Crash Algorithms for Autonomous Cars: How the Trolley Problem Can Move Us Beyond Harm Minimisation

Dietmar Hübner and Lucie White

Pre-print version. Published in *Ethical Theory and Moral Practice* (2018).
The final version is available at https://doi.org/10.1007/s10677-018-9910-x

Abstract

The prospective introduction of autonomous cars into public traffic raises the question of how such systems should behave when an accident is inevitable. Due to concerns with self-interest and liberal legitimacy that have become paramount in the emerging debate, a contractarian framework seems to provide a particularly attractive means of approaching this problem. We examine one such attempt, which derives a harm minimisation rule from the assumptions of rational self-interest and ignorance of one's position in a future accident. We contend, however, that both contractarian approaches and harm minimisation standards are flawed, due to a failure to account for the fundamental difference between those 'involved' and 'uninvolved' in an impending crash. Drawing from classical works on the trolley problem, we show how this notion can be substantiated by reference to either the distinction between negative and positive rights, or to differences in people's claims. By supplementing harm minimisation with corresponding constraints, we can develop crash algorithms for autonomous cars which are both ethically adequate and promise to overcome certain significant practical barriers to implementation.

1. Introduction

Self-driving cars have great potential to drastically reduce the number of crashes on our roads, particularly as factors relating to human error, including speeding, negligence, drink driving, and distracted driving, play a crucial role in a significant number of crashes globally (World Health Organization 2015). A national survey conducted in the USA estimated that the "critical reason" leading to a crash could be attributed to the driver in 94 percent of cases (National Highway Traffic Safety Administration 2015). However, even in a perfectly designed and absolutely failproof self-driving system, some crashes will be unavoidable, particularly because autonomous vehicles will almost certainly be introduced alongside cars with human drivers, and in areas also occupied by pedestrians and cyclists. Although manufacturers may prefer to focus solely on the considerable safety potential of this technology without wishing to engage in debates as to who

should survive and who should be sacrificed in a disaster, an autonomous vehicle will react in *some* way in the case of an unavoidable accident. The problem of *how* autonomous cars should crash, whether explicitly accounted for by the vehicle's programming or just implicitly determined by its overall driving control setup, is thus inexorable. Consequently, philosophers have begun to focus on the ethical problem of crash algorithms, and, to a more limited extent thus far, to suggest ethical frameworks that might help us to dictate how self-driving cars should behave in inevitable accidents.

Though this debate is still in its nascent stages, two concerns have become prominent. The first is the need to take the self-interest of potential consumers into account (Bonnefon, Shariff and Rahwan 2016; Gogoll and Müller 2017; Nyholm and Smids 2016). The second is the well-accepted liberal tenet that any mandated rule in this domain must be one that no member of society could reasonably reject (Gogoll and Müller 2017; Lin 2015; Mladenovic and McPherson 2016). Against this background, contractarian approaches to crash algorithms for autonomous cars appear particularly attractive, as they, by their very setup, take people's rational self-interest as their systematic starting point, while, at the same time, aiming at a reconciliation of these individual preferences into some common solution that all rational agents should be willing to accept. In this article, we will discuss a specific example of such a contractarian approach which ultimately endorses a harm minimisation algorithm as the optimal solution to the abovementioned concerns. We will argue, however, that harm minimisation, and contractarian approaches to the problem in general, are flawed. More precisely, crash algorithms must go beyond harm minimisation by recognising a fundamental distinction between those 'involved' and 'uninvolved' in a crash situation. Neglect of this distinction will result in algorithms that are both ethically inadequate, and that face practical difficulties concerning implementation.

In order to flesh out this distinction and reveal the need for it, we will turn to some of the first treatments of a much discussed and maligned thought experiment in the context of crash algorithms: the trolley problem. We will draw from discussions by Philippa Foot and Judith Jarvis Thomson to flesh out two ways in which the distinction between 'involved' and 'uninvolved' can be understood: as either demarcating those who hold positive rights to assistance from those who have negative rights to non-interference, or as distinguishing between those who hold stronger and weaker claims against interference. The abovementioned concerns with the self-interest of the car's occupants and the necessity of finding a solution acceptable to all, as well as concerns about real-life applicability, have led to widespread scepticism concerning

the use of the trolley problem to make progress on crash algorithms. But we will demonstrate that these discussions of rights and claims reveal shortcomings with contractarian attempts at devising crash algorithms, and a strategy of harm minimisation, drawing our attention to crucial ethical factors that these approaches cannot recognise or incorporate. Furthermore, we will show that sensitivity to the distinction between 'involved' and 'uninvolved' produces an approach which is better able to address the concerns of the sceptics than a contractarian solution.

Our argument will take the following form: after sketching the current landscape of the philosophical debate, with a particular focus on the repudiation of the analogy between crash algorithms for autonomous cars and the trolley problem (2), we show why a contractarian perspective on crash algorithms seems an attractive development, and how such an account arrives at a minimise harm rule (3). We then argue that contractarian approaches fail to provide us with an intuitively plausible and morally acceptable framework for dealing with crash situations for autonomous cars, as they are insensitive to, and unable to incorporate, certain factors that prove to be ethically crucial in some crash scenarios. We will draw out and elucidate these factors by returning to the earliest accounts of the trolley problem: Foot's differentiation between negative and positive rights (4) and Thomson's focus on claim-based aspects of dilemma situations (5) both emphasise distinctions that may be highly relevant to the construction of adequate crash algorithms for autonomous cars. We contend that these frameworks offer conceptual tools for drawing an essential distinction between involved and uninvolved persons (6), which is both ethically and pragmatically superior to a sole standard of harm minimisation (7).

2. Autonomous Cars and the Trolley Problem

In some academic treatments of crash algorithms for autonomous cars (see Bonnefon, Shariff and Rahwan 2015; Lin 2015), and particularly in the media (see Aschenbach 2015; Doctorow 2015; Millar 2014; Windsor 2015; Worstall 2014), much has been made of the possible parallels between crash behaviour of autonomous vehicles and the iconic 'trolley problem'. In the classical variation of this thought experiment, a third-party observer sees an out-of-control trolley heading towards five unsuspecting people on the track, who will certainly be killed if nothing is done. The observer is standing next to a switch, with which she can divert the trolley onto a second track. Unfortunately, there is one person standing on this second track, who will be killed if she intervenes and turns the switch.

At first glance, there is a striking resemblance between this situation, and the algorithm designer of an autonomous vehicle deciding whether, for example, a self-driving car facing an unavoidable crash should continue along its path, or swerve into another lane to minimise overall fatalities. Accordingly, the task of devising appropriate crash algorithms for autonomous cars may seem to be a more or less straightforward real-life application of the trolley problem. More precisely, the following three assumptions suggest themselves: (a) solutions for crash scenarios can be informed by and derived from assessments of trolley problems; (b) this is because the essential structures of crash scenarios and trolley cases are largely analogous; (c) in particular, the issue of crash algorithms must be viewed from the same fundamental normative perspectives and conform to the same general normative standards as the problem of trolley cases. Recently, however, several scholars have questioned this analogy. In fact, all three mentioned assumptions have been challenged: (a) some authors doubt that discussions of the trolley problem can lead us to solutions concerning crash algorithms; (b) some stress essential differences between real-life crash scenarios and idealised trolley thought experiments; (c) others emphasise deep incongruities in the normative demands suggested by each scenario. These challenges have played a fundamental role in shaping recent debate and inspiring alternative approaches to the problem.

(a) Patrick Lin (2015) endorses the broad analogy between the trolley problem and crash algorithms. However, he points out that discussion of this thought experiment, far from establishing an insightful ethical consensus that might settle the problem of crash algorithms, has resulted in intractable ethical disagreement. In fact, trolley problems are often used to illustrate irreconcilable incongruities between competing ethical approaches: an adherent of straightforward consequentialist views, such as act utilitarianism, will hold that the trolley should be turned to the alternative track, thus minimising harm (provided that there are no exceptional circumstances suggesting that the balance of utilities might actually be higher by letting the five die and sparing the one). A deontologist, conversely, might refuse to flip the switch, focusing on the nature of the act and stressing that killing is worse than letting die (irrespective of the number of persons doomed in each scenario). Accordingly, instead of providing a solution to the dilemma, trolley problems just seem to reproduce an inexorable controversy of antagonistic outlooks, each of which might be reasonably rejected. Adopting any one of them would thus seem to be incompatible with the standards of a liberal society, which forestall the imposition of any normative stances on other persons that cannot be expected to freely agree to them (see Gogoll and Müller 2017; Loh and Loh 2017; Mladenovic and McPherson 2016). It is plausible

that this problem has formed part of the reason that philosophers, thus far, have often been unwilling to make specific recommendations concerning how self-driving vehicles should crash, or at least tried to avoid any reference to specific normative frameworks when advancing their position (see Gogoll and Müller 2017; Goodall 2014).

(b) Sven Nyholm and Jilles Smids (2016) draw out several important differences between trolley problems as typically conceived, and real-life crash situations. Primarily, they are concerned that trolley problems are highly abstracted, and constitute acute, individual, simplified choice situations. This means that they are not sufficiently connected to crucial questions about moral and legal responsibility, failing to make room for and thus account for the concrete obligations and liabilities of different persons or groups involved. Furthermore, they have little resemblance to how crash algorithms will and should be designed: in a prospective manner, by a plurality of stakeholders, and with regard to a multitude of potentially relevant parameters. Finally, they note that in a fictitious trolley situation there is complete certainty concerning the outcome of each option, whereas realistic crash scenarios will be saturated with risks and probabilities.

(c) Jan Gogoll and Julian Müller (2017) maintain that trolley problems and crash situations each evoke profoundly different normative demands. Whereas a standard trolley problem just focuses on the perspective of an unaffected third-person agent deciding about the lives and deaths of others, a crash algorithm should take into account the perspectives of those affected by an accident, including drivers and their passengers. This is not just a pragmatic issue, insofar as crash algorithms must correspond to these persons' preferences if they are to have any prospects on the market, leading to the uptake and success of the technology (see Bonnefon, Sharif and Rahwan 2016). It is also an ethical issue, as these personal interests of potential buyers are fundamentally legitimate, and thus need to be adequately addressed by any crash algorithm supposed to provide a fair balance of all interests involved.

3. A Contractarian Perspective

We will revisit Nyholm and Smids' concerns about real-life complexity until section 7. For the moment, we will focus on the problems of acceptability and accommodation of interests. Gogoll and Müller (2017) propose that we approach crash algorithms for autonomous cars from a contractarian perspective. This does seem to be an attractive way of meeting these challenges: contractarianism is explicitly designed to derive standards acceptable to all rational agents, eliminating the problem of intractable ethical disagreement, and takes self-interest as its starting

point, ensuring that the concerns of the occupants are given appropriate weight in the equation. Based on this type of reasoning, Gogoll and Müller ultimately endorse 'minimise harm' as the optimal algorithm for self-interested agents.

Gogoll and Müller first consider whether liberal neutrality on crash rules might favour a "personal ethics setting" (PES) in which every driver could make her own choice concerning the crash behaviour of her autonomous vehicle based on her individual normative commitments, rather than imposing uniform ethical rules on all self-driving cars. However, they quickly conclude that this system would lead to unsatisfying results, as participants would find themselves in a classical 'prisoner's dilemma'. For simplicity, Gogoll and Müller imagine a system in which all participants are either 'selfish agents' or 'moral agents', with the choice between two possible algorithms, namely egoistic 'maximum protection', which prioritises the welfare of the car's occupants above all else, and altruistic 'harm minimisation', which attempts to achieve the optimal net outcome for all persons potentially affected. Under these conditions, opting for egoism over altruism is advantageous for any single participant whichever rule the other participants select: egoism is likely to grant the participant the *best outcome* in an accident if the other cars follow altruism, and will usually *fare better* than altruism if the other cars follow egoism as well, whereas altruism will tend to burden the participant with the *worst outcome* in an accident if the other cars follow egoism, and will normally *do worse* than egoism if the other cars follow altruism too. Thus being the 'dominant strategy', egoism will, in due course, be chosen by all participants. More precisely, it will be favoured by those who are exclusively motivated by self-interest, but also by those who would be willing to sacrifice themselves for the greater good, but not to become the prey of others' selfishness. That is, as long as the moral agents do not wish to adhere to 'harm minimisation' at all costs, but only under the condition that others are willing to act ethically, too, they will eventually be 'crowded out' by the selfish agents, gradually turning to 'maximum protection' as well. This result, however, is rationally suboptimal, as general egoism will lead to a worse outcome for every individual than general altruism would have procured. Even selfish rational agents will be disappointed by this result, and will wish that the altruistic rule was mandated from the beginning.

Gogoll and Müller take this as a conclusive reason for the agents to endorse a "mandatory ethics setting" (MES), i.e. a general regulation for crash algorithms. In order to spell out more clearly what this mandatory setting should amount to, they move from their game-theoretic scenario of clashing uncoordinated choices to a decision-theoretic analysis of finding an optimal common

rule. More specifically, they utilise an argument from probability, combining it with a specific decision-theoretic strategy: as an agent cannot know what position he will occupy in a future accident, Gogoll and Müller contend he should anticipate occupying any possible position with equal probability. Given this, a rational, self-interested agent interested in optimising his own prospects of avoiding harm should choose harm minimisation: clearly, this general rule will give him the best chances of surviving and not being injured. Indeed, this provides an explication of why the general implementation of harm minimisation was judged to be *rationally superior* to the universal choice of maximum protection in the above analysis. Now it becomes apparent that this rule is *rationally optimal* under the constraints of prudence and ignorance. Thus, by moving from a suboptimal "personal ethics setting" (PES) to a superior "mandatory ethics setting" (MES), Gogoll and Müller arrive at a general rule, without reference to a substantive ethical theory (in particular, without explicit commitment to utilitarian thought). Rather, they have argued solely from rational self-interest and its prudential reconciliation into a mutually optimal system (following classical lines of contractarian reasoning).

The idea that we can derive a utilitarian rule from egoistic rationality, presupposing ignorance of one's own position and aiming for maximisation of expected values, is not new to ethics. In fact, the reasoning employed by Gogoll and Müller runs largely parallel to an argument espoused by John Harsanyi, in which he, similarly referring to a self-interested, rational choice under risk, claims that a just society should distribute resources according to utilitarian standards. Harsanyi defends this claim by advancing a so-called "equiprobability model", in which every participant must assume the same probability 1/N of being allotted any single position within a community of N persons. He supposes that, when selecting a distribution of goods for this community, a rational participant would try to maximise her "expected utility" EU, defined as the sum of the utilities U_i attached to all possible positions, each multiplied by her probability 1/N of attaining this very position, i.e. $\Sigma_i (1/N \cdot U_i)$. This expected utility EU of a given distribution, furthermore, is equal to that distribution's average utility AU, defined as the sum of the utilities of all positions, divided by the number N of positions available, i.e. $(\Sigma_i U_i)/N$. Rational maximisation of expected utility thus yields average utilitarianism as the optimal distribution (Harsanyi 1953; 1982).

The parallels between Gogoll and Müller's and Harsanyi's argumentation are immediately apparent. In both cases, the rational self-interest of the participants and the ignorance of their own positions are assumed as boundary conditions, and in both cases, a utilitarian rule is derived.

The motivation behind the selection of these boundary conditions does differ between the two approaches: in Harsanyi's model they come in as moral constraints, meant to deliver a just distribution for general society (his account is contractualist in nature), whereas Gogoll and Müller take them as factual conditions, reflecting the supposed self-interest of each participant and the impossibility of anticipating one's future position (their account is explicitly contractarian). Nonetheless, the argumentative structure is essentially analogous: rational self-interest and ignorance of one's own position, when maximising expectation values, yield utilitarian optimisation.

These fundamental parallels with Harsanyi's eminent argument might lend a veneer of plausibility to Gogoll and Müller's theory. There is, however, another, more profound difference between the two approaches: while Harsanyi is concerned with the distribution of resources, Gogoll and Müller are concerned with killing and injuring people. Their theory, that is, pertains to the distribution of harms rather than goods. When we recognise this, a less palatable parallel springs to mind: Gogoll and Müller's argument echoes the reasoning in John Harris' notorious 'survival lottery' argument.

4. Negative Rights and Positive Rights

In his survival lottery thought experiment, Harris asks us to imagine a society in which organ transplantation procedures have been perfected. Every member of this society is given a lottery number. Whenever a doctor has two or more patients whose lives could be saved through transplantation of different organs, and there are no suitable organs available to save them, a computer system will pick a number at random. The person with this number will be killed, and her organs will be harvested to rescue the patients. Harris notes that this scheme could be refined in various ways to ward off potential difficulties concerning implementation, e.g. by ensuring just age distribution, or by excluding those from the system who have brought their organ failure upon themselves. Either way, on principled grounds at least, Harris contends that utilitarians should be enthusiastic proponents of this notion: the overall number of deaths will be minimised, and fewer people will die prematurely, thus maximising net utility. Furthermore, he maintains, it would be rational for everybody to endorse the scheme on prudential grounds: under this system, after all, one will be more likely to have one's life saved through the receipt of an organ than to be killed for donation. Harris does note that despite these arguments, we retain the intuition that it is simply unacceptable to sacrifice one healthy person in order to avoid the

deaths of two or more ill persons, and the confidence that there is a morally relevant difference here between killing and letting die, obscured by a sole focus on harm minimisation (1975).

What might provide us with the theoretical basis for elucidating this distinction, and avoiding the gruesome conclusion of the survival lottery? We can find one possible answer in the paper in which Philippa Foot first introduces the trolley problem. Foot presents the trolley scenario as part of a suite of examples, among which is the case of a doctor, considering whether to sacrifice one of his patients and distribute his organs to save five others. She compares this to a situation in which one patient requires a huge dose of a life-saving drug, with which five further patients could be saved instead. Foot contends that, in the latter example, we would be justified in denying the one patient the drug, even though he will die, to save the five. At the same time, she insists, we are clearly not justified in killing the one patient and removing his organs in order to save the five. Foot suggests that the best way of getting to the heart of our opposing assessments of the two situations is by distinguishing between "what one does" (actively) and "what one allows" (passively). There is, she argues, a crucial distinction, built into our moral system, between our duties not to interfere with others, and our duties to provide someone with aid (1967, 10).

This distinction can be framed in the jurisprudential terms of negative and positive rights and duties: a negative right is a right not to be impaired or harmed, and corresponds to a negative duty on others of non-interference, while a positive right is a right to be assisted or benefitted, and entails a corresponding positive duty on he from whom assistance is required. Negative rights are generally thought to 'trump' positive rights, particularly where the stakes are comparable, irrespective of the number of persons in each of the conflicting parties. Foot suggests that this distinction best captures the morally relevant features, and crucial moral distinctions between the pair of scenarios outlined above, as well as the other cases considered in her paper: in the drug example, we are "weighing aid against aid". Only positive rights are at stake, and it is thus acceptable, perhaps even required, to pursue the course of action that will minimise overall harm. To kill a man in order to distribute his organs, however, is to do him "injury to bring others aid". This would be to violate his negative rights, in order to satisfy our positive duties towards five others. Exactly on these grounds, Foot argues, the action is unacceptable (1967, 13).

Foot's distinction between negative and positive rights can be used to pinpoint what is wrong with the survival lottery: killing one innocent person in order to save the lives of others, in accordance with simple harm minimisation or overall utility maximisation, infringes a negative right on behalf of a positive right, while the stakes of both parties are the same. Such an action is ethically wrong, regardless of the number of persons killed or saved. This general distinction has promise in highlighting a constraint that, we contend, is also ethically crucial for crash algorithms: a self-driving car should not be allowed to rope uninvolved bystanders into crash situations, even where this would minimise overall deaths. For instance, an autonomous vehicle should not plough into a pedestrian on the sidewalk, or into a cafe on the roadside, in order to protect the occupants of the car or other cars on the road. Those who would not otherwise be hit, or non-participants in traffic, deserve special protection, prohibiting the introduction of their lives in an autonomous car's balance sheet when trying to minimise the number of victims. These contentions might be adequately captured by the notion that certain persons in a crash scenario might have (higher ranking) negative rights not to be killed, while others may have (lower ranking) positive rights to be saved.

In fact, Gogoll and Müller seem to be open to such additional constraints. Based on their contractarian argument, they propose the general maxim, "*Minimize the harm for all people affected!*" (2017, 695). Yet, in a footnote they add that it needs still to be clarified, "who should count as 'all people affected'" – particularly, whether this formula includes only motorised traffic participants or pedestrians too (2017, 695). Unfortunately, their approach does not give us a consistent means of incorporating such a qualification. Their contractarian theory supposes that a rational person, given uncertainty concerning her position in a possible future accident, opts for the system in which the fewest people overall are harmed, thus maximising her chance of avoiding injury or death. There is no principled way to leave bystanders out of the calculation – we are, after all, just as likely to occupy a bystander's position as any other.

Foot's approach, in contrast, gives us a natural means of expressing this distinction: to divert the car into a pedestrian on the sidewalk or a bystander in a cafe is to violate her negative right to not be interfered with or injured, whereas those already on the road might be seen as having positive rights to aid and assistance. Refusing to turn the car and kill the uninvolved bystander could thus be seen as akin to refusing to kill an innocent person and distribute her organs. Contractarian derivations miss this point by the very structure of their reasoning: the distinction between negative and positive rights is not a relevant parameter in an optimising decision based

on rational self-interest. This, however, only demonstrates the ethical shortcomings of such approaches, and their need for corresponding supplementation.

5. Differences in Claims

Yet, there is reason to think that Foot might take issue with our above analysis: in her version of the trolley problem, the decision-maker is the driver of the runaway trolley, and what he faces is solely "a conflict of negative duties" (1967, 12). From his position, both courses of action plausibly present themselves as 'doing' rather than 'allowing'. He has a negative duty to avoid injuring the one man on the side track, and a negative duty to avoid injuring the five men on his current path. As this situation does not involve a hierarchy of rights (negative over positive), Foot finds it apparent that the driver should make his decision on the basis of harm minimisation (thus killing the one rather than the five) (1967). One might argue that crash situations concerning autonomous cars should be conceived of in the same terms: even if we are to endorse the priority of negative over positive rights, following Foot, this hierarchy might simply not apply to autonomous cars, just as it does not apply, according to Foot, to the trolley problem. Rather, we have duties of noninterference concerning every potential crash victim, and should act (as Gogoll and Müller suggest) to minimise harm (or, in other words, to ensure that as few negative rights are violated as possible).

We may, then, have reason to abandon Foot's notion of negative and positive rights as a means of elucidating a distinction between 'involved' and 'uninvolved' when it comes to crash algorithms for autonomous cars. This, however, does not mean that we need to abandon the distinction entirely. We can find an alternative theoretical basis for this distinction in Judith Jarvis Thomson's well-known account of the trolley problem. Thomson presents the trolley case in its most famous form, envisioning the decision maker as a third-person bystander, rather than the train driver. This situational shift casts the case in a different light: turning the trolley still seems a clear instance of 'doing', but refusing to pull the lever now seems to be a more clear-cut case of 'allowing'. However, Thomson thinks that, on the face of things, it is still permissible (though not required) to turn the trolley in her modified scenario (1985). At the same time, she agrees that it is clearly impermissible for a surgeon to cut up a man and distribute his organs to save five others. Because both of these cases now represent instances of doing vs. allowing, Thomson argues that Foot's distinction between negative and positive rights does not get to the crucial ethical difference here. Instead, Thomson proposes, the relevant distinction between these two

cases will best be elucidated by looking at the special, context-dependent claims that the various parties have against each other (1976).[1]

In the trolley scenarios sketched by Foot and herself, no such difference in people's claims can be ascertained: the five on the main track as well as the one on the side track appear to have the same claims to being spared or saved, whether by the driver or a bystander. However, Thomson suggests alternative scenarios in which potential differences in claims become apparent: for instance, the five on the main track could be railway workmen, having been warned of the danger of the location and receiving higher wages to compensate for the prospective risks, whereas the one on the side track could be an unsuspecting picnicker, having been invited by the mayor to have lunch there and having been guaranteed that there is no danger lurking on this part of the tracks. Additionally, the person to decide on the runaway trolley's path could be the mayor himself. Given this constellation, Thomson argues, the picnicker has stronger claims against the mayor than the five workmen, so that the five should die rather than him, contrary to the demands of harm minimisation (1976).

Thomson's claim-based framework is capable of absorbing a multitude of situational factors: in her own example, we may point to varying awareness of impending hazards, possible compensation for given risks, or explicit promises, trust, and special obligations. Many other factors may be associated with increased or reduced claims, including carelessness or disregard of warnings or prohibitions, the giving or withholding of consent, and various well-recognised rights such as property rights.[2] Several of these aspects, with their corresponding implications concerning claims, may also apply to crash situations including autonomous cars and provide us with another way to differentiate between persons involved and uninvolved. A pedestrian on a sidewalk or a person in a cafe, for example, may reasonably expect to be safe where they are, did not voluntarily enter a risk situation with motorised vehicles, do not share the advantages of self-driving cars, and may even object to motorised traffic in general. As a result, they might be said to have stronger claims against being killed by an autonomous vehicle than those participating in traffic. On the same grounds, they might be deemed 'uninvolved' rather than 'involved' in an imminent accident with an autonomous car. And again, the specific lens of contractarian reasoning is systematically incapable of accounting for these situational aspects: a rational choice

[1] In her 1985 work in which she presents the bystander variant of the trolley problem, Thomson frames similar concerns in terms of rights. However, in order to maintain a clear distinction between her and Foot's views, and because we will primarily draw from the 1976 paper in what follows, we stick to her 1976 usage of the term 'claims' in discussing these considerations.

[2] Thomson relies on the latter to account for why the surgeon may not distribute the patient's organs (1976).

meant to maximise one's expected survival will ignore all these additional factors and instead restrict itself to straightforward optimisation of chances.

6. Involved vs. Uninvolved

We should not be too quick, however, to abandon the notion of negative and positive rights as a potentially helpful tool for analysis of these matters. If, as Thomson suggests, we conceive of the trolley problem in its now classical form, i.e. as involving a bystander, rather than the driver, making the decision, it is plausible to understand this scenario as a case of doing (flipping the switch), and thus violating the negative rights of the one, versus allowing (failing to flip the switch), and thus refusing positive assistance to the five. Perhaps, then, both negative and positive rights are in fact involved in car crash scenarios as well. Indeed, the algorithm designer is arguably in a more similar position to Thomson's third-person bystander than to Foot's driver. If we adopt this perspective on the situation, and understand our distinction between uninvolved and involved in terms of negative and positive rights (adhering to the priority of the former over the latter), the terms 'uninvolved' and 'involved' acquire an *action-theoretic meaning*, based on the difference between doing and allowing: 'uninvolved' means 'would be unaffected if no action was taken' (and so has a negative right against being killed), while 'involved' means 'would be affected if no action was taken' (and thus has a positive right to be saved). This approach may suggest a rule forbidding the car from swerving into neighbouring lanes, sidewalks, buildings or other adjacent occupied areas: such a manoeuvre would violate the negative rights of persons who would otherwise survive (the 'uninvolved'), in order to satisfy the positive rights of persons who would be, through the manoeuvre, saved from the impending crash (the 'involved').

Alternatively, if we, as Foot might, doubt that the distinction between negative and positive rights applies to autonomous cars or deny, like Thomson, that this distinction is ethically important, we can still point to differences in persons' claims to insist that crash algorithms should prioritise the safety of the uninvolved over the involved. The terms 'uninvolved' and 'involved' now take on a *situation-contextual meaning*, based on people's expectations, engagements, individual decisions and social interactions: 'uninvolved' would mean 'may legitimately assume to be safe where she is, does not benefit from self-driving vehicles, etc.' (and so has stronger claims against being killed), while 'involved' would mean 'has voluntarily entered a traffic situation, ought to be aware of its inherent dangers, etc.' (and so has weaker claims against being killed). This perspective might prompt a rule prohibiting the car from driving into areas whose inhabitants should not be regarded as traffic participants freely accepting the corresponding

hazards: these persons (the 'uninvolved'), not partaking in traffic, with a fair expectation of safety etc., would have stronger claims to protection than other persons (the 'involved') who have entered the traffic zone and can be supposed to have agreed upon its inherent risks. Both rights and claims thus provide us with powerful analytical instruments for expounding upon the distinction between those involved and uninvolved in crash scenarios for autonomous cars, and lend support to the contention that the uninvolved require special protection. The precise meaning and content of this distinction, however, differs depending on which line of reasoning we follow.

It is not our intention here to advance an argument regarding which of the above senses of 'involved' vs. 'uninvolved' should be preferred, or to explicate more precisely how this distinction might translate into concrete rules for autonomous cars. Our aim, rather, is to introduce these elements into the conversation, as we are convinced that some sort of distinction between 'involved' and 'uninvolved' must be incorporated into an ethically adequate crash algorithm, and considerations of this sort have not, thus far, featured in the emerging philosophical debate on these issues.

Indeed, the need for such a distinction is often obscured by current approaches to the topic. For example, sketches and diagrams designed to illustrate arguments or gauge opinions concerning ethical crash algorithms fail to include scenarios in which people unambiguously uninvolved in a situation might be hit: generally, all people in such images are already on the road, or stepping onto it, so that all of them should arguably count as involved, at least on a claims-reading, and possibly also on a rights-reading of the term (see e.g. Bonnefon, Shariff and Rahwan 2016; MIT Media Lab 2017). By focusing only on situations in which there is no clear, discernible difference in the claims and rights of the persons affected and, consequently, no supplementary concerns besides harm minimisation need to be invoked, the insufficiency of this rule and the need for additional considerations is not made apparent.

7. Ethical and Pragmatic Advantages

We noted, in section 2, that much of the contemporary debate concerning crash algorithms for autonomous cars has been motivated by various objections against using the trolley problem as a theoretical tool for making progress on this real-life challenge. At the same time, we have drawn from the earliest discussions of trolley scenarios to bring out our distinction between 'involved' and 'uninvolved', which, we maintain, is necessary for an ethically adequate approach to crash

algorithms. Now we will return to the sceptical arguments in section 2. We will try to demonstrate that our proposed distinction is capable of addressing the concerns raised, perhaps better than the contractarian-based harm minimisation alternative that these reservations have engendered.

(a) Lin (2015) is correct that discussions of the trolley problem have often generated intractable ethical disagreement. This does not mean, however, that some considerations raised in trolley discussions, when deployed in a sufficiently sophisticated manner, cannot be of use in making progress on crash algorithms. In particular, referring to these considerations does not mean that we must adopt one normative perspective, to the exclusion of all others. Quite the contrary: bringing rights or claims into the equation does not imply abandoning harm minimisation altogether, but rather amending it with additional side constraints. Uninvolved persons should be afforded higher protection in crash scenarios, but when it comes to those already involved, a minimise harm rule is still appropriate. So, rather than just reproducing a confrontation between irreconcilable ethical perspectives, we can integrate elements from various approaches into a 'mixed' algorithm which is capable of accounting for a variety of ethical concerns. And in virtue of this, as we will explore below, it may be more likely to actually achieve widespread acceptance in society.

(b) Nyholm and Smids (2016) maintain that trolley cases are unable to provide us with an adequate basis for approaching questions of moral and legal responsibility, account for various stakeholder perspectives, and deal with inevitable risks and probability-based calculations. Regarding their first concern, the qualitative difference between negative and positive rights is highly essential to legal discourse. Any rule insensitive to this differentiation is thus likely to contradict the most basic standards of jurisprudential thinking. The concept of claims, in turn, is well suited to incorporate elementary demands of moral relationships, including trust, accountability, voluntariness and consent. Rights and claims are also better suited to capturing the concerns of various stakeholders, due to their sensitivity to position and perspective. Finally, by introducing rule-based constraints which are more independent from calculations of consequences, our approach remains more stable in the face of unavoidable uncertainties than recourse to a sole principle of harm minimisation.

(c) Gogoll and Müller (2017) suggest that crash algorithms must take into account the legitimate self-interest of all persons affected by a crash. We do not reject this claim, but rather go further:

an adequate crash algorithm must respect the legitimate self-interest of both those involved and uninvolved in a crash. This has an ethical and a pragmatic dimension. Ethically, just as a car's occupants can legitimately expect that their interests will not be sacrificed without appropriate consideration, uninvolved persons may legitimately invoke their special rights or claims against being unjustly involved in an accident. The supplementary concerns we advocate just present a fuller representation of the various interests of all agents that should be taken into account, given differences in their positions. Pragmatically, just as the technology will not succeed on the market unless it gives due consideration to drivers' legitimate self-interest, it will not be accepted by society at large if it fails to properly account for the rights and claims of all persons affected, including those uninvolved in traffic.

Gogoll and Müller contend that harm minimisation would in fact be embraced by society at large, appealing to empirical evidence suggesting that research subjects were "generally comfortable with utilitarian autonomous cars", programmed to minimise harm in crash situations (2017, 695). However, in the study cited by Gogoll and Müller, the authors note that this "utilitarianism is qualified by a self-preserving bias" (Bonnefon et al 2015, 7). In an expanded version of their work, they further emphasise this contention, framing it as the central dilemma for crash algorithms: although most people praised utilitarian vehicles that would sacrifice the car's occupants in order to minimise casualties, a majority also indicated that they would not buy a car programmed in this manner, and would object to regulations mandating harm minimisation for autonomous cars (Bonnefon et al 2016). The authors conclude that legislation enforcing a minimise harm algorithm "could substantially delay the adoption of AVs, which means that the lives saved by making AVs utilitarian may be outnumbered by the deaths caused by delaying the adoption of AVs altogether" (2016, 1575-6). Evidently, actual public preferences do not track Gogoll and Müller's argument, nor, we venture, is the argument itself likely to shift public opinion, particularly as it runs parallel to the famously counterintuitive survival lottery argument.

A framework containing distinctions between involved and uninvolved persons, being more in line with accepted moral and legal standards, may have a better chance of gaining widespread acceptance in society. In particular, supplementing harm minimisation with additional protections against being sacrificed for the greater good appeals to both legitimate rational self-interest and impersonal ethical intuitions. Regarding the rational dimension, such an approach not only represents the interests of those who could occupy any position in an accident with equal probability, but also of those who want to make sure that in some positions they will never

be targeted as possible victims. Regarding the ethical dimension, such a combination appeals not only to certain ideals of altruism (which always runs into the problem of appearing morally praiseworthy to adopt for oneself but morally suspect when imposed upon others), but also counterbalances these with basic ideas of justice (which has an inherent affinity to the concepts of rights and claims).[3]

One final pragmatic advantage of this distinction is worth mentioning in passing: this approach has the potential to increase transparency, which has been emphasised as a crucial factor in the emerging literature (Gerdes and Thornton 2015; Lin 2015). A car programmed to act according to fixed constraints makes the relationship between its behaviour and its programming much more apparent than a car that pursues unrestricted optimisation, following complicated calculations of costs and benefits, which may make it difficult to work out just why the car acted as it did. This increased intelligibility does not only assist in the retrospective assessment of an autonomous car's behaviour: it also facilitates its integration into existing traffic, by better allowing human drivers to predict its movements in real time. An autonomous car programmed to, for example, avoid swerving into neighbouring lanes, or to avoid certain zones, will be much easier to anticipate and react to than a sophisticated optimiser.

8. Conclusion

Doubts concerning the analogy between trolley cases and autonomous cars bring out the appeal of contractarian approaches to crash algorithms. However, contractarian perspectives are not sensitive to distinctions that must be incorporated to deliver an ethically and pragmatically adequate algorithm. These elements can be drawn out by abandoning a contractarian attempt at derivation, moving beyond a sole principle of harm minimisation, and turning to the notions of negative and positive rights and to the concept of claims, highlighted in the earliest discussions of the trolley problem. Doing so allows us to draw out an essential distinction between those involved and uninvolved in a crash situation. This approach is sufficiently sophisticated to deal with problems that, at first glance, seemed to speak against referring to trolley discussions. It integrates elements from various ethical perspectives to move towards a more pluralist and thus acceptable solution. It can give us a way of dealing with questions of moral and legal responsibility, accounting for a plurality of stakeholders and reducing problems concerning risk and uncertainty. And it is better able to acknowledge the legitimate interests of participants and

[3] Unfortunately we cannot, as of yet, point to empirical evidence establishing the actual appeal of this approach, because, as we noted in section 6, existing surveys do not unambiguously differentiate between involved and uninvolved persons. We hope that future research will begin to incorporate and emphasise these factors.

non-participants in traffic by taking their relevant differences into account. Such an approach is more likely to gain widespread acceptance in society, thus leading to greater uptake of the technology. Finally, it may also be advantageous in terms of behavioural transparency, thus facilitating integration into existing traffic.

We hope that the considerations we have presented demonstrate not just the need to acknowledge the distinction between 'involved' and 'uninvolved', but will also spur further discussion about how this distinction is best conceived. As we have demonstrated, it might be substantiated through differences between negative and positive rights, or through differences in claims based on information received, promises given, awareness, voluntariness, etc. This corresponds to the question of how and to which cases this distinction translates: whether, for example, we should make a distinction between those who would be affected by inaction and those who would be affected by action, or between those participating in traffic and those not taking part in it. Although this just represents the beginning of bringing these elements into the equation, we will need, in some way or another, to include them, and to move beyond harm minimisation.

Acknowledgments

Many thanks to Markus Ahlers, Sven Nyholm, and two anonymous referees from Ethical Theory and Moral Practice for their useful comments on an earlier draft of this paper.

References

Aschenbach J (2015) Driverless cars are colliding with the creepy trolley problem. The Washington Post. https://www.washingtonpost.com/news/innovations/wp/2015/12/29/will-self-driving-cars-ever-solve-the-famous-and-creepy-trolley-problem Accessed 4 October 2017

Bonnefon J, Shariff A, Rahwan I (2015) Autonomous vehicles need experimental ethics. Are we ready for utilitarian cars? Computing Research Repository. https://arxiv.org/abs/1510.03346v1 Accessed 4 October 2017

Bonnefon J, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. Science 352:1573-1576

Doctorow C (2015) The problem with self-driving cars: who controls the code? The Guardian. https://www.theguardian.com/technology/2015/dec/23/the-problem-with-self-driving-cars-who-controls-the-code Accessed 4 October 2017

Foot P (1967) The problem of abortion and the doctrine of double effect. Oxford Review 5:5-15

Gerdes J, Thornton S (2015) Implementable ethics for autonomous vehicles. In: Maurer M, Gerdes J, Lenz B, Winner H (eds) Autonomous driving: technical, legal and social aspects, Springer, Berlin, pp 87-102 http://link.springer.com/book/10.1007/978-3-662-48847-8

Gogoll J, Müller J (2017) Autonomous cars: in favor of a mandatory ethics setting. Sci Eng Ethics 23:681-700

Goodall N (2014) Ethical decision making during automated vehicle crashes. Transp Res Record 2424:58–65

Harris J (1975) The survival lottery. Philosophy 50:81-7

Harsanyi J (1953) Cardinal utility in welfare economics and in the theory of risk-taking. J Polit Econ 61:434-5

Harsanyi J (1982) Morality and the theory of rational behaviour. In: Sen AK, Williams BAO (eds) Utilitarianism and beyond, Cambridge University Press, Cambridge, pp 39-62

Lin P (2015) Why ethics matters for autonomous cars. In: Maurer M, Gerdes J, Lenz B, Winner H (eds) Autonomous driving: technical, legal and social aspects, Springer, Berlin, pp 69-85 http://link.springer.com/book/10.1007/978-3-662-48847-8

Loh W and Loh J (2017) Autonomy and responsibility in hybrid systems – the example of autonomous cars. In: Lin P, Abney K, Jenkins R (eds) Robot ethics 2.0. From autonomous cars to artificial intelligence, Oxford University Press, New York, pp 35-50

Millar J (2014). An ethical dilemma: when robot cars must kill, who should pick the victim? Robohub. http://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim Accessed 4 October 2017

MIT Media Lab (2017) Moral machine. http://moralmachine.mit.edu Accessed 4 October 2017

Mladenovic M, McPherson T (2016) Engineering social justice into traffic control for self-driving vehicles? Sci Eng Ethics 22:1131-49

National Highway Traffic Safety Administration (2015) Critical reasons for crashes investigated in the national motor vehicle crash causation survey. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115 Accessed 4 October 2017

Nyholm S, Smids J (2016) The ethics of accident-algorithms for self-driving cars: an applied trolley problem? Ethical Theory Moral Pract 19:1275–89

Thomson J (1976) Killing, letting die and the trolley problem. Monist 59:204-17

Thomson J (1985) The trolley problem. Yale Law J 94:1395-1415

Windsor M (2015) Will your self-driving car be programmed to kill you if it means saving more strangers? Science Daily. https://www.sciencedaily.com/releases/2015/06/150615124719.htm Accessed 4 October 2017

World Health Organization (2015) Global status report on road safety 2015. http://www.who.int/violence_injury_prevention/road_safety_status/2015/en Accessed 4 October 2017

Worstall T (2014) When should your driverless car from google be allowed to kill you? Forbes. https://www.forbes.com/sites/timworstall/2014/06/18/when-should-your-driverless-car-from-google-be-allowed-to-kill-you Accessed 4 October 2017