A Critique of the Moral Machine Critique

Francisca Vasconcelos

November 2019

In her paper, *Why the Moral Machine is a Monster*, Abby Jaques makes several arguments as to why the MIT Media Lab's Moral Machine experiment is misguided in determining the ethics of self-driving cars. In particular, she emphasizes the importance of one flaw in the logic, which she deems to be "fundamental." The Moral Machine asks users to select one option or another, such as hitting a homeless man versus a businessman, in a series of trolley-problem type case studies. In doing so, the experiment seems to suggest that individual decisions, made in overly-simplified scenarios, can be programmed into a generalized decision-making algorithm for self-driving cars. However, Jaques argues that the decision of whether to hit an older or younger person extends far beyond the lines of code it is implemented in. By embedding these sorts of decisions in autonomous vehicles, engineers would establish a new policy, affecting the rights of citizens in the society in which these self-driving cars ride in the streets. She gives a vivid example in which engineers choose to spare the young over the old. In this world, according to Jaques, grandparents would be unable to walk in the grocery stores parking lots, in constant fear of being run over by a car that is swerving to prevent hitting a child. Meanwhile, kids would run recklessly through the streets, knowing that the cars will do everything not to hit them. Although this is an exaggerated example, it does raise an interesting point about the implications of the ethics of self-driving vehicles, which is clearly missed in the Moral Machine experiment. Jaques claims that even if you do believe the young should be saved over the old, the Moral Machine has no way of assessing whether or not you agree with societal implications and potential injustices that would emerge from the *policy* established by writing code to do so.

Although Jaques' argument aims to provide insight into the effects of a Moral Machine-type policy, her overall view of the matter is misguided in much the same way as the Moral Machine itself. Throughout this paper, I will argue that both Jaques and the Moral Machine are focused on the wrong problem. Rather than worrying about who to kill in extreme edge cases, which will be negligible relative to the total number of lives saved by self-driving cars, we should instead focus on making the technology more robust and less biased. I will begin by (1) challenging the plausibility of Jaques' examples and policy argument. I will then (2) argue why people will not need to explicitly encode a killer decision-making system for self-driving cars.

1 Killer Self-Driving Cars: An Exaggeration

Jaques presents two main examples to illustrate the idea that decision-making code has large policy implications for society. The first is the previously-described example of grandparents and children in grocery store parking lots. The second example targets jaywalkers, who are technically committing a crime and, thus, more likely to be deserving of the blame. In this case, however, Jaques argues that jaywalking is not a crime worthy of the death penalty and we should not live in a regime in which people are extremely afraid of committing such a small felony. Although demonstrative of Jaques' point, these examples are contrived and unrealistic, weakening her overall claim.

In the case of jaywalkers, she makes it seem as though the self-driving cars are intentionally running over these "law-breaking" pedestrians, in order to penalize them for breaking the law. In reality, jaywalking laws are established primarily to protect the pedestrians, aiming to keep them out of dangerous situations in which the driver does not expect them. If someone does choose to jaywalk, they are intentionally putting themselves in a more dangerous situation and thus making it more likely that a car, either human or autonomously driven, would hit them. However, in this circumstance, a self-driving car is actually much better equipped to see the pedestrian with enough time to make an abrupt halt. Thus, in the case of jaywalking, we would expect cars to be less of an authoritarian regulating body than their human counterparts.

In the grocery store situation, Jaques makes it seem as though self-driving cars will be whizzing around, killing people left and right. In reality, self-driving cars will follow speed-limits and be even more law-abiding than human drivers. In fact, as previously described, the whole transition to self-driving cars is largely motivated by the belief that they will be safer and cause less crashes than human drivers, a result that Jaques even cites in her own paper [3]. Given that this is the case, would grandparents truly fear walking through the grocery store parking lot? The likelihood of getting hit by a self-driving car is already less than with a human driver. Furthermore, the Moral Machine made clear that, even in the case of some strange trolley problem type situation, a human driver would generally choose to save the child. Thus, a grandparent's odds of survival could only be the same or improved in such a contrived situation.

In conclusion, the policy effects Jaques is pushing for are not as rampant as she makes them out to be. In a world of self-driving cars, even those biased against jaywalkers or grandparents, we would argue that these two groups should actually feel all the more confident and safe in roaming the streets. Now, you may still be thinking that in the grocery-store case of a child-biased algorithm (in which grandparents are favored), the child's odds of survival are worsened and they should be afraid. However, I would like to once again point out that far more children's lives would be saved, simply because self-driving cars are better than human drivers. The total number of children dying because of a self-driving car's policy would be far outnumbered by the number of children dying because of an unbiased human driver. Thus, children would still be safe to roam the streets (but maybe should be a little less reckless).

2 Technology to Avoid Decision Making

Putting aside the improved safety statistics of self-driving cars, there still lies the displeasing notion of a car choosing to kill one person over another for superficial reasons, as promoted in the Moral Machine. In describing her structural view of the matter, Jaques hints at the notion that we should be putting less focus on these trolley-problem type edge cases and more on improving the actual safety of the vehicle.

> But recognizing these as structural issues, and thus looking at them more broadly, suggests ways to refuse the choice [of who to kill] as presented. We are likely to think about how we can minimize the chance that cars will be unable to stop when they detect a pedestrian... We might think that the right rule will be *chancy*, since once we've done all we can to minimize these situations it seems unfair to legislate winners and losers in the deeply luck-driven cases that remain... Even with a structural analysis there will still be hard choices. But when we talk about the choices as policy choices, we'll generally be better positioned

> to avoid the kind of basic errors the Moral Machine makes inevitable. [8]

She even acknowledges the fact that we if we put enough focus and regulation into the safety of these systems, the number of trolley-problem type scenarios (which could really only be caused by several major failures of brakes and swerving mechanisms, as well as their monitoring systems) would practically be zero. I agree with all these statements.

However, she claims that in the rare trolley-problem type situation, we could either chose to use a *chancy* type decision-making function or still would ultimately need to encode a choice. She also claims that the Moral Machine is harmful in the types of policy decisions that would need to be made surrounding such an argument. I disagree with both of these statements.

The problem of coding a self-driving car is extremely difficult and requires mimicking human-like intelligence. Although recent developments in machine learning, such as neural networks, have allowed us to train computers to perform certain tasks far better than humans, these algorithms are largely black-boxes. We have little understanding of what is going on under the hood and rely purely on the absurd amounts of data used to train these neural networks that they will learn to do the right thing. It is already hard enough to determine whether the neural network can actually recognize that there is a human near the car, let alone gauge its thoughts on saving a child versus a grandparent.

Our sole means of trying to better understand the system is by providing targeted input and seeing how it responds, or trying to understand and correct for known biases in the data. In such a situation, the Moral Machine actually becomes an invaluable tool. Rather than using the Moral Machine for it's intended purpose of trying to dictate the types of decisions the car should make, we can exploit the fact that it is overly simplified and extremely biased. The biases inherent to the Moral Machine are a strong indication of biases among the population that will almost surely be reflected in the data collected to train the neural networks. Thus, the Moral Machine makes us aware of improper moral and social biases that plague human drivers. This information can be provided to engineers, such that they can remove cases of bias from the training set.

Thus, Jaques should not be attacking the Moral Machine, nor trying to develop a new game for self-driving vehicle policy. Instead, she should leverage the so-called "fundamental flaw" of the Moral Machine to help engineers develop the bias-free moral policies she envisions.