Why the moral machine is a monster

Abby Everett Jaques

Somewhat startlingly for philosophers, a bit of philosophy has recently both gone viral and been treated as of significant real-world import. I mean, of course, the trolley problem. Originated by Foot (1967) and elaborated by Thomson (1976), this thought experiment has populated Intro to Philosophy classrooms and persisted in certain corners of ethical theorizing, but it only became meme material when the scenario seemed to come to life, thanks to the arrival of self-driving cars.

In this paper, I argue that this is, alas, a very bad thing. Despite appearances, the trolley problem is precisely the wrong tool for thinking about the question at hand, because it *misstates* the question at hand. There are lots of ways that's true; what I'll explore here is the deepest one. At bottom, the trouble is that the trolley problem frames the question as if all we need to do is figure out what an individual ought to do while driving, and then make that the rule for autonomous vehicles. As a long tradition of feminist philosophy and social science has shown, this kind of methodological individualism is never the right way to approach social phenomena.

In what follows, I'll bring out the problem and suggest what a better approach would look like. In §1, I'll describe the most famous version of the trolley problem approach to the ethics of autonomous vehicles: the Moral Machine experiment, created by researchers at the MIT Media Lab. In §2, I'll explain how the Moral Machine goes astray. In §3, I'll consider what's needed for a productive approach to these questions. In §4, I'll take stock.

1 The Moral Machine experiment

The Moral Machine project (Awad et al. 2018), is a game-like online platform that poses binary choices in scenarios involving self-driving cars that are going to crash (say, because the brakes have failed). Users must decide if the car should continue on its path or swerve to change course, where doing one or the other will affect how many people are killed; whether pedestrians or passengers are protected; whether people of different ages or social positions are favored; etc. The idea is that gathering data about people's choices can

inform the programming of autonomous vehicles, turning them into, as it says on the tin, moral machines.

In essence, then, the Moral Machine project seeks to *crowdsource* guidelines for the programming of autonomous vehicles. Users are presented with approximately 13 scenarios, and asked to choose one of two outcomes in each. The scenarios have the flavor of the *trolley problem*, the philosophical thought experiment in which one must choose whether to allow a runaway trolley to strike five workers on the track, or divert it to a different track where only one worker will be killed. The classic version of the trolley problem (a massive number of variants have been introduced in the half-century since the problem's introduction) involves two variables: the number of people killed, and the relative moral import of killing versus letting die. The Moral Machine incorporates quite a few additional variables:

- Saving more lives versus fewer
- Favoring passengers versus pedestrians
- Penalizing jaywalkers versus not doing so
- Intervening versus not intervening
- Favoring men versus women
- Favoring humans versus non-human animals
- Favoring the young versus the old
- Favoring 'fit' people versus 'large' people
- Favoring people based on 'social value' versus not doing so (where 'social value' is represented by identifying people as doctors, businesspeople, homeless people, or criminals)
- Favoring pregnant women versus non-pregnant people

The researchers aggregate users' responses and track the results. Millions of people, in hundreds of countries, have participated. According to the results recently published in *Nature*, the data indicates that people support minimizing loss of life and protecting children, favoring the fit and wealthy, and sacrificing people who are old, overweight, or homeless.

2 How the Moral Machine goes off the rails

As the researchers acknowledge, the preferences they have recorded diverge substantially from the principles endorsed by professional policymakers, ethicists, and clergy. And I take it to be clear that these results are seriously in error, morally speaking; many of them are actively repugnant. (No, we should not spare rich and target the poor. Trust me, I'm a professional ethicist.) So what's going wrong? To some extent, the answers are (I hope) obvious. But the point of this paper is to show that there are less obvious issues that are in fact crucial.

As I said, I am not the first to think that the trolley problem may be unhelpful for thinking about what autonomous vehicles should do. For example, the engineers actually building autonomous vehicles complain that this framing focuses on extremely rare edge cases rather than more pressing, because more likely, situations; and that talking this way is likely to frighten the public in ways that will impede adoption of a technology that should in fact make the roads safer overall.[1] And of course, the Moral Machine, collecting as it does descriptive information about what people say they prefer, is not obviously relevant at all to the question of what we *should* prefer. There are good reasons why mere aggregate majority preferences are not typically taken to be reliable indicators of what should be done, so that for example the U.S. Constitution enshrines protections for minorities under the democratic structure. I endorse those critiques and caveats. But they don't capture the central problem.

The Moral Machine, and its autonomous trolley problem kin, don't just overstate the frequency of such scenarios, or have potentially worrying side-effects. They *guarantee* that respondents will answer wrongly, because they ask the wrong questions.

To see why, first notice that the transactional, individualistic framing of the Moral Machine makes it tempting to answer that the car should, say, swerve to avoid hitting a doctor, sacrificing a homeless person instead. When you're presented with the highly schematized question of whether to swerve or stay the course, you'll latch onto *whatever* features are presented to try to find a differentiator. And the Moral Machine very often offers up categories that purport to capture 'social value' with no other distinguishing features.

1. See Iagnemma 2018 for a representative discussion.

And notice what this involves. The categories that are supposed to capture social value rank doctors at the top, businesspeople just below them, unflagged people in the middle ground, homeless people nearly at the bottom, and 'criminals' lowest of all. But why should we attribute increased social value to 'businesspeople' over those who are not identified by occupation? Is any old businessperson more socially valuable than, say, every teacher? Many homeless people are combat veterans; shouldn't we honor their sacrifice at least enough to resist the impulse to kill them in order to save a draft-dodging real estate bro? And what about this undifferentiated class of 'criminals?' Are we to understand that someone who has a conviction for marijuana use—who is far more likely to be a person of color than a white person, even though white people use drugs at higher rates—thereby becomes socially worthless, and to the same degree as a neo-Nazi mass shooter? These categories are morally contentless, and don't actually track anything about the 'social value' of particular individuals. Larry Nassar is a doctor; he also molested hundreds of young gymnasts.

And of course, if what we mean to capture with this talk of 'social value' is who has the more positive effect on the world, there is in fact no way for us to know that. To see this, consider a case adapted from James Lenman's (2000) work. Suppose we choose to spare a young doctor. She may be a kind and dedicated practitioner who saves many lives. But if one of the lives she saves is that of a man who goes on to become a new Hitler, then the world would have been better—in the kind of consequentialist way that talk of 'social value' is supposed to speak to—if the kind young doctor had died, because then Hitler II would have as well. And to *really* track down any person's social value, you'd have to know about every causal chain in which they participate, all the way to the heat death of the universe. Given an infinite universe and some chaos theory, this isn't just beyond our ken, it's beyond the reach of computation.

So the categories the Moral Machine is using just aren't suited to capturing 'social value;' social value isn't calculable anyway; and, as noted above, neither social value itself nor the particular features we're hoping to use as proxies for it are the kind of thing you can tell by looking at someone. There are cancer researchers and artists who dress like homeless people, and serial killers in Savile Row suits; appearance won't even give you the inadequate categories the machine is using, much less differentiate the businessperson who is financing microloans to women in the Global South from the hedge-

funder who's just thrown thousands of people out of work on his way to crashing the global economy. Humans are very bad at judging people based on appearances—we are swayed by physical beauty to an appalling degree, for example. (See Hamermesh 2011.) But there is no reason to think autonomous vehicles could do it at all.

In short, if we wanted to favor those of 'higher social value' in our scenarios, we are not equipped to implement such a strategy. However, it's also just misguided to think that we should try. There is a reason that the moral equality of persons is table stakes in moral and political theorizing.[2] Which makes it all the more concerning that the Moral Machine gives users basically nothing else to go on: it thereby ensures that at best their choice will be morally arbitrary; more likely (and borne out by the data) is that they'll choose from bias. But the fact that the bias here exists antecedently in the users doesn't mean the Moral Machine itself is neutral, a mere pass-through. As I've been saying, in identifying choices in terms of peoples' fitness or lack thereof, homelessness, profession, and so on, making those features salient in the context of choice and denying other materials for practical reasoning, the Moral Machine suggests and even advocates for the moral relevance of those features. When the Machine asks users whether to swerve and kill the large woman or go straight and kill the fit man, it *invites* users to express their biases. So the underdescribed scenario, in which choice is forced and the only resources are irrelevant, inevitably leads to bad results.

But this is still not the fundamental problem. Suppose we remove obviously morally irrelevant (and/or imperceptible) features from the description of the situation. We might still think we need to decide if the car should, for example, choose to go straight and hit a pedestrian in a crosswalk or swerve and hit a bystander. We might even think it's relevant whether or not the pedestrian is jaywalking. Part of what we're wondering is what features really are relevant to our decision—so even if we think some of those presented by the Moral Machine are definitely not to be considered, we have to make decisions about what features do count, morally. So what's the real issue, and how do we avoid it?

In a slogan, the problem is that an algorithm isn't a person, it's a policy. And you don't get policy right by just assuming that an answer that might be fine in an individual case will generalize. You have to look at the overall

2. Kymlicka (1990, 5) provides a representative discussion.

structure you create when you aggregate the individual transactions. But a trolley problem methodology like the Moral Machine's makes the structural features of one's answer invisible. The right question isn't what would I do if I were forced to choose between swerving and going straight. The right question is what kind of world will I be creating if this is the *rule*. What will the *patterns* of advantage and disadvantage be if this preference is encoded in our autonomous vehicles?

The Moral Machine and its trolley-loving cousins go wrong because they obscure the actual choice. They make the individual transaction a stand-in for the rule, but looking at a single transaction is not adequate to perceive the relevant properties of the scenario, just as looking at a single tree is not adequate to characterize a forest. What is the actual choice that's being obscured? We can see it if we move from thinking about individual transactions—single instances of swerving or not—to thinking about the structure of the world under particular rules.

As Anderson (2012, 164) helpfully puts it:

> The cumulative effects of a series of transactions, each of which satisfies the local criteria of justice, and which begins from a just starting point, may be disastrous... A structural theory supplies criteria for assessing global properties of a system of rules that govern transactions, and imposes constraints on permissible rules with an eye toward controlling the cumulative effects of individual transactions that may be innocent from a local point of view.[3]

Anderson's point is supported and strengthened by the discussion in Garfinkel (1981); he shows that looking only at the level of individual transactions can *never* yield adequate accounts of social phenomena. His argument is complex, but very roughly the issue is that capturing features at the individual level, even aggregated, will fail to provide a complete account because those features are not *independent*—so, for example, peoples' preferences depend on what others' preferences are, and this relative property can't be captured at the individual level. Social phenomena, he argues, always involve this kind of lack of independence, and thus are not amenable to individualistic, or reductivist, accounts.

3. A paradigmatic structural theory of the kind Anderson has in mind is Rawls' (1999) theory of justice.

This is all a bit obscure when discussed in the abstract. So to see what we gain by abandoning the individualistic, transactional, trolley problem approach of the Moral Machine and thinking structurally instead, let's take a look at what's revealed when we shift levels.

3 The solution: a structural approach

One claim I'm making is that it's the transactional, individualistic framing of the Moral Machine that makes it tempting to answer that the car should, say, swerve to avoid hitting a baby in a stroller, sacrificing an elderly person instead—and that the question would look very different if we took it up at the level of structural analysis. At that level, remember, we would be asking: what kind of world would we be creating if this preference was encoded in our autonomous vehicles, if it was our *policy*?

And at that level of analysis, we can see that encoding this rule would mean creating a world in which the elderly lose access to the public square. Consider what a supermarket parking lot would be like: a place where children dart around with impunity, knowing they won't be harmed, but where the elderly are therefore almost certain to be run over, as all the autonomous vehicles veer wildly through the space trying to avoid the children. At the structural level, it's easy to see that making some personal feature like wealth or fitness or age the basis for choice is a non-starter. You just have to notice how impossible it is to claim that people don't have a right to, say, go to the grocery just like anyone else.

But structural analysis doesn't just help with the features we (hopefully) already knew were irrelevant. It also helps us think through what it means if we take other features to be decisive. Suppose we're wondering whether our AV should be programmed to avoid pedestrians in the street even if it means swerving onto a sidewalk. If that's our rule, then it will make sense for *all* pedestrians to walk in the street, because being on the sidewalk will be too risky. So that can't be right.

Given that, maybe we want to say that it matters if the pedestrian is jaywalking. So a pedestrian is to be avoided if she's crossing in a crosswalk, with the light, etc; but not if she's flouting the rules. Structural analysis reveals that even if that seemed tempting in an individual transaction, it too looks bad when you realize as a policy it amounts to implementing the death

penalty for jaywalking.

It's worth noting that in the kinds of urban environments in which jaywalking is common—like Manhattan—traffic is so dense that if pedestrians *don't* jaywalk they may never be able to cross the street at all.[4] But suppose we grant that jaywalking is an infraction. It is patently not the kind of infraction for which the death penalty is appropriate. People do small wrong things; they do not deserve to die for them. What we are theorizing about are how to manage traffic accidents. There is not going to be a way to calculate who deserves to die in one, because no one does.

What this discussion brings out is that structural analysis won't always make the right answer immediate or easy. But unlike the trolley problem approach, it enables the necessary thinking. Because sometimes, when you think structurally, you see that what needs to change *is* the structures.

Structural analysis will rightly make us worried that sparing all pedestrians incentivizes walking in the street. It will likewise make us worry that *targeting* all pedestrians is indefensible. But recognizing these as structural issues, and thus looking at them more broadly, suggests ways to refuse the choice 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—perhaps by adding guardrails to prevent crossing outside of crosswalks (and of course making sure that there are *adequate, well-placed* crosswalks) and ensuring excellent sightlines around them. 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. In short, we'll be inclined to notice that the *social and material environment* are factors under our control, and we'll be more apt to recognize fruitful and justifiable options that undermine the original choice scenario instead of grasping at any feature within that situation that happens (or purports) to be perceptible.

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. Where the Moral Machine's authors talk about dilemmas and force a choice, at the

4. And that historians will tell us that the concept of jaywalking was invented by automakers, as Stromberg 2015 explains.

structural level we see that more features than we might suppose are up to us.

4 Conclusion

The Moral Machine is a good-faith effort to help us think about how autonomous vehicles should be programmed. But unfortunately it's also deeply unhelpful, all the moreso because it received the imprimatur of *Nature* and then went viral, so that it now drives the discussion about the ethics of AVs. If we want to get the discourse back on track, we may well need not just a way of framing the question that avoids the Moral Machine's methodological errors, but also one that shares the Moral Machine's clickbaity features. Can we make a gamified interface for the structural questions? Can we make it as much fun to think about the richer, more complex issues that a structural analysis reveals? I don't yet know the answer to that question, but I'm working on it.

References

Anderson, Elizabeth. (2012) "Epistemic justice as a virtue of social institutions." *Social Epistemology* 26(2): 163–173.

Awad, Edmond, et al. (2018) "The Moral Machine experiment." *Nature* 563: 59-64.

Foot, Philippa. (1967) "The problem of abortion and the doctrine of double effect." *Oxford Review* 5.

Garfinkel, Alan. (1981) *Forms of Explanation*. New Haven, CT: Yale University Press.

Hamermesh, Daniel. (2011) *Beauty Pays: Why attractive people are more successful*. Princeton: Princeton University Press.

Iagnemma, Karl. (2018) "Why we have the ethics of self-driving cars all wrong." World Economic Forum Annual Meeting. Available online at https://medium.com/world-economic-forum/why-we-have-the-ethics-of-self-driving-cars-all-wrong-92566f282733

Kymlicka, Will. (1990) *Contemporary Political Philosophy*. Oxford: Clarendon Press.

Lenman, James. (2000) "Consequentialism and cluelessness." *Philosophy & Public Affairs* 29 (4): 342–70.

Rawls, John. (1999) *A Theory of Justice*. Cambridge, MA: Belknap Press.

Stromberg, Joseph. (2015) "The forgotten history of how automakers invented the crime 'jaywalking." *Vox.com*. Available online at https://www.vox.com/2015/1/15/7551873/jaywalking-history

Thomson, Judith J. (1976) "Killing, letting die, and the trolley problem." *Monist 59: 204–17*.