Autonomous Vehicles, Business Ethics, and Risk Distribution in Hybrid Traffic

Brian Berkey

I. Introduction

It is widely agreed that the introduction of autonomous vehicles onto our roadways can be expected to significantly reduce the number of injuries and deaths resulting from vehicle accidents (Goodall 2014a, p. 58; Nyholm & Smids 2016, p. 1275; Etzioni & Etzioni 2017, p. 415; Fleetwood 2017, p. 532; Gogoll & Müller 2017, p. 682; Gurney 2017, p. 51; Kelley 2017, p. 167; Bergmann et. al. 2018, p. 1; Hübner & White 2018, p. 685; Wolkenstein 2018, p. 163, 165, 167; Faulhaber et. al. 2019, p. 400, 412). To the extent that autonomous vehicles will in fact reduce injuries and deaths on the road, there are strong reasons to favor replacing human-driven vehicles with autonomous ones. The introduction of autonomous vehicles, however, is likely to be a gradual process, with the replacing of human-driven vehicles occurring over the course of many years (van Loon & Martens 2015; Nyholm & Smids forthcoming). We can, therefore, reasonably predict that there will be a period of time in which autonomous vehicles share the road with human-driven vehicles, that is, a period that will be characterized by *hybrid traffic* (Goodall 2014a, p. 59; Hübner & White 2018, p. 686; Nyholm & Smids forthcoming). Because of this, firms that produce autonomous vehicles will have to make decisions about how to program those vehicles to behave in potential conflict situations involving human-driven vehicles.[1]

[1] It is important to note that these decisions will almost certainly not take the form of discrete choices regarding programming for circumstances with highly specific sets of features. This is because autonomous vehicles are being designed such that machine learning mechanisms will determine how they will come to behave in new types of conflict situations (as well as more generally). This complicates how we must think about the ethics of the relevant programming decisions somewhat, although in my view it does not affect the fundamental normative structure of these decisions. This is because whatever values would rightly guide the direct programming of autonomous vehicles were they to be designed in a way that made that possible ought, as much as possible, to guide the

My aim in this chapter is to examine the morally relevant dimensions of these programming choices, and to argue that, in addition to the generally accepted aim of reducing traffic-related injuries and deaths as much as possible, a principle of fairness in the distribution of risk should inform our thinking about how firms that produce autonomous vehicles ought to program them to respond in conflict situations involving human-driven vehicles.[2] This principle, I will claim, rules out programming autonomous vehicles to systematically prioritize the interests of their occupants over those of the occupants of other vehicles, including human-driven vehicles. Given that a recent academic study (Bonnefon, Shariff, & Rahwan 2016) indicates that most consumers would prefer to purchase autonomous vehicles that do systematically prioritize the interests of occupants over those of others,[3] if my argument is correct it generates a substantial ethical restriction on firms' efforts to gain market share in the initial stages of the introduction of autonomous vehicles onto the road.

Complying with this restriction would require decision-makers at firms producing autonomous vehicles to refrain from satisfying some of the preferences that potential consumers have regarding the design of those vehicles, and to do so despite the fact that this might be expected to reduce the profits that the vehicles will generate. There are reasons to be skeptical that firms will actually comply with any ethical restriction of this kind, since their decision-making structures are typically designed to respond to consumer preferences in order to maximize profits. In many cases it is unproblematic that firms produce products that aim to satisfy consumer preferences as much as possible. In the case of autonomous vehicles, however,

---

programming of the relevant machine learning mechanisms as well. I am grateful to John Basl and Jeff Behrends for a helpful discussion of this issue.

[2] This principle is also relevant to programming choices involving other kinds of conflicts, for example those with pedestrians or cyclists, although for reasons that I will suggest briefly in section III, its implications regarding how much risk each party should bear plausibly differ significantly.

[3] This study was limited to U.S. residents (Bonnefon, Shariff, & Rahwan, p. 1573). It is, of course, possible that the results would differ at least somewhat in different societies. For the results of surveys conducted across the globe on a range of questions about autonomous vehicle ethics, including cross-cultural variation, see Awad et. al. (2018).

both the potential impacts of their use on others, including those who will continue to be occupants of human-driven vehicles, and the difference in increased safety that would be achieved for everyone if autonomous vehicles are, in the long run, programmed to minimize harm rather than always protecting occupants, provide strong reasons to think that firms ought to comply with the restriction that I will defend.

I will proceed in the remainder of the chapter as follows. First, in section II, I will describe some of the ways in which the interests of occupants of autonomous vehicles might come into conflict with the interests of occupants of human-driven vehicles. This will involve considering particular cases that have roughly the structure of "trolley" cases,[4] but also examining the ways in which different candidate approaches to programming autonomous vehicles to behave in the relevant conflict situations would be likely to systematically affect the interests of different groups of people, including in particular the types and degrees of risk to which they would be subjected on the road. Next, in Section III, I will argue that, as an ethical

[4] There has been a fair bit of debate about the usefulness of trolley-style cases for thinking about some of the ethical issues raised by the development and introduction onto the road of autonomous vehicles. In my view, the more critical discussions tend to overstate the reasons to be concerned about the ways in which such cases have been employed by some contributors to the discussion of autonomous vehicle ethics, and to neglect the ways in which reflecting on the cases can helpfully inform our thinking about the ethical programming of autonomous vehicles, even granting the claim, which many critics emphasize (e.g. Nyholm & Smids 2016, pp. 1280-1286), that the most widely discussed types of trolley cases are quite far from being analogous to autonomous vehicles programming choices in all relevant respects. One of the disanalogies that Nyholm & Smids point to is that in typical trolley cases, facts about what will happen if an agent chooses each option are stipulated and assumed to be certain, whereas in cases involving autonomous vehicles the outcomes of the different available courses of action can only be estimated to in a probabilistic manner, so that issues of risk and uncertainty must be considered (Nyholm & Smids 2016, pp. 1284-1286). In my view, however, this is not a fundamental distinction, and does not make the use of trolley cases objectionable, for two reasons. First, it can sometimes be useful to consider what ought to be done in cases in which certainty about the outcomes of different actions is assumed before reflecting on what ought to be done in cases that are similar in many respects but also involve risk and/or uncertainty. And second, it is not difficult to design trolley-style cases in a way that includes the dimensions of risk and/or uncertainty that will generally characterize cases involving autonomous vehicles. Indeed, the cases that I will discuss in section II stipulate facts about the risks that individuals would be subject to given different courses of action by an autonomous vehicle (see also, e.g., Keeling 2018, p. 425). For critical discussions of the use of trolley cases in discussions of autonomous vehicle ethics, see Coeckelbergh (2016, p. 749); Nyholm & Smids (2016); Etzioni & Etzioni (2017, pp. 415-416); Liu (2017, p. 200, 202); Himmelreich (2018); Davnall (forthcoming). For employment of and/or defense of employment of trolley cases, see Lin (2015, pp. 78-79); Leben (2017); Martin (2017, p. 951-952); Millar (2016 and 2017, p. 23); Santoni de Sio (2017); Hübner & White (2018); Keeling (forthcoming). More mixed assessments of the value of trolley cases for the ethics of autonomous vehicles can be found in Goodall (2016); Wolkenstein (2018); and Borenstein, Herkert, & Miller (2019, pp. 384, 386).

matter, the programming of autonomous vehicles for circumstances involving hybrid traffic ought to be guided, as much as possible, by a principle of fairness in the distribution of the unavoidable risks of the road. I will describe the central features of the kind of principle that I favor, and offer some reasons for thinking that it is the right kind of principle to guide autonomous vehicle programming for hybrid traffic. In particular, I will claim that it captures the importance of avoiding an outcome in which the benefits of increased safety that autonomous vehicles are likely to generate are enjoyed disproportionately by wealthier members of society, and the risks of the road are borne primarily by the less well off.[5] In section IV, I will consider two objections to the claim that the principle that I offer in section III ought to guide the programming of autonomous vehicles for hybrid traffic, and defend the view that firms that produce autonomous vehicles ought to be guided by that principle against those objections. I will conclude, in section V, by noting the limitations of the principle's applicability to business decisions involving the programming of autonomous vehicles. In particular, I will argue that while every firm has an obligation not to be the first to offer autonomous vehicles programmed in a way that is inconsistent with my principle to consumers, and in particular obligated not to be the first to offer vehicles programmed to systematically prioritize the interests of occupants, if one firm in fact does so, others are at least most likely permitted to follow suit.

[5] It is already the case that, at least to some extent, wealthier people have access to more of the benefits of vehicle transportation, while those who are worse off bear a significant portion of the risks. Most would agree, however, that this is objectionable, and that these disparities ought to be remedied to the extent that this is possible. In addition, the introduction of autonomous vehicles that are programmed to systematically prioritize the interests of occupants would exacerbate these disparities, and so there are reasons to think that producing such vehicles would be objectionable even if the current distribution of the risks of the road could be defended.

II. Conflicts of Interests in Hybrid Traffic

Despite the fact that autonomous vehicles can be expected to substantially reduce the frequency of serious traffic accidents, circumstances will undoubtedly still arise in which injuries and even deaths on the road are unavoidable (Marchant & Lindor 2012; Goodall 2014a, pp. 58-59; Goodall 2014b; Hevelke & Nida-Rümelin 2015, p. 620; Lin 2015, pp. 71-72; Nyholm & Smids 2016, p. 1275; Fleetwood 2017, p. 534; Gogoll & Müller 2017, p. 682; Kelley 2017, p. 171, 179; Millar 2017, p. 22; Bergmann et. al. 2018, p. 2; Coca-Vila 2018, p. 60; Hübner & White 2018, p. 685-686; Nyholm 2018, p. 1; Borenstein, Herkert, & Miller 2019, pp. 388-389). There are good reasons to think that this will be the case even if, at some point in the future, all vehicles on the road are autonomous. Whether or not this is correct, however, it will certainly be the case during the period in which autonomous vehicles share the road with human-driven vehicles. And in many of the circumstances in which accidents involving, or potentially involving, both autonomous and human-driven vehicles are unavoidable, it is likely that the programming of the autonomous vehicles will play a significant role in determining exactly how the relevant accidents play out, and therefore who will suffer which resulting injuries and deaths.

Consider the following two cases:

> ***Cliffside Road***[6]: The driver of a standard (i.e. non-autonomous) bus traveling on a narrow and lightly traveled two-lane cliffside road swerves from the cliffside lane into the inner lane in order to avoid an animal in the road. An autonomous car traveling in the inner lane comes around a sharp curve and recognizes that the bus is in its lane just ahead. Based on data on the behavior of human drivers that is available to the autonomous system, it estimates that if the autonomous car continues in its lane, it is virtually certain that the bus driver will attempt to swerve back into cliffside lane; and assuming that the bus driver does attempt to swerve back, there is an approximately 90% probability[7] that he will lose control of the bus, and the bus will go over the cliff, killing the driver and all 30 passengers. The autonomous car's only other option is to drive into the cliff wall,

[6] This case is based loosely on a case given by Patrick Lin (2015, pp. 76-77).
[7] This probability, and the others noted in this and the following case, are the estimates that the autonomous vehicle's system assigns on the basis of all of the relevant data available to it.

which would carry an approximately 10% risk of death for the vehicle's single occupant, and, conditional on survival, a 50% risk of serious injury. If the autonomous vehicle does this, there is an approximately 99% probability that the bus will continue on safely and avoid any injuries or deaths.

Country Road: The driver of a standard (i.e. non-autonomous) small car traveling with one passenger on a moderately traveled two-lane country road swerves from the northbound lane toward the southbound lane in order to avoid an animal in the road. The driver brakes hard, and the car turns partially and comes to an abrupt stop. Its front half is in the southbound lane, with the passenger side directly exposed to oncoming traffic, and its rear half is still in the northbound lane. An autonomous large sport utility vehicle (SUV) is traveling in the southbound lane, and does not have time to stop in its lane before hitting the front half of the stopped car ahead. The autonomous SUV has three options. First, it can continue in its lane and hit the stopped car, in which case the passenger in the stopped car will almost certainly be killed, and the driver of that car will very likely suffer a moderate to serious injury, but survive. The single occupant of the autonomous SUV would most likely suffer only a minor injury, but there is a small risk (approximately 10%) that he would suffer a moderately serious injury. Second, the autonomous SUV can swerve to the right, into a concrete wall, in which case its occupant would face an approximately 30% risk of suffering a serious injury, and an approximately 2% risk of death, while the occupants of the car would certainly be uninjured. Finally, the autonomous SUV can swerve across the northbound lane onto that lane's shoulder, in which case its passenger would have an approximately 90% chance of avoiding any injury, and an approximately 10% chance of suffering a minor injury. There is a large standard truck approaching in the northbound lane, and while the autonomous system calculates that there is very likely enough time for the autonomous SUV to cross that lane onto the shoulder without being hit even if the truck driver continues proceeding straight (there is a small chance that the truck will hit the rear of the SUV, in which case the occupant of the SUV might suffer a minor injury, and the driver of the truck would have an approximately 30% chance of suffering a moderately serious injury), it is fairly likely that the truck's driver will be caused by the SUV's swerving to swerve toward the southbound lane and hit the stopped car. If the autonomous SUV takes this option, there is an approximately 70% chance that the truck will hit the stopped car, and if it does there is an approximately 90% chance that both the driver and passenger of the car will be killed, and an approximately 50% chance that the driver of the truck will suffer a serious injury.

In order to recognize the distinctive risk distribution issues raised by the prospect of hybrid traffic scenarios, it is important to notice that if the other vehicles involved in these cases were also autonomous, their occupants could be subject to significantly lower risks of injury and death, while the occupants of the autonomous vehicles could be at greater risk. For example, if

the bus in *Cliffside Road* were autonomous, its system could recognize that its occupants' safety would be best protected by moving back into the cliffside lane more slowly than a human driver would likely attempt to move. This would, we can imagine, certainly result in a collision with the autonomous car, thus guaranteeing that its occupant is at least seriously injured, if not killed; but it would also ensure that the bus's occupants suffer at most very minor injuries. And in *Country Road*, if the large truck were autonomous, its system could recognize that swerving toward the southbound lane is the most dangerous option for its occupant. We can even imagine that the truck could best protect its occupant by swerving toward the shoulder, which would put the SUV's occupant at greater risk of serious injury were the SUV to attempt to swerve to the shoulder.

These cases are complex, but the general point that they help to highlight is, I think, fairly clear, and on reflection should not be surprising. This point is that because autonomous vehicle systems will have access to massive amounts of data that human drivers cannot employ in their necessarily split-second decision making in conflict situations on the road, and will be capable of using that data to determine what they will do, autonomous vehicles could, in principle, be programmed in ways that would make it the case that occupants of human-driven vehicles are systematically subject to greater risks of injury and death on the road than are occupants of autonomous vehicles. The autonomous car in *Cliffside Road*, for example, could be programmed in a way that will make it very likely that the bus will go over the cliff and kill all of the people onboard, despite the fact that it could instead have been programmed in a way that would ensure that its occupants are subjected to somewhat more risk when this is necessary in order to prevent a greater number of people in a human-driven vehicle from being subject to more extensive and more serious risks. Similarly, the autonomous SUV in *Country Road* could be programmed in a

way that will make it quite likely that the driver and passenger in the car will be killed, and that the driver of the truck will be seriously injured, despite the fact that it could instead have been programmed in a way that ensured that none of the occupants of the other vehicles are injured by allowing its own occupant to be subject to somewhat greater and more serious risks.

If autonomous vehicles are programmed in ways that systematically prioritize protecting their occupants from risks and harms as much as possible, then the result, in hybrid traffic conditions, will be that when circumstances arise involving both autonomous and human-driven vehicles in which an accident of some kind is unavoidable, occupants of human-driven vehicles will systematically suffer more harms, and more serious harms, than occupants of autonomous vehicles. This is because autonomous vehicles' programming will play a very large role in determining how the accidents play out, and programming those vehicles to systematically prioritize the interests of occupants would ensure that the risks of the road are distributed, as much as possible, away from those occupants, and therefore onto those in other vehicles.

This should concern us a great deal, since it seems very likely that, at least for a significant period of time, there will be a correlation between wealth and autonomous vehicle ownership and use. Like other new and heavily anticipated products the development of which requires large research and development costs, autonomous vehicles seem likely to be a luxury item, available primarily to wealthier people, at the time of their initial release on the market, and most likely for a while thereafter.[8] If this is in fact the case, and autonomous vehicles are

[8] Some analysts predict that the development of autonomous vehicles will dramatically reduce rates of vehicle ownership, and that users will instead typically purchase individual rides (Edelstein 2018; Huffman 2018; Standage 2018). Even if this is true, however, it seems likely that wealthier individuals might choose to purchase their own autonomous vehicle, whereas this option will be out of reach for the less well off. It also seems likely that, at least for a period of time, purchasing individual rides in autonomous vehicles for all of one's transportation needs will be more expensive than owning and using one's own standard vehicle. I do not think, then, that a trend away from vehicle ownership brought on, at least in part, by the release of autonomous vehicles onto the roads, would prevent the concerns about potential inequalities in the risks to which road users are subjected from arising. If, however, the most optimistic predictions about how much the introduction of autonomous vehicles will reduce the cost of

programmed to systematically prioritize the interests of their occupants in circumstances in which an accident is unavoidable, then the result will be that less well off individuals, who will mostly continue to drive standard vehicles, will systematically be at greater risk of injury and death on the road than the wealthier people who own and ride in autonomous vehicles.

These differences in the risks to which the wealthy and the less well off might be subjected could be much less substantial, if autonomous vehicles are programmed in ways that refrain from prioritizing the interests of their occupants so heavily. For example, they could be programmed in a way that would ensure that in cases like *Cliffside Road*, occupants are subjected to the risks associated with driving into the cliff wall, in order to protect all of the occupants of buses from very likely death. Similarly, they could be programmed in a way that would ensure that in cases like *Country Road*, occupants are subjected to the risks associated with driving into the concrete wall, in order to protect the occupants of human-driven vehicles from much greater risks of death and other serious injuries. These alternative ways of programming autonomous vehicles would at least limit the extent to which the very likely less well off occupants of human-driven vehicles are subject to greater and more serious risks on the road than are the likely better off occupants of autonomous vehicles.

In my view, it would clearly be objectionable if the benefits of increased safety that, as a general matter, autonomous vehicles are likely to provide end up drastically disproportionately protecting wealthier individuals rather than being distributed more evenly among all of those on the road. If this is correct, then it is important to think about what principle or principles ought to

---

purchasing individual rides, how quickly it will have this effect, and how quickly the purchase of rides in autonomous vehicles will become widely available are correct, then it is at least possible that the effects of the introduction of autonomous vehicles on the distribution of the risks of the road would be less unequal than I have suggested even if the vehicles are generally programmed to prioritize the interests of occupants.

guide the programming of autonomous vehicles, given that we can expect them to share the road with human-driven vehicles for at least a period of time.

III. The Fair Distribution of the Risks of the Road

The type of principle that I favor holds that, consistent with the aim of reducing traffic-related injuries and deaths as much as possible, firms ought to program autonomous vehicles in ways that aim to ensure that the safety-related benefits of those vehicles are distributed as fairly as possible among all of those who could potentially be harmed as a result of the use of motor vehicles.[9] This type of principle governs not only the ways in which the programming of autonomous vehicles can permissibly affect the distribution of risks among vehicle users, but also the risks that it is acceptable to impose on, for example, pedestrians and cyclists in cases in which their safety might come into conflict with the safety of vehicle occupants

What, then, would a fair distribution of the risks of the road look like? A reasonable starting point, it seems to me, is to think that these risks should be distributed as evenly as possible, consistent with the aim of minimizing the total amount of harm that will be caused by traffic-related accidents. Taking this as a starting point, we can then ask when it is either permissible or required for firms producing autonomous vehicles to deviate from aiming at an

[9] The aim of reducing injuries and deaths as much as possible should, on my view, take priority over distributing risks equally when there is a conflict. The principle that I defend, then, is not subject to the levelling down objection (for discussion see Parfit 2000, pp. 98-99, 110-115; Temkin 2000). Furthermore, in practice, the degree of unavoidable conflict between the aim of minimizing total injuries and deaths and the aim of distributing the risks of the road equally should not be especially large for firms programming autonomous vehicles. This is because the distribution of overall risks should be assessed from an ex ante perspective, prior to any knowledge about which conflict situations any particular people will actually find themselves in. And programming autonomous vehicles to minimize total harm in conflict situations will tend, in effect, to distribute ex ante overall risks as equally as possible. There will be exceptions, such as cases in which particular individuals regularly travel along routes that are such that minimizing total harm in conflict situations will tend to require autonomous vehicles to behave in ways more likely to harm them than others. But this will tend to occur only in unusual cases, such as when an individual regularly travels along a particular road only in one direction (i.e. the direction that is such that minimizing total harm will tend to require harming those traveling in that direction).

equal distribution of the risks. Here, then, is an initial formulation of a fair risk distribution principle that I find plausible:

> ***Fair Risk Distribution***: Autonomous vehicles ought to be programmed so that, to the greatest extent possible, the risks of the road are distributed equally among all of those who might be harmed as a result of the use of motor vehicles, unless there is a morally compelling reason for deviating from this aim.

Determining what, more specifically, a principle of this type should be thought to imply with regard to the programming of autonomous vehicles requires thinking about what, on reflection, might count as a morally compelling reason to deviate from aiming at an equal distribution of the risks of the road.

The clearest example of a compelling reason to deviate from aiming at an equal distribution of the risks of the road is that doing so is necessary in order to reduce the absolute level of risk that everyone is subject to. In addition, it seems to me that, at least in general, deviating from aiming at an equal distribution of risks will be justified when this is necessary to reduce the total amount of harm that can be expected to be caused by traffic accidents, even if not everyone would benefit (ex ante) from this deviation.

Furthermore, although I cannot defend this position in detail in this chapter, it seems to me that a potentially significant deviation from an equal distribution of risks between, on the one hand, occupants of vehicles, and on the other, pedestrians and cyclists, might be required. This is because while pedestrians do not expose others to any risks at all just in virtue of using the road (or sidewalk), and cyclists expose others only to relatively small and (on average) less serious risks, those who choose to ride in vehicles expose others to (much more) substantial risks of harm, including risks of serious injury and death, just in virtue of their use of vehicles. It seems plausible that those who choose to introduce risks of this kind in order to enjoy the benefits of

the activities that unavoidably involve these risks, should, where possible, at least bear a greater share of the risks than those who are not engaged in the activities that impose them. If this is correct, then it may be impermissible for autonomous vehicles to be programmed in ways that will lead them to, for example, swerve into a single pedestrian when this would risk causing her significant harm, even when this is the only way to protect multiple occupants from the risk of very serious harms.

What reasons might be offered in defense of the claim that deviation from aiming at an equal distribution of the risks of the road among road users can be justified? One argument that might be made begins by noting that autonomous vehicles will be significantly safer than human-driven vehicles. Because this is the case, anyone who transitions from driving a standard vehicle to using an autonomous vehicle will reduce the total amount of risk to which road users are subject. It might be claimed that their role in reducing the overall risks of the road entitles autonomous vehicle users to a greater share of the benefits of that risk reduction than those who continue to drive standard vehicles.

In my view, this argument would be compelling if everyone had at least roughly equal access to use of the safer alternative, and so could equally avoid imposing greater overall risks on road users. In a world, for example, in which purchasing and/or using an autonomous vehicle were no more expensive than purchasing and/or using a standard vehicle, users of autonomous vehicles would have a legitimate claim to have their vehicles programmed in ways that, at least to some extent, prioritize their safety over that of occupants of human-driven vehicles. In that world, those who choose to drive standard vehicles rather than using safer autonomous vehicles would be in a position analogous to those who choose to use a vehicle of any kind rather than walking or cycling. They would be imposing greater risks than they could have chosen to

impose, presumably in order to obtain what they perceive as some benefit for themselves, such as the enjoyment that they get from driving. Choices of this kind make it permissible for others to, where possible, ensure that the people who make them bear a greater share of the risks of the road than those who choose safer alternatives.

In a world, however, in which access to the safer alternative of autonomous vehicles is strongly correlated with wealth, it is not legitimate for those who are fortunate enough to have access to those vehicles to also insist that they benefit significantly more from the reduction in the overall risks of the road than those who simply cannot afford to switch to using them. And since it is likely that there will be substantial inequality in access to autonomous vehicles, at least for a period of time, we should reject the claim that those who use them during this period are entitled to an unequal share of the benefits of increased overall safety that they can be expected to provide.

In order to highlight more clearly why we should find the view that autonomous vehicles can permissibly be programmed to systematically prioritize the interests of their occupants over the interests of occupants of human-driven vehicles problematic, it will be helpful to consider what the aggregate effects of this type of programming might be with regard to the distribution of the harms that result from traffic accidents involving both autonomous and human-driven vehicles.[10] For reasons that I described briefly in the previous section, it seems clear that if autonomous vehicles are programmed to systematically prioritize the interests of their occupants, then accidents involving both autonomous vehicles and human-driven vehicles will, on average,

[10] One thing that is worth noting here is that occupants of human-driven vehicles will, in conditions of hybrid traffic, be subject to greater overall risks of injury and death than occupants of autonomous vehicles, even if autonomous vehicles are programmed in the way that my principle requires. This is because there will be many more accidents involving only human-driven vehicles than accidents involving only autonomous vehicles. I am inclined to think that this might also be objectionable, and that there might be obligations applying to, for example, governments, vehicle manufacturers, and even individuals, to promote equality in access to autonomous vehicles so that these inequalities in the risks to which individuals are subjected are at least more limited than they would otherwise be.

result in occupants of human-driven vehicles suffering more, and more serious, injuries than occupants of autonomous vehicles. In addition, autonomous vehicles will be, on average, more likely than human-driven vehicles to avoid being involved in accidents, and in some cases will play a role in bringing about more serious harms to occupants of human-driven vehicles than would have otherwise occurred, precisely by adopting courses of action that prevent them from being involved, strictly speaking, in an accident at all (this is, for example, the most likely outcome in *Country Road* if the autonomous SUV swerves across the northbound lane).

How large the inequalities in the risks to which occupants of autonomous vehicles and human-driven vehicles would be subject depends on how effective autonomous vehicles programmed to systematically prioritize the interests of occupants would be at distributing the risks of the road away from their occupants. Imagine, I think plausibly, that they would be quite effective – specifically, imagine that autonomous vehicles programmed to systematically prioritize the interests of occupants would, on average, be involved in 70% fewer accidents than human-driven vehicles (in part because they would sometimes avoid involvement by taking actions that will likely result in an accident involving only human-driven vehicles), and that when accidents did occur involving both autonomous and human-driven vehicles, occupants of human-driven vehicles would be killed five times more often than occupants of autonomous vehicles, and would be seriously injured three times more often. If these disparities are at least primarily explained by the fact that autonomous vehicles are programmed to systematically prioritize the interests of their occupants, when they could instead have been programmed to distribute the risks more equally (without thereby causing there to be more total injuries and deaths), the disparities seem intuitively objectionable. If we add to our description of the case that the average occupant of an autonomous vehicle has, say, a six-figure income, whereas the

average occupant of a human-driven vehicle has a below-average income, as well as that these differences in income are by far the most important factor explaining why people use the types of vehicles that they do, it should seem even clearer that the fact that occupants of human-driven vehicles are subject to much greater risks of injury and death is unacceptable.

In addition, it is important to note that, independent of the concerns about the distribution of risk that I have been focusing on, the aim of minimizing total harm itself counts strongly against the permissibility of programming autonomous cars to systematically prioritize the interests of their occupants as well. For example, in both *Cliffside Road* and *Country Road*, the autonomous vehicles would cause it to be likely that significantly more, and more serious harms would result if they are programmed to systematically prioritize the interests of their occupants, than would likely occur if they were instead programmed in a way that is consistent with distributing the risks of the road as equally as possible.

This should not be surprising. Programming autonomous vehicles to systematically prioritize the interests of occupants would result in those vehicles behaving, in many cases, in ways that cause greater harm to others than occupants would have suffered if they were programmed to minimize to minimize total harm. On the other hand, because it is generally not predictable ex ante which people will be situated where in the conflict situations that they will, or might, find themselves in on the road, programming autonomous vehicles to behave in ways that can be expected to minimize total harm will at least strongly tend, in effect, to distribute the (ex ante) risks of the road equally among the occupants of all vehicles.

On the basis of these considerations, we should, it seems to me, conclude that there is a strong initial case for thinking that autonomous vehicles should be programmed with the aim of generating an equal distribution of the risks of the road between users of those vehicles and users

of standard vehicles. Relatedly, we should think that there is a substantial burden that must be met by any purported justification for programming autonomous vehicles so as to systematically prioritize the interests of their occupants in conditions of hybrid traffic. In the next section I will consider the prospects of arguments aimed at meeting this burden.

IV. Defending Fair Risk Distribution

One potential objection to the view that Fair Risk Distribution should govern firms' decision-making about the programming of autonomous vehicles for hybrid traffic is that it is inconsistent with existing and widely accepted safety-related practices of vehicle-producing firms. These firms already produce vehicles that prioritize the safety of occupants as much as possible, offer more extensive safety packages on more expensive vehicles, and, perhaps most importantly, offer vehicle designs that are safer for occupants and more dangerous for those in other vehicles in many of their more expensive products (e.g. large SUVs). It is not generally thought that firms that produce large SUVs are violating any moral requirements merely by offering these vehicles to consumers (or if they are, it is typically thought that the reasons have to do with high greenhouse gas emissions, and not with the resulting unequal distribution of the risks of the road). But Fair Risk Distribution appears to imply that firms that produce large SUVs, the designs of which make them particularly safe for occupants and notably more dangerous than, for example, typical cars to others on the road, are acting wrongly, at least so long as the SUVs are priced in a way that makes them available primarily to wealthier members of society, and inaccessible to most others.

I have two related responses to this objection. The first is that it does seem to me that there is something morally troubling about a world in which firms produce vehicles, such as

large SUVs, that are marketed primarily to wealthier consumers and unaffordable for the less well off, given that driving those vehicles increases the risks faced by those driving typical cars in comparison with a scenario in which they are not produced at all. We should not simply accept that it is permissible for firms to facilitate the wealthy in distributing the risks of the road away from themselves and onto those who cannot afford the more expensive, safer products that they can produce. Relative safety on the road should not, in my view, be in effect for sale on the market, with the result that, on average, the richer one is, the less risk she is likely to face on the road, at least if she makes her vehicle purchasing decisions with her own safety as a central concern.[11]

I suspect that there are at least two important reasons why we generally fail to recognize the reasons to be concerned about the effects on the distribution of the risks of the road of the production and sale of vehicles like large SUVs mainly to wealthier consumers. First, we tend to think that consumers are entitled to be concerned about their own safety when they are purchasing products, and to think that firms are doing something good when they make their products safer for consumers. In most cases this is clearly correct, since most products are such that making them safer for their consumers does not make them more dangerous for others. Vehicles at least can be an exception to this, however, since, for example, for occupants of typical car, a crash with a large SUV will, on average, cause more harm than a crash with another typical car. I suspect that many people's view about the ethics of producing large SUVs might change at least somewhat if they were to attend more clearly to this fact.

[11] Many people have a similar view about the unacceptability of a correlation between wealth and risk in other dimensions of life. For example, it is widely thought to be unacceptable for access to medical care to be largely determined by wealth, with the result that the wealthy tend to live longer and healthier lives than the less well off. I am grateful to Ryan Jenkins for this example.

The second reason why I suspect that we fail to recognize the grounds for concern about the production and sale of vehicles like large SUVs is that no individual firm refraining from producing such vehicles would be likely to make a difference to the overall distribution of the risks of the road, since consumers who can afford them would simply buy vehicles produced by a different firm. If we ask whether it is wrong for a particular firm to produce such vehicles, given that other firms are already producing them, it is difficult to see what reason there could be for thinking that it is. If we imagine, however, that a firm is deciding whether to be the first to produce and market a vehicle of a type that would have the effect of redistributing some of the risks of the road away from its (mostly wealthy) occupants and onto other users of the road, it will, I think, seem much more plausible that there is at least some moral reason for the firm to refrain.

At this point it might be objected that even if I am correct that there is something objectionable about producing vehicles that increase the risks faced by others while reducing those faced by occupants, autonomous vehicles, even if they are programmed to systematically favor the interests of occupants, would not have this effect. Instead, because of how much safer they will be than human-driven vehicles, they would reduce the risks faced by all road users, but reduce them significantly more for occupants than for those who continue to use standard vehicles. Since everyone would benefit, relative to the status quo, from the production of such vehicles, it might be claimed that no one has a compelling objection to their production. Furthermore, it might be added, since there are reasons to think that consumers would not purchase autonomous vehicles that are not programmed to prioritize their interests (Bonnefon, Shariff, & Rahwan 2016), the result of firms refraining from producing vehicles programmed in ways that are inconsistent with Fair Risk Distribution might be that the benefits that everyone

could enjoy relative to the status quo would simply be forgone. And it certainly cannot be the case that this would be preferable, morally speaking, to maintaining the status quo.

In response, I agree that if we were really faced with a choice between reducing risks for everyone, but in a way that is inconsistent with Fair Risk Distribution in that the wealthy, on average, would benefit significantly more, and simply maintaining the status quo, we should prefer to reduce risks for everyone. But there do not seem to me to be especially compelling reasons to think that we are actually faced with this choice. If potential consumers can be made aware that autonomous vehicles are significantly safer than human-driven vehicles, then they should be willing to purchase and use them, even they are not programmed to systematically prioritize the interests of occupants. For firms to respond to consumers' preference for vehicles that systematically prioritize their interests, rather than attempting to educate them about the overall safety benefits of autonomous vehicles for occupants, even if those vehicles are programmed in a way that is consistent with Fair Risk Distribution, would be to facilitate the wealthy in distributing the safety gains that autonomous vehicles will provide disproportionately to themselves. The fact that everyone would gain relative to the status quo is not, in my view, a sufficient justification for permitting such a distribution of the benefits, so long as there are available alternatives in which at least as much aggregate benefit is produced and the benefits are distributed more fairly.

V. The Limits of Fair Risk Distribution

If my argument to this point is correct, then firms have strong moral reasons to aim at as equal a distribution of the risks of the road among vehicle users as possible in the programming of their autonomous vehicles. The most important implication of the argument is that every firm

is obligated not to be the first to offer vehicles programmed to systematically prioritize the interests of occupants to consumers. So long as no firm offers such vehicles to consumers, it will be possible for the industry as a whole to operate in a way that is consistent with Fair Risk Distribution.

If even one firm does offer autonomous vehicles programmed to systematically prioritize the interests of occupants, however, it would appear that, given reasonable predictions about consumer behavior, other firms cannot be obligated to refrain from following suit. This is because at least most consumers can be expected to purchase vehicles programmed in that way rather than other vehicles programmed in ways that are more consistent with Fair Risk Distribution.[12] A particular firm refraining from offering vehicles programmed to prioritize occupants' interests, then, would likely have no effect on the overall distribution of the risks of the road between those who can afford to purchase autonomous vehicles and those who cannot. Because of this, it seems to me that once at least one firm has violated the requirements of Fair Risk Distribution, it is likely that the principle no longer implies that other firms are prohibited from offering vehicles programmed to prioritize the interests of occupants.

If this is correct, then the issue of the effects on the distribution of the risks of the road of autonomous vehicle programming for hybrid traffic constitutes a kind of collective action problem. Every firm will likely take itself to have significant business reasons to offer vehicles programmed in the way that evidence suggests will make the vehicles easiest to sell, namely to systematically prioritize the interests of occupants. Morally speaking, however, there are powerful reasons for firms to refrain from promoting an outcome in which the safety-related benefits of autonomous vehicles are enjoyed disproportionately by the wealthy. Furthermore, the

[12] I take no position here on whether it would be morally permissible for individual consumers to purchase autonomous vehicles programmed to systematically prioritize the interests of occupants, if there were alternatives on the market that were programmed in ways more consistent with Fair Risk Distribution.

potential economic gains of acting against these moral reasons are likely greatest when one's firm stands to be the first to do so, and as soon as one firm does so, there is likely nothing any other firm can do to prevent the objectionable outcome.

All of this means that firms producing autonomous vehicles must resist the temptation to seek to gain a large share of the market early in the process of autonomous vehicles being integrated onto the roads by programming them in the way that will make them most attractive to the likely initial consumer base. This is far from a trivial constraint on firms' pursuit of profit in the autonomous vehicle market, and I do not expect that many firms will be inclined to abide by it. But if and when the first firm releases an autonomous vehicle programmed to systematically prioritize the interests of occupants onto the market, I believe that this will constitute a very serious moral violation, and will virtually ensure a quite disproportionate and unfair distribution of the safety benefits of autonomous vehicles.[13,14]

References

Awad, E. et. al. 2018. "The Moral Machine Experiment." *Nature* 563: 59-64.

Bergmann, L.T. et. al. 2018. "Autonomous Vehicles Require Socio-Political Acceptance: An Empirical and Philosophical Perspective on the Problem of Moral Decision Making." *Frontiers in Behavioral Neuroscience* 12: 1-12.

Bonnefon, J.F., Shariff, A., & Rahwan, I. 2016. "The Social Dilemma of Autonomous Vehicles." *Science* 352: 1573-1576.

Borenstein, J., Herkert, J.R., & Miller, K.W. 2019. "Self-Driving Cars and Engineering Ethics: The Need for a System Level Analysis." *Science and Engineering Ethics* 25: 383-398.

Coca-Vila, I. 2018. "Self-Driving Cars in Dilemmatic Situations: An Approach Based on the Theory of Justification in Criminal Law." *Criminal Law and Philosophy* 12: 59-82.

Coeckelbergh, M. 2016. "Responsibility and the Moral Phenomenology of Using Self-Driving Cars." *Applied Artificial Intelligence* 30: 748-757.

Davnall, R. Forthcoming. "Solving the Single-Vehicle Self-Driving Car Trolley Problem Using

[13] Ideally, firms would be legally required to refrain from programming autonomous vehicles to systematically prioritize the interests of occupants, and required to program them to, at least roughly, minimize harm. In the absence of this kind of policy solution, however, it will up to individual firms to comply with the ethical requirement that I have defended.

[14] I am grateful to Ryan Jenkins and Tomáš Hříbek for valuable comments.

Risk Theory and Vehicle Dynamics." *Science and Engineering Ethics*.
Edelstein, S. 2018. "Self-Driving Cars and the End of Car Ownership." *The Drive*: https://www.thedrive.com/tech/20533/self-driving-cars-and-the-end-of-car-ownership.
Etzioni, A. & Etzioni, O. 2017. "Incorporating Ethics into Artificial Intelligence." *Journal of Ethics* 21: 403-418.
Faulhaber, A.K. et. al. 2019. "Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles." *Science and Engineering Ethics* 25: 399-418.
Fleetwood, J. 2017. "Public Health, Ethics, and Autonomous Vehicles." *American Journal of Public Health* 107: 532-537.
Gogoll, J. & Müller, J.F. 2017. "Autonomous Cars: In Favor of a Mandatory Ethics Setting." *Science and Engineering Ethics* 23: 681-700.
Goodall, N.J. 2014a. "Ethical Decision Making During Automated Vehicle Crashes." *Transportation Research Record* 2424: 58-65.
_____. 2014b. "Machine Ethics and Automated Vehicles." In Meyer & Beiker (eds.), *Road Vehicle Automation*. Dordrecht: Springer.
Gurney, J.K. 2017. "Imputing Driverhood: Applying a Reasonable Driver Standard to Accidents Caused by Autonomous Vehicles." In Lin, Jenkins, & Abney (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York: Oxford University Press.
Hevelke, A. & Nida-Rümelin, J. 2015. "Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis." *Science and Engineering Ethics* 21: 619-630.
Himmelreich, J. 2018. "Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations." *Ethical Theory and Moral Practice* 21: 669-684.
Hübner, D. & White, L. 2018. "Crash Algorithms for Autonomous Cars: How the Trolley Problem Can Move Us Beyond Harm Minimization." *Ethical Theory and Moral Practice* 21: 685-698.
Huffman, M. 2018. "Report: Rise of Autonomous Vehicles Will Reduce Car Ownership." *Consumer Affairs*: https://www.consumeraffairs.com/news/report-rise-of-autonomous-vehicles-will-reduce-car-ownership-030118.html.
Keeling, G. 2018. "Legal Necessity, Pareto Efficiency, & Justified Killing in Autonomous Vehicle Collisions." *Ethical Theory and Moral Practice* 21: 413-427.
_____. Forthcomimg. "Why Trolley Problems Matter for the Ethics of Automated Vehicles." *Science and Engineering Ethics*.
Kelley, B. 2017. "Public Health, Autonomous Automobiles, and the Rush to Market." *Journal of Public Health Policy* 38: 167-184.
Leben, D. 2017. "A Rawlsian Algorithm for Autonomous Vehicles." *Ethics and Information Technology* 19: 107-115.
Lin, P. 2015. "Why Ethics Matters for Autonomous Cars." In Maurer, Gerdes, Lenz, & Winner (eds.), *Autonomes Fahren: Technische, Rechtliche und Gesellschaftliche Aspekte*. Berlin: Springer.
Liu, H.Y. 2017. "Irresponsibilities, Inequalities and Injustice for Autonomous Vehicles." *Ethics and Information Technology* 19: 193-207.
Marchant, G.E. & Lindor, R.A. 2012. "The Coming Collision Between Autonomous Vehicles and the Liability System." *Santa Clara Law Review* 52: 1321-1340.
Martin, D. 2017. "Who Should Decide How Machines Make Morally Laden Decisions?" *Science and Engineering Ethics* 23: 951-967.

Millar, J. 2016. "An ethics Evaluation Tool for Automating Ethical Decision-Making in Robots and Self-Driving Cars." *Applied Artificial Intelligence* 8: 787-809.

_____. 2017. "Ethics Settings for Autonomous Vehicles." In Lin, Jenkins, & Abney (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York: Oxford University Press.

Nyholm, S. 2018. "The Ethics of Crashes with Self-Driving Cars: A Roadmap, I." *Philosophy Compass* 13: 1-10.

Nyholm, S. & Smids, J. 2016. "The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?" *Ethical Theory and Moral Practice* 19: 1275-1289.

_____. forthcoming. "Automated Cars Meet Human Drivers: Responsible Human-Robot Coordination and the Ethics of Mixed Traffic. *Ethics and Information Technology*.

Parfit, D. 2000. "Equality or Priority." In Clayton & Williams (eds.), *The Ideal of Equality*. New York: Palgrave Macmillan.

Santoni de Sio, F. 2017. "Killing by Autonomous Vehicles and the Legal Doctrine of Necessity." *Ethical Theory and Moral Practice* 20: 411-429.

Standage, T. 2018. "Why Driverless Cars Will Mostly be Shared, not Owned." *The Economist*: https://www.economist.com/the-economist-explains/2018/03/05/why-driverless-cars-will-mostly-be-shared-not-owned.

Temkin, L. 2000. "Equality, Priority, and the Levelling Down Objection." In Clayton & Williams (eds.), *The Ideal of Equality*. New York: Palgrave Macmillan.

van Loon, R.J. & Martens, M.H. 2015. "Automated Driving and its Effect on the Safety Ecosystem: How do Compatibility Issues Affect the Transition Period?" *Procedia Manufacturing* 3: 3280-3285.

Wolkenstein, A. 2018. "What Has the Trolley Dilemma Ever Done for Us (and What Will it Do in the Future)? On Some Recent Debates about the Ethics of Self-Driving Cars." *Ethics and Information Technology* 20: 163-173.