AI and Ethics (2021) 1:151–155
https://doi.org/10.1007/s43681-020-00018-z

ORIGINAL RESEARCH

MIT's moral machine project is a psychological roadblock to self-driving cars

Heidi Furey[1] · **Scott Hill**[2]

Received: 29 July 2020 / Accepted: 23 September 2020 / Published online: 6 October 2020

Abstract

In the moral machine project, participants are asked to form judgments about the well-known trolley example. The project is intended to serve as a starting point for public discussion that would eventually lead to a solution to the social dilemma of autonomous vehicles. The dilemma is that autonomous vehicles should be programed to maximize the number of lives saved in trolley-style dilemmas. But consumers will only purchase autonomous vehicles that are programed to favor passenger safety in such dilemmas. We argue that the project is seriously misguided. There are relevant variants of trolley to which the project's participants are not exposed. These variants make clear that the morally correct way to program autonomous vehicles is not at odds with what consumers will purchase. The project is hugely popular and dominates public discussion of this issue. We show that, ironically, the project itself is largely responsible for the dilemma.

Keywords Moral machine project · Trolley problem · Autonomous vehicles

In the *moral machine project*[1], participants are asked to form judgments about variations of this well-known example:

> *Trolley*: There is a runaway trolley. If you do nothing, the trolley will hit and kill five people. If you pull a lever, the trolley will be diverted onto another track and kill one person.

The variations replace trollies with autonomous vehicles and people on tracks with passengers and pedestrians. The project is intended to serve as a starting point for public discussion that would eventually lead to a solution to the *social dilemma of autonomous vehicles*. The dilemma is that autonomous vehicles should be programed to maximize the number of lives saved in *trolley*-style dilemmas. But consumers will only purchase autonomous vehicles that are programed to favor passenger safety in such dilemmas. As the project's searchers (2016, p. 1575–6) put it:

> Moral algorithms for AVs create a social dilemma (18, 19). Although people tend to agree that everyone would be better off if AVs were utilitarian (in the sense of minimizing the number of casualties on the road), these same people have a personal incentive to ride in AVs that will protect them at all costs. Accordingly, if both self-protective and utilitarian AVs were allowed on the market, few people would be willing to ride in utilitarian AVs, even though they would prefer others to do so. Regulation may provide a solution to this problem, but regulators will be faced with two difficulties: First, most people seem to disapprove of a regulation that would enforce utilitarian AVs. Second—and a more serious problem—our results suggest that such regulation could substantially delay the adoption of AVs, which means that the lives saved by making AVs utilitarian may be outnumbered by the deaths caused by delaying the adoption of AVs altogether. Thus, carmakers and regulators alike should be considering solutions to these obstacles.

We argue that the project is seriously misguided. There are relevant variants of *trolley* to which the project's participants are not exposed. These variants make clear that the morally correct way to program autonomous vehicles is not at odds with what consumers will purchase. The project is hugely popular and dominates public discussion of this

[1] For representative papers, please see [1,2,3,11,14].

✉ Scott Hill
hillscottandrew@gmail.com

[1] Department of Philosophy, Manhattan College, 4513 Manhattan College Parkway, Riverdale, NY 10471, USA

[2] CU Boulder Philosophy, Hellems 169 UCB 232, Boulder, CO 80309-0232, USA

issue. We show that, ironically, the project itself is largely responsible for the dilemma.

> Consider a well-known variant of *trolley*:
> *Large man*: There is a runaway trolley. If you do nothing, the trolley will pass through a tunnel, come out the other end, and hit and kill five people stuck to the tracks. You are standing above the entrance to the tunnel right behind a very large man. If you push the large man onto the tracks before the trolley enters the tunnel, it will hit can kill him. But it will stop before exiting the tunnel and the five will be spared.'

Most philosophers think it is wrong to push in this case. Non-philosophers tend to agree. There is much controversy about *why* it is wrong to push. But not *whether* it is wrong. Next, consider:

> *Children*: There is a runaway trolley. If you do nothing, it will kill five strangers. If you pull a lever, it will be diverted onto another track and kill your four children.

Most people agree that it is permissible for you to refrain from killing your children. This example elicits the intuition that we have stronger obligations to family than to strangers. Consider another example[2]:

> *Seatbelt*: An autonomous vehicle is unable to stop. If the programing does nothing, it will crash into another vehicle containing a passenger who is not wearing a seat belt. If the programming diverts, it will crash into a vehicle containing a passenger that is wearing a seatbelt.

Programming an autonomous vehicle in such a way as to maximize the number of lives saved would require programming it to divert into the passenger wearing a seatbelt. That seems immoral. *Seatbelt* elicits the intuition that acting in such a way that discourages recklessness and criminality matters. These examples show that manufacturers and policy-makers need not choose between autonomous vehicles that behave morally and autonomous vehicles that the public will accept. Moral programming does not coldly aim to maximize the number of lives spared without taking into account the relevance of how lives are spared, who is spared, and whether recklessness is encouraged.

There are other psychologically relevant factors ignored by the project.[3] Order of presentation changes how subjects assess *trolley*-type cases.[4] But the project does not control for order. Details change how subjects assess *trolley*-type cases.[5] But the project leaves details ambiguous. For example, in one *trolley* variant, participants are shown a figure with a medical symbol. Without further information, some participants may think that the figure is a doctor while other participants may think the figure is a patient. Indeed, one of us asked our relatives to participate in the experiment. And exactly this confusion arose causing the relatives to arrive at different judgments about what to do.

The project is closely associated in the public imagination with autonomous vehicles. It is the only exposure some manufacturers and policy-makers have to this topic. Unfortunately, the project makes *trolley*-style scenarios more salient in the public imagination than ordinary causes of vehicle accidents. And it only makes salient *trolley* variants that elicit the intuition that maximizing the number of lives spared matters. However, very few current accidents are the result of *trolley*-style scenarios. Most accidents now, without the widespread adoption of autonomous vehicles, are the result of human error. If autonomous vehicles were universally adopted, most of the relevant human decisions would be replaced by much more reliable automated decisions. Even if autonomous vehicles were programed in such a way that they make the wrong decision in *trolley*-style cases, the number of lives saved by the implementation of automated decision-making would vastly outweigh the lives lost as a result of widespread adoption of autonomous vehicles. By making *trolley*-style scenarios so unrepresentatively salient, and by focusing only on one narrow and misleading class of such scenarios, the Project has caused the public to think they must choose between purchasing an autonomous vehicle that protects their family and one that is moral. This has created unnecessary public resistance to autonomous vehicles and, if left uncorrected, will cause unnecessary deaths.

1 Objections and replies

First objection: We seriously misunderstand (or take too seriously) the moral machine experiment, which was meant to only help start a wider conversation about the many value judgments involved with programming automated cars. It was not meant to arrive at any conclusions about how these vehicles should be programmed, since ethics is more than a survey of preferences. For example: a population might overwhelmingly support genocide, slavery, torture, etc., but that fact by itself does not make those atrocities ethical.

[2] This example is inspired by an example used in Lin [9].

[3] The literature covering these issues is surveyed in Chapter 2 of Machery [11].

[4] See Petrinovich and O'Neill [15], Lanteri et al. [7], Lombrozo [10], Wiegman et al. [18], and Liao et al. [8].

[5] As well as the work discussed in the previous footnote, see Nadelhoffer and Feltz [12], Cikara et al. [5], Strohminger et al. [17], Pastotter et al. [14].

Reply: Members of the project have developed an algorithm that takes the data set generated by the project and delivers judgments about what autonomous machines should do. As they (2017, p. 20) put it:

> The design of intelligent machines that can make ethical decisions is, arguably, one of the hardest challenges in AI. We do believe that our approach takes a significant step towards addressing this challenge. In particular, the implementation of our algorithm on the moral machine dataset has yielded a system which, arguably, can make credible decisions on ethical dilemmas in the autonomous vehicle domain.

The purpose of their algorithm is to take surveys of public opinion as input and give moral decisions as output. As they put it in the abstract of their paper on this topic:

> We present a general approach to automating ethical decisions, drawing on machine learning and computational social choice. In a nutshell, we propose to learn a model of societal preferences, and, when faced with a specific ethical dilemma at runtime, efficiently aggregate those preferences to identify a desirable choice. We provide a concrete algorithm that instantiates our approach; some of its crucial steps are informed by a new theory of swap-dominance efficient voting rules. Finally, we implement and evaluate a system for ethical decision making in the autonomous vehicle domain, using preference data collected from 1.3 million people through the moral machine website.

It is clear that researchers behind the project think the purpose of the data is, at least in part, to help create AI that will make moral decisions.

Second objection: We pin most of the blame for public attitudes (or anxiety and confusion) about automated cars on the moral machine experiment, yet that experiment (launched in 2016) was not close to being among the first conversations about the ethics of automated vehicles, which have been going on globally since at least 2012: https://www.newyorker.com/news/news-desk/moral-machines. The experiment did help to raise awareness about the issues, though, but not nearly to the extent implied in this paper.

Reply: Two points. First, if the objector is correct, then this only increases the interest of our paper. We have argued the project is based on misunderstandings about the relevance of and the ethics of *trolley*-style dilemmas. If these misunderstanding were already widespread before the project, and if others have the same misunderstandings, then all the more reason for us to step in and correct such misunderstandings. Almost all of the popular interest in self-driving cars focuses on *trolley*-style dilemmas. It is a confusion that needs to be corrected whatever its source. Second, although the discussion about self-driving cars has been going on globally for some time. Attention skyrocketed right around the time the project began in 2016. The top Google search on the ethics of self-driving cars was the MIT moral machine project (and it's still in the top 3–4).

Third objection: There is a pervasive and unhelpful confusion throughout the paper in how the terms 'moral' and 'utilitarian' are used. We slip from one to the other leaving the reader not quite sure how the authors are defining the terms and if the authors regard moral and utilitarian as perhaps entirely synonymous. We say that "The Dilemma is that autonomous vehicles should be programmed to maximize the number of lives saved….But consumer will only…." The word 'should' gives the strong impression that we are proclaiming that the simple utilitarian approach which "maximizes the number of lives saved" is what 'should' be programmed and is 'morally correct'.

Reply: We didn't mean to endorse utilitarianism or to suggest that utilitarianism is the only viable theory. Although both of us think that utilitarianism is an important moral theory and is worthy of exploration and development, neither of us accepts it. What we meant to do in stating the dilemma is to simply report what the project takes to be a dilemma. Our considered view is that the dilemma is not genuine because people do not have (simple) utilitarian intuitions about trolley cases. but the *project* is set up in such a way that it illicits simple utilitarian intuitions. So we reject simple utilitarianism. But we think the reason consumer preferences and moral design of autonomous vehicle do not come apart is because moral design of autonomous vehicle does not match simple utilitarianism. It only seems like it matches simple utilitarianism when we focus on the narrow range of cases that the project exposes us to.

Fourth objection: If we are not utilitarians, then what is our view? We owe an explanation.

Reply: When doing applied ethics, we think it is a mistake to adopt and endorse one particular normative ethical theory. Here we adopt the approach advocated by Brennan [4], and Temikin [17]. As Furey et al. [6] put it:

> The best two theories in physics are General Relativity and Quantum Mechanics. General Relativity is good at explaining the behavior of big things like planets and stars. It has implications for the behavior of very small things like subatomic particles. But it is much less reliable in that domain. Quantum mechanics is very good at explaining the behavior of subatomic particles. It has implications about the behavior of big things like planets and stars. But it is less reliable than it is about big things.
>
> General Relativity and Quantum Mechanics are inconsistent with one another. They describe the universe in radically different ways. Physics use both theories but for different purposes. No one has been able to

> unify them into a single theory even though they have tried. In light of such failure, physicists simply use each theory where it is strongest.
>
> Some philosophers have suggested that we should think about moral theories in the way that physicists think about their theories. None of our moral theories are good enough to count as the One True Moral Theory. But we have consequentialist theories that are pretty good at explaining some aspects of morality. And we have deontological theories that are pretty good at explaining other aspects of morality. If we are interested in applied ethics, we should just use each theory where it is most successful. Moral theories are useful tools. But no single moral theory perfectly captures morality.

And that is what we think about the case at hand. Right now we do not have a moral theory that is good enough to capture all aspects of morality. But different moral theories are useful for different purposes. One of the things that concerns us about the project is this: it only employs examples that illicit utilitarian intuitions. But there are other examples that illicit non-utilitarian intuitions. Imagine a physicist defended general relativity by appealing only to observations that support general relativity and ignoring all the observations that support quantum mechanics. That would be a mistake. We think that is exactly the sort of mistake that the *project* makes. Let us be clear, we think it is worth continuing to do theoretical ethics and to look for the one true moral theory. This is just as it is worth continuing to do theoretical physics and looking for the one true physical theory. But we do not think anyone has found that theory yet. And we don't think the right way to do applied ethics is to just assume one of the theories is true and run with it. It is better to use each theory where it is strongest. And better to take account of the variety of intuitions about morality that do not fit with any one particular theory.

Fifth objection: We have shown that there is a gap between a utilitarian solution which 'maximizes the number of lives saved' and a moral decision that people will recognize and accept. What we have not shown or discussed is how such nuances could be incorporated into a split second algorithmic decision. Do we think it is in theory possible?

Reply: We think this objection raises a number of interesting issues. The issues are very big for us to fully address here. And neither of us are computer scientists. So we are not competent to answer the question of how the details of programming autonomous vehicles might work. Nevertheless, we are happy to speculate a bit.

One way of taking the objector's question is this: If we adopt the relevant form of utilitarianism advocated by the *project*, then one might think there is a simple way to program autonomous vehicles. There is just one simple rule: *maximize the number of lives saved*. On the other hand, if one agrees with us that that simple rule fails to match what is moral in a bunch of other cases, then is there another simple rule that the AI could be given in place of simple utilitarianism? We do not think there is such a simple rule. Or, at least, we don't know what one would be. However, programs can do more than follow a single simple rule. Programs can also follow ugly, disjunctive, and complex rules. So we would recommend, instead of a simple rule, that the program follows an ugly disjunctive rule.

Sixth objection: We say that 'Moral programming does not coldly aim to maximize the number of lives spared without taking into the relevance of how lives are spared, who is spared, and whether recklessness is encouraged.' Whose moral programming are we referring to?

Reply: The aim of the project is to get at what ordinary people think about the morality of autonomous vehicles. We think the project fails to do this. There are *trolley*-like examples about which philosophers have formed a consensus. And work by psychologists shows that that consensus is matched by a consensus among ordinary people about what to do in such examples. The consensus of philosophers and the consensus of ordinary people discovered by psychologists does not match what the researchers behind the project claim is the consensus of the project's participants.

Now, there are difficult questions about moral epistemology here. And it is really hard to say how we (philosophers and ordinary people) can come to know moral truths. (Although, see the introduction to [anonymized] for our own take on this topic). But as difficult as these questions are, we think we can side step such issues here. The project claims to have discovered what ordinary people think about such cases and alleges that what ordinary people think conflicts with what they will buy. We think the project doesn't accurately capture what ordinary people think. And what such people think is moral doesn't conflict with what they will purchase.

Seventh objection: What is our original contribution here? As we point out, the *trolley* variants we use have been discussed in the literature already. So what do we add?

Reply: Our original contribution is not to theoretical ethics. As the objector points out, we have not come up with a new variant of *trolley* or anything like that. Instead, we think our contribution is this: the project claims that what consumers will purchase is different than what consumers think is moral. But we show that this is not true. Lots of people have this same misunderstanding. We think it is valuable for us to point this mistake out.

Eighth objection: We need to be more forthright in saying what 'will bring more harm' or 'seriously misguided' means.

Reply: We think that the project is seriously misguided in the following way: it only illicits intuitions about a narrow range of *trolley* variants. Go back to the analogy with

a physicist who defends general relativity by considering only the data that supports general relativity but not being attentive to data that instead supports quantum mechanics. It would be fair to say that such a physicist is seriously misguided. In the same way, it is fair to say that the project is seriously misguided.

We think the project is potentially harmful in the following way: It makes salient only *trolley* variants that illicit utilitarian intuitions. And it makes it seem like there is a conflict between the cars people want and the cars people think are moral when there really is not. Without correction, that would make manufactures, policy-makers, and consumers unnecessarily weary of autonomous vehicles. That would delay the adoption of such vehicles. And that would lead to more deaths than if people knew there was really no conflict between what consumers will purchase and what they think is moral.

Funding This material is based upon work supported by the National Science Foundation under Grant No. SES–1734521.

References

1. Shariff, A., Bonnefon, J.F., Rahwan, I.: Psychological roadblocks to the adoption of self-driving vehicles. Nat. Hum Behav. **1**, 694–696 (2017)
2. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., Rahwan, I.: The moral machine experiment. Nature **563**, 59–64 (2018)
3. Bonnefon, J.F., Shariff, A., Rahwan, I.: The social dilemma of autonomous vehicles. Science **352**(6293), 1573–1576 (2016)
4. Brennan, J.: The Best Moral Theory Ever: The Merits and Methodology of Moral Theorizing. Dissertation. University of Arizona (2007)
5. Cikara, M., Farnsworth, R.A., Harris, L.T., Fiske, S.T.: On the wrong side of the trolley track: Neural correlates of relative social valuation. SocCognit Affect Neurosci **5**, 404–413 (2010)
6. Furey, H., Hill, S., Bhatia, S.: Beyond the Code: A Philosophical Guide to Engineering Ethics. Routledge, London (2021)
7. Lanteri, A., Chelini, C., Rizzello, S.: An experimental investigation of emotions and reasoning in the trolley problem. J. Bus. Ethics **83**, 789–804 (2008)
8. Liao, S.M., Wiegmann, A., Alexander, J., Vong, G.: Putting the trolley in order: experimental philosophy and the loop case. PhilosPsychol **25**, 661–671 (2012)
9. Lin, P.: The ethical dilemma of self-driving cars [Video file]. Retrieved from https://www.ted.com/talks/patrick_lin_the_ethical_dilemma_of_self_driving_cars?language=en#t-140814 (2015)
10. Lombroso, T.: The role of moral commitments in moral judgment. CognitSci **33**, 273–286 (2009)
11. Machery, Edward: Philosophy within its Proper Bounds. Oxford University Press, Oxford (2017)
12. Nadelhoffer, T., Feltz, A.: The actor–observer bias and moral intuitions: adding fuel to Sinnott-Armstrong's fire. Neuroethics **1**, 133–144 (2008)
13. R. Noothigattu, S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, A. D. Procaccia (2017). A voting-based system for ethical decision making. (arXiv)
14. Pastötter, B., Gleixner, S., Neuhauser, T., Bäuml, K.H.T.: To push or not to push? Affective influences on moral judgment depend on decision frame. Cognition **126**, 373–377 (2013)
15. Petrinovich, L., O'Neill, P.: Influence of wording and framing effects on moral intuitions. Ethol. Sociobiol. **17**, 145–171 (1996)
16. Strohminger, N., Lewis, R.L., Meyer, D.E.: Divergent effects of different positive emotions on moral judgment. Cognition **119**(2), 295–300 (2011)
17. Temkin, L.: Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning. Oxford University Press, Oxford (2012)
18. Wiegmann, A., Okan, J., Nagel, J.: Order effects in moral judgment. PhilosPsychol **25**, 813–836 (2012)