AI & SOCIETY (2021) 36:405–415
https://doi.org/10.1007/s00146-020-00964-6

ORIGINAL ARTICLE

Safety requirements vs. crashing ethically: what matters most for policies on autonomous vehicles

Björn Lundgren[1,2]

Received: 27 December 2019 / Accepted: 3 March 2020 / Published online: 6 April 2020

Abstract
The philosophical–ethical literature and the public debate on autonomous vehicles have been obsessed with ethical issues related to crashing. In this article, these discussions, including more empirical investigations, will be critically assessed. It is argued that a related and more pressing issue is questions concerning safety. For example, what should we require from autonomous vehicles when it comes to safety? What do we mean by 'safety'? How do we measure it? In response to these questions, the article will present a foundation for a continued discussion on these issues and an argument for why discussions about safety should be prioritized over ethical concerns related to crashing.

Keywords Autonomous vehicles · Self-driving vehicles · Ethical crashing · Trolley problem · Safety argument · Vision zero

1 Introduction

It is widely presumed that autonomous (or self-driving) vehicles will be safer than human-driven vehicles. This, in turn, is often recognized as an important argument for the future implementation of autonomous vehicles. Indeed, many authors have argued that autonomous vehicles' potential to be safer than ordinary vehicles provides strong ethical reasons to develop and then transition to using such vehicles. I will refer to this, and similar ideas, as 'the safety argument'. According to Daniel J. Hicks, versions of this "*safety argument* is perhaps the most widely cited argument in favor of the rapid development and widespread adoption of" autonomous vehicles (2018, p. 63). However, in the philosophical–ethical literature and in the public debate on autonomous vehicles, most papers discuss the issue of crashing—with a focus either on *how we should crash* (i.e., ethical crashing) or *who is responsible* in the event of a crash (see, e.g., Doctorow 2015; Hern 2016; Jaipuria 2017; Leben 2017; Lin 2014, 2015; Simon 2017; Wolkenstein 2018; see also Nyholm 2018a, b for overviews). In the debate on ethical crashing, there seems to be an implicit belief that since autonomous vehicles will be extremely safe, the issue of safety requirements will be of less importance than the issue of ethical crashing.

In this article, I will first critically assess the discussion on ethical crashing to argue that there are serious flaws in the discussion and that there is a further need to evaluate the safety argument. Next, I will critically evaluate the safety argument to illustrate that there are fundamental policy issues that need to be sorted out in relation to this argument, issues that are more pressing than ethical crashing. I am setting aside the issue of responsibility for crashes because two manufacturers recently declared that they will take responsibility for the accidents (Atiyeh 2015; Maric 2017)—if this trend continues the question of responsibility will (from a policy perspective) be less pressing. Questions of forward-looking responsibility may still be important from a policy perspective, but—as my argument will indicate—they relate strongly to the safety argument.[1]

Before turning to the arguments, I should mention the limitations and scope of the arguments in this article. First, I am concerned with a *technologically near* (*or close*) *future*. By 'technologically near future', I am not referring to a specific time, but rather a future in which autonomous vehicles start to be implemented broadly and a future in which there

✉ Björn Lundgren
bjorn.lundgren@iffs.se

[1] Institute for Futures Studies, Stockholm, Sweden

[2] Department of Philosophy, Stockholm University, Stockholm, Sweden

[1] This does not mean that it is not philosophically interesting to discuss who is responsible. It just means that the policy question is practically resolved.

is mixed traffic (i.e., traffic including both autonomous and human-driven vehicles). Thus, the issues that I will discuss in this article do not concern a *technologically further future* in which autonomous vehicles will have taken over; what I am concerned with is the issue of what policies should guide us to (or away from) such a future. Second, as already indicated, I am interested in applied normative questions. That is, ethical concerns that are relevant for policies. Thus, I am less concerned with normative evaluations of science-fiction. I am also less interested in more theoretical debates on right and wrong. That is, while some considerations are theoretical in nature, they should be policy relevant. It is in this light that I will criticize the current focus in the ethical discussion on autonomous vehicles.

The rest of the article will be structured as follows. In Ethical crashing, I will critically assess the discussion on ethical crashing. In The safety argument, I will turn to the safety argument. Finally, I will conclude and summate my findings.

2 Ethical crashing

2.1 What is wrong with the discussion on ethical crashing?

In this section, I will critically assess the discussion on ethical crashing, in particular the "methodological" focus of this debate, which is inspired by the so-called 'trolley problem' (Foot 1967; Thompson 1985). Thus, the focus is on what Sven Nyholm and Jilles Smids (2016) call 'applied trolley problems' (i.e., binary choice situations of how to crash in a situation when an accident is unavoidable). This has, arguably, been the most common focus in the philosophical and public debate on the ethics of crashing (see Nyholm 2018a for an overview). While it is fair to say that there is no consensus in the literature, I will refine some older arguments and introduce some new ones in support of the position that holds that the 'trolley methodology' is mistaken in some sense (e.g., because the applied trolley problems are irrelevant or misleading for the issue of ethical crashing). Despite broad criticism, application of the trolley methodology has been defended as recently as this year by Geoff Keeling (2020) and became broadly well-known because of the so-called *Moral Machine experiment* (Awad et al 2018).

According to Edmond Awad et al., consumers will only switch from human-driven vehicles to autonomous vehicles if they understand the origins of the ethical principles that are programmed into these vehicles (p. 59). This, according to Awad et al., implies that:

> even if ethicists were to agree on how autonomous vehicles should solve moral dilemmas [i.e., these applied trolley problems], their work would be useless if citizens were to disagree with their solution, and thus opt out of the future that autonomous vehicles promise in lieu of the status quo. Any attempt to devise artificial intelligence ethics must be at least cognizant of public morality (p. 59; my addition within brackets).[2]

Hence, Awad et al. attempted to investigate the public's preferences of decision-making in applied trolley problems (i.e., of "unavoidable accident scenarios with two possible outcomes"), proposing that "these preferences can contribute to developing global, socially acceptable principles for machine ethics" (ibid).[3]

Like most ethicists discussing ethical crashing, Awad et al. discusses binary choice situations of unavoidable accidents; accidents involving, for example, the choice between the unavoidable killing of a man and a woman. While discussions of such examples have been broadly popular, engineers have argued that they have yet to encounter a trolley problem and "if we did see a scenario like that, usually that would mean you made a mistake a couple of seconds earlier" (Hern 2016).[4]

So why would the focus on these binary choices make sense? Contrary to the engineer's supposition of the vehicles' faultlessness, autonomous vehicles will inadvertently crash (see, e.g., Goodall 2014a, 2014b; Lin 2015). Given that all crashes, arguably, include trade-offs, the vehicle must be prepared to crash in an ethical way. For example, Patrick Lin uses various types of trolley-like situations where the vehicles must choose between, for example, crashing into an 8-year old girl and an 80-year old grandmother (2015, p. 70). This type of example is supposed to illustrate the trade-off that is at the core of the trolley problem. Lin also thinks that these examples relevantly illustrate the need to program the vehicles to make ethical choices in situations where harm is unavoidable (ibid). But do they? There are several reasons to be critical—in particular—of the trolley methodology, but also—in general—of the discussion of the ethics of crashing.

Most of the problems I want to discuss relates to different problems of idealization. First, the application of trolley problems with scenarios involving two possible outcomes are highly idealized. But reality is not. And when you add uncertainty to a situation you are arguably changing the

[2] Casey (2017) argues that the problem should be resolved by lawyers instead, by making use of liability regulations.

[3] Recently Harris (2020) have criticized the Moral Machine experiment for, for example, conflating preferences with morality, calling the work of Awad et al. "useless" (pp. 74–75).

[4] That we should focus on avoiding trolley problems, has been defended more extensively by Alexander G. Mirnig and Alexander Meschtscherjakov (2019).

normative analysis of it. Previously, Nyholm and Smids (2016; cf., e.g., Goodall 2016 Himmelreich 2018), building on the work by Sven Ove Hansson (2013, cf. 2003) have criticized the usage of trolley problems, for example, for ignoring the fact that the machine decision-making involves risk (i.e., decision under known probabilities) and uncertainty (i.e., decision under unknown probabilities). Henceforth, I will sometimes—for simplicity—use 'risk' or 'uncertainty' to indicate both risk and uncertainty (cf. Hansson 2003).

As argued by Hansson (2003), there has been a flawed division of labor in philosophy, in which ethics deals with idealized and well-determined situations under the supposition that when the ethics of these idealized situations are resolved, decision-theory can deal with any uncertainties. However, Hansson argues that risks and uncertainties must be normatively evaluated (i.e., risk and uncertainty itself presents ethical problems that are not reducible to idealized examples), so risk and uncertainty cannot be dealt with by decision-theory alone. Supposing Hansson is right, then it is fair to say that there is "a categorical difference between trolley-ethics and accident-algorithms for AVs" (Nyholm and Smids 2016).

Recently, Keeling (2020) attempted to counter this argument by showing that the difference between choices in scenarios with absolute descriptions and standard decision-making under risk are not sufficiently different to warrant the claim of a categorical difference (pp. 299–300). However, this seems to miss the point made by Hansson (2003) that risky decision requires normative evaluation beyond what standard risk analysis offers. Hansson argues that standard normative theories currently do not address risks in a satisfactory way and his proposal ends up being very different from a standard risk analysis, which is merely about expected utility maximation. What Hansson proposes is that we have a *prima facie*—or *pro tanto*—right not to be exposed to risks, a right that may be overridden under specific conditions (2003).[5]

Although such arguments are not universally accepted (e.g., objective utilitarians would deny this), it is still a substantial question how we can translate ethical evaluations of absolute outcomes with perfect information to situations involving probabilities and uncertainty. As pointed out by Adam Bjorndahl et al (2017), Decisions that are easy to make under certainty can become much more difficult and morally fraught under uncertainty.

For the empirical methods of Awad et al., the situation is worse, since they are dealing with people's preferences, they cannot rely on a normative theory of how to aggregate from preferences about choices in scenarios with absolute descriptions to scenarios involving risk and uncertainty. The problem is that it is not evident that people's preferences in idealized scenarios, with certain outcomes of well-defined harm, can be perfectly converted to preferences in real situations involving risks of uncertain harms. For example, it is well-known that risk preferences cannot be presumed to uniformly match with standard models of expected utility (see, e.g, Kahneman and Tversky 1979), and it has been shown that even factors such as time can influence risk preferences (Andreoni and Sprenger 2012). Arguably, to enable a perfect conversion, we would have to presume the truth of the independence axiom, which sometimes fail to hold with generality under experimental settings (see, e.g., Chandler 2017 for a brief overview). The independence axiom allows us to deduce choice-preferences for more granular and complex situations from preferences of choices in simpler situations.[6]

Second, another form of idealization involves the problem of human–machine incongruence. Simply put, it is not evident that human preferences can be translated into rules for a machine. This is because choice-descriptions from a human and a machine perspective differs and may be incongruent. Indeed, the machines may both lack information humans have and vice versa, or the machine descriptions may be incompatible with human descriptions of reality, possibly making a translation impossible. Thus, it is not obvious that we can construct machines rules that satisfy the surveyed preferences, which potentially would provide a problem for policies based on such preferences. (cf. Lundgren 2020b).

A related argument is presented by Johannes Himmelreich (2018), who argues that reliance on trolley problems assumes "a top-down" design approach, which implements rules rather than allowing the system to learn (pp. 675–676). More to the point, conclusions about what should be done in singular trolley situations are hardly helpful, since the examples are too few to enable training data for so-called 'machine-training'. This is true even for the 26 million possibilities considered in the *Moral Machine Experiment* survey (Awad et al 2018, see the complimentary method description). Part of the problem is, again, the focus on binary choices, ignoring all forms of situational complexities and risk and uncertainty. Hence, the information is not representative for the complexity of reality. Alternatively, if we instead try a top-down approach, we could attempt to program the vehicles based on the results of Awad et al. However, as previously noted, it is unclear what the results

[5] Hansson uses the term '*prima facie*' following a long-standing tradition in ethics. However, as pointed out by me Lundgren (2020a), it would be more sensible to talk of a *pro tanto* right in this case, since as argued by Shelly Kagan (1989), '*prima facie*' is an epistemic concept (i.e., something that appears to be have genuine weight), while '*pro tanto*' indicates something that has genuine weight but may be overridden.

[6] I want to thank Erik Angner for a helpful conversation on the last issue in the paragraph.

are. For example, even if a majority prefers to save women over men, how can we translate that conclusions from binary choice situations to a broader principle ready for programming? We, certainly, cannot—in practice—encode each and every possible situation and if we were to attempt that, these 26 million possibilities are way too few.[7]

Keeling (2020, pp. 301–303) attempts to counter these types of arguments by noting that discussions of trolley problems may nevertheless be useful in that sense that analysis of them provide an answer to the question what we ought to do. While this may motivate ethicist to focus on the applied trolley problems, it offers no argument for applying the more empirical methods of Awad et al. More importantly, when doing *applied* ethics, it is arguably important to contribute with something that can actually be applied. I am not denying that trolley problems can be an important tool for normative ethics, but if we want our conclusions about applied ethical issues to be useful in practice (i.e., for real policies), then something must be added. Keeling does not seem to address this issue, which is arguably the real problem (or at least the argument that I and others are concerned with).

Third, the problem of human–machine incongruence is also related to what we may call the 'science–fiction presumption' (or at least presumptions that fall outside of the scope of a technologically near future). The problem is that in the discussion of ethical crashing, idealization does not only apply to the situations, but also to the type of information that the machines will be able to access—or extract from reality—instantly, while making a choice on how to crash. To illustrate my point, consider some of the vivid examples used in the applied trolley problems. For example, Lin's previously mentioned example—that is, of choosing between crashing into an 8-year-old girl and an 80-year-old grandmother—would require both instant face-recognition capabilities and retrieval of personal information. However, the problem also extends outside of the trolley methodology. For example, Derek Leben (2017), in arguing for a Rawlsian algorithm—based on a normative evaluation of utility in terms of the likelihood of survival, which we can question in its own right—"assume[s] that it is possible for an autonomous vehicle to estimate the likelihood of survival for each person in each outcome" (p. 110). Such abilities are far from the current and near future autonomous vehicles. Some of these examples would require an ability for autonomous vehicles to perform instant and complicated object identification (sometimes not only for types, but for tokens) and information retrieval in a time-limited accident situation. Similarly, evaluating the likelihood of survival with any precision would require tremendous capabilities not yet available.[8]

One may argue that science–fiction discussions, like Leben's examples, are still valuable. That is, one may argue that if the normative argument holds, which we can question, we can potentially use these idealized arguments as a guide on what to do in situation in a technologically closer future (similar to the argument from Keeling considered previously above).[9] However, the question, again, would be how we can abstract from these idealized (science–fiction) situations to fit with the way that current and near future technology does or will function.

Fourth, the discussion on human–machine incongruence also raised the issue that the accident situations are too idealized (i.e., beyond the issue of probabilities and uncertainties). As previously indicated, traffic situations like those envisioned in binary choice situations are arguably rare (cf., e.g., Hern 2016). Thus, even if discussions of applied trolley problems could give us guidance about how machines should act and be programmed or trained in binary choice situations of unavoidable crashes, it is not evident how we can extrapolate moral choices for any type of traffic accident from preferences or moral choices in trolley-based traffic accidents or a small sub-set of traffic accidents. Crashes in normal traffic are often more complex and involve many more choices (cf., e.g., Borenstein et al 2019), so it is not evident how we can abstract from moral choices in simple situations to moral choices in more complex situations.

Furthermore, the ethical choice of an autonomous vehicles, even in a crashing situation, cannot be designed in isolation. We must take the whole infrastructure into consideration (a similar point is made by Borenstein et al 2019; cf. also Nyholm and Smids 2016). Indeed, take—as an illustrative example—the Vision Zero policy, which aims to remove fatal and serious injuries (see, e.g., Belin et al 2012). It addresses the whole infrastructure. More importantly, it contradicts the engineer who previously argued

[7] I am not saying that this cannot be given an answer (e.g., we can assign a specific priority to women over men in accordance with the mean preference), I am saying that it is not clear whether this captures the actual preferences.

[8] In connection with this argument it is worth to mention that despite enthusiasm amongst some engineers and companies, there have been an increased skepticism about when, or even if, we can achieve a level 5 autonomous vehicle (see, e.g., Tibken 2018; Murray 2019; Henry 2020). Level 5 standardly implies full automation under all road conditions (see, e.g., SAE 2018). It is easy to see why this would be problematic if we take level 5 to include the ability to use as information input the kind of bodily expressions that pedestrians use to communicate with human drivers when, for example, passing a street.

[9] The normative presumptions are problematic for several reasons, for example, since this may exclude serious harm from which a person is likely to survive. See Keeling (2018) for a detailed criticism of the normative ideas underlying Leben's argument (including whether it is actually Rawlsian).

that accidents are an error, since avoiding fatal and serious injuries can sometimes require *more* non-serious traffic accidents. For example, a round-about would normally have more accidents than an intersection, but the accidents in a round-about would mostly be non-serious, while accidents in intersections are often serious (ibid).[10] What we should infer from this example is that we cannot evaluate ethics of any machine or machine-decision in isolation, but only as part of a larger system. The ethics of machine decisions need to take this system approach into consideration.

Lastly, there is a problem that only concerns the Moral Machine experiment. It is the problem of ethically bad preferences. That is, although Awad et al. claims that consumers will only shift technology if the ethical choices underpinning these machines respects the public opinions,[11] it is not evident that preferences *should* guide moral action (cf. Harris 2020). Indeed, suppose that a majority has preferences for racist policies, what guidance should that give us? Arguably, none. Thus, even if the survey reflected actual preferences of more realistic situations, which could be translated into rules for a machine, it is questionable if it should give us guidance in deciding upon such rules.

2.2 Why should we turn to the safety argument?

As I have argued there are various problems with the discussion on ethical crashing. However, that does not necessarily imply that we should direct our focus to the safety argument. Alternatively, it may imply that we need to revise how the issue of ethical crashing is discussed. Thus, before turning the discussion of the safety argument, I will briefly defend the idea that the safety argument is a more pressing policy concern, and that discussions of the safety arguments are relevant for further discussions of the ethics of crashing.

To answer this question, we should first look to the arguments in favor of the importance of the ethics of crashing. For example, Awad et al. argues that there is something special about a situation in which machines will make decision about who lives or dies (2018, p. 59). However, it is not clear why that requires more attention from a policy perspective, than choices in which human's make determinations about lives. Nor is it clear why it is more important *who* dies from a machine than whether the machine imposes serious risks to people's life or the quality of their lives.[12]

Nyholm (2018a) provides another form of argument. Based on three recent examples (from 2016 to 2018) of accidents involving autonomous vehicles, in which the failure was (at least in part) due to the machines rather than other human drivers, he argued that:

> These incidents in 2016 and 2018 illustrate that crashes involving self-driving cars are not merely material for hypothetical thought experiments. This is a real-world issue. It requires a serious response from both society and the developers of self-driving cars. Human lives are at risk. Accordingly, the new and developing topic of the ethics of crashes with self-driving cars is a very important one. (pp. 1–2).

Yet, these—arguably anecdotal—examples are insufficient to establish how important this issue will be. Even if crashes—in a technologically near future—remain relatively common, the ethics of crashing is only relevant for a subset of all accidents (i.e., those involving substantial choices). That subset is likely much smaller than the amount of people we can save by appropriate safety requirements. Thus, it is arguably more pressing to consider what we should accept when it comes to accidents and safety policies.[13]

Furthermore, as pointed out already by Bryant Walker Smith (2015, cf. Thierer 2015), under the presumption that

[10] This argument will apply to autonomous vehicles if they, like humans, have a behavior that is partly imperfect, as it relates to the given safety-goal. While it should be held true that AI applications will be partly imperfect, it is likely that the autonomous vehicles will result in different kinds of errors than those of human drivers.

[11] The implicit claim that consumers would not purchase (or use) an autonomous vehicle without influence over the ethics settings is not supported by Awad et al. and there is *prima facie* evidence to the contrary. You can look to products or services on the market with unpopular policies, but more importantly at least one survey asked the question of "Who should determine how the car responds to the Tunnel Problem?" (the Tunnel Problem is an applied trolley problem in which the choice is between killing you—the passenger—or a child). Respondents answer: Passenger (44%), Lawmakers (33%), Manufacturer/designer (12%), and Other (11%) (Moon et al 2014). Although there are methodological limits to this survey, we should note that even if we take "Other" to be "the public", it is only a weak majority that answers in a way that may support Awad et al. Moreover, the survey do not ask what you would require to buy the vehicle, but what you would prefer. We have no reason to think that everyone who would prefer to set their own settings would also require such a function to buy the vehicle.

[12] An illustrative example is the case of the Ford Pinto. According to Ibo van de Poel and Lambèr Royakkers, Ford knew that the car could explode under special circumstances. They could also make adjustments that would protect against it. Ford opted not to do so, based on a cost–benefit analysis of the societal costs and benefits. In the end the vehicle exploded with a couple of teenagers inside (2011: 65–70).

[13] Now, of course, it may seem as if I have not only re-introduced the trolley problem, but also argued that the answer is simple: prioritize more people's lives over fewer. However, the argument above lacks an important element in order for it to match with the standard trolley problem. In the trolley problem we have a choice between action and inaction, to interject in an ongoing event. Here we have a choice what to focus our research endeavors on. That some may already have focused on applying the trolley problems to autonomous vehicles is certainly no reason to keeping doing it. (We could potentially argue that the same argument applies to how the trolley problem is used in the discussion—i.e., that it is not really a trolley problem.)

autonomous vehicles have a potential to reduce the death toll from accidents substantially, we should ask: "what is the proper balance between caution and urgency in bringing these systems to the market? How safe is safe enough?" These issues are not only more important than the ethics of crashing, but the importance of ethical crashing also depends on these issues. If autonomous vehicles cannot be justified, then the ethics of crashing is just a theoretical problem of little or no practical concern.[14] Thus, these issues are related, and the relevance of ethical crashing depend on settling the questions related to the safety argument.

Presumably, however, autonomous vehicles can be justified. Nevertheless, the issues of justification and safety requirements are more important, because it is about the fundamental question whether we should use the technology at all and if so, how? That is, under which conditions should it be allowed, relative to safety requirements, to broadly use autonomous vehicles and under which conditions should we switch from human driven vehicles to autonomous vehicles? While the debate on ethics of crashing and the responsibility of crashes seems to presume that "Self-driving cars hold out the promise of being much safer than regular cars" (Nyholm 2018a, p. 2; cf., e.g., Hevelke and Nida-Rümelin 2015, p. 620), I will argue in "The safety argument" that this claim is more complicated than it *prima facie* seems and that the safety argument requires further analysis. Hence, I will now turn to address issues relating to these questions.[15]

3 The safety argument

3.1 Specifying the *safety argument*

As implied by Hicks (2018), one of the main reasons to favor an autonomous vehicle over a human-driven vehicle would be that the former is supposed to be safer than the latter. In this part I will critically asses this argument (i.e., the safety argument). To do so, I will introduce a *prima facie* reasonable specification of the argument in the form of a justification-criterion: *a necessary criterion of justification for the broad usage of autonomous vehicles is that they should be at least as safe as human-driven vehicles*. Henceforth I will call this the 'safety-criterion'.

The aim here is to argue that there are complications related to the safety-criterion that deserve further attention from an ethical perspective as policy considerations are concerned. Note that while I have stipulated the thesis I want to consider (i.e., the safety-criterion), the discussion will not depend on accepting the safety-criterion as such. Even if we reject this thesis, or argument, the discussion will still be relevant for other versions of the safety argument more broadly. That is, the main point is to illustrate that the discussion on these issues deserve more attention. In particular, I will aim to specify what we need to discuss and introduce some preliminary suggestions on how this discussion should proceed.

The discussion will require a degree of conceptual analysis and it involves some empirical issues, which are partly normative. More importantly, satisfying the safety-criterion is further complicated by the fact that there are policy proposals that we out to enact, which would improve the safety of human-driven vehicles; thus further pushing the demands on the level of safety that autonomous vehicles must achieve.

3.2 What do we mean by 'safety'?

To determine whether autonomous vehicles are as safe as human-driven vehicles, we must first qualify what we mean by 'safety'. Such a qualification is not, generally, as straight forward as it may seem. Indeed, while 'safety' standardly, in a technical context, is thought of as the inversion of risk, the concept is arguably more complex (Möller et al 2006). What we may call the traditional view of traffic safety, matches close with the standard technical conception of safety. Traditionally traffic safety is defined as the absence of accidents. In some more modern traffic safety policies, such as the Vision Zero policy, safety is defined as the absence of severe or lethal accidents (see, e.g., Belin et al 2012).

As previously noted, these two ideas about what safety is (or what the goal of safety is) yields different policy proposals. This is because there are trade-offs between different forms of accidents. Again, an illustrative example of this are roundabouts, which have a higher accident rate than four-way crossings. However, with a round-about the accidents are mainly non-severe examples of vehicles brushing into each other. Comparatively, while four-way crossings have fewer than round-abouts, when accidents occur—in four-way crossings—they are usually of a more severe kind (such as full-frontal or frontal-side collisions). Thus, if we want to avoid as many accidents as possible, then a four-way crossing is better than a round-about; if we want to avoid severe and lethal accidents, then a round-about is better than a four-way crossing.

Although the Vision Zero policy have been criticized (see Abebe et al 2020 for an overview), we have reason to settle

[14] Autonomous vehicles are, of course, already on our roads. But it is not impossible that we may conclude that autonomous vehicles should not be broadly used, and this decision can be made before the implementation of any crashing algorithms. If so, then the ethics of crashing is just a theoretical problem.

[15] Of course, justification certainly depend on other issue than traffic safety (such as climate effects—see Kopelias et al 2019 for a recent review article on environmental impacts and climate effects of autonomous vehicles). For simplicity I will set those issues aside in part of the upcoming discussion to show that issues related to safety require further normative analysis.

for the Vision Zero policy, rather than the traditional view, if we think of lethal or severe accidents as unacceptable in road traffic. Such a view may, for example, be supported by Hansson's analysis of the ethics of risk. As previously noted, according to Hansson we have a *prima facie*—or, *pro tanto*[16]—right not be exposed to risks.[17] Hansson argues that this right can only be overridden under some specific circumstances such that:

> Exposure of a person to a risk is acceptable if and only if the total benefits that the exposure gives rise to outweigh the total risks, measured as the probability-weighted disutility of outcomes. (2003, p. 306).

A potential problem, however, is that engineers seems to develop autonomous vehicles in accordance with the traditional traffic safety view (cf., e.g., Hern 2016; Mirnig and Meschtscherjakov 2019). Of course, it may turn out that the best strategy to achieve autonomous vehicles that reduces severe and lethal accidents is to reduce accidents in general. Nevertheless, traffic planners, should still approach the traffic system with an intent to minimize severe and lethal accidents (even if the best practices for achieving that may change if autonomous vehicles differ in substantial ways from human-driven vehicles).

3.3 Should safety of autonomous vehicles be reduced merely to accident-related safety?

When we talk of safety in this context, we implicitly seem to think of accident-related safety (as implied by the previous section). However, in western societies, exhaustion and noise from traffic have a more substantial effect on human lives than accidents. For example, in the US, more Americans die from pollutions from vehicles than from traffic accident (Caiazzo et al 2013, p. 207). Furthermore, according to The World Health Organization, traffic noise is second only to air pollution when it comes to health effects.[18]

This clearly implies that traffic policies need to take a broader scope of issues into consideration. In line with Vision Zero we should accept a zero policy for road traffic deaths and severe harm more broadly, not just in relation to accidents.

These factors may, to some extent, depend on switching to autonomous vehicles. For example, switching to autonomous vehicles may affect number of vehicles or the total travelled distance (see, e.g., Soteropoulos et al 2018). Nevertheless, for simplicity I will set these issues aside to—in the next sections—further investigate the safety-criterion and safety-argument relative to accident-related safety-requirements.

3.4 Problems of measuring the safety of autonomous vehicles

Supposing we have settled for the relevant measurement and/or conception of safety (or that we can deal with different conceptions and measurements at the same time), then the issue of determining the safety level of autonomous vehicles still remain. According to Nidhi Kalra and Susan M. Paddock, autonomous vehicles would need to drive 275 million miles "without failure to demonstrate with 95% confidence that their failure rate is at most" 1.09 deaths per 100 million miles (2016, p. 191). Of course, as previously noted, failures have already occurred. Thus, in order "to demonstrate with 95% confidence and 80% power that their failure rate is 20% better than the human driver failure rate of" 1.09 deaths per 100 million miles would require 11 billion miles driven. According to Kalra's and Paddock's estimates, this would take 500 years. Comparatively, Hicks estimates that it would go much faster, only 84 years (2018, p. 64). Thus, even with more favorable estimates, using the best available observational data, it will take a lot of time to determine if autonomous vehicles are indeed as safe as or safer than human-driven vehicles.[19]

However, although companies developing autonomous vehicles use many alternative methods, many of these have limits as well. For example, results from simulations and mathematical proofs are merely as certain as the assumptions that they are based on (Hicks 2018, pp. 64–65). However, most developers of autonomous vehicles probably use a combination of technologies which may allow them to test and revise such assumptions. For example, while Waymo uses conventional road tests they have also "driven "tens of billions" of miles through computer simulations" (Nieva 2020). Another type of combined effort is used by Tesla. When the vehicle is controlled by a human, the automated

[16] See footnote 5.

[17] This, of course, does not imply that competing ideas (such as cost–benefit analysis) are necessarily wrong when they claim that non-severe accidents are more costly than severe or lethal accidents. It just means that we have settled the conceptual issue of the main safety concern (i.e., that it is about is severe physical harm to human lives).

[18] https://www.transportenvironment.org/what-we-do/vehicle-noise.

[19] It is worth to point out that, for example, for Waymo (Google's autonomous vehicle company) "It took 10 years for the first 10 million, then a little over a year for the next 10 million" (Nieva 2020). At a speed of 10 million miles per year it would take Waymo along 1100 years to do 11 billion miles. However, that is just *one* company and what these number shows is an incredible increase in numbers of miles driven per year. For example, if we have 20 companies doing the same mileage that gives us 55 years (excluding any yearly increase in mileage). At the time of writing this Waymo will soon start mapping routes in Texas and New Mexico, which will then be followed by truck driving (Woodyard 2020). These and other trends illustrate that the increase in resources continue in this area, which will increase mileages per year.

system will run in the background and compare its decisions against the human decisions. However, although that provides us with lots of relevant data, it does not tell us if alternative strategies by the autonomous system would have been successful (Hicks 2018, pp. 64–65).

An alternative option is the possibility of performing experiments with the aim of testing autonomous vehicles in extreme and difficult situations. Indeed, Neil McBride (2016) suggests a driver test for autonomous vehicles (i.e., requiring a driver license for the vehicle). Testing has been broadly suggested before (see, e.g., Koopman and Wagner 2017, p. 93, for several references). Of course, such test must be designed with a high level of variation to both ensure that the vehicle can manage all types of road conditions (for an early attempt to develop a framework for safety validation, see Koopman and Wagner 2018) and to ensure that vehicle manufactures do no designing their vehicles simply for the test, rather than for natural situations, similar to how Volkswagen cheated on emission tests.[20] Relatedly, there is also the question whether the source code should be evaluated by an independent organization (Holstein et al 2018).

Furthermore, we can also aggregate from different methods. Indeed, if different ways of measuring the safety of autonomous vehicles converges, then that strengthens the evidence. Unless there is a systematic error for each or all testing methods, multiple methods can be used to gain reliable results since it is possible to statistically exclude the small chance that they converge because of coincidence or random errors.

These empirical questions are important because we must ask ourselves how much empirical evidence that we should require to satisfy the safety-criterion or any specific safety requirement. However, we should also recognize that we currently experience similar challenges with new vehicle models. There is uncertainty as to whether they will be as safe as vehicles already available. Yet, once a new type of vehicle is released, statistical data would more rapidly accumulate.[21] Nevertheless, we should recognize that new models are reasonably closer in kind to what is already available.[22]

3.5 Safe as comparative to what?

The issue of comparing autonomous vehicles and human-driven vehicles also depend on applying a correct comparison. Arguably, comparing with current death rates of accidents (or other current accident rates), as Karla and Paddock (2016) does, is normatively misleading. That is, although autonomous vehicles may offer a promise to provide a safer option than human-driven vehicles—simply because human error is the major cause of accidents (approximately 93% in the USA)—it needs to be noted that a third of the human errors are due to intoxication, 30% is speeding, and 20% is due to distracted drivers (Fagnant and Kockelman 2015). Thus, by installing alcohol locks, speed controls (suppose, e.g., that speed limits will be wirelessly transferred to the vehicles from the traffic operatives), and technology to evaluate driver focus (see, e.g., Sandle 2017; Szeszko 2017), we could avoid, or at least reduce, most of human errors. Thus, when comparing the safety of autonomous vehicles against the alternative of human-driven vehicles, we need to compare accident rates for *future* human-driven vehicles, not accidents rates using old technology. Indeed, perhaps the most promising option may turn out to be AI-assisted human-driven vehicles.[23]

However, although alcohol interlocks have been commercially available for decades, no country has a policy of requiring all new vehicles to be sold with alcohol interlocks. Although a EU vehicle safety standards proposals from 2018 included a requirement that "All new vehicles sold in the EU will feature a standardised interface to enable the fitment of aftermarket alcohol interlock devices" (ETSC 2018), we should recognize that efficient policies are not always enacted. Therefore, autonomous vehicles may, all-things-considered, turnout to be safer than AI-assisted human-driven vehicles, even in a scenario where an AI-assisted human-driven vehicle would be safer under the optimal policy requirements. Conversely, Nyholm (2018b) argues that "the introduction of self-driving cars might put some pressure on people to either try to make their conventional cars safer or switch over to self-driving cars instead" (p. 6). So, autonomous vehicles may fast-track an improvement of human-driven vehicles as well.

Lastly, we need to keep in mind that just as human-driven vehicles can be improved by technology, autonomous vehicles will also improve over time. For example, in the beginning it may turn out that the safety of autonomous vehicles are improved relative to human-driven vehicles (with or without speed controls, alcohol locks, etc.) only in some ways, which technological development may or may not overcome in a slightly more distant near future. These issues are to a large extent technical questions, but a technically informed ethical analysis of how we should balance these trade-offs is needed.

[20] See, e.g., https://en.wikipedia.org/wiki/Volkswagen_emissions_scandal for an overview.

[21] As pointed out to me in conversation by Sven Ove Hansson.

[22] It may be illustrative to compare with the problems related to the Boeing 737 MAX.

[23] Requiring speed limiters and alcohol locks have previously been defended (Smids 2018; Grill and Nihlén Fahlquist 2012).

When setting a safety-requirement for autonomous vehicles, we need to consider what we should compare with. But we also need to consider if we have a broader responsibility to enact related policies for human-driven vehicles, which may postpone the use of autonomous vehicles. This may create complicated policy considerations, since such a postponement may imply a slower implementation of the technology. It is possible that a slower implementation of autonomous vehicles over time would result in a large loss of life, even if an earlier implementation of the technology would result in a larger loss of life now. Arguably, this illustrates quite well why the safety argument is more important than the ethics of crashing, because whether we should broadly implement autonomous vehicles depends firstly on consideration relating to the safety-criterion and safety considerations of both autonomous and human-driven vehicles, not the ethics of crashing.

4 Conclusion and summation

In "Ethical crashing", I argued that the focus on the issue of ethical crashing is problematic for two reasons. First, there are serious methodological challenges with the way that the discussion is currently being performed, both in the philosophical-ethical literature, in the empirical literature, and in the public debate. Second, the debate relates and is secondary to the more important issues of safety requirements and the safety argument.

In "The safety argument", I turned to the safety argument to argue that there are lots of considerations that need more attention from a policy perspective. We need to settle the conceptual debate on what we mean by safety and how broadly we should apply the concept. We also need to settle the issue of what we should require from safety validation and testing. Most importantly, we need to have a serious discussion about the justification of autonomous vehicles and address the normative question on requirements of safety-levels, and the—as I have argued—related issue of safety policies for human-driven vehicles. All of which shows that there is a lot to be done.

Acknowledgement Open access funding provided by Stockholm University. I want to thank two anonymous reviewers for *AI & Society* for their helpful comments. I also gratefully acknowledge that this article was written with funding from the *Swedish Transport Administration* (*Trafikverket*), and that open access funding was provided by *Stockholm University*.

References

Abebe HG, Edvardsson Björnberg K, Hansson SO (2020) Arguments against vision zero: a literature review. Unpublished manuscript.

Andreoni J, Sprenger C (2012) Risk preferences are not time preferences. Am Econ Rev 102(7):3357–3376. https://doi.org/10.1257/aer.102.7.3357

Atiyeh C (2015) Volvo will take responsibility if its self-driving cars crash. Car and Driver. Retrieved from 8 Oct https://blog.caranddriver.com/volvo-will-take-responsibility-if-its-self-driving-cars-crash/

Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Sharif A, Bonnefon J-F, Rahwan I (2018) The moral machine experiment. Nature 563(7729):59–64. https://doi.org/10.1038/s41586-018-0637-6

Belin M, Tillgren P, Vedung E (2012) Vision zero—a road safety policy innovation. Int J Injury Control Saf Promot 19(2):171–179. https://doi.org/10.1080/17457300.2011.635213

Bjorndahl A, London, AJ, Zollman KJS (2017) Kantian decision making under uncertainty: dignity, price, and consistency. Philos Imprint 17(7):1–22. https://hdl.handle.net/2027/spo.3521354.0017.007

Borenstein J, Herkert JR, Miller KW (2019) AVs and engineering ethics: the need for a system level analysis. Sci Eng Ethics 25(2):383–398. https://doi.org/10.1007/s11948-017-0006-0

Caiazzo F, Ashok A, Waitz IA, Yim SHL, Barrett SRH (2013) Air pollution and early deaths in the United States. Part I: quantifying the impact of major sectors in 2005. Atmos Environ 79:198–208. https://doi.org/10.1016/j.atmosenv.2013.05.081

Casey B (2017) Amoral machines, or: how robotics can learn to stop worrying and love the law. Northwest Univ Law Rev 112:1–20. https://scholarlycommons.law.northwestern.edu/nulr/vol111/iss5/4/

Chandler J (2017) Descriptive decision theory. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (Winter 2017 Edition). https://plato.stanford.edu/archives/win2017/entries/decision-theory-descriptive/

Doctorow C (2015) The problem with self-driving cars: who controls the code? The Guardian. Retrieved from https://www.theguardian.com/technology/2015/dec/23/the-problem-with-self-driving-cars-who-controls-the-code

European Transport Safety Council (2018) EU vehicle safety proposals to require standardised alcohol interlock interface. ETSC. Retrieved from 23 June, https://etsc.eu/eu-vehicle-safety-proposals-to-require-standardised-alcohol-interlock-interface/

Fagnant DJ, Kockelman K (2015) Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. Transp Res Part A 77:167–181. https://doi.org/10.1016/j.tra.2015.04.003

Foot P (1967) The problem of abortion and the doctrine of the double effect. Oxf Rev 5:5–15

Goodall NJ (2014a) Ethical decision making during automated vehicle crashes. Transp Res Record 2424(1):58–65. https://doi.org/10.3141/2424-07

Goodall NJ (2014b) Machine ethics and automated vehicles. In: Meyer G, Beiker S (eds) Road vehicle automation lecture

notes in mobility. Springer, Cham, pp 93–102. https://doi.org/10.1007/978-3-319-05990-7_9
Goodall NJ (2016) Away from trolley problems and toward risk management. Appl Artif Intell 30(8):810–821. https://doi.org/10.1080/08839514.2016.1229922
Grill K, Nihlén Fahlquist J (2012) Responsibility, paternalism and alcohol interlocks. Public Health Ethics 5(2):116–127. https://doi.org/10.1093/phe/phs015
Hansson SO (2003) Ethical criteria of risk acceptance. Erkenntnis 59(3):291–309. https://doi.org/10.1023/A:1026005915919
Hansson SO (2013) The ethics of risk: ethical analysis in an uncertain world. Palgrave Macmillan, Basingstoke
Harris J (2020) The immoral machine. Camb Q Healthc Ethics 29(1):71–79. https://doi.org/10.1017/S096318011900080X
Henry J (2020) VW Exec: level 4 self-driver may be as good as it gets. *WardsAuto*. Retrieved from 9 Jan https://www.wardsauto.com/ces/vw-exec-level-4-self-driver-may-be-good-it-gets
Himmelreich J (2018) Never mind the trolley: the ethics of autonomous vehicles in mundane situations. Ethical Theory Moral Pract 21:669–684. https://doi.org/10.1007/s10677-018-9896-4
Hern A (2016) Self-driving cars don't care about your moral dilemmas. The Guardian. Retrieved from 22 Aug https://www.theguardian.com/technology/2016/aug/22/self-driving-cars-moral-dilemmas
Hevelke A, Nida-Rümelin J (2015) Responsibility for crashes of autonomous vehicles: an ethical analysis. Sci Eng Ethics 21(3):619–630. https://doi.org/10.1007/s11948-014-9565-5
Hicks DJ (2018) The safety of autonomous vehicles: lessons from philosophy of science. IEEE Technol Soc Mag 37(1):62–69. https://doi.org/10.1109/MTS.2018.2795123
Holstein T, Dodig-Crnkovic G, Pelliccione P (2018) Ethical and social aspects of self-driving cars. arXiv preprint. https://arxiv.org/pdf/1802.04103.pdf
Jaipuria T (2017) Self-driving cars and the trolley problem. The Blog. The Huffington Post. Retrieved from https://www.huffingtonpost.com/tanay-jaipuria/self-driving-cars-and-the-trolley-problem_b_7472560.html
Kagan S (1989) The limits of morality. Clarendon, Oxford
Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47(2):263–292. https://doi.org/10.2307/1914185
Kalra N, Paddock SM (2016) Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? Transp Res Part A 94:182–193. https://doi.org/10.1016/j.tra.2016.09.010
Keeling G (2018) Against Leben's Rawlsian collision algorithm for autonomous vehicles. In: Müller V (ed) Philosophy and theory of artificial intelligence 2017. PT-AI 2017. Studies in applied philosophy, epistemology and rational ethics, vol 44. Springer, Cham, pp 259–272. https://doi.org/10.1007/978-3-319-96448-5_29
Leben D (2017) A Rawlsian algorithm for autonomous vehicles. Ethics Inf Technol 19:107–115. https://doi.org/10.1007/s10676-017-9419-3
Keeling G (2020) Why trolley problems matter for the ethics of automated vehicles. Sci Eng Ethics 26:293–307. https://doi.org/10.1007/s11948-019-00096-1
Koopman P, Wagner M (2017) Autonomous vehicle safety: an interdisciplinary challenge. IEEE Intell Transp Syst Mag 9(1):90–96. https://doi.org/10.1109/MITS.2016.2583491
Koopman P, Wagner M (2018) Toward a framework for highly automated vehicle safety validation. SAE Tech Pap. https://doi.org/10.4271/2018-01-1071
Kopelias P, Elissavet D, Vogiatzis K, Skabardonis A, Zafiropoulou V (2019) Connected & autonomous vehicles—environmental impacts—a review. Sci Total Environ 712:135237. https://doi.org/10.1016/j.scitotenv.2019.135237
Lin P (2014) Here's a terrible idea: robot cars with adjustable ethics settings. Wired. Retrieved from 18 Aug https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/
Lin P (2015) Why ethics matters for autonomous cars. In: Maurer M, Gerdes J, Lenz B, Winner H (eds) Autonomes Fahren. Springer, Berlin, pp 69–85. https://doi.org/10.1007/978-3-662-45854-9_4. Also available in 2016-edition in English (*Autonomous Driving*)
Lundgren B (2020a) Against AI-improved personal memory. In: Haltaufderheide J, Hovemann J, Vollmann J (eds) Aging between participation and simulation: Ethical dimensions of socially assistive technologies in elderly care. De Gruyter, Berlin, pp 223–233
Lundgren B (2020b) Ethical machine-decisions and the input-selection problem. *Manuscript under review*.
Maric P (2017) Audi to take full responsibility in event of autonomous vehicle crash. Car Advice. Retrieved from 11 Sep https://www.caradvice.com.au/582380/audi-to-take-full-responsibility-in-event-of-autonomous-vehiclecrash/
McBride N (2016) The ethics of driverless cars. SIGCAS Comput Soc 45(3):179–184. https://doi.org/10.1145/2880000/2874265/p179-mcbride.pdf
Mirnig AG, Meschtscherjakov A (2019) Trolled by the trolley problem: on what matters for ethical decision making in automated vehicles. CHI '19: Proceedings of the 2019 CHI conference on human factors in computing systems, Paper No 509:1–10. https://doi.org/10.1145/3290605.3300739
Moon A, Millar J, Bassani C, Fausto F, Rismani S (2014) Robohub. Retrieved from 23 June https://robohub.org/if-a-death-by-an-autonomous-car-is-unavoidable-who-should-die-results-from-our-reader-poll/
Murray C (2019) Automakers are rethinking the timetable for fully autonomous cars. Plastics Today. Retrieved from 17 May https://www.plasticstoday.com/electronics-test/automakers-are-rethinking-timetable-fully-autonomous-cars/93993798360804
Möller N, Hansson SO, Peterson M (2006) Safety is more than the antonym of risk. J Appl Philos 23:419–432. https://doi.org/10.1111/j.1468-5930.2006.00345.x
Nieva R (2020) Waymo driverless cars have driven 20 million miles on public roads. CNET. Retrieved from 6 Jan https://www.cnet.com/news/waymo-driverless-cars-have-driven-20-million-miles-on-public-roads/
Nyholm S, Smids J (2016) The ethics of accident-algorithms for AVs: an applied trolley problem? Ethical Theory Moral Pract 19(5):1275–1289. https://doi.org/10.1007/s10677-016-9745-2
Nyholm S (2018a) The ethics of crashes with AVs: a roadmap. I Philos Compass 13(7):e12507. https://doi.org/10.1111/phc3.12507
Nyholm S (2018) The ethics of crashes with AVs: a roadmap, II. Philos Compass 13(7):e12506. https://doi.org/10.1111/phc3.12506
van de Poel I, Royakkers LMM (2011) Ethics, technology, and engineering: an introduction. Wiley-Blackwell, Chichester
SAE (2018) Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Ground vehicle standard J3016_201806. SAE Int. https://doi.org/10.4271/J3016_201806
Sandle T (2017) Artificial intelligence helps to keep tired drivers awake. Digit J. Retrieved from 6 Aug https://www.digitaljournal.com/tech-and-science/technology/artificial-intelligence-helps-to-keep-tired-drivers-awake/article/499369

Simon M (2017) To make us all safer, robocars will sometimes have to kill. Wired. Retrieved from https://www.wired.com/2017/03/make-us-safer-robocars-will-sometimes-kill/

Smids J (2018) The moral case for intelligent speed adaptation. J Appl Philos 35:205–221. https://doi.org/10.1111/japp.12168

Smith, B. W. 2015. Slow down that runaway ethical trolley. *CIS Blog* January 12. Retrieved from: https://cyberlaw.stanford.edu/blog/2015/01/slow-down-runaway-ethical-trolley

Soteropoulos A, Berger M, Ciari F (2018) Impacts of automated vehicles on travel behaviour and land use: an international review of modelling studies. Transp Rev. https://doi.org/10.1080/01441647.2018.1523253

Szeszko E (2017) Technology against drowsy driving. Medium. Retrieved from 30 Sep https://medium.com/vorm/technology-against-drowsy-driving-72ede9265b84

Thierer A (2015) Making sure the "Trolley Problem" doesn't derail life-saving innovation. The Technology Liberation Front. Retrieved from Jan 13 https://techliberation.com/2015/01/13/making-sure-the-trolley-problem-doesnt-derail-life-saving-innovation/

Tibken S (2018) Waymo CEO: autonomous cars won't ever be able to drive in all conditions. CNET. Retrieved from Nov 13 https://www.cnet.com/news/alphabet-google-waymo-ceo-john-krafcik-autonomous-cars-wont-ever-be-able-to-drive-in-all-conditions/

Thompson JJ (1985) The trolley problem. Yale Law J 94(5):1395–1515. https://doi.org/10.2307/796133

Woodyard C (2020) Self-driving big-rig trucks coming soon? Waymo set to begin mapping interstates in Texas, New Mexico. USA TODAY. Retrieved from 23 Jan https://eu.usatoday.com/story/news/nation/2020/01/23/waymo-texas-new-mexico-mapping-self-driving-big-rigs/4546366002/