In 2011, when Sebastian Thrun, the father of self-driving cars, claimed that "It's a no-brainer that 50 to 60 years from now, cars will drive themselves" [Allen, 2011], he clearly did not foresee the rapid pace at which the technology would advance. Merely 9 years later, companies like Cruise, Waymo, Aurora and others have already received permits to transport passengers in fully automated robotaxis in the state of California. [Dickey, 2020] While the industry pushes ahead, it is important that not only do regulatory frameworks keep apace with technological advancements, but also that these regulatory frameworks are underpinned by thorough ethical analysis.

In this essay, I explore two influential schools of thought in normative ethics — utilitarianism and contractarianism — that have had tremendous staying power in contemporary society and analyze how well they translate to the autonomous vehicle (AV) era. In particular, I use the following two policy questions — whether AVs should optimize for passenger safety or global human welfare and whether AVs that jeopardize the welfare of certain groups of road users while increasing global human welfare should be deployed on roads — as vehicles to illustrate the relative merits and pitfalls of the aforementioned frameworks. I find that while utilitarianism is appealing intuitively, it is inadequate as an ethical framework for two reasons: it leads to contradictions when employed in the real world situations that autonomous vehicles face and that it fails to take seriously the distinctions between persons. I then argue that these flaws are resolved more adequately by Rawlsian contractarianism and hence the latter theory should serve as the framework within which AV regulators operate.

Consider the following thought experiment: An AV is travelling on the right-most lane at the posted speed limit on a road with no crosswalk in sight when some pedestrians begin jaywalking right in front of the car. The AV cannot possibly decelerate enough to avoid a fatal collision with the pedestrians. Hence, it has the following two choices: stay on course or swerve to the right and crash into the wall of a nearby building. If the AV stays on course, it will result in the death of the pedestrians but avert the passengers of the vehicle from suffering any harm whereas if it swerves right, the pedestrians would be spared but at the expense of the passengers' lives. Assume that no other road users will be affected by the choice made by the autonomous agent.

Classical utilitarianism, would suggest that the right decision is the one that maximizes global utility or equivalently, according to Sidgwick [1], inflicts the least harm on the stakeholders. Hence, we must consider the number of passengers seated in the car and the number of pedestrians in front of the car

[1] Sidgwick defines utility as "the greatest possible surplus of pleasure over pain, the pain being conceived as balanced against an equal amount of pleasure, so that the two contrasted amounts annihilate each other for purposes of ethical calculation" [Sidgwick, 1871], implying happiness and misery exist in a zero-sum continuum where minimizing one leads to maximizing the other and vice versa

and accordingly make the decision that spares the most lives and inflicts the least injury. The utilitarian calculus is further improved by using a metric such as quality-adjusted life years (QALY) to measure utility instead of just the raw numbers of lives. Doing so allows us to account for more specific situations where the passengers might be children or the pedestrians might be terminally ill, etc., but fundamentally the objective still remains the same: maximize global utility, regardless of the metric used to define utility.

However, there are certain practical constraints to using utilitarianism as the north star in autonomous decision making. The first is that while using a metric like QALY is undoubtedly more utilitarian than a coarse metric such as raw number of lives, it would be near impossible for an AV to gather all the relevant data necessary to compute the QALYs for each individual stakeholder in the scenario. This is because the data that the AV has at it's disposal is limited to what it can collect from it's sensors and cameras. While some reasonable inferences about an individual's characteristics such as their age and gender can perhaps be made from this data, short of having access to their comprehensive medical records, it would be impossible to precisely compute QALYs and make an accurate utilitarian calculation.

It can be argued that imperfect information is not an indictment against utilitarianism for one can simply use whatever information they have at their disposal to best make a utilitarian calculation and act accordingly. While a do-the-best-that-you-can argument has a certain appeal, it is contradicts the consequentialist philosophy that underpins utilitarianism which judges an action purely based on it's outcomes, not it's intentions [Sinnott-Armstrong, 2019]. If, in the example above, the passengers were four terminally ill seniors who would die within the week and the pedestrians were three children who had their whole lives ahead of them, a utilitarian AV with imperfect information would choose to spare the seniors even though the consequences of this decision are that it led to lower global utility than if the AV had chosen otherwise. The justification for the AV's decision does not factor into the utilitarian's judgement since the utilitarian is committed to evaluating the decision based solely on it's outcomes.

Another similar practical constraint for utilitarianism is the inability for decision-making agents, whether human or autonomous systems, to perfectly forecast all the downstream consequences of an action. Consider the following example that elucidates this limitation: An AV with four passengers is trailing a cargo truck when suddenly the truck's cargo comes loose and falls in the path of the AV. There are two bicyclists on either side of the AV; the one on the left is wearing a helmet while the one on the right is not. Assume that all the stakeholders involved in this scenario are identical. Since there are four passengers in the car, a utilitarian AV cannot stay on course and jeopardize the safety of the four for that of one bicyclist. Hence, the AV must choose to swerve either to the left or the right, and injure one bicyclist in the process.

Utilitarianism would require the AV to swerve left and injure the bicyclist wearing the helmet since the helmet would mitigate some of the harm caused to the bicyclist and hence this option leads to the least global suffering. However, the AV ended up penalizing the bicyclist wearing the helmet despite the fact that she was behaving more responsibly than the reckless bicyclist not wearing a helmet. Even though the AV made the choice that led to the outcome with the least suffering in the situation above, the bicyclist wearing the helmet was worse off than she would have otherwise been if she had not worn the helmet. In such a world, bicyclists maximizing their own self-interest would all choose to forgo wearing helmets as they would rather be the bicyclist spared by the AV. However, if all bicyclists make this decision, there will be more suffering in the world since the next time an AV is in a similar predicament as earlier, it will have to choose between two bicyclists not wearing a helmet and no matter which bicyclist the AV chooses, worse injuries would be sustained.

The key paradox being highlighted here is that when an agent acts myopically in a utilitarian manner, the very framework that they used to make their decision condemns that decision post-hoc. In a perfect world with perfect information, this would not be an issue because the utilitarian calculation when making a decision would always be correct and hence the actual outcome would be the intended outcome. However, AVs do not operate in this perfect world, and hence would arrive at a contradiction whenever their utilitarian calculations with imperfect information lead to outcomes that utilitarianism itself condemns. Hence, to avoid such contradictions, it becomes necessary to seek an alternate ethical theory, such as Rawlsian contractarianism, to provide a logically consistent framework that AVs can use to ground their decisions.

In his seminal work *Theory of Justice (1971)*, Rawls argues that fair decisions are the ones that could be agreed upon by individuals behind the "veil of ignorance". If one does not know anything about their personal situation, it is in their best interest to devise rules that would disadvantage no one since otherwise there would be a possibility that they could be the person who would be disadvantaged. In the context of AVs, if one does not know which road user they would be — the passenger or the pedestrian — they would probably devise rules that balance the interests of both parties since there is an equal likelihood that they could be the pedestrian or the passenger. These rules would probably resemble current road rules such as having to cross a road at a pedestrian crossing when the traffic light permits it. The person in the original position would not want to let pedestrians cross anywhere anytime for if they were a passenger, they would be annoyed by the consequent lower speeds cars would have to maintain and similarly, they would not want there to be no pedestrian crossings at all, for if they were a pedestrian, they would want a place where they could safely cross the road. Hence, the social contract that currently exists between road users in the form of road rules can be deemed procedurally fair.

Thus, let's evaluate the jaywalking example articulated earlier in the paper through the lens of contractarianism. The fact that the pedestrians were jaywalking and not legally crossing at a crosswalk is morally relevant now. This is because, according to the social contract that exists between users of the road, pedestrians forsake certain liberties such as the liberty to cross the road wherever they please in order to be guaranteed other rights such as the right to life when they legitimately cross the road at a pedestrian crossing. When jaywalkers cross the road at a non-designated crossing, they are in violation of this social contract and hence, by extension, they are forsaking their right to life while crossing the road. It is thus fair then for the AV to prioritize the safety of its passengers — who have upheld their end of social contract and not committed any wrongdoings — over that of the pedestrians.

However, if the pedestrians happened to have been legally crossing, the scales tip unfavorably against the passengers. This is because in the absence of any breaches of the social contract, the implicit indemnification agreement assumed by passengers takes precedence. This is because of the two classes of stakeholders involved (the passengers and the pedestrians), it is the passengers who make the choice to ride in the AV and hence must accept the risks of doing so, including vehicular malfunction. When the situation arises where either the passengers or the pedestrians must be harmed, it would thus be unjust to inflict harm upon a class of persons who had no say whatsoever on the presence of the AV on the road and did not consent to any of the risks associated with the use of the AV. Hence, in this situation, it is only fair that the AV prioritize the safety of the pedestrians at the expense of the passengers.

Hence, on the question of whether an AV should be programmed to optimize for passenger safety or global human welfare, utilitarianism and Rawls' contractarianism provide divergent answers. Utilitarianism would unequivocally demand that AVs strive to optimize for global human welfare. However, it is unlikely that this ideal could ever be achieved due to the limited information that AVs have at their disposal about the downstream consequences of their actions. Contractarianism on the other hand does not suffer from such practical limitations. It would require that AVs make decisions that respect the social contracts that govern road use. Contractarianism has the advantage of being able to provide uncontradictory ethical justifications for decisions that AVs will have to make in an imperfect world unlike utilitarianism whose ethical justifications would only hold water for AVs that operate in utopic conditions, not the real world.

Another area where utilitarianism falls short is on the question of whether AVs that improve global human welfare but harm certain specific sub-populations (such as bicyclists, for example) should be deployed or not. Utilitarianism would resoundingly answer in the affirmative. However, the fatal flaw with this approach is what Rawls describes as utilitarianism's failure to "take seriously the

distinction between persons" [Rawls, 1971]. This is because "The striking feature of the utilitarian view of justice is that it does not matter, except indirectly, how th[e] sum of satisfactions is distributed among individuals" [Rawls, 1971]. In other words, a salient feature of utilitarianism is the principle of aggregation, i.e. lumping individuals together and allowing individual rights to be trampled to provide greater aggregate benefits to others [Ashford and Mulgan, 2018]. However, individuals, by virtue of being human beings, have certain inalienable rights that resist aggregation seeking to deprive them of their individual liberties.

To understand how contractarianism responds to the above question more satisfactorily, consider the following example: An AV company has developed two AV technologies. In tests, the first one results in a 50% reduction of motor-related deaths but consistently kills bicyclists at a higher rate than the status quo. The second technology has the same 50% reduction in motor-related deaths, but in tests, it randomly kills one group of road users at a higher rate than the status quo. Utilitarianism would not have qualms with the company rolling out either version of their AV technology since the outcomes have the same overall utility, regardless of the distribution [Ashford and Mulgan, 2018]. However, the contractarian would object to the deployment of the first technology because from the original position, one would realize that they have equal odds at being a bicyclist as they have at being any other road user and hence would prefer the second situation where harms are distributed evenly because there every individual's rights are respected. Hence, contractarianism prevails over utilitarianism for even though it accounts for some of the consequentialist philosophy of utilitarianism, it does not do so at the expense of individual natural rights.

Thus, through the above examples, I hope that I have convinced the reader that as humanity prepares itself for the advent of the autonomous era, it requires an ethical framework that is logically consistent and leads to moral outcomes. For the reasons articulated above, I believe that this framework is contractarianism and not utilitarianism.

References

[Allen, 2011] Allen, F. E. (2011). Name You Need to Know: Sebastian Thrun.

[Ashford and Mulgan, 2018] Ashford, E. and Mulgan, T. (2018). Contractualism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2018 edition.

[Dickey, 2020] Dickey, M. R. (2020). Cruise can now transport passengers in self-driving cars in CA.

[Rawls, 1971] Rawls, J. (1971). *A Theory of Justice*. Belknap Press.

[Sidgwick, 1871] Sidgwick, H. (1871). *The Methods of Ethics*. Macmillan.

[Sinnott-Armstrong, 2019] Sinnott-Armstrong, W. (2019). Consequentialism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition.