To Be or Not to Be: An Ethical Framework for the Use of Autonomous Vehicles

It is rare to drive around Silicon Valley without seeing a white Waymo minivan, decked out with dozens of sensors and a human relaxing in the "driver's seat." These autonomous vehicles (AVs) hold immense promise to both eliminate accident-causing human error and raise fuel efficiency when deployed on a large scale, benefitting most people in society.[1] Nevertheless, these innovations are not without cost. Pedestrians have already been killed in the testing of AVs, indicating that safety for some may come at the expense of others.[2] And some worry that AVs will render human driving obsolete, resulting in the loss of livelihoods and a cherished pastime.

At the heart of both these concerns lies a question: When most people in society stand to gain from a decision, how should we prioritize the interests of minority groups who fare worse? In this paper, I will use John Rawls's ideas of justice as fairness and the veil of ignorance to argue that we cannot sacrifice the disadvantaged in the name of the "good of society." Specifically applied to AVs, I argue that if AVs harm cyclists while protecting pedestrians and drivers, the technology should not be deployed because cyclists are already a disadvantaged group. Meanwhile, in the case that AVs do not disproportionately harm cyclists or others from a disadvantaged group, it is acceptable to limit the freedom to drive by disabling a "human pilot mode" for some drivers, although only to the extent that safety is improved substantially and consistent with the demands of value pluralism. People who enjoy driving are not, in practice, a disadvantaged group, so their preferences should not be inherently prioritized.

On the surface, it seems intuitive that AVs should be deployed if they save more lives relative to their non-deployment. Upon deeper reflection, it is evident that a utilitarian net safety increase is an insufficient standard on which to make a decision; it matters which groups become

[1] "Case Study: Autonomous Vehicles," n.d., 4.
[2] "Case Study: Autonomous Vehicles," 19.

more or less safe. As captured in John Rawls's *A Theory of Justice*, "Social and economic inequalities…are just only if they result in compensating benefits for everyone, and in particular for the least advantaged members of society. These principles rule out justifying institutions on the grounds that the hardships of some are offset by a greater good in the aggregate."[3] To throw Rawls' "justice as fairness" argument into stark relief, society has benefitted from medical knowledge derived through human experimentation on Black Americans and on European Jews during the Holocaust. However, there is broad consensus that such practices are unacceptable and unjust on account of their effects on these historically disadvantaged groups, despite the benefits to society and even despite the benefits to other historically disadvantaged groups like women. Though admittedly an extreme example, these instances of human experimentation serve to illustrate that net societal benefit alone is inadequate.

In the case of autonomous vehicles, we first must ask, who are the disadvantaged groups? The answer is not so clear-cut: drivers encompass both workers who live in the poor outskirts of cities and those who can afford Teslas. Cyclists, similarly, include those too young or too destitute to have a car and the wealthy who use cycling for exercise and leisure. In an economic sense, it is not immediately clear which group is more privileged. However, we must consider that most cities are configured for drivers more so than for cyclists, and accidents between drivers and cyclists are far riskier for the cyclists. Moreover, people who are unable to drive have few affordable alternatives but to bike, owing to the poor state of public transportation worldwide. Cyclists thus here constitute the "least advantaged members of society" that Rawls referred to, and as such, must be given special consideration in deploying AVs. Importantly, even if pedestrians are also disadvantaged and may benefit from AVs, we cannot harm one disadvantaged group even if

[3] John Rawls, *A Theory of Justice*, Rev. ed (Cambridge, Mass: Belknap Press of Harvard University Press, 1999), 13.

another nominally benefits; as with the example of human experimentation, there can be no weighing of net lives saved. When AVs harm cyclists in order to benefit drivers and pedestrians, this does not constitute justice and such a policy cannot be implemented.

Some may argue that it is unfair to label cyclists as a "disadvantaged group" in any setting. Unlike being Black or Jewish, being a cyclist is not an inflexible trait. Nobody is born a cyclist, nor is cycling absolutely necessary for sustenance, especially in cities with robust public transportation alternatives for cyclists who feel unsafe. Meanwhile, some may reject factoring in disadvantaged groups at all. Notably, in 2016, the German government established a national ethics committee for automated driving, and one of the ethical guidelines from this committee's report states, "In the event of unavoidable accident situations, any distinction based on personal features…is prohibited. It is also prohibited to offset victims against one another."[4] Implicit in this statement is that cyclists should not be given special consideration: on the road, all lives are equal.

I agree that "disadvantaged groups" must be defined in their local context, and there may be places in which cyclists are not disadvantaged. However, even if we consider all groups as equal, AVs should not be deployed under these circumstances. In making a decision for a policy that benefits some people at the cost of others, assuming nobody harmed is of a disadvantaged group, we can use Rawls's idea of the "veil of ignorance" to evaluate the policy. Under the veil of ignorance, we do not know if we are a driver, pedestrian, or cyclist.[5] How would we evaluate AVs? If we were a driver or pedestrian, AVs that harm only cyclists seem nothing but beneficial. Cyclists appear less safe, but they can choose to purchase a car or use public transportation. Perhaps alternative commuting methods are more inconvenient or less enjoyable than cycling, but

[4] Christopher Luetge, "The German Ethics Code for Automated and Connected Driving," *Philosophy & Technology*, n.d.
[5] Rawls, *A Theory of Justice*, 11.

ultimately, cyclists can take steps to use safer transportation forms. If we do not consider cyclists a disadvantaged group, then it seems intuitive under the veil of ignorance that we should prefer implementing AVs.

However, this calculus ignores the imminent threat of climate change, which will kill thousands and displace millions starting as early as 2030 if emissions are not cut sharply.[6] In light of this fact, it is alarming to note that road vehicles were responsible for 17% of 2008 global greenhouse gas emissions.[7] According to the Center for American Progress, the number of vehicle miles traveled would increase as AVs are phased in because the opportunity cost of driving is reduced; yet the increased efficiency of AVs is not enough to offset the higher number of cars on the road.[8] Still worse, AVs that increase safety for car passengers and decrease safety for cyclists discourage people from using sustainable transportation, further exacerbating the problem. Once climate change is considered, AV implementation will actually decrease human safety of *all* people because they increase the risk of displacement or death—especially for members of the Global South, an undoubtedly disadvantaged group. Thus, even if cyclists are not a disadvantaged group, the veil of ignorance demonstrates that it is still unethical to deploy AVs under these circumstances.

It is reasonable to expect that the debate over AVs will expand to whether anyone should be able to drive. If AVs can be proven to make all people—including cyclists—safer in terms of accidents, society may want to eliminate human driving altogether. Is this just, given that many

---

[6] "Quantitative Risk Assessment of the Effects of Climate Change on Selected Causes of Death, 2030s and 2050s" (World Health Organization, n.d.).
[7] "Climate Impacts," World Health Organization, accessed February 27, 2020, http://www.who.int/sustainable-development/transport/health-risks/climate-impacts/en/.
[8] Myriam Alexander-Kearns, Miranda Peterson, and Alison Cassady, "The Impact of Vehicle Automation on Carbon Emissions," Center for American Progress, accessed February 27, 2020, https://www.americanprogress.org/issues/green/reports/2016/11/18/292588/the-impact-of-vehicle-automation-on-carbon-emissions-where-uncertainty-lies/.

people structure their lives around driving and derive pleasure from it? First, we note that people who enjoy driving are not an underprivileged group: nothing in particular places these drivers at a disadvantage. Thus, we must evaluate this policy decision by, again, using the veil of ignorance. Under this framework, people who enjoy driving would be harmed by any policy that limits their freedom to do so, while *all people* would be harmed by allowing such a policy.[9] The pleasure of some becomes secondary to the survival and wellbeing of humans across the globe; we can only argue over pleasure if we are alive to do it.

Yet one may argue that it is unjust to remove the ability to drive under all circumstances, and that if we do deploy AVs, we should maintain a human pilot mode. Value pluralism holds that there are many equally important and competing values to consider in making any decision.[10] If safety were the only goal, we might also choose to eliminate all forms of personal privacy so that no person is ever the victim of a crime. We do not do this because we have multiple important considerations in addition to safety, including freedom and enjoyment.

Such an objection has merit here; we must assess values that may tip in favor of human driver intervention but limit this intervention to situations in which safety is improved substantially and consistent with the demands of value pluralism. Under the veil of ignorance, I would not want a policy that prioritizes only safety of all at the cost of the pleasure of many, nor would I want a policy that aims to satisfy the whims of one group of people while disregarding the safety of society; in the first scenario, I would be unhappy if I were a person who likes to drive, and under the second, I would unhappy if I were anybody else.

---

[9] Citizens of the Global South are actually disproportionately harmed, but all people in the world will suffer consequences of displacement, food shortage, natural disasters, and so on. In any case, climate change surely does not discriminate based on who likes to drive and who does not.

[10] Isaiah Berlin, *The Crooked Timber of Humanity: Chapters in the History of Ideas*, Second Edition (Princeton; Oxford: Princeton University Press, 2013), 11.

This dilemma is not unique. Societies face this exact question in deciding at what ages people can drink, smoke, and drive. No limitations on such activities put many at risk, while extensive limits threaten people's pleasure and autonomy. For these problems, society decides to set minimum ages because young people are shown to be particularly risky when they drive, drink, or smoke, so limiting their driving substantially improves safety.[10] This solution makes many youths unhappy, but may also not satisfy those who worry that any drinking or smoking is dangerous for society. However, this is the solution that satisfies the veil of ignorance and respects value pluralism. Similarly, it is reasonable to limit human pilot mode to those who are sufficiently safe drivers. This may entail placing minimum and maximum age restrictions on humans who can use pilot mode, or granting licenses based on performance on reaction-time tests and the completion of extensive training.

Through the democratic process, society can decide what threshold is a "substantial" improvement in safety, seeing as we have readily available data about the relationship between accident rates and age or reaction time. If the technology for fully autonomous vehicles exists, there can surely be verification technology to only enable human pilot mode when licensed individuals are driving. Stepping under the veil of ignorance, most of those who value driving will be able to do so, while those who value safety can be assured that there is a minimum performance of human drivers. Although nobody is fully satisfied, no value has a "trump card" over any other and everyone has an aspect of their concern addressed.

---

[10] McCartt, Anne T., Laurie A. Hellinga, and Bevan B. Kirley. 2010. "The Effects of Minimum Legal Drinking Age 21 Laws on Alcohol-Related Driving in the United States." *Journal of Safety Research*. https://doi.org/10.1016/j.jsr.2010.01.002.
Notably, researchers found that "there is no evidence that alcohol education can even partially replace the effect" of raising the drinking age. In this case, deaths from car crashes dropped significantly, and there is no alternative policy to raising the drinking age that comes close to achieving this outcome.

Autonomous vehicles promise many benefits but are accompanied by a host of ethical problems. I have created a framework to address only two of such problems, though it can be employed in other scenarios. Under the theory of justice, we must protect disadvantaged groups when considering a policy that promises to improve society at the cost of the less fortunate. In other cases, when groups are similarly situated and we are under the veil of ignorance, we must consider value pluralism to create a viable solution that addresses at least some concerns of each group where possible while safety is improved.

Ultimately, as the debate about how to deploy AVs rages on, we must not forget that we have a far larger problem on our hands. AVs are useless if climate change persists and no humans exist to use them. Though I have only grazed the surface of the issue in this paper, it is important that society consider an ethical framework for how we prioritize technological development. Rather than arguing about the implementation of AVs, it is possible that AVs may not be the project worth funding.