aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
1–5

DOI: 10.1177/1745691620922201
www.psychologicalscience.org/PPS
SAGE

Doubting Driverless Dilemmas

Julian De Freitas[1], Sam E. Anthony[2], Andrea Censi[3], and George A. Alvarez[1]
[1]Department of Psychology, Harvard University; [2]Perceptive Automata, Inc., Palo Alto, California; and [3]Department of Mechanical and Process Engineering, ETH Zürich

Abstract
The alarm has been raised on so-called driverless dilemmas, in which autonomous vehicles will need to make high-stakes ethical decisions on the road. We argue that these arguments are too contrived to be of practical use, are an inappropriate method for making decisions on issues of safety, and should not be used to inform engineering or policy.

Keywords
moral judgment, autonomous vehicles, driverless policy

A trolley is on course to kill five unsuspecting workers on the tracks unless you redirect it to another track with only one worker on it. What would you do? The philosopher Phillipa Foot (1967) invented the trolley dilemma to try to figure out when it is morally permissible to harm someone, normatively speaking. This same question led the philosopher Judith Jarvis Thomson (1984) to construct a famous variant of the dilemma in which one can save the five workers either indirectly (via a switch that redirects the train to the other track) or directly (by pushing a heavy man off a footbridge and onto the track, thereby bringing the train to a gruesome halt). More broadly, trolley dilemmas have been used to contrast different moral philosophies, especially deontology (which appeals to principles, e.g., "don't deliberately harm someone") as opposed to utilitarianism (which focuses on outcomes, e.g., "minimize the total amount of harm"; Greene, 2016a; Kagan, 2016; Thomson, 1984). Trolley dilemmas have also been imported by psychologists to study how people typically make moral judgments (De Freitas, DeScioli, Nemirow, Massenkoff, & Pinker, 2017; Greene, 2016a). By forcing people to make difficult moral choices, the dilemma exposes underlying moral preferences, which provides insights into how moral psychology works.

Trolley Dilemmas for Driverless Cars?

Authors of recent prominent articles within academia and beyond (Awad et al., 2018; Bonnefon, Shariff, & Rahwan, 2016; Donde, 2017; Edmonds, 2018; Gogoll & Müller, 2017; Greene, 2016b; Johnson, 2018; Lester, 2019; Lin, 2016; Markoff, 2016; Noothigattu et al., 2018; Nowak, 2018; Shariff, Rahwan, & Bonnefon, 2016) have warned that autonomous vehicles (AVs) will face trolleylike moral dilemmas and so should be programmed with ethical principles. A subset of this work proposes to solve this purported necessity by asking people on the web to consider simple scenarios in which an AV faces a two-alternative forced choice of whom to kill or save—for example, a driver or a pedestrian, a homeless man or a skilled workman. Researchers have asked people to choose on the AV's behalf and have then aggregated these choices to assemble a "global-preference scale," which they have argued should inform AV policy (Awad et al., 2018).

Many of these projects are impressive in scope, ambition, and creativity; contribute valuable cross-cultural data sets on people's moral intuitions; and serve as good conversation starters for machine ethics. That said, here we argue that the scenarios that have been employed are too contrived to be of practical use, are an inappropriate method for making decisions on issues of safety, and cannot be treated as a direct indicator of the public's opinion.

Corresponding Author:
Julian De Freitas, Department of Psychology, Harvard University, 33 Kirkland St., Cambridge, MA 02139
E-mail: defreitas@g.harvard.edu

Real-Road Trolley Dilemmas Are Highly Unlikely, Hard to Detect, and Hard to Act On

Trolley dilemmas can be useful considerations only if they (a) occur with some reasonable frequency, (b) can be detected with confidence, and (c) can be negotiated by an AV system with controlled actions. However, all three of these conditions are unmet and probably never can be met.

Trolley dilemmas are incredibly unlikely to occur on real roads

The point of the two-alternative forced choice in the thought experiments is to simplify real-world complexity and expose people's intuitions clearly. But such situations are vanishingly unlikely on real roads. This is because they require that the vehicle will certainly kill one individual or another, with no other location to steer the vehicle, no way to buy more time, and no steering maneuver other than driving head-on to a death. Some variants of the dilemmas also assume that AVs can gather information about the social characteristics of people (e.g., whether they are criminals or contributors to society). Yet many of these social characteristics are inherently unobservable. You cannot ethically choose whom to kill if you do not have information about whom you are choosing.

Lacking in these discussions are realistic examples or evidence of situations in which human drivers have had to make such choices. This makes it premature to consider them as part of any practical engineering endeavor (Dewitt, Fischhoff, & Sahlin, 2019). The authors of these articles acknowledged this point, writing, for example, that "it is extremely hard to estimate the rate at which human drivers find themselves in comparable situations," yet nevertheless, "regardless of how rare these cases are, we need to agree beforehand how they should be solved" (Awad et al., 2018, p. 59). We disagree. Without evidence that such situations occur and that the social alternatives in the thought experiments can be identified in reality, it is unhelpful to consider them when making AV policies or regulations.

Trolley dilemmas cannot be reliably detected by any real-world perception system

For the purposes of a thought experiment, it is simple to assume that one is already in a trolley dilemma. But on real roads, the AV would have to detect this fact, which means that it would first need to be trained how to do this perfectly. After all, because the overwhelming majority of driving is not a trolley dilemma, drivers should *choose to hit* someone only if they are definitely in a trolley dilemma. The problem is that it is nearly impossible for drivers to robustly differentiate when they are in a true dilemma that forces them to choose between whom to hit (and possibly kill) as opposed to an ordinary emergency that does not require such a drastic action. Accurately detecting this distinction would require unrealistic capabilities for technology in the present or near future, including (a) knowing all relevant physical details about the environment that could influence whether less deadly options are viable (e.g., the speed of each car's braking system; road conditions), (b) accurately simulating all the ways the world could unfold to confirm that one is in a true dilemma no matter what happens next, and (c) anticipating the reactions and actions of pedestrians and drivers so that their choices can be taken into account.

Trying to teach AVs to solve trolley dilemmas is thus a risky safety strategy because the AV must optimize toward solving a dilemma that is incredibly challenging to detect in the first place. Finally, if we take a learning approach to this problem, then these algorithms may need to be exposed to a large number of dilemmas. Yet the conspicuous absence of such dilemmas from real roads means that they would need to be simulated and multiplied within any data set, potentially introducing unnatural behavioral biases when AVs are deployed on real roads (e.g., "hallucinating" dilemmas where there are none).

Trolley dilemmas cannot be reliably acted on by any real-world control system

Driverless dilemmas also assume a fundamental paradox: An AV has the freedom to make a considered decision about which of two people to harm yet does not have enough control to instead take some simple action, like swerving or slowing down, to avoid harming anyone altogether (Himmelreich, 2018). In reality, if an AV is in such a bad emergency that it has only two options left, it is unlikely that these options will neatly map onto two options that require a moral rule to arbitrate between. Likewise, even if an AV does have a particular moral choice planned, the more constrained its options are, the less likely it is to have the control to successfully execute a choice—and if it cannot execute a choice, then there is no real dilemma.

Thus, Trolley Dilemmas Are an Engineering and Policy Distraction

We do not teach humans how to drive by telling them whom to kill if faced with a forced choice. This is

because planning for an unlikely, undetectable, and uncontrollable situation would be a distraction from the goal we *do* teach novice drivers: Minimize harming anyone. The goal of avoiding harm enjoys several distinct advantages over planning for a trolley dilemma: (a) It is a generic principle that covers both common driving situations and rarer emergencies, (b) it is cognitively tractable because it does not require superhuman abilities for identifying dilemmas but humanlike ones for minimizing harm, and (c) it is less risky because if a driver falsely identifies a "must-kill" state and steers the vehicle toward someone, the driver has committed a catastrophic error.

The same goes for teaching machines to drive safely. The main safety goal for any driver—human or machine—is to minimize 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 have nothing to do with trolleylike dilemmas.

None of the four 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 trolleylike dilemmas despite driving AVs on real roads every day. (We asked, "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 "What percent of your budget would you say is devoted specifically to this problem?") In the words of nuTonomy cofounder and CSO Emilio Frazzoli, "I consider 'trolley problems' one of the red herrings plaguing the world of AVs, distracting from the real issues." And recently, the CEO of May Mobility, Edwin Olson, published a piece titled "Trolley Folly" (Olson, 2018). Senior engineers and directors from all companies expressed the view that teaching AVs to solve trolleylike dilemmas would be foolhardy and irrelevant to AV safety. They dismissed the idea of choosing whom to kill on the basis of a person's social category because of the associated practical, ethical, and legal problems.

Recent articles have managed to raise the alarm on so-called driverless dilemmas by resonating with the public's understandable yet unfounded tendency to moralize new technologies because of their scary unfamiliarity (Assis, 2018). Instead of reinforcing these fears with distracting thought experiments, we should empower safety engineers to continue improving at the real 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. As just one example, there is plenty of work to do in getting AVs to recognize when dangerous situations *may* happen in order to avoid them in the first place (Ji, Khajepour, Melek, & Huang, 2016; Lefèvre, Vasquez, & Laugier, 2014; Shalev-Shwartz, Shammah, & Shashua, 2017). This requires feeding AVs more useful information and developing behavior plans to minimize the chances of such situations—for example, modulating the vehicle's speed and chosen path so that (a) if a maneuver must happen, it is less of a risky emergency maneuver and (b) if an emergency maneuver must happen, the AV has enough time and control to execute it safely.

The Ethics Challenge of AVs

Although we have poured skepticism on the trolley-dilemma approach to AVs, this does not mean that there is no ethics challenge to AVs. AV technology is an inherently ethical endeavor, given that it aims to reduce causalities on the road. Failing to solve this problem effectively will clearly have ethical consequences.

Yet solving it is also complex for several reasons. First, ethics solutions need to arbitrate among the various stakeholders working within the industry's technical, regulatory, and social spheres, each of whom has a stake in the ethics problem (e.g., manufacturers must satisfy public expectations of safety and ethics, lawyers and legal scholars care about how AV regulations fit within the current legal framework, and insurance companies will need to quantify ethically wrong behavior as estimates of riskiness that determine insurance tariffs).

Second, ethical solutions must grapple with the pragmatic issues of getting AVs to make decisions on real roads, including notions of uncertainty (probabilities of outcomes, contingent on different possible actions), states of information (perceptual false positives and negatives), trade-offs among multiple objectives (e.g., safety vs. compliance vs. efficiency), and rule violations (e.g., briefly crossing a white line to overtake a large parked truck).

Third, solutions must clarify the ethics goal of AVs, including to what extent this notion can remain stable across cultural borders. Here we have made a high-level argument for the superiority of harm avoidance over specifically planning for trolley dilemmas. Yet in practice, there are several possible variants of harm avoidance (e.g., avoid all harm vs. immediate harm vs. immediate harm at fault), each with its own ethical implications. To add to this complexity, dilemmas will

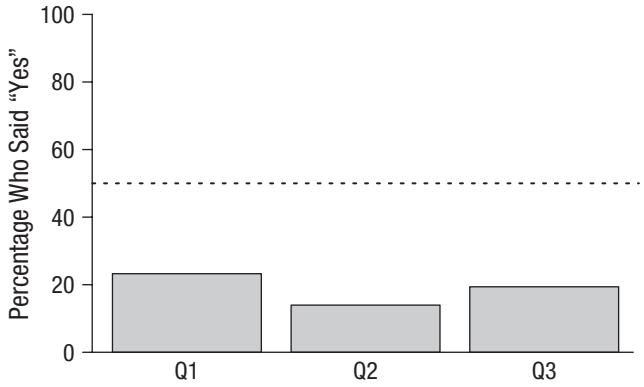

Fig. 1. Percentage of participants answering "yes" to questions about the following scenario: "Imagine that a driverless vehicle is about to have an inevitable accident, and it must decide whom of two people to kill or save." (Q1) "Should humans preprogram the vehicle to have a bias toward saving certain 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 Fig. 2b from Awad et al., 2018, which depicts a scale of social preferences and clarified, "Note: the scale below is just an example. The exact ordering of the scale could be different.") Participants were recruited from Amazon's Mechanical Turk, an online crowdsourcing platform.

arise if different goals, such as minimizing harm and preserving the passenger at all costs, are allowed to operate concurrently (e.g., people may prefer to buy a car that preserves them at all costs even if it endangers other pedestrians; Bonnefon et al., 2016).

Finally, to the extent that public preferences may play a role in solving the ethics challenge, we expect that data will have to go beyond hypothetical surveys to assessments of actual AV behavior (simulated or real) and existing human-driving patterns. Preferences from these studies will likely need to be vetted by ethically informed technical specialists (Savalescu, Kahane, & Gyngell, 2019), who can determine whether the ethical choices are computationally implementable in machines while ensuring that these choices do not obviously violate existing ethical principles.

Addendum: 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 informed and responsible one. For instance, Awad et al. (2018) assumed that their global-preference scale provided "essential topics to be considered by policymakers" (p. 60). Yet this scale is derived from contrived, two-alternative, forced-choice questions that corner participants into picking an option even if they disagree with the entire premise of the experiment. When we asked web participants (N = 129, mean age = 36, 49% female) if they thought AVs should use social preference scales to solve moral dilemmas, fewer than 20% said yes (Fig. 1).

Even if most had answered "yes," we think 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, cartoon scenarios provide a sound basis for policy governing AVs in the real world. These people are unlikely to be morally consistent (Bonnefon et al., 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.

Transparency

Action Editor: Laura A. King
Editor: Laura A. King
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iD

Julian De Freitas https://orcid.org/0000-0003-4912-1391

Acknowledgments

For incredibly helpful comments, we thank Jamie Horder, Emilio Frazzoli, Edwin Olson, Talia Konkle, members of the Harvard Vision Sciences Laboratory, and especially Daniel L. K. Yamins, Steven Pinker, and Peter DeScioli.

References

Assis, S. (2018, May 22). Seven out of 10 U.S. drivers fear self-driving cars, AAA says. *Market Watch*. Retrieved from 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., . . . 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.

De Freitas, J., DeScioli, P., Nemirow, J., Massenkoff, M., & Pinker, S. (2017). Kill or die: Moral judgment alters linguistic coding of causality. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *43*, 1173–1182.

Dewitt, B., Fischhoff, B., & Sahlin, N. (2019). 'Moral machine' experiment is no basis for policymaking [Letter]. *Nature*, *567*, 31.

Donde, J. (2017, September 21). Self-driving cars will kill people. Who decides who dies? *WIRED*. Retrieved from

https://www.wired.com/story/self-driving-cars-will-kill-people-who-decides-who-dies/

Edmonds, D. (2018, November 14). Cars without drivers still need a moral compass. But what kind? *The Guardian*. Retrieved from 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. (2016a). Our driverless dilemma. *Science*, *352*, 1514–1515.

Greene, J. D. (2016b). Solving the trolley problem. In J. Sytsma & W. Buckwalter (Eds.), *A companion to experimental philosophy* (pp. 173–189). Hoboken, NJ: Wiley.

Himmelreich, J. (2018). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice*, *21*, 669–684.

Ji, J., Khajepour, A., Melek, W. W., & Huang, Y. (2016). Path planning and tracking for vehicle collision avoidance based on model predictive control with multiconstraints. *IEEE Transactions on Vehicular Technology*, *66*, 952–964.

Johnson, C. Y. (2018, October 24). Self-driving cars will have to decide who should live and who should die. Here's who humans would kill. *The Washington Post*. Retrieved from https://www.washingtonpost.com/science/2018/10/24/self-driving-cars-will-have-decide-who-should-live-who-should-die-heres-who-humans-would-kill

Kagan, S. (2016). Solving the trolley problem. In R. Rakowski (Ed.), *The trolley problem mysteries* (pp. 151–168). New York, NY: Oxford University Press.

Lefèvre, S., Vasquez, D., & Laugier, C. (2014). A survey on motion prediction and risk assessment for intelligent vehicles. *ROBOMECH Journal*, *1*(1), 1–14.

Lester, C. A. (2019, January 24). A study on driverless-car ethics offers a troubling look into our values. *The New Yorker*. Retrieved from https://www.newyorker.com/science/elements/a-study-on-driverless-car-ethics-offers-a-troubling-look-into-our-values

Lin, P. (2016). Why ethics matters for autonomous cars. In M. Maurer, J. Gerdes, B. Lenz, & H. Winner (Eds.), Autonomous driving (pp. 69–85). Berlin, Germany: Springer.

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

Noothigattu, R., Gaikwad, S. S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A. D. (2018). A voting-based system for ethical decision making. In *Thirty-Second AAAI Conference on Artificial Intelligence* (pp. 1587–1594). Palo Alto, CA: AAAI Press. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/17052

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

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

Savulescu, J., Kahane, G., & Gyngell, C. (2019). From public preferences to ethical policy. *Nature Human Behaviour*, *3*, 1241–1243.

Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2017). On a formal model of safe and scalable self-driving cars. arXiv preprint arXiv:1708.06374.

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

Thomson, J. J. (1984). The trolley problem. *Yale Law Journal*, *94*, 1395–1415.