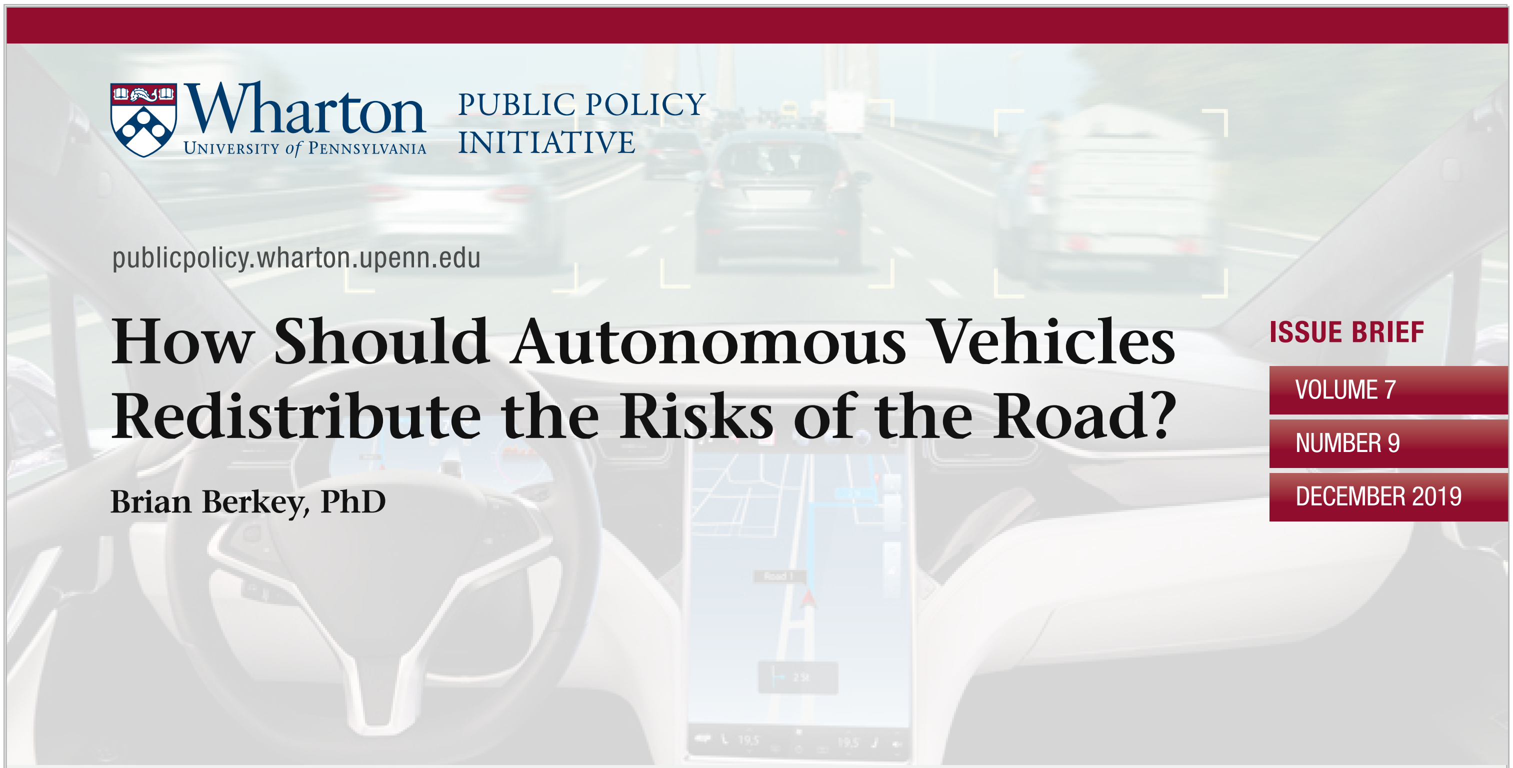

Wharton
UNIVERSITY of PENNSYLVANIA
PUBLIC POLICY
INITIATIVE

publicpolicy.wharton.upenn.edu

How Should Autonomous Vehicles Redistribute the Risks of the Road?

Brian Berkey, PhD

ISSUE BRIEF
VOLUME 7
NUMBER 9
DECEMBER 2019

There is a consensus among researchers who study autonomous vehicles that the introduction of these machines onto roadways could significantly reduce the number of injuries and deaths from vehicle accidents.[1]

If this expectation materializes, there are strong reasons to favor replacing human-driven vehicles with autonomous ones. The advent of autonomous vehicles, however, likely will be gradual, with the replacement of human-driven vehicles occurring over the course of many years.[2] 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*.[3] Consequently, companies 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.[4]

These programming decisions will affect the lives and health of the public. Therefore, autonomous vehicle companies, elected officials, regulators, and independent experts must examine, through a transparent process of open dialogue, the morally relevant dimensions of the machine learning mechanisms that direct the actions of autonomous vehicles operating in hybrid traffic situations.

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 companies that

SUMMARY

- This Issue Brief considers the principles that should govern how companies that produce autonomous vehicles should program them to behave in potential conflict situations with vehicles controlled by human drivers.
- Research shows that consumers would prefer to purchase autonomous vehicles that are programmed to prioritize the safety of their occupants. But doing so means that in hybrid conditions, occupants of human-driven vehicles would systematically suffer more harms.
- This disparity should be of great concern, as it is likely that there will be a correlation between autonomous vehicle usage and wealth, since the large R&D costs that go into the making of autonomous vehicles will probably make them a luxury item, at least early on.
- The Issue Brief therefore proposes a Fair Risk Distribution principle to govern the programming of autonomous vehicles, and lays out the moral obligations of all manufacturers to not be the first to offer vehicles programmed to systematically prioritize the interests of their occupants.
- But can businesses that make autonomous vehicles be expected to uphold this moral obligation, and to resist the temptation to profit from the systematic prioritization of their occupants' interests? Policymakers need to have an open discussion now as to whether legislation or regulation may be needed in order to achieve the fair distribution of risk.

produce autonomous vehicles ought to program them to respond in conflict situations involving human-driven vehicles.[5] In this Issue Brief, I argue that this principle 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[6] indicates that most consumers would prefer to purchase autonomous vehicles that *do* systematically prioritize the interests of occupants to those of others, my argument generates a substantial ethical restriction on companies' efforts to gain market share in the initial stages of the autonomous vehicle life cycle.

CONFLICTS OF INTEREST IN HYBRID TRAFFIC

Circumstances undoubtedly will arise in which injuries and deaths on the road are unavoidable, even if, at some point in the future, all vehicles on the road are autonomous. At the very least, this will certainly be the case during the period of hybrid traffic. 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.

There are several ways in which the interests of occupants of autonomous vehicles might come into conflict with the interests of occupants of human-driven vehicles, particularly in situations that have the rough structure of "trolley" cases.[7] Consider the following case:[8]

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 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, 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.

In order to recognize the distinctive risk distribution issues raised by the prospect of hybrid traffic scenarios, it is important to note that if the bus were also autonomous, its occupants could be subject to significantly lower risks of injury and death, while the occupant of the autonomous vehicle could be at greater risk. For example, the bus's 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, 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

NOTES

[1] For a sampling of this research, see the Introduction of my paper, "Autonomous Vehicles, Business Ethics, and Risk Distribution in Hybrid Traffic," upon which this Issue Brief is based.

[2] van Loon, R.J. & Martens, M.H. (2015). "Automated Driving and its Effects on the Safety Ecosystem: How Do Compatibility Issues Affect the Transition Period?" Procedia Manufacturing 3:3280-85.

[3] Goodall, N.J. (2014). "Ethical Decision Making During Automated Vehicle Crashes." Transportation Research Record 2424, p. 59; and 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, p. 686.

[4] These decisions almost certainly will not take the form of discrete choices—that is, companies will not be programming autonomous vehicles 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 not fundamentally. Whatever values would rightly guide the direct programming of autonomous vehicles ought, as much as possible, to guide the 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.

[5] This principle is also relevant to programming choices involving other kinds of conflicts, for example those with pedestrians or cyclists, although its implications regarding how much risk each party should bear plausibly differ

very minor injuries.

This case is complex, but the general point it helps to highlight is fairly clear and, on reflection, should not be surprising. 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 since they will be capable also of using that data to determine what they will do—autonomous vehicles could, in principle, be programmed in ways that would ensure 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 the above case, for example, could be programmed in a way that would make it very likely that the bus will go over the cliff and kill all of the people onboard, despite the fact that it instead could have been programmed in a way that would ensure that its occupant is subjected to more risk when necessary in order to prevent a greater number of people in the human-driven bus from being subject to more extensive 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 occupants of human-driven vehicles will systematically suffer more harms, and more serious harms, than occupants of autonomous vehicles in circumstances in which an accident is unavoidable. And this is the case not only for those in human-driven vehicles, but also for pedestrians, cyclists, and road workers.

> "autonomous vehicles could, in principle, be programmed in ways that would ensure 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."

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 for which development requires large R&D costs, autonomous vehicles seem likely to be a luxury item, at least initially, available primarily to wealthier people. If this occurs, then 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. These differences in risk exposure could be much less substantial if autonomous vehicles are programmed in ways that refrain from prioritizing the interests of their occupants so heavily.

In the face of such a conflict, it is important for there to be a principle (or principles) guiding the programming of autonomous vehicles.

NOTES

significantly.

[6] Bonnefon, J.F., Shariff, A., & Rahwan, I. (2016). "The Social Dilemma of Autonomous Vehicles." Science 352: 1573-76.

[7] 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, 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 (as in traditional trolley cases) before reflecting on what ought to be done in cases that are similar in many respects but also involve risk and/or uncertainty. Furthermore, 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.

[8] This case is based loosely on a case given by Patrick Lin (2015). "Why Ethics Matters for Autonomous Cars." In Maurer et al. (eds.), Autonomous Fahren: Technische, Rechtliche und Gesellschaftliche Aspekte. Berlin: Springer, pp. 76-77).

[9] As matter of public policy, it is worth considering whether there might be obligations applying to, for example, governments, vehicle manufacturers, and even individuals, to promote equality in access to autonomous vehicles so that the inequalities in the road risks to which individuals are subjected are at least more limited than they would otherwise be.

[10] Consumers tend to believe both that they are entitled to be concerned about their own safety when they are purchasing products and that companies are doing something good when they make their products safer for consumers. In most cases this is clearly correct, since most products

THE GUIDING PRINCIPLE

As an ethical 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. This principle would capture the importance of avoiding an outcome in which wealthier members of society disproportionately enjoy the benefits of increased safety generated by autonomous vehicles, while the less well off—as well as pedestrians, cyclists, and road workers—disproportionately bear the risks of the road.

What would a fair distribution of the risks of the road look like? A reasonable starting point 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 companies producing autonomous vehicles to deviate from aiming at an equal distribution of the risks. Here, then, is an initial formulation of a plausible Fair Risk Distribution principle:

Autonomous vehicles ought to be programmed so that, to the greatest extent possible consistent with the aim of minimizing traffic-related injuries and deaths, 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.

> “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.”

Determining what may count as a morally compelling reason to deviate from aiming at an equal distribution of the risks of the road would require dialogue among companies, relevant regulators, and elected officials. Two arguments for deviation come to mind immediately:

1. SPECIAL PROTECTIONS FOR PEDESTRIANS AND CYCLISTS?

It may be the case 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. It seems reasonable that those who choose to introduce risks like serious injury or death on and near roadways in order to enjoy the benefits of the activities (driving) that unavoidably involve these risks, should, where possible, at least bear a greater share of the risks than those (pedestrians and cyclists) 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

NOTES

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 belief, 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. 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.

occupants from the risk of very serious harms.

2. SPECIAL BENEFITS FOR AUTONOMOUS VEHICLE USERS?

Another argument begins by noting that autonomous vehicles will be significantly safer than human-driven vehicles. Because of this, 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. This argument for deviating from an equal distribution of risk should be rejected.

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. However, when 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 insist that they also benefit significantly more from the reduction in the overall risks of the road than those who simply cannot afford to switch to using them.

We should conclude that, in the programming of autonomous vehicles, there is no clear justification for deviating from aiming at an equal distribution of the risks of the road between users of autonomous vehicles and users of standard vehicles in conditions of hybrid traffic.[9] We should not simply accept that it is permissible for companies to facilitate the wealthy in distributing these risks 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 be, in effect, for sale on the market.[10]

There are, however, limitations of the principle's applicability to business decisions involving the programming of autonomous vehicles.

THE LIMITS OF FAIR RISK DISTRIBUTION

If my argument to this point is correct, then companies 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 is that every company has an obligation not to be the first to offer autonomous vehicles programmed in a way that is inconsistent with a fair distribution of risk and, in particular, obligated not to be the first to offer vehicles programmed to systematically prioritize the interests of occupants.

If even one company does this, however, it would appear that, given reasonable predictions about consumer behavior, other companies cannot be obligated to refrain from following suit. This constitutes a kind of collective action problem. The business case for making a moral decision that runs counter to the Fair Risk Distribution principle is simply that the potential economic gains available are too great, especially for a company that has the opportunity to be the first to market. But companies producing autonomous vehicles must resist this temptation to profit from the systematic prioritization of their occupants' interests. Doing so would constitute a very serious moral violation, and it would virtually ensure a quite disproportionate and unfair distribution of the safety benefits of autonomous vehicles.

Other collective action problems are typically resolved through legislation and/or regulation. Whether or not either policy approach is appropriate in this instance, there should, at the very least, be an open dialogue about the concerns discussed in this Issue Brief between autonomous vehicle producers and federal policymakers while a desirable outcome—in terms of public health and fairness—is still achievable.

ABOUT THE AUTHOR

BRIAN BERKEY, PHD

Assistant Professor of Legal Studies & Business Ethics,
The Wharton School

In addition to his appointment at the Wharton School of the University of Pennsylvania, Professor Berkey is an associated faculty member in the Department of Philosophy at Penn, and an affiliated faculty member of the Institute for Law and Philosophy at Penn Law. He works in moral and political philosophy (including business ethics and environmental ethics), in particular on questions about the demandingness of morality, individual obligations of justice, ethical issues arising with regard to climate change, and the relationship between ideal and non-ideal theory.

He is also interested in the notion of collective obligations and their relationship to individual obligations, as well as in methodological issues in ethics and political philosophy, including the appropriate role of appeals to intuitions. His work has appeared in journals such as Mind, Philosophical Studies, Canadian Journal of Philosophy, Utilitas, and Journal of Applied Philosophy. Professor Berkey earned his PhD in Philosophy at UC-Berkeley in 2012, and was a postdoctoral fellow at the Center for Ethics in Society at Stanford before moving to Penn. During the 2018-19 academic year, he was a Berggruen Fellow at the Edmond J. Safra Center for Ethics at Harvard University.

ABOUT THE WHARTON PUBLIC POLICY INITIATIVE

The Wharton Public Policy Initiative (PPI) is a hub for research and education, engaging faculty and students across the University of Pennsylvania and reaching government decision-makers through independent, practical, timely, and nonpartisan policy briefs. With offices both at Penn and in Washington, DC, the Initiative provides comprehensive research, coverage, and analysis, anticipating key policy issues on the horizon.

ABOUT WHARTON PUBLIC POLICY INITIATIVE ISSUE BRIEFS

Wharton PPI publishes issue briefs at least once a month, tackling issues that are varied but share one common thread: they are central to the economic health of the nation and the American people. These Issue Briefs are nonpartisan, knowledge-driven documents written by Wharton and Penn faculty in their specific areas of expertise.

CONTACT THE WHARTON PUBLIC POLICY INITIATIVE

At Penn
Steinberg Hall-Dietrich Hall, Room 201
Philadelphia, PA 19104-6302
+1.215.898.1197

In Washington, DC
777 6th Street, NW
Washington, DC 20001
1+202-870-2655

For additional copies, please visit the Wharton PPI website at **publicpolicy.wharton.upenn.edu.**
Follow us on Twitter: **@WhartonPPI**