Doubting driverless dilemmas

Julian De Freitas[1*], Sam E. Anthony[2], and George A. Alvarez[1]

[1] Department of Psychology, Harvard University

[2] Perceptive Automata, Inc.

[*] Corresponding author

Abstract

The alarm has been raised on so-called ‘driverless dilemmas’, in which autonomous vehicles will need to make ethical decisions on the road. We argue that these ideas are too contrived to be of practical use, represent an incorrect model of proper safe decision making, and should not be used to inform policy.

Keywords: Moral judgment; autonomous vehicles; driverless policy

Recent prominent articles within academia and beyond (Awad et al., 2018; Bonnefon, Shariff, & Rahwan, 2016; Bonnefon, Shariff, & Rahwan, 2018; Donde, 2017; Edmonds, 2018; Greene, 2016; Gogoll & Müller, 2017; Lester, 2019; Lin, 2013; Markoff, 2016; Nowak, 2018; Noothigattu et al., 2018; Shariff, Rahwan, & Bonnefon, 2016) warn that autonomous vehicles (AVs) will face moral dilemmas, in which they "need to decide how to divide up the risk of harm between the different stakeholders on the road" (Awad et al., 2018). This work proposes to solve the problem by asking people on the web to consider simple thought experiments (traditionally known as trolley dilemmas; Foot, 1967) in which an AV faces a two-alternative forced-choice between whom to kill or save — e.g., a driver vs. a pedestrian; a group of cats vs. a group of humans; a homeless man vs. a skilled workman. They ask people to choose on the AVs behalf, then they aggregate these choices to assemble a 'global preference' scale, which they argue should inform AV policy.

Many of these projects are impressive in scope, ambition, and creativity, contribute valuable cross-cultural datasets on people's moral intuitions, and have served as good conversation starters for machine ethics. That said, the thought experiments they employ are too contrived to be of practical use, represent an incorrect model of proper safe decision making, and do not reflect the publics' opinion.

Trolley Dilemmas Are Utterly Unlikely

The whole point of the two-alternative forced-choice in the thought experiment is to simplify real world complexity. But such situations are vanishingly unlikely in the real world. This is because they require a perfect 50-50 chance of killing each individual in the same amount of time, with no other location to steer the vehicle, and no other possible steering maneuver but driving head-on to a death. Further, these dilemmas assume a fundamentally paradoxical situation: An AV has ample freedom to make a considered decision about whom of two pedestrians to hit, yet does not have enough control to instead take some simple action — like swerving or slowing down — to avoid hitting either pedestrian.

Lacking in these discussions are realistic examples or evidence of situations where human drivers had to make such choices, making it immature to consider them as part of any practical engineering endeavor. Even the authors of these papers seem aware of this, admitting, for example, that "it is extremely hard to estimate the rate at which human drivers find themselves in comparable situations" yet they nevertheless say, "Regardless of how rare these cases are, we need to agree beforehand how they should be solved" (Awad et al., 2018). We disagree. Without evidence (i) that such situations occur, and that (ii) planning for them improves overall safety, it is irresponsible to consider them when making AV policies or regulations.

Trolley Dilemmas are a Distraction

We don't teach humans how to drive by telling them whom to kill if faced with a forced choice, since the idea that one would specifically plan for such a situation is absurd. Instead of distracting would-be drivers with such irrelevant theoretical considerations, we intensively focus on teaching them how to avoid harm in the first place.

The same goes for teaching machines to drive safely. The main safety goal for any driver —human or machine — is to avoid harm. In fact, engineers at AV companies are *already* focused on achieving this goal (Olson, 2018). Unfortunately, both humans and today's best computer systems are imperfect at it. Even so, the substantial improvements that we rightfully expect from future AV systems are utterly unlikely to come from considering trolley-like dilemmas.

None of the 4 AV companies we polled (May Mobility, nuTonomy, Perceptive Automata, and a global automobile company that asked to remain anonymous) have teams or budgets devoted to solving trolley-like dilemmas, despite driving AVs on real roads every day. (We asked: (1) “Do you have anybody at [Company name] working specifically on how to resolve ‘trolley problem’ type dilemmas? If so, are you doing this based on the social category to which a person belongs?” and (2) “What percent of your budget would you say is devoted specifically to this problem?”). In the words of nuTonomy co-founder and CTO, Emilio Frazolli: “I consider ‘trolley problems’ one of the red herrings plaguing the world of AVs, distracting from the real issues”. And just last month the CEO of May Mobility, Edwin Olson, published a piece entitled “Trolley Folly” (Olson, 2018). All companies said that teaching AVs to solve trolley-like dilemmas would be foolhardy and irrelevant to AV safety— worse, it could be actively counterproductive, especially if enforced by officials ignorant of the real engineering challenges. They also rolled eyes at the glaring practical, ethical, and legal problems of choosing whom to kill based on a person’s social category.

Recent articles have managed to raise the alarm on so-called driverless dilemmas by capitalizing on the public’s understandable, yet unfounded, tendency to moralize new technologies because of their scary unfamiliarity (Assis, 2018). Instead of stoking these flames with distracting thought experiments, we should empower safety engineers to continue improving at the main goal of minimizing harm. As science communicators, we should reassure the public that safety engineers are already working on the *correct* safety goal, while being guided by a combination of professional safety codes and strong incentives to safeguard the reputation and legal liability of their companies. As scientists, we should focus on the relevant, concrete challenges remaining in the path toward fully safe AVs.

The Experiments Do Not Accurately Reflect Anyone's Opinion

Even if trolley dilemmas were relevant to real-world safety concerns, experiments involving them should not inform policy because they do not represent anyone's "opinion" — certainly not an expert one. For instance, Awad et al., (2018) assume that their 'global preference' scale provides "essential topics to be considered by policymakers". Yet, this scale is derived from contrived, two-alternative force-choice questions that corner participants into picking an option, even if they disagree with the entire premise of the experiment. When we plainly asked the same sorts of web participants (N = 129, M_{age} = 36, 49% female) if they thought AVs should use social preference scales to solve moral dilemmas, fewer than 20% said yes (Figure 1). And even if most had said yes, it is misguided to assume that the gut feelings of a group of people on the web who give a few seconds of thought to exotic, cartoonified scenarios provides a sound basis for policy governing AVs in the real world. These people are unlikely to be morally consistent (Bonnefon, Shariff, & Rahwan, 2016), have given little thought to the issues, know nothing about the legal, moral, and practical complexities, and are not responsible for the consequences of any policy they might recommend.

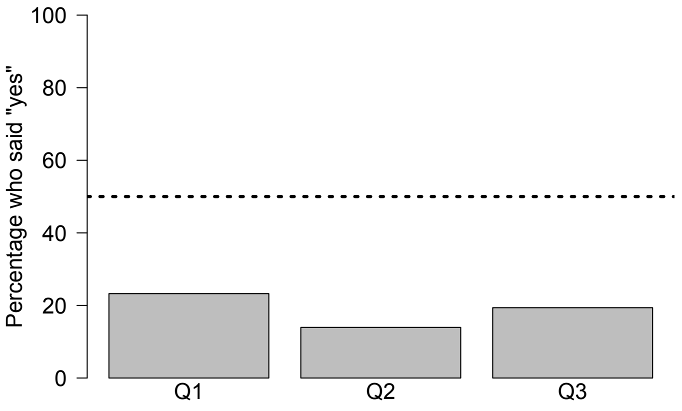

Figure 2. Percentage of participants answering “yes” to the following questions: “Imagine that a driverless vehicle is about to have an inevitable accident, and it must decide whom of two people to kill or save”. (Q1) “Should humans pre-program the vehicle to have a bias toward saving certain people over others?”, (Q2) “Should the vehicle make the decision of whom to kill or save based on the social category to which a person belongs, e.g., their race, age, gender, social class, or criminal status?”, (Q3) “Should the vehicle make the decision of whom to kill or save by using a preference scale like the one below, e.g., favor a girl over a boy, or a large woman over a homeless person, etc.?” (We presented Fig2b from Awad et al., (2018), a social preference scale, and clarified, “Note: the scale below is just an example. The exact ordering of the scale could be different.”) Participants were recruited from the online crowdsourcing platform, Amazon’s Mechanical Turk.

Acknowledgments

For incredibly helpful comments, thanks to Steven Pinker, Daniel L. K. Yamins, Emilio Frazzoli, Edwin Olson, Talia Konkle and members of the Harvard Vision Sciences Laboratory.

References

Assis, C. (2018). Seven out of 10 U.S. drivers fear self-driving cars, AAA says. *Market Watch.* https://www.marketwatch.com/story/seven-out-of-10-us-drivers-fear-self-driving-cars-aaa-says-2018-05-22.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J-F., & Rahwan, I. (2018). The moral machine experiment. *Nature,* 563*,* 59–64.

Bonnefon, J-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*, 1573–1576.

Bonnefon, J-F., Shariff, A., & Rahwan, I. (2018). Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars?. arXiv preprint arXiv:1510.03346.

Donde, J. (2017). Self-driving cars will kill people. Who decides who dies? *Twitter.* https://www.wired.com/story/self-driving-cars-will-kill-people-who-decides-who-dies/.

Edmonds, D. (2018). Cars without drivers still need a moral compass. But what kind? https://www.theguardian.com/commentisfree/2018/nov/14/cars-drivers-ethical-dilemmas-machines.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5–15.

Gogoll, J. & Müller, J. F. (2017) Autonomous cars: in favor of a mandatory ethics setting. *Science and Engineering Ethics, 23*, 681–700.

Greene, J. D. (2016) Our driverless dilemma. *Science, 352*, 1514–1515.

Lin, P. (2013) The ethics of autonomous cars. *The Atlantic.* https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/.

Lester, C. (2019). A study on driverless-car ethics offers a troubling look into our values. *The New Yorker.* https://t.co/8q47XEZ6mW.

Markoff, J. (2016). Should your driverless car hit a pedestrian to save your life? *New York Times.* https://www.nytimes.com/2016/06/24/technology/should-your-driverless-car-hit-a-pedestrian-to-save-your-life.html.

Noothigattu, R., Gaikwad, S. N. S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A. D. (2017). A voting-based system for ethical decision making. arXiv preprint arXiv:1709.06692.

Nowak, P. (2018). The ethical dilemmas of self-driving cars. *The Globe and Mail.* https://www.theglobeandmail.com/globe-drive/culture/technology/the-ethical-dilemmas-of-self-drivingcars/article37803470/.

Olson, E. (2018). Trolley folly. *Medium.* from https://medium.com/may-mobility/trolley-folly-fcbd181b7152.

Shariff, A., Rahwan, I., & Bonnefon, J-F. (2016). Whose life should your car save? *New York Times.* https://www.nytimes.com/2016/11/06/opinion/sunday/whose-life-should-your-car-save.html.