I write "emerge" and "stable" to indicate that the game is played repeatedly. Even if players will not cooperate in one-shot games, the prospects for achieving widespread cooperation look much better when PD is played repeatedly.

payoffs and transform the game into a different one. That driving in traffic is a PD thus cannot be taken as an exogenous fact.[26]

Second, unlike what the best interest argument suggests, cooperation might indeed emerge. Although it is true that rational players always defect in a one-shot PD, traffic and driving are not adequately modelled by a *one*-shot PD. The game seems to be played iteratively without any discernable or foreseeable end. Moreover, driving involves many players simultaneously. Hence, traffic might better be modelled as an iterative game with more than two players (or changing pairs of players). Even if there is no cooperation in a two-person one-shot PD, cooperation *does* evolve in variations of iterated PDs, even with many players.

Whether cooperation evolves in the iterated two-player prisoners dilemma depends crucially on the structure of the interaction, the presence of players that initially cooperate, and the ability of players to remember and recognize others and their behavior (Alexander 2007, chap. 3). Moreover, the celebrated supergame strategy Tit-for-Tat — which starts with cooperation and imitates what an opponent did on a previous move — does remarkably well in the iterative prisoners dilemma (Axelrod 2009).

Seen this way, cooperation could thrive specifically when *self-driving cars* play the traffic game. Insofar as the emergence of cooperation improves with the length of memory that players have of past interactions, self-driving cars are well-disposed to cooperate because self-driving cars could remember almost *all* interactions with other cars and drivers. Moreover, self-driving cars can communicate and share the learnings from past interactions within a cooperating fleet of cars. In addition, cars might share a blacklist of "bad drivers" as a collective defense.

In sum, the best interest argument presents false alternatives. MES are not necessary to attain an outcome that is in the best interest of society. There are other options. First, policymakers can regulate the strategic interaction in traffic such that traffic is not a PD. Second, even if traffic resembles a PD, cooperation can emerge. MES are hence not necessary in order to attain the outcome that is in the best interest of society.

[26] It could be said that the traffic game is embedded in other games within the political structure.

However, although the best interest argument also fails, the discussion offers two important upshots. First, it reiterates the importance of the desideratum that I called "safety despite strategy." Second, the discussion suggests that problems of strategic interaction in traffic, which may lead to suboptimal overall health outcomes, can be addressed by policy makers or in the design of self-driving cars, a point to which I will return shortly.

3 Passenger Ethics Settings: Desiderata, Dimensions and Design

The existing arguments either have an overly narrow purview of autonomy (autonomy and moral proxy argument) or they disagree about the locus and the effects of strategic behavior (social dilemma and interest of society argument). Despite their shortcomings, each argument turns on important considerations and reasons. Autonomy, safety, and the effects of strategic behavior matter. This is captured by the desiderata to respect autonomy and to achieve safety despite strategic behavior.

In addition to these two desiderata, a third desideratum is that a policy governing the motion planning of self-driving cars should respond to occupants' individual preferences, given a reasonable diversity of such preferences. We can call this desideratum *responsiveness to pluralism*. Today, with manually operated cars, driving styles differ, sometimes for good reason. Some might run late for an appointment, some face an emergency, some are just impatient. Such a pluralism of driving styles raises the question: "how should we proceed given widespread normative disagreement about the appropriate ethics setting of autonomous cars?" (Gogoll and Müller 2017, 687).

This catalogue of three desiderata — respect for autonomy, safety despite strategy and responsiveness to pluralism — amounts to a conditional case for PES. Responsiveness to pluralism speaks largely in favor of PES.[27] It is hard to see how MES would satisfy this desideratum to the same degree. I will argue that PES moreover meets the other desiderata. But my case, so far, in favor of PES is conditional. PES meet the desiderata if and only if the specific proposal is designed in such a way as to also respect the autonomy of outsiders and to address the threat of strategic behavior to undermine the

[27] Of course, also MES could incorporate a concern for pluralism. But, arguably, PES are more responsive to occupants' preferences. On PES, the average distance between behavior and preference will likely be narrower than on MES.

safety benefits of self-driving cars. In other words, whether the desiderata can be met depends on how PES are designed.

For the rest of the paper, I defend what such a PES design could look like. I begin with a more general exercise. I examine the relevant parameter space from which passengers can adjust their cars driving behavior. I propose that the relevant parameter space is multidimensional in that it represents at least two distinct value conflicts. I then illustrate that such a multi-dimensional parameter space can be used to limit the extent to which passengers can choose settings. Instead of allowing passengers to set parameters in the multidimensional space independently in each dimension, the choice of ethics settings is limited by making the parameters in the different dimensions interdependent and thereby creating a tradeoff between the different parameter dimensions. Finally, I will address the issue of strategic behavior.

3.1 Dimensions: Mobility–Safety and Self-interest–Other-interest

In the debate around PES, some have suggested that a central underlying normative issue consists in the conflict between self-interest and the interests of others (Contissa, Lagioia, and Sartor 2017). Here is an example of self-interest in traffic: In a four-way intersection with four stop signs, you could assertively proceed to cross through the intersection even when it is not yet your turn, knowing that the others will yield their right of way and not risk a collision. Of course, the same self-interest strategy could be followed by a self-driving car. Just as human drivers might face situations in which the interest of the passenger and the interest of outsiders conflict, so will self-driving cars.[28]

In addition to this well-known value conflict between self-interest and other-interest, another value conflict that has so far been overlooked in the debate about PES/MES is that between mobility and safety.[29] Mobility and safety conflict in a variety of driving situations large and small (cf. N. Goodall 2019). Consider four examples.

---

[28] Another illustration of this conflict between others' interest and your interest is, of course, in trolley cases and collision scenarios such as in the Tunnel Problem where a car needs to choose between running over a pedestrian or running the car into the wall of a tunnel (Millar 2014a).

[29] By "mobility" I understand the time required to get to a destination. By "safety" I understand the absence of risk, defined as a function of the probability of a hazardous event and the harm to the occupants and others. It should be noted that I understand both "mobility" and "safety" impartially as everyone's mobility and safety and not just those of vehicle occupants.

First, a conflict between mobility and safety is evident in unprotected left turns. These maneuvers are relatively risky — especially when the left turn is unnecessary. Often, instead of making a left turn, an alternative route is available of driving around the block with three right turns.

Second, mobility and safety conflict in overtaking maneuvers, for example, in the decision of whether you should pass a slower car on a two-lane street with oncoming traffic (Sezer 2018), in overtaking maneuvers in urban traffic in cases of occlusion (Bouton et al. 2018; Gerdes, Thornton, and Millar 2019), or in lane-changing maneuvers (Moridpour, Sarvi, and Rose 2010). For example, imagine that on a two-lane street your view of a pedestrian crosswalk is blocked by a parked car. You need to pass the parked car that occludes the view fast enough in order to avoid a collision with oncoming traffic in the other lane. The faster you are willing to pass the parked car, the sooner will you find a gap in the oncoming traffic. But, of course, you would at the same time increase the risk of having to avoid colliding with someone trying to cross the street.

Third, mobility and safety conflict in determining lateral safe passing distances (N. Goodall 2019). The width that is effectively available in a lane could be reduced, for example, by construction work or by other traffic participants, such as a parked delivery truck. Imagine that a car to the left of you drives on the far-right end of its lane. At the same time, your car may have to pass a cyclist that is in the same lane as you ahead of you.[30] A self-driving car will have to trade-off lateral distances: How close should you get to the car on the left and how much space should you leave to the cyclist on the right? Or should you slow down to the cyclists' speed to avoid passing them in the lane?

Finally, mobility and safety conflict in macro path planning, that is, in determining which route to take to a destination (Gerdes, Thornton, and Millar 2019). Suppose you can choose between two routes on your morning commute. One option takes the highway and avoids residential neighborhoods but turns out to be slower. The other option is quicker and takes you through residential areas and passes schools at a time when you know that children will be on their way to the first period. The first route is

[30] Assume also that this situation occurs in a location that does not prescribe a minimum lateral distance for safe passing.

safer but takes you longer. The second route is less safe but offers greater mobility in return. Such situations that give rise to a conflict between mobility and safety may arise frequently in planning routes to a destination already today in the routes determined by mapping applications.

This conflict between safety and mobility seems ubiquitous in driving and is not newly introduced with self-driving cars. Yet, self-driving cars make this trade-off more salient and make theorizing this conflict more urgent insofar as engineers or policymakers could now, in principle, regulate in greater detail driving decisions that were impossible to regulate before.

3.2 Design: One Dial not Two

As we have seen in the discussion of the autonomy argument for PES, PES should be limited. That is, not all technically possible driving styles should be on the menu of options from which occupants can choose. Although all authors who defend PES agree that PES should be limited, relatively little attention has been paid to the precise method of how such a limit should be conceptualized and implemented.

One approach would be to give passengers two dials, one for each tradeoff. That is, one dial would be for the mobility–safety tradeoff and the other for the self-interest–other-interest tradeoff. The driving parameters for each of these value conflicts would be set independently.

By contrast, I propose to give passengers *one* dial. That is, the two trade-offs are be made interdependently. The one dial is used to adjust settings on both parameters together — on the mobility–safety and on the self-interest–other interest trade-off. An increase in mobility will lead to an increase in other interest. And inversely, an increase in safety will lead to an increase of self-interest. Figure 1 illustrates this idea.

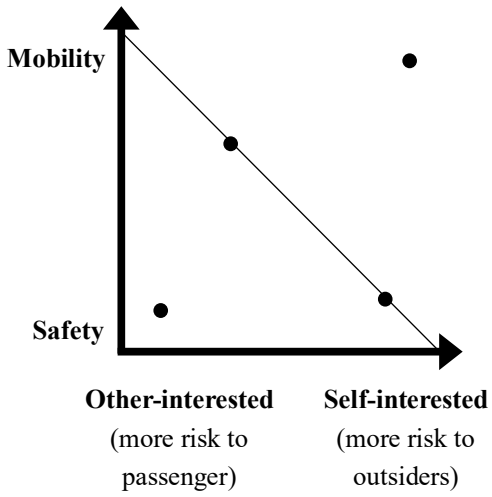

Figure 1: Occupants use PES by selecting a point on the diagonal line.

The two-dimensions are made dependent on another by reducing the trade-off to one dimension. The argument for this is straight-forward. Allowing passengers to make settings along the two trade-offs independently from one another would be patently immoral by any reasonable standard. For example, a passenger could choose greater mobility and also choose greater self-interest (in Figure 1 this setting is represented by the point in the upper right-hand corner). In result, the gains in mobility, which accrue mostly to the passenger, will lead to risks that are, as chosen by the passenger, borne by outsiders. In other words, occupants could choose to have others pay the price for their expensive preference. This would violate the desideratum of respect for autonomy — the outsiders' autonomy in this case.[31]

One might object that this way of trading-off the two value conflicts faces a problem in practice. In practice, a high mobility setting will have great benefits for the occupants whereas the accompanying high other-interest setting will have only comparatively small benefits to outsiders. This is because the situations in which you can gain mobility are many, whereas situations in which you can respect the interests of outsiders are few.

This objection rests on an empirical assumption, namely, about the asymmetric relative frequencies of situations that allow mobility gains vs. gains in the satisfaction of outsiders' interests. Of course, this assumption cannot be assessed here. Notice, however, that I have left open the exact "exchange rate" between mobility and other-interest. Depending on the relative frequencies of the situations, this rate of substitution, which is fixed by the design, between the two value conflict pairs can be adjusted.[32] But, moreover, my claim is not that having one dial achieves a moral or

[31] Of course, the details of this would have to be worked out by operationalizing these value conflicts and by studying the user interaction design (cf. Thornton et al. 2019).

[32] This is a matter of how the one dial trades off between the mobility–safety conflict and the other for the self-interest–other-interest conflict. How the one dial makes this tradeoff — the path of the

welfare equivalence. That is, I do not say the gains achieved by greater consideration of the interests of outsiders are equivalent to the moral or welfare costs associated with the stronger occupant preference for mobility.[33]

Further, one might object that this way of limiting PES is too restrictive. Specifically, it does not make room for altruism. Some passengers might have a preference for safety and also want to have a setting that weights greatly the interest of others (in Figure 1 this setting is represented by the point in the lower left-hand corner).

This objection can be easily accommodated. One option would be a separate altruism option (Gogoll and Müller 2017, 698). Another option would be to change the shape of the curve depicted in Figure 1. The line segment could "bend" towards the lower left corner and would hence "cover" such altruistic preferences.

This proposal, so far, meets the desideratum of responsiveness to pluralism (in virtue of being a PES) and it meets the desiderata of respect for autonomy with a novel way of limiting the parameter space from which passengers can choose driving settings that allows policy makers to balance the autonomy of passengers and outsiders.

3.3 Signaling and strategic considerations

I will now return to the problem of strategic interactions and the desideratum of safety despite strategy. As I alluded to above, policymakers can change the environment in which strategic interactions play out in a way that avoids violating this desideratum. For example, policymakers can impose taxes and influence insurance rates to incentivize certain behavioral settings in cars. I focus instead on an incentive that is not imposed by policy but that can result from a design choice: I suggest that signaling functions can be incorporated into PES to this effect.[34]

---

indifference curve though the space of parameter combinations — is an important question for ethics and design.

[33] Another problem with this objection is that it considers frequency but not stakes. It might be true that there are more opportunities for mobility and few for safety. But the stakes for safety might be much higher than those for mobility: Safety is about avoiding injuries and physical harms but mobility only about getting to a destination faster.

[34] Shariff et al. (2017) discuss the importance of "virtue signalling", however, not in the context of PES but instead as a psychological mechanism to exploit (in advertisement and communication) to increase AV adoption.

Suppose you have a self-driving car that has a PES and imagine that a light outside of your car indicates what driving settings you have chosen. The light shines in red if you chose a setting that values mobility highly but not other-interest. By contrast, the light shines in green if you value safety and other-interest highly. Moreover, imagine that your car is also equipped with a transponder that, using vehicle-to-vehicle (V2V) communication, broadcasts an ID of your vehicle and your chosen PES to nearby vehicles. This would allow the cars around you to "recognize" your car and "anticipate" how your car will behave in traffic. In other words, the first device — the light — signals your PES settings in a way easily accessible to humans, the second device — the transponder — signals your PES settings in a way that is easily accessible to other vehicles. Such a light or a transponder, or both, could be mandated by policy.

The effect be of such a light and a transponder would be to increase "cooperation" in traffic. First, the light would lead to greater cooperation insofar as there is social desirability for safety. In other words, individuals tend to want to be seen as cooperative and as caring about the safety of others, especially in context where they can be recognized (i.e. mainly in urban traffic). Second, the transponder would lead to greater cooperation because other cars can anticipate the driving maneuvers, team up with other cooperators, or even penalize defectors (probably mainly on highway traffic).

Technically, such signaling devices change the payoffs of a game. That people want to be seen as cooperators and that non-cooperation can be perfectly detected changes the "calculation" that underwrites how rational individuals chose their driving styles. Hence, this specific proposal of designing PES also meets the condition of the "safety despite strategy" desideratum.

Moreover, this PES design involving signaling devices also meets the respect for autonomy desideratum. One paradigmatic violation of moral autonomy is that agents have relevant information withheld from them. For example, if a doctor does not inform a patient about the lethal condition the patient find themselves in, this violates the patient's moral autonomy (Arpaly 2004, 120). Likewise, how a car drives that a pedestrian or cyclist may encounter in traffic is relevant information for the pedestrian or cyclist in question. By visually indicating what driving style occupants have chosen, outsiders are provided with information about decisions that others have taken. Assuming that the decisions of driving styles is suitably limited to a set of reasonable

driving styles, this achieves a balance of respecting the normative autonomy of occupants as well as outsiders. A PES design proposal such as this one would hence meet all three desiderata and insofar seem like a very attractive answer to the question of who should decide how a self-driving car should drive.

4 Conclusion

Much of the debate on the ethics of self-driving cars has concentrated on trolley situations and what would be the right decision for a car or a passenger to take. This paper has instead concentrated on the central normative *policy* issue of self-driving cars, namely the question: Who should decide how self-driving cars should drive? I have built on a critical review of existing arguments and a discussion of their shortcomings to make a case in favor of passenger ethics settings (PES).

The existing literature offers two main arguments for PES. One is based on autonomy, the other is based on a social dilemma. The literature also offers one main argument for MES based on strategic considerations to favor collective health and safety. All these arguments fail, as I have argued here, but each of them brings out an important consideration which I captured by formulating three desiderata: respect for autonomy, safety despite strategy and responsiveness to pluralism. These desiderata amount to a conditional case in favor of PES. That is, if a specific PES proposal achieves to respect the autonomy of outsiders and if this PES attains safety benefits of self-driving cars despite strategic behavior, then the desiderata speak in favor of this specific PES proposal. I have then sketched what such a specific PES proposal could look like that meets these desiderata.

Although the surface topic here is the public policy of self-driving cars, the issues are much broader. Reflections on personal autonomy in traffic, on how traffic can be conceptualized in game-theoretic terms, when risks of harm can be permissibly imposed, or design ideas that encourage cooperative behavior in traffic — these are all issues that could inform a more general ethics of driving. Normally, when we talk about individual drivers, the stakes in an ethics of driving are low. Not so with self-driving cars. Now the algorithm is the driver. Proposals to regulate who should decide

how a self-driving car drives are already being drafted today. As an area of political philosophy, traffic and transportation are only bound to become more relevant.[35]

5 References

Alexander, J. McKenzie. 2007. *The Structural Evolution of Morality*. Cambridge: Cambridge University Press.

Arpaly, Nomy. 2004. *Unprincipled Virtue: An Inquiry Into Moral Agency*. Oxford: Oxford University Press.

AutoPacific. 2022. "FADS — AutoPacific Insights." AutoPacific. August 18, 2022. https://www.autopacific.com/autopacific-insights/tag/FADS.

Awad, Edmond, Sohan Dsouza, Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2020. "Crowdsourcing Moral Machines." *Communications of the ACM* 63 (3): 48–55. https://doi.org/10.1145/3339904.

Axelrod, Robert. 2009. *The Evolution of Cooperation: Revised Edition*. New York: Basic Books.

Basl, John, and Jeff Behrends. 2020. "Why Everyone Has It Wrong about the Ethics of Autonomous Vehicles." In *Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2019 Symposium*. National Academies Press. https://doi.org/10.17226/25620.

Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. 2016. "The Social Dilemma of Autonomous Vehicles." *Science* 352 (6293): 1573–76. https://doi.org/10.1126/science.aaf2654.

———. 2020. "The Moral Psychology of AI and the Ethical Opt-Out Problem." In *Ethics of Artificial Intelligence*, edited by S. Matthew Liao, 109–26. Oxford: Oxford University Press.

Borenstein, Jason, Joseph R. Herkert, and Keith W. Miller. 2017. "Self-Driving Cars and Engineering Ethics: The Need for a System Level Analysis." *Science and Engineering Ethics*, November, 1–16. https://doi.org/10.1007/s11948-017-0006-0.

Bouton, Maxime, Alireza Nakhaei, Kikuo Fujimura, and Mykel J. Kochenderfer. 2018. "Scalable Decision Making with Sensor Occlusions for Autonomous Driving."

[35] I am grateful for thoughts and comments I received from Johanna Thoma and Sebastian Köhler, from students at Sonoma State University, and from participants and the audience at the Automated Vehicles Symposium 2019 in Orlando.

In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2076–81. https://doi.org/10.1109/ICRA.2018.8460914.

Brändle, Claudia, and Michael W. Schmidt. 2021. "Autonomous Driving and Public Reason: A Rawlsian Approach." *Philosophy & Technology* 34 (4): 1475–99. https://doi.org/10.1007/s13347-021-00468-1.

Choi, Jong Kyu, and Yong Gu Ji. 2015. "Investigating the Importance of Trust on Adopting an Autonomous Vehicle." *International Journal of Human–Computer Interaction* 31 (10): 692–702. https://doi.org/10.1080/10447318.2015.1070549.

Cohen, Tom, and Clémence Cavoli. 2019. "Automated Vehicles: Exploring Possible Consequences of Government (Non)Intervention for Congestion and Accessibility." *Transport Reviews* 39 (1): 129–51. https://doi.org/10.1080/01441647.2018.1524401.

Contissa, Giuseppe, Francesca Lagioia, and Giovanni Sartor. 2017. "The Ethical Knob: Ethically-Customisable Automated Vehicles and the Law." *Artificial Intelligence and Law* 25 (3): 365–78. https://doi.org/10.1007/s10506-017-9211-z.

Crawford, Kate, and Ryan Calo. 2016. "There Is a Blind Spot in AI Research." *Nature News* 538 (7625): 311. https://doi.org/10.1038/538311a.

Cunneen, Martin, Martin Mullins, Finbarr Murphy, Darren Shannon, Irini Furxhi, and Cian Ryan. 2020. "Autonomous Vehicles and Avoiding the Trolley (Dilemma): Vehicle Perception, Classification, and the Challenges of Framing Decision Ethics." *Cybernetics and Systems* 51 (1): 59–80. https://doi.org/10.1080/01969722.2019.1660541.

Dietrich, Manuel, and Thomas H. Weisswange. 2019. "Distributive Justice as an Ethical Principle for Autonomous Vehicle Behavior beyond Hazard Scenarios." *Ethics and Information Technology* 21 (3): 227–39. https://doi.org/10.1007/s10676-019-09504-3.

Epting, Shane. 2019. "Automated Vehicles and Transportation Justice." *Philosophy & Technology* 32 (3): 389–403. https://doi.org/10.1007/s13347-018-0307-5.

Etzioni, Amitai, and Oren Etzioni. 2017. "Incorporating Ethics into Artificial Intelligence." *The Journal of Ethics* 21 (4): 403–18. https://doi.org/10.1007/s10892-017-9252-2.

Fraade-Blanar, Laura, Marjory S. Blumenthal, James M. Anderson, and Nidhi Kalra. 2018. "Measuring Automated Vehicle Safety." Research Reports. Santa

Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_reports/RR2662.html.

Gabriel, Iason. 2020. “Artificial Intelligence, Values, and Alignment.” *Minds and Machines* 30 (3): 411–37. https://doi.org/10.1007/s11023-020-09539-2.

———. 2022. “Towards a Theory of Justice for Artificial Intelligence.” *Dædalus* 151 (2): 218–31.

Gerdes, J. Christian, Sarah M. Thornton, and Jason Millar. 2019. “Designing Automated Vehicles Around Human Values.” In *Road Vehicle Automation 6*, edited by Gereon Meyer and Sven Beiker, 39–48. Lecture Notes in Mobility. Cham: Springer Nature Switzerland.

Gill, Tripat. 2021. “Ethical Dilemmas Are Really Important to Potential Adopters of Autonomous Vehicles.” *Ethics and Information Technology* 23 (4): 657–73. https://doi.org/10.1007/s10676-021-09605-y.

Gogoll, Jan, and Julian F. Müller. 2017. “Autonomous Cars: In Favor of a Mandatory Ethics Setting.” *Science and Engineering Ethics* 23 (3): 681–700. https://doi.org/10.1007/s11948-016-9806-x.

Goodall, Noah. 2019. “More Than Trolleys: Plausible, Ethically Ambiguous Scenarios Likely to Be Encountered by Automated Vehicles.” *Transfers* 9 (2): 45–58. https://doi.org/10.3167/TRANS.2019.090204.

Goodall, Noah J. 2014. “Machine Ethics and Automated Vehicles.” In *Road Vehicle Automation 6*, edited by Gereon Meyer and Sven Beiker, 93–102. Lecture Notes in Mobility. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-319-05990-7_9.

———. 2016. “Away from Trolley Problems and Toward Risk Management.” *Applied Artificial Intelligence* 30 (8): 810–21. https://doi.org/10.1080/08839514.2016.1229922.

———. 2017. “From Trolleys to Risk: Models for Ethical Autonomous Driving.” *American Journal of Public Health* 107 (4): 496–496. https://doi.org/10.2105/AJPH.2017.303672.

Harper, Corey D., Chris T. Hendrickson, Sonia Mangones, and Constantine Samaras. 2016. “Estimating Potential Increases in Travel with Autonomous Vehicles for the Non-Driving, Elderly and People with Travel-Restrictive Medical Conditions.” *Transportation Research Part C: Emerging Technologies* 72 (November): 1–9. https://doi.org/10.1016/j.trc.2016.09.003.

Himmelreich, Johannes. 2018. “Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations.” *Ethical Theory and Moral Practice* 21 (3): 669–84. https://doi.org/10.1007/s10677-018-9896-4.

———. 2020. “Ethics of Technology Needs More Political Philosophy.” *Communications of the ACM* 63 (1): 33–35. https://doi.org/10.1145/3339905.

JafariNaimi, Nassim. 2017. “Our Bodies in the Trolley’s Path, or Why Self-Driving Cars Must Not Be Programmed to Kill.” *Science, Technology, & Human Values*, July, 0162243917718942. https://doi.org/10.1177/0162243917718942.

Kalra, Nidhi, and Susan M Paddock. 2016. “Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?” Research Reports. Santa Monica, CA: RAND Corporation. https://doi.org/10.7249/RR1478.

KBB Editors. 2022. “How to Buy a New Car in 10 Steps.” *Kelley Blue Book* (blog). March 23, 2022. https://www.kbb.com/car-advice/10-steps-to-buying-a-new-car/.

Keeling, Geoff. 2020. “Why Trolley Problems Matter for the Ethics of Automated Vehicles.” *Science and Engineering Ethics* 26 (February): 293–307. https://doi.org/10.1007/s11948-019-00096-1.

Keeling, Geoff, Katherine Evans, Sarah M. Thornton, Giulio Mecacci, and Filippo Santoni de Sio. 2019. “Four Perspectives on What Matters for the Ethics of Automated Vehicles.” In *Road Vehicle Automation 6*, edited by Gereon Meyer and Sven Beiker, 49–60. Lecture Notes in Mobility. Cham: Springer Nature Switzerland.

Liljamo, Timo, Heikki Liimatainen, and Markus Pöllänen. 2018. “Attitudes and Concerns on Automated Vehicles.” *Transportation Research Part F: Traffic Psychology and Behaviour* 59 (November): 24–44. https://doi.org/10.1016/j.trf.2018.08.010.

Lin, Patrick. 2014. “Here’s a Terrible Idea: Robot Cars With Adjustable Ethics Settings.” *Wired*, August 18, 2014. https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/.

———. 2017. “Robot Cars And Fake Ethical Dilemmas.” Forbes. April 3, 2017. https://www.forbes.com/sites/patricklin/2017/04/03/robot-cars-and-fake-ethical-dilemmas/.

Liu, Peng, Run Yang, and Zhigang Xu. 2019. “How Safe Is Safe Enough for Self-Driving Vehicles?” *Risk Analysis* 39 (2): 315–25. https://doi.org/10.1111/risa.13116.

Milakis, Dimitris, Bart van Arem, and Bert van Wee. 2017. “Policy and Society Related Implications of Automated Driving: A Review of Literature and Directions for Future Research.” *Journal of Intelligent Transportation Systems* 21 (4): 324–48. https://doi.org/10.1080/15472450.2017.1291351.

Millar, Jason. 2014a. “Proxy Prudence: Rethinking Models of Responsibility for Semi--Autonomous Robots.” SSRN Scholarly Paper ID 2442273. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2442273.

———. 2014b. “You Should Have a Say in Your Robot Car’s Code of Ethics.” *Wired*, September 2, 2014. https://www.wired.com/2014/09/set-the-ethics-robot-car/.

———. 2015. “Technology as Moral Proxy: Autonomy and Paternalism by Design.” *IEEE Technology and Society Magazine* 34 (2): 47–55. https://doi.org/10.1109/MTS.2015.2425612.

———. 2017. “Ethics Settings for Autonomous Vehicles.” In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, edited by Patrick Lin, Ryan Jenkins, and Keith Abney, 20–34. New York: Oxford University Press.

Mladenovic, Milos N., and Tristram McPherson. 2016. “Engineering Social Justice into Traffic Control for Self-Driving Vehicles?” *Science and Engineering Ethics* 22 (4): 1131–49. https://doi.org/10.1007/s11948-015-9690-9.

Moridpour, Sara, Majid Sarvi, and Geoff Rose. 2010. “Lane Changing Models: A Critical Review.” *Transportation Letters* 2 (3): 157–73. https://doi.org/10.3328/TL.2010.02.03.157-173.

NHTSA. 1995. “Synthesis Report: Examination of Target Vehicular Crashes and Potential ITS Countermeasures.” DOT HS 808 263. Cambride, MA: United States Department of Transportation.

———. 2017. “Automated Driving Systems 2.0: A Vision for Safety.” Washington, D.C.: United States Department of Transportation.

Nunes, Ashley. 2019. “Driverless Cars: Researchers Have Made a Wrong Turn.” *Nature*, May. https://doi.org/10.1038/d41586-019-01473-3.

Nyholm, Sven. 2018. “The Ethics of Crashes with Self-Driving Cars: A Roadmap, I.” *Philosophy Compass* 13 (7): e12507. https://doi.org/10.1111/phc3.12507.

Nyholm, Sven, and Jilles Smids. 2016. “The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?” *Ethical Theory and Moral Practice* 19 (5): 1275–89. https://doi.org/10.1007/s10677-016-9745-2.

———. 2020. “Automated Cars Meet Human Drivers: Responsible Human-Robot Coordination and the Ethics of Mixed Traffic.” *Ethics and Information Technology* 22: 335–44. https://doi.org/10.1007/s10676-018-9445-9.

O’Connor, M. R. 2019. “The Fight for the Right to Drive,” April 30, 2019. https://www.newyorker.com/culture/annals-of-inquiry/the-fight-for-the-right-to-drive.

Rahwan, Iyad. 2018. “Society-in-the-Loop: Programming the Algorithmic Social Contract.” *Ethics and Information Technology* 20 (1): 5–14. https://doi.org/10.1007/s10676-017-9430-8.

Ravid, Orly. 2014. “Don’t Sue Me, I Was Just Lawfully Texting & Drunk When My Autonomous Car Crashing into You.” *Southwestern Law Review* 44 (1): 175–208.

Rodríguez-Alcázar, Javier, Lilian Bermejo-Luque, and Alberto Molina-Pérez. 2021. “Do Automated Vehicles Face Moral Dilemmas? A Plea for a Political Approach.” *Philosophy & Technology* 34 (4): 811–32. https://doi.org/10.1007/s13347-020-00432-5.

Roy, Alex. 2018. “This Is the Human Driving Manifesto.” The Drive. March 5, 2018. https://www.thedrive.com/opinion/18952/this-is-the-human-driving-manifesto.

Ryan, Mark. 2020. “The Future of Transportation: Ethical, Legal, Social and Economic Impacts of Self-Driving Vehicles in the Year 2025.” *Science and Engineering Ethics* 26 (June): 1185–1208. https://doi.org/10.1007/s11948-019-00130-2.

Schoettle, Brandon, and Michael Sivak. 2014. “Public Opinion about Self-Driving Vehicles in China, India, Japan, the U.S., the U.K., and Australia.” Technical Report. University of Michigan, Ann Arbor, Transportation Research Institute. http://deepblue.lib.umich.edu/handle/2027.42/109433.

Sezer, Volkan. 2018. “Intelligent Decision Making for Overtaking Maneuver Using Mixed Observable Markov Decision Process.” *Journal of Intelligent Transportation Systems* 22 (3): 201–17. https://doi.org/10.1080/15472450.2017.1334558.

Shariff, Azim, Jean-François Bonnefon, and Iyad Rahwan. 2017. “Psychological Roadblocks to the Adoption of Self-Driving Vehicles.” *Nature Human Behaviour* 1 (10): 694. https://doi.org/10.1038/s41562-017-0202-6.

———. 2021. “How Safe Is Safe Enough? Psychological Mechanisms Underlying Extreme Safety Demands for Self-Driving Cars.” *Transportation Research*

Part C: Emerging Technologies 126 (May): 103069. https://doi.org/10.1016/j.trc.2021.103069.

Soltanzadeh, Sadjad, Jai Galliott, and Natalia Jevglevskaja. 2020. "Customizable Ethics Settings for Building Resilience and Narrowing the Responsibility Gap: Case Studies in the Socio-Ethical Engineering of Autonomous Systems." *Science and Engineering Ethics* 26 (5): 2693–2708. https://doi.org/10.1007/s11948-020-00221-5.

Susskind, Jamie. 2018. *Future Politics: Living Together in a World Transformed by Tech*. Oxford: Oxford University Press.

Tesla, dir. 2021. *Tesla AI Day*. https://www.youtube.com/watch?v=j0z4FweCy4M.

Thoma, Johanna. 2022. "Risk Imposition by Artificial Agents: The Moral Proxy Problem." In *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives*, edited by Silja Vöneky, Philipp Kellmeyer, Oliver Müller, and Wolfram Burgard. Cambridge University Press. https://philarchive.org/rec/THORIB-2.

Thornton, Sarah M., Benjamin Limonchik, Francis Eugene Lewis, Mykel Kochenderfer, and J. Christian Gerdes. 2019. "Towards Closing the Loop on Human Values." *IEEE Transactions on Intelligent Vehicles*, 1–1. https://doi.org/10.1109/TIV.2019.2919471.

Xu, Zhigang, Kaifan Zhang, Haigen Min, Zhen Wang, Xiangmo Zhao, and Peng Liu. 2018. "What Drives People to Accept Automated Vehicles? Findings from a Field Experiment." *Transportation Research Part C: Emerging Technologies* 95 (October): 320–34. https://doi.org/10.1016/j.trc.2018.07.024.

Zimmermann, Annette, Elena Di Rosa, and Hochan Kim. 2020. "Technology Can't Fix Algorithmic Injustice." *Boston Review*, January 9, 2020. http://bostonreview.net/science-nature-politics/annette-zimmermann-elena-di-rosa-hochan-kim-technology-cant-fix-algorithmic.