# Who Should Bear the Risk When Self-Driving Vehicles Crash?

Antti Kauppinen

University of Helsinki

Forthcoming in *The Journal of Applied Philosophy*

As the adage goes, accidents will happen. When vehicles traveling at speed are involved, there's a significant risk that people will suffer harm if they get in the way. This gives rise to important ethical questions if it is possible to influence who must bear the risk. These questions take on a particular urgency in the context of designing and programming self-driving vehicles guided by artificial intelligence, since this makes it possible to implement in advance ethical principles for distributing risk, and not only retrospectively assess decisions made in the heat of the moment.

In this paper, I'm going to argue, first, that when an accident becomes inevitable, we must take into account not only the magnitude and probability of harm that is likely to result from alternative trajectories, but also the degree of moral responsibility that people bear for creating the risky situation. As lessons from recent work on self-defense suggest, people who voluntarily engage in risky behavior can become morally liable to harm in proportion to their degree of moral responsibility, so that other things being equal, they should suffer the lion's share of the bad consequences of their actions instead of those who get in harm's way merely because of bad brute luck.

Second, I'll argue that this thesis about the just distribution of accidental harm has significant implications for how self-driving vehicles should be programmed to behave in accident situations, and indeed for the moral permissibility of using them. In particular, while there are scenarios in which vehicles should be programmed to minimize total expected harm, morally optimal behavior is different when the risky situation results from reckless or

culpable behavior of some of those potentially affected. (Unlike earlier discussions of moral responsibility in this context, such as Hevelke and Nida-Rümelin 2015, my focus is thus on the role that responsibility for risk plays in determining *what the right action is*, not on who is to *blame* for wrongful harm.) In situations in which questions about liability arise, humans with their fallible capacity to make quick, context-sensitive judgments about responsibility have an advantage over AI systems when it comes to accessing morally relevant information. This yields a pro tanto reason to restrict AI guidance in environments in which significant interaction with humans, whether drivers or pedestrians, can be expected. We must balance the likely reduction in overall risk of harm against a likely increase of unjust harm in certain accident situations. I'll also argue that taking the moral importance of liability into account entails that by default, designers of self-driving vehicles should program them so as to ensure that a higher chance of harm falls on the user rather than on an innocent outsider.

## 1. Automated Vehicles and the Problem of Distributing Risks

At the time of this writing, numerous companies are developing vehicles guided by artificial intelligence drawing on real-time data from various sensors, and consequently capable of navigating in traffic without direct human input (henceforth, self-driving vehicles). Automating the transport of people and goods promises many benefits beyond avoiding the hassle of commuting. Among them is increased safety as a result of avoiding the numerous accidents that currently result from human error or vice. According to some optimistic estimates, this could mean saving over a million lives worldwide each year (Howard 2013).

However, AI-guided vehicles are nevertheless bound to get into accidents when in widespread use. Propelling an object weighing several tons at high speeds on roads that are shared with other drivers, cyclists, and pedestrians will always pose a potential threat to people (Goodall 2014). Important ethical questions arise when a risk of some kind of harm to

someone is unavoidable, and the vehicle still has the chance to react in a way that affects the type of harm or who must bear it. The term 'crash optimization' is sometimes used for this programming task (Jenkins 2017), which is an important part of the broader project of managing the risk of such new technologies (Goodall 2016).

It has been popular to liken such situations to moral dilemmas that arise in so-called trolley cases (e.g. Wallach and Allen 2008, 14), in which an agent must choose between either allowing a runaway trolley to kill many people or intervening one way or another to bring about the death of one.[1] However, while there are certainly parallels, there are also important disanalogies. First, in trolley cases the outcome of each choice is stipulated to be certain, while in real life the outcome of a choice is virtually always uncertain (Nyholm and Smids 2015; Goodall 2016). To highlight this, I will use the language of posing a *risk* rather than causing harm in what follows. Second, as Nyholm and Smids (2015) emphasize, trolley cases involve an agent's response in a once-off real-time situation, while the question for autonomous cars concerns pre-set general algorithms, so that agency is exercised in advance of any emergency situation, and by those who program the car to react in a certain way. Relatedly, Johannes Himmelreich (2018) observes that while trolley cases concern individual morality, regulating autonomous vehicles requires political choices. Finally, and most importantly for my purposes, in trolley scenarios each potential victim is entirely innocent in the sense that by stipulation, they neither deserve nor are liable to be killed. As I'll argue, the fact that this is not always the case in unavoidable crash situations can make a major moral difference.

Given the disanalogies between real-life accidents and trolley cases, I believe it is better to start thinking about crash optimization from a more general debate in normative

[1] These scenarios were introduced by Philippa Foot and Judith Thomson to investigate the relevance of different causal relations between agents and victims to moral permissibility. There are very many variations (see Kamm 2015).

ethics. This is the question of *how unavoidable harms should be distributed*. The simplest view is a quasi-utilitarian principle of impersonally minimizing total expected harm. But this idea has well-known problematic consequences. I will assume that there are situations in which we shouldn't impose serious harm to one person in order to avoid even a much greater sum of minor harms to many people (Scanlon 1998). For example, if a driver must choose between an equal risk of permanently damaging a cyclist's arm, on the one hand, and causing a rock slide that will destroy a thousand people's parked cars without physical damage to people, on the other, they should let the rocks roll.

To accommodate the idea that there are limitations on aggregating harm, here is a better, if still rough version of a risk-minimizing principle inspired by Alex Voorhoeve's (2014) work:

> *Minimize Relevant Harms (MRH)*
>
> Other things being equal, minimize the sum of morally relevant expected harms. Harms are morally relevant in a choice situation only when they ground moral claims that are sufficiently strong compared to competing harm-based claims.

MRH says that it's wrong for the driver to damage the cyclist's arm, since the harm-based claims of the car-owners are not relevant, because they are not sufficiently strong relative to the cyclist's claim not to suffer significant bodily harm. Applying MRH will require non-trivial judgments about the comparative strength of harm-based claims (including categorization of certain harms as roughly equally serious), but I will not attempt to make it more precise, since my focus in this paper is largely on cases in which the principle doesn't apply.

## 2. Rights, Liability, and the Just Distribution of Harm

Commonsense morality is resolutely non-utilitarian, and holds that other things are often not equal in the sense required for MRH to apply. In particular, it's sometimes not morally permissible to act in ways that would minimize relevant harms when doing so would violate someone's *rights*.[2] In this paper, I'm going to assume that this is correct, and work out some of the consequences for just distribution of harm. The role of rights in ethics is precisely to protect people against being used to minimize harm or maximize benefit. To be sure, sometimes we find ourselves in situations in which all available options involve transgressing equally stringent rights of different people. In such cases, the idea of minimizing harm is plausible even in the context of infringing rights. For example, if a driver must choose between driving over one innocent person or three innocent people, the right thing to do is surely to minimize harmful infringements of rights by choosing the lesser evil of steering towards the one.

However, as long as people have rights, there are also possible situations in which it is not permissible to minimize harm. In the cases that are pertinent here, this is because someone has an intact right not to be harmed, and the only way to avoid violating it causes (or risks) greater or equal harm to someone who has lost their right not to be harmed that way. For example, if three robbers are trying to kill one innocent person to steal her wallet, it is morally permissible to kill all of them if necessary to save the one (even if they would afterwards become upright citizens), because they have forfeited their right not to be harmed, while the innocent person hasn't. Following Jeff McMahan (2005; 2009) and others (e.g. Frowe 2015), I'll say that when a person *lacks a right* against being harmed in a particular way due to something she has done, she is *liable* to such harm, and isn't *wronged* if she is so harmed. Justly convicted criminals are harmed by being imprisoned, but they are not wronged thereby. It is important to bear in mind, however, that one may lose a right against harm

[2] In the following, I will work with nonconsequentialist language, but I believe everything I say could be accommodated by sophisticated forms of consequentialism (see Portmore 2011).

without *deserving* harm, or without being *blameworthy*. Further, people's rights are not the only morally relevant considerations, so it may sometimes be all-things-considered permissible to harm someone who *isn't* liable (for example, in order to avert a catastrophe).

What makes someone liable to harm, then? In the next section, I will develop an account of liability in accident situations. My inspiration will be Jeff McMahan's well-known Responsibility Account of liability to defensive harm (although I won't assume its truth). McMahan's starting point is dissatisfaction with Judith Thomson's (1991) account of self-defense, according to which, roughly, someone is liable to harm if they will otherwise violate a right that is stringent enough for the harm to be proportionate, even if they're not responsible for their action or are fully excused. But many find it implausible that *nonresponsible* threats, who have not done anything to lose their right not to be harmed (like someone pushed off a cliff), could be liable to harm. Similarly, while *excused* threats have indeed engaged in action that ends up posing a threat to another, some of them do so in virtue of some causal link they couldn't have anticipated. For example, if a terrorist has unbeknownst to me rigged up the light switch in my bathroom so that flipping it causes a bomb to explode somewhere, my action does pose a threat to someone, but I have a full excuse for it, so it's not plausible that I have lost my right against harm.

According to McMahan, the reason why such nonresponsible and excused threats are not liable to harm is that they are not *morally responsible* for the unjust threat they pose. On his view, then, the criterion for liability to defensive harm is "moral responsibility, through action that lacks objective justification, for a threat of unjust harm to others" (2005, 394). Here an agent lacks *objective justification*, when there is, as a matter of fact, sufficient reason for her not to perform the act, whether or not she is aware of this. If you believe on good but misleading evidence that someone is trying to kill you, you have subjective justification to harm them if necessary, but no objective justification. And the threat of harm is *unjust* (or

wrongful) if the victim has an intact right not to be harmed and there is no sufficient reason to override it.

The most interesting condition of defensive liability for my purposes is *moral responsibility*. According to McMahan, to be morally responsible for a threat, it is necessary to have voluntarily engaged in action that foreseeably poses a risk to others. This rules out both the man pushed off a cliff (who doesn't act voluntarily) and the person innocently turning on the light that has been rigged (who isn't in a position to know what they're doing). However, it's worth emphasizing with Helen Frowe (2015, 73-76) that meeting these conditions doesn't suffice for minimal moral responsibility, since responsibility depends also on what the *alternatives* are – if any other option would be unreasonably costly to the agent or others, she's not responsible for voluntarily posing a threat. Whether a cost is unreasonable depends on the degree of prospective harm to the agent and victim, respectively, as well as possibly the causal role of the agent in bringing it about (ibid., 74).

Importantly, on this picture, responsibility for a threat does *not* require bad intent. Consider McMahan's most controversial example:

> *Conscientious Driver*
>
> A person keeps his car well maintained and always drives cautiously and alertly. On one occasion, however, freak circumstances cause the car to go out of control. It has veered in the direction of a pedestrian whom it will kill unless she blows it up by using one of the explosive devices with which pedestrians in philosophical examples are typically equipped. (McMahan 2005, 393)

McMahan holds that the driver, in spite of his precautions, is liable to being killed in self-defense, because "he voluntarily engaged in a risk-imposing activity and is responsible for the consequences when the risks he imposed eventuate in harms" (2005, 394). It's important to

bear in mind here the distinction between liability and blameworthiness: no one claims that the conscientious driver deserves some kind of moral criticism. It is also important to emphasize, as McMahan has done in his later work, that liability is a matter of degree, which makes a difference to what kind of harm is proportional to the threat. The fact that the driver is conscientious is a partial excuse, so his degree of liability is low. This means that if there is a way to divide the burden or risk of harm, it is morally optimal for the innocent victim to share some of it (McMahan 2009, 161). For example, if the pedestrian could somehow save her life also by redirecting the conscientious driver's car so that he crashes into a tree and loses a leg, she should do so, even if it meant that she lost a finger herself. Further, the driver will be fully excused if he is, for example, providing the only means of transporting a heart attack victim to a hospital in time, since refraining from driving would impose an unreasonable cost on others. In that case, he is plausibly not liable to defensive harm at all.

The deeper rationale for linking liability to moral responsibility is the thought that it is *unfair* if an innocent party has to suffer harm when it could fall instead on the person who brought about the wrongful threat. As Jonathan Quong (2012) emphasizes, the issue is one of a kind of local distributive fairness: if someone engages in behavior that can foreseeably cause harm to someone, and this risk eventuates, then if it is possible to make a difference to who will bear the cost, it's fair that the harm should be distributed on the shoulders of the agent who created the risk in the first place. This idea is reminiscent of luck egalitarian views in political philosophy, according to which, very roughly, the distribution of benefits and harms should not reflect people's unchosen circumstances, but should be sensitive to the choices that they're responsible for (Dworkin 2000). Kerah Gordon-Solmon links these ideas as follows:

> The relevant egalitarian intuition says that there is something unfair about the (full) costs of one person's choices being offloaded onto an innocent 3rd party. The

> parallel intuition, in the context of preventive justice, says that there is likewise something unfair about the (full) harm resulting from one person's risk-imposing activity being offloaded onto an innocent 3rd party. (Gordon-Solmon 2018, 551)

Even if the conscientious driver does not *wrong* the unlucky pedestrian threatened by his malfunctioning car, as Quong (2012) argues, he has by his voluntary choice brought it about that a cost must fall on someone, knowing that there was a non-negligible chance that this would happen. And this does indeed seem to make for a moral asymmetry between the two that is relevant to how the risk should be distributed, even if no one is wronged by the act.

**3. Responsibility and Liability to Harm in Accident Situations**

I started with the natural thought that other things being equal, self-driving vehicles should be programmed to react so as to minimize the sum of relevant expected harms. But I'll argue that people's actions can make them *liable to accidental harm* when they are responsible for creating a situation in which risk has to be distributed among non-threats, in which case it may be wrong to minimize relevant harms. To begin with, consider the following case:

> *Reckless*
>
> Tom, Dick, and Harry are on their way back home after a great party at another frat house. They're in high spirits in more than one sense. They decide to race each other back to the house, and all three burst into a run. Alas, in the course of doing so, they neglect to pay any attention to traffic, and appear all of a sudden in front of Ricky's truck, having jumped over a wall. Given the men's appearance and the time of the year, Ricky realizes what's going on, but it's too late to brake. Since there's a wall on the left side of the one-way street, the only way she can avoid hitting the three men is steering the truck to the footpath on the right, where Sven is walking back to

work after his lunch. Ricky has to choose whether to risk the life of one or three people.

In Reckless, I believe, it would be wrong for Ricky to steer the truck towards Sven in order to save Tom, Dick, and Harry. Why? The evident reason is that Tom, Dick, and Harry are morally responsible for creating the risky situation in the first place. It is their recklessness that forces Ricky to make a choice. The cost for their bit of fun should not fall on Sven, who has done nothing to make himself liable to be placed in risk of injury or death. Tom, Dick, and Harry, in contrast, have made themselves liable for such risk. Again, when I say that Tom, Dick, and Harry are liable, I don't mean that they *deserve* to be placed in risk of being killed. It's not as if they're mass murderers, but just kids having fun. But it would not wrong them or violate their rights to make them bear the risk of harm that results from their reckless choice.

This view of liability to accidental harm is supported by its mesh with liability to defensive harm. Were Ricky to steer towards Sven, it would intuitively be permissible for Sven to save himself with a ray-gun that would pulverize Ricky and her truck, if it was the only way for him to rescue himself. It would also be permissible for Sven to save himself by deflecting the truck towards Tom, Dick, and Harry by using a special shield, if that was necessary for him to save himself. In contrast, it would intuitively *not* be permissible for Tom, Dick, or Harry to redirect the truck toward Sven to save themselves.

Note that although the view I defend here is indebted to responsibility accounts of liability to *defensive* harm, it is not identical with them. In Reckless, the agents who are liable to harm do not themselves pose a (significant) risk of harm to anyone. Consequently, the issue isn't whether they are liable to be *defensively* harmed. What they do is rather make it the case that *someone* is faced with a risk. If Tom, Dick, and Harry didn't jump in front of Ricky's

truck, there would be no risk to distribute. Some might argue that this difference suffices to undermine the case for liability. But the luck egalitarian ideas inspiring the Responsibility Account are not in any way restricted to the context of self-defense – indeed, if anything, they fit more easily in the context of distributing inevitable harms. Recall the line from Gordon-Solmon: there is something unfair about the harm resulting from one person's risk-imposing activity being offloaded onto an innocent third party. That is exactly what would happen if Ricky steered toward Sven.

This suggestion relies on distinguishing between different kinds of luck (Dworkin 2000, 73). If Tom, Dick, and Harry get hit by a truck, they have bad *option luck*: they took a gamble, and lost it. If Sven gets hit by a truck while engaged in activity that can't reasonably be expected to be risky, he has bad *brute luck*. Of course, someone might object that walking on a sidewalk is a risky choice, since it is after all always possible that a vehicle will veer off the road. But then again, it is always possible that a meteor will crash into your house. Nothing we can do is completely risk-free. Insofar as there is a difference between option luck and brute luck, as there seems to be (even if it's not always clear-cut), it is morally relevant in the context of distributing harm. Indeed, considering the following variant of Reckless should make it clear that it matters whether the bad luck is option luck or brute luck:

> *Lost Runners*
>
> Abe, Ben, and Constantine are competing in the Stockholm marathon. They believe correctly and with excellent justification that the route is shut off from traffic. Alas, because of misleading signage, they take a wrong turn without any way of realizing it. Believing with justification that they are on the correct route, they pay little attention to traffic around them, and appear all of a sudden in front of Billy's car, having jumped over a traffic barrier. Given the way they're dressed and his knowledge of the marathon, Billy realizes what's going on, but it's too late to brake.

> Since there's a barrier on the left side of the one-way street, the only way he can avoid hitting the three men is steering the car to the right, where Ned is walking back to work after his lunch. Billy has to choose whether to risk the life of one or three people.

Here, again, Abe, Ben, and Constantine are *causally* responsible for creating a risky situation. But they could not reasonably anticipate doing so by participating in a supposedly well-organized race. Their bad luck is brute. Their position is similar to the person who flicks a light switch that, unbeknownst to them, triggers a bomb. So they are not *morally* responsible for the need to distribute risk. That's why their rights are intact, and their moral situation with Ned is symmetrical (it wouldn't, for example, be right for Ned to deflect the car towards the three). Thus, steering towards Ned would minimize infringement of equally stringent rights.

What exactly does it take to be sufficiently morally responsible for creating a situation in which someone has to bear a risk? My two cases highlight the *epistemic* condition: one must be in a position to know that one's action creates a non-negligible risk of harm. Whether a risk is negligible depends both on its probability and the magnitude of the potential harm – even a minute risk of a nuclear meltdown, say, is non-negligible. But as Frowe emphasized, we should also include a *reasonable avoidability* condition: there must be something else the agent could have done instead (like sleep it off) without a disproportionately high cost. And the action must be (sufficiently) *voluntary*, however exactly that is cashed out (I'm not trying to give a general theory of moral responsibility here). I'll summarize these conditions in the following:

> *Minimal Moral Responsibility for Risky Situations*
>
> A is at least minimally morally responsible for creating, by F-ing, a situation S in which risk of harm H is imposed on someone if a) she was in a position to know that

F-ing would under the circumstances create a non-negligible risk of a situation like S (that is, the risk of a risky situation), b) she could have performed a different action without an unreasonable cost to herself or others, and c) she voluntarily F-d.

The Lost Runners, unlike the Reckless, don't meet condition (a). If Tom, Dick and Harry were running around heedlessly because they were fleeing from a terrorist attack, they wouldn't meet condition (b), and would also be excused.

The reasons I've given in support of intuitive verdicts about accident cases can be summed up in the following principle linking liability to accidental harm and moral responsibility in the above sense:

*The Responsibility Account of Liability to Accidental Harm (RALAH)*

Agent A is liable to harm in the case of an inevitable accident if and only if A is at least minimally morally responsible for creating a situation in which someone must suffer (risk of) harm, and it is possible to shift the risk of harm to A.

I focus here on risk of creating a risky situation, since that is what's relevant in the traffic cases. Take Reckless, for example. Tom, Dick, and Harry don't directly pose a risk to Sven by running on the road. But they have nevertheless created a situation in which someone risks being harmed, and are at least minimally morally responsible for doing so. By RALAH, they're consequently liable to harm.

Again, to say that the agent is liable to harm H is to say that she lacks a right not to be harmed that way. So as such, RALAH doesn't say anything about what is morally permissible or obligatory. But with the notion of liability, we can formulate a principle for distributing risk:

*Required Risk-Shifting (RRS)*

> Presumptively, an agent who is capable of doing so is morally required to shift a risk of harm from a non-liable party to a liable party (ideally, in a way that reflects the degree of liability), and not to shift the risk from a liable to a non-liable party.

I formulate RRS in presumptive terms, since it is possible that other morally relevant considerations override liability – for example, given special obligations, you might not be morally required to shift the risk towards your own child, even if they're liable to it.

I'll call the combination of RALAH and RRS the Responsibility-Focused Account of Accidental Harm (for short, 'the responsibility-focused account') to distinguish it from related views of defensive harm. It (unsurprisingly) entails the intuitive verdicts for the two cases I've discussed: if possible, risk must be directed to the runners in Reckless, but not to the runners in Lost Runners. The principles also have seemingly correct implications for other kinds of accident situation. Consider, for example, Pushers, in which two people push another in the path of a car in response to a perceived insult. In the absence of special circumstances, the two are liable according to RALAH, and RRS requires the driver to shift the risk to them, if possible, instead of minimizing harm by hitting the one person on the road.

In addition to appealing to considerations that seem to make a normative difference in some paradigmatic cases, the responsibility-focused account could be defended within at least some leading moral frameworks, such as contractualism and rule consequentialism. While I lack the space for a full discussion here, I want to make a brief suggestion about how a contractualist justification might go.[3] For Scanlon (1998), an act is permissible roughly when and because it is allowed by a principle for the general regulation of behavior that is justifiable to others – when no one could reasonably reject it while seeking a mutually

[3] A rule-consequentialist story (cf. Hooker 2000) would hold that were people to internalize and transmit a moral code that ignored responsibility for risk in distributing accidental harm, there would be less of an incentive to avoid creating risky situations, which would lead to greater overall harm than internalizing a responsibility-focused account.

acceptable arrangement. This formula focuses attention on objections that could be made to proposed principles from individual perspectives without aggregating across individuals. It is important that the grounds for reasonable rejection are not limited to effects that permitting certain act types have on welfare. Most relevantly for my purposes, in his discussion of the value of choice, Scanlon emphasizes that the strength of a person's objection to permitting a potentially harmful policy depends on whether it allows them a *choice* to avoid the risk – for example, even if a person is in fact harmed as a result of transporting hazardous material to a safe place, they may lack a complaint if they received ample warning and had a genuine opportunity to avoid harm (Scanlon 1998, 256-260).

Compare, then, possible principles for distribution of accidental harm that are responsibility-focused with those that are not. *Prima facie*, anyone has an objection to a principle that would allow directing the risk of great harm to them. But a person who is morally responsible for the existence of a risky situation has a much weaker complaint than those who aren't. After all, they could have chosen, without unreasonable cost, to do something that wouldn't have placed anyone in harm's way. In contrast, those who have chosen to act in a way that doesn't pose a (non-negligible) risk to others have a strong complaint against having to bear the cost of the choices of others. So it seems that insofar as choice has value, principles for distributing accidental harm that don't take responsibility into account could be reasonably rejected, while the kind I have defended couldn't be.

The responsibility-focused account can thus be defended on more theoretical grounds as well. Before moving on to its implications for self-driving vehicles, I want to consider a special case, namely its implications for the *driver's* responsibility, as in the following case:

*Conscientious Driver With a Choice*

Sherry keeps her car well maintained and always drives cautiously and alertly. One Sunday when she's driving to check out a new lunch place, however, freak

> circumstances cause the car to go out of control and careen off the road towards Peter, who is sunbathing in a park by the road. Sherry realizes that she has only two options: hit Peter, or steer the car into a ravine at mortal risk to herself.

If RALAH is correct, Sherry is liable to accidental harm (assuming normal background conditions), while Peter is not. That's because she is at least minimally morally responsible for creating the risky situation, since she knows or should know that going for a drive involves a non-negligible risk of doing so, and she decided to do so, while it would not have been unreasonably costly to refrain.[4] Since she has taken steps to minimize the risk, however, her degree of responsibility, and consequently liability, is low. (Indeed, I think that given Sherry's conscientiousness, she may *not* be liable to *defensive* harm, contrary to McMahan.) RRS, then, calls for Peter to share some of the burden, if possible. However, by stipulation, it is not possible in this case. And although Sherry has a partial excuse, it is because of her choice that someone will now face mortal danger, as it turns out. So if other things are equal, RRS requires her to steer into the ravine.

But is Sherry morally required to sacrifice herself? Not necessarily. Most nonconsequentialists believe in agent-centered prerogatives or options, which allow for certain amount of privileging one's own interests over those of others (Scheffler 1994). Here the moral asymmetry between doing and allowing plays a role. It's common to believe that we have a prerogative to let someone die rather than risk our own life, while it's not permissible to kill a bystander to save ourselves. However, Quong (2009) argues that we also have a prerogative to *do harm* to a non-liable person who poses a mortal threat to us, because

[4] Is Peter also responsible for creating the risky situation, since it wouldn't exist without him being there, and he was in a position to know this might happen? No, because the risk of his sunbathing in a park forcing a choice between life and death was *negligible*, unlike the risk involved in operating a fairly large machine moving at speed in the vicinity of unprotected people, however carefully. (I thank a reviewer for this journal for raising this concern about symmetry.)

morality would require too much of us if we had to give up our own lives in such a situation. Sherry occupies an interesting intermediate position here, since she has set the threat in motion, but she hasn't directed it toward Peter. It could be argued that while it wouldn't be permitted for her to steer *towards* Peter if necessary to save herself, morality would be too demanding if she had to steer *away* at the expense of mortal risk (even though Peter himself poses no threat). The presumption of RRS would then be defeated in such situations.

I'll leave it open here whether Sherry has such a prerogative. Either way, it's worth considering what a neutral third party should do, if they had the chance to direct the car either way (say by remote control). The implication of my account is that *they* would be morally required to direct the car so as to impose the risk on Sherry, since she is the only party liable to accidental harm in the situation. There may thus be an asymmetry between what's permissible from first-person and third-person perspectives, which will be relevant in the following.

**4. Implications for the Programming and Use of Self-Driving Vehicles**

In the previous section, I considered what human drivers should do in various traffic accident scenarios, taking into account who is liable to harm. I'll now turn to what this means for the ethics of programming or using self-driving vehicles.

A preliminary question to address here is how crash optimization algorithms should relate to the ethical choices of a human driver in a corresponding situation. My assumption is that automated vehicles should be programmed to behave only in ways that would be permissible for human drivers. If it is not permissible for a human driver to crash into the Lost Runners, for example, then it is not permissible to program a self-driving vehicle to do so. (However, I will argue below that that there are some things it is permissible for a driver to do that it is not permissible to program a vehicle to do.)

If the arguments of the previous section are correct, vehicles must then be built and programmed to act as if they are guided by RRS and MRH, when applicable. If MRH were the only relevant principle, the task might be feasible with conceivable near future technology. We could start by categorizing injuries into, say, five different classes of seriousness.[5] Perhaps death is a Class 1 injury, serious permanent disability Class 2, and so on until harm to replaceable property in Class 5. With enough data, one could come up with probabilities for each kind of harm, given a particular type of collision (taking into account what part of a person or property is hit, at what speed, from what angle etc.). The next step would be calculating the probability of such collisions given each feasible trajectory of the vehicle in the situation (determined by the effectiveness of brakes, steering possibilities, trajectories of possible neighboring vehicles and potential targets, and so on). The task for the vehicle, then, would be to choose the trajectory that minimizes the probability of harm in a lexical ordering – first make sure no one is killed; if the probability of causing someone's death is negligible, minimize the total sum of disabilities caused; and so on.[6]

However, there is also RRS, which takes precedence over MRH when it applies. Programming a vehicle to abide by RRS requires checking of each potential victim if they are liable according to RALAH – roughly, are they responsible for acting in a way that foreseeably created the risky situation? And here it does not suffice to tell who is *causally* responsible, since one may be causally responsible without being liable, as is the case with the Lost Runners. Instead, in addition to knowledge of causal facts, sophisticated mind-reading capabilities are needed to tell whether someone has acted recklessly or intentionally created a

[5] As a reviewer noted, there are carefully constructed medical scales for assessing the severity of injuries, such as the Abbreviated Injury Scale (https://www.aaam.org/abbreviated-injury-scale-ais/).

[6] Jenkins (2017) develops a fairly detailed proposal along these lines, though without incorporating a relevance criterion. Himmelreich (2018) points out the complication that in mundane, low-stakes situations other values such as efficacy and environmental impact of driving algorithms are also ethically relevant, so that trade-offs with risk-minimization may be required.

risky situation. This is something that humans tend to be pretty good at (though of course fallible), as long as they have the right kind of information. We make such judgments every day, and they make a difference to our attitudes, even if not always behavior. We resent the person in the subway car who pushed us because he wasn't paying attention, but not the one who got caught off balance; we honk at the cyclist who speeds onto the road simply assuming we'll give way, but not the one who just didn't notice us. There's no reason to think these verdicts are systematically mistaken. For present and near future automated vehicles, however, the task of catching such subtle cues in real time appears to be quite impossible. So it doesn't seem to be feasible to program them to perform required risk-shifting.

I'll come back to the implications of this likely failure in the concluding section. But before that, I want to look at an interesting and controversial special case that doesn't hang on mind-reading technology: how to treat the *user* of an automated vehicle. For simplicity, let's focus on a self-driving car with a single user. It is a much debated issue whether the vehicle should give special weight to the safety of its user relative to non-users. If the user is liable to harm, this would be a mistake. To highlight the issue of liability, let's construct a case that is parallel to Conscientious Driver With a Choice:

> *Conscientious User*
>
> Jerry keeps his self-driving car well maintained and updated, and it has always driven him to his destination without harming anyone. On one occasion, however, freak circumstances cause the car to go out of control and careen off the road towards Janey, who is sunbathing in a park by the road. Jerry's car's computer calculates that there are only two realistic trajectories, one of which will result in hitting Janey, and the other crashing into a ravine with Jerry. The computer estimates that each trajectory involves high risk of Class 1 injury to a person.

What should Jerry's car be programmed to do in this kind of trade-off situation? Here is an argument to the effect that it should be programmed to steer into the ravine:

1. Even a conscientious user of an automated vehicle is at least minimally morally responsible for choosing to act in a way that she knows or should know may create a situation in which risk of serious harm is imposed on someone who is not liable to be harmed in that way (assuming alternatives are not unreasonably costly).
2. If an agent is at least minimally morally responsible for choosing to act in a way that may foreseeably create a situation in which someone has to bear a risk of serious harm, the agent is liable to harm to some degree. (Derived from RALAH.)
3. So, even a conscientious user of an automated vehicle is liable to harm (to a low degree), if the foreseeable risk resulting from her responsible choice is realized.
4. A person sunbathing in a recreational area is not liable to traffic-related harm.
5. If risk of harm to either a liable party or a non-liable party is inevitable, and other moral considerations (such as catastrophic consequences or special obligations) are absent, and a third party can direct the risk to either, the third party must direct the risk toward the liable party. (Derived from RRS.)
6. The programmer of an automated vehicle is and should be a third party with respect to the user and the sunbather.
7. So, the programmer of an automated vehicle must ensure that in a conflict situation, the vehicle will by default direct (a larger share of) the risk towards the user.

The most obvious way to reject this argument is to deny the first premise. For example, Hevelke and Nida-Rümelin (2015) argue that users of self-driving vehicles are not responsible for the risk that their vehicle poses. Their argument is that "the person whose autonomous vehicle crashes did not do anything different from any other user of autonomous cars; he was simply unlucky" (2015, 627), and that such bad moral luck doesn't render an agent blameworthy, given that they didn't do anything wrong. The problem with this argument is that while bad moral luck may well defeat *culpability*, it doesn't follow that it defeats *liability*. No one says that conscientious drivers or users are *blameworthy* if things go wrong. The claim is rather based on fair distribution of risk, and here it matters that the fact that someone has to bear the risk is a foreseeable consequence of the driver's actions. A conscientious user does not seem to me to be any less responsible for posing a risk to non-liable others than a conscientious ordinary driver. As Nyholm (2018) argues, it is also pertinent here that users of self-driving vehicles are akin to supervisors or managers, who share in the responsibility of those who work for them, even if they don't exercise direct agency. Such supervisory responsibility is arguably sufficient for liability in other contexts as well, so why not here?

The second point at which the argument might be questioned is the sixth premise. This is the crucial difference from the Conscientious Driver With a Choice case. In that case, I argued, Sherry may be morally permitted not to sacrifice herself, even though she is liable, if she has an agent-centered prerogative not to do so. But if Jerry's car is programmed by Patty, who has no special obligations towards either Jerry or Janey, RRS requires her to ensure that the unavoidable risk is borne by Jerry rather than Janey. Here, however, technology opens up an intriguing possibility. The arguments I've made earlier suggest that since Jerry's degree of liability is low, the burden of accidental harm should be shared between him and Janey, if possible. In the scenario as described, only one of them must bear the entire cost. But there is another way to get the vehicle's behavior to reflect the small difference in the liability of the

two potential victims: the car could be programmed to steer one way or another with a certain *probability*. Perhaps there should be, say, a 4/5 chance that Jerry's life is risked and 1/5 chance that Janey's is. This kind of randomization is not possible for a human driver, but it would be easy to implement by default in a self-driving vehicle.

Another way to reject the sixth premise is to hold that crash algorithms should be up to the *user* (Contissa et al. 2017). However, even if we allowed users to implement settings that result in supererogatory self-sacrifice, it wouldn't be morally acceptable to let the user ensure their own survival in accident situations.[7] What, after all, could ground such a prerogative to do harm, which it would amount to? I said that a Conscientious Driver with a Choice might be permitted to save her own life by allowing her car to hit a pedestrian, but not to steer towards one to avoid falling in the ravine. Nor is it morally too demanding to require conscientious users to accept a risk of injury or death in rare accident situations rather than program their vehicle to harm an innocent party. It's one thing to ask someone to accept an almost certain death to save a stranger and another to ask them to accept a minute chance of dying to spare another.

## 5. Conclusion

Automating all or much of traffic would almost certainly save lives with near-future technology. But it would not necessarily save the *right* lives, if what I've been arguing is along the right lines. Conversely, technologies that are not capable of responding to the occasionally subtle considerations that determine liability would end up violating many people's intact right against being harmed.

Many authors, such as Howard and Hevelke and Nida-Rümelin, argue that increased safety gives rise to a moral imperative to develop and introduce self-driving vehicles. At the

[7] This is compatible with giving users authority for low-stakes ethics settings, where there's only risk of minor harms (Millar 2017, 30).

same time, Hevelke and Nida-Rümelin themselves accept that "A violation of some person's fundamental rights cannot be legitimized on the basis of benefits for others, no matter how large" (2015, 622). If we took this absolutist stance, it would follow from what I've said that self-driving vehicles should not be allowed on the roads until they have the capability to distinguish between liable and non-liable parties and distribute risk accordingly. Of course, humans are also apt to make mistakes on these issues, so the baseline for introduction shouldn't be perfect compliance, but something close to human-level achievement. Nevertheless, taking the absolutist stance would likely result in a long delay in introducing automated vehicles – at the time of this writing, they occasionally struggle to identify whether a traffic light is green or red, and noticing stop signs, so we can hardly expect them to assess recklessness or culpability of agents.[8]

However, I don't want to draw such a dramatic conclusion. There are various kinds of compromise that we could make. On the moral side, we sometimes accept some amount of inevitable rights-violations as the price of reducing overall harm, so if the kind of situations I've discussed are sufficiently rare and the benefits of automatization sufficiently large, the moral risk might be acceptable. Nevertheless, what I want to emphasize is that we must take the increased risk of injury to non-liable parties into account, which has not been the case in public conversation so far. The second kind of possible compromise is a practical one: perhaps we should require human drivers to take control of automated vehicles in areas where risk of collision with other people is high and mind-reading capabilities are needed in addition to obeying traffic rules.[9] This compromise would probably mean that the total number of accidents is higher, but the use of human discriminatory capacities would save some innocent

[8] See thedrive.com, 'California Reports Highlight Autonomous Cars' Shortcomings' (http://www.thedrive.com/tech/20561/california-reports-highlight-autonomous-cars-shortcomings), accessed May 3, 2018.

[9] The general point that the complementary strengths of AI and human intelligence favor a developing a hybrid of the two is made by Mindell 2015. (I owe this reference to a reviewer for this journal.)

lives. Either way, issues of responsibility and liability to risk will have an important place in the ethics of automated vehicles and robots in general.[10]

**References**

Contissa, Giuseppe, Lagioia, Francesca, and Sartor, Giovanni 2017. The Ethical Knob: Ethically-Customisable Automated Vehicles and the Law. *Artificial Intelligence and Law* 25 (3): 365–378.

Dworkin, Ronald 2000. *Sovereign Virtue*, Cambridge MA: Harvard University Press.

Frowe, Helen 2015. *Defensive Killing*. Oxford: Oxford University Press.

Goodall, Noah 2014. Machine Ethics and Automated Vehicles. In G. Meyer and S. Beiker (eds.), *Road Vehicle Automation*. Dordrecht: Springer, 93–102.

Goodall, Noah 2016. Away from Trolleys and Toward Risk-Management. *Applied Artificial Intelligence* 30(8): 810–821.

Gordon-Solmon, Kerah 2018. What Makes a Person Liable to Defensive Harm? *Philosophy and Phenomenological Research* 97 (3): 543–567.

Hevelke, Alexander and Nida-Rümelin, Julian 2015. Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. *Science and Engineering Ethics* 21 (3): 619–630.

Himmelreich, Johannes 2018. Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations. *Ethical Theory and Moral Practice* 21: 669–684.

Howard, Don 2013. Robots on the Road: The Moral Imperative of the Driverless Car. *Science Matters*. Online at http://donhoward-blog.nd.edu/2013/11/07/robots-on-the-road-the-moral-imperative-of-the-driverlesscar/#.U1oq-1ffKZ1. Accessed June 28, 2018.

[10] I want to thank Raul Hakli, Pekka Mäkelä, Lilian O'Brien, and Johan Sandelin for helpful discussions and comments. I owe a special debt to Sven Nyholm, whose critical but encouraging comments on my original sketch significantly shaped the paper.

Jenkins, Ryan 2017. The Need for Moral Algorithms in Autonomous Vehicles. In Philip Otto and Eike Gräf (eds.), *3TH1CS*. iRIGHTS Media.

Kamm, Frances 2015. *The Trolley Problem Mysteries*. New York: Oxford University Press.

McMahan, Jeff 2005. The Basis of Moral Liability to Defensive Killing. *Philosophical Issues* 15: 386–405.

McMahan, Jeff 2009. *Killing in War*. Oxford: Oxford University Press.

Millar, Jason. 2017. Ethics Settings for Autonomous Vehicles. In Patrick Lin, Ryan Jenkins, and Keith Abney (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York: Oxford University Press, 20–34.

Mindell, David 2015. *Our Robots, Ourselves*. New York: Viking.

Nyholm, Sven and Smids, Johan 2016. The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem? *Ethical Theory and Moral Practice* 19(5): 1275–1289.

Nyholm, Sven 2018. Attributing Agency to Automated Systems: Reflections on Human-Robot Collaboration and Responsibility-Loci. *Science and Engineering Ethics* 24(4): 1201–1219.

Portmore, Douglas 2011. *Commonsense Consequentialism*. New York: Oxford University Press.

Quong, Jonathan 2009. Killing in Self-Defense. *Ethics* 119 (3): 507–537.

Quong, Jonathan 2012. Liability to Defensive Harm. *Philosophy and Public Affairs* 40 (1): 45–77.

Scanlon, Thomas 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Scheffler, Samuel 1994. *The Rejection of Consequentialism*. Rev ed. Oxford: Clarendon Press.

Thomson, Judith 1990. *The Realm of Rights*. Cambridge, MA: Harvard University Press.

Thomson, Judith 1991. Self-Defense. *Philosophy and Public Affairs* 20 (4): 283–310.

Voorhoeve, Alex 2014. How Should We Aggregate Competing Claims? *Ethics* 125 (1): 64–87.

Wallach, Wendell and Allen, Colin 2009. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.