ETHICS

The social dilemma of autonomous vehicles

Jean-François Bonnefon,[1] Azim Shariff,[2]* Iyad Rahwan[3]†

Autonomous vehicles (AVs) should reduce traffic accidents, but they will sometimes have to choose between two evils, such as running over pedestrians or sacrificing themselves and their passenger to save the pedestrians. Defining the algorithms that will help AVs make these moral decisions is a formidable challenge. We found that participants in six Amazon Mechanical Turk studies approved of utilitarian AVs (that is, AVs that sacrifice their passengers for the greater good) and would like others to buy them, but they would themselves prefer to ride in AVs that protect their passengers at all costs. The study participants disapprove of enforcing utilitarian regulations for AVs and would be less willing to buy such an AV. Accordingly, regulating for utilitarian algorithms may paradoxically increase casualties by postponing the adoption of a safer technology.

The year 2007 saw the completion of the first benchmark test for autonomous driving in realistic urban environments (*1*, *2*). Since then, autonomous vehicles (AVs) such as Google's self-driving car covered thousands of miles of real-road driving (*3*). AVs have the potential to benefit the world by increasing traffic efficiency (*4*), reducing pollution (*5*), and eliminating up to 90% of traffic accidents (*6*). Not all crashes will be avoided, though, and some crashes will require AVs to make difficult ethical decisions in cases that involve unavoidable harm (*7*). For example, the AV may avoid harming several pedestrians by swerving and sacrificing a passerby, or the AV may be faced with the choice of sacrificing its own passenger to save one or more pedestrians (Fig. 1).

Although these scenarios appear unlikely, even low-probability events are bound to occur with millions of AVs on the road. Moreover, even if these situations were never to arise, AV programming must still include decision rules about what to do in such hypothetical situations. Thus, these types of decisions need be made well before AVs become a global commodity. Distributing harm is a decision that is universally considered to fall within the moral domain (*8*, *9*). Accordingly, the algorithms that control AVs will need to embed moral principles guiding their decisions in situations of unavoidable harm (*10*). Manufacturers and regulators will need to accomplish three potentially incompatible objectives: being consistent, not causing public outrage, and not discouraging buyers.

However, pursuing these objectives may lead to moral inconsistencies. Consider, for example, the case displayed in Fig. 1A, and assume that the most common moral attitude is that the AV should swerve. This would fit a utilitarian moral doctrine (*11*), according to which the moral course of action is to minimize casualties. But consider then the case displayed in Fig. 1C. The utilitarian course of action, in that situation, would be for the AV to swerve and kill its passenger, but AVs programmed to follow this course of action might discourage buyers who believe their own safety should trump other considerations. Even though such situations may be exceedingly rare, their emotional saliency is likely to give them broad public exposure and a disproportionate weight in individual and public decisions about AVs. To align moral algorithms with human values, we must start a collective discussion about the ethics of AVs—that is, the moral algorithms that we are willing to accept as citizens and to be subjected to as car owners. Thus, we initiate the data-driven study of driverless car ethics, inspired by the methods of experimental ethics (*12*).

We conducted six online surveys (n = 1928 total participants) between June and November 2015. All studies were programmed on Qualtrics survey software and recruited participants (U.S. residents only) from the Amazon Mechanical Turk (MTurk) platform, for a compensation of 25 cents each. Studies described in the experimental ethics literature largely rely on MTurk respondents, with robust results, even though MTurk respondents are not necessarily representative of the U.S. population (*13*, *14*). A possible concern with MTurk studies is that some participants may already be familiar with testing materials, particularly when these materials are used by many research groups. However, this concern does not apply to our testing materials, which have never been used in a published MTurk study to date.

In all studies, participants provided basic demographic information. Regression analyses (see supplementary materials) showed that enthusiasm for self-driving cars was consistently greater for younger, male participants. Accordingly, all subsequent analyses included age and sex as covariates. The last item in every study was an easy question (e.g., how many pedestrians were on the

[1]Toulouse School of Economics, Institute for Advanced Study in Toulouse, Center for Research in Management, CNRS, University of Toulouse Capitole, Toulouse, France. [2]Department of Psychology, University of Oregon, Eugene, OR 97403, USA. [3]The Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
*Present address: Department of Psychology and Social Behavior, 4201 Social and Behavioral Sciences Gateway, University of California, Irvine, Irvine, CA 92697, USA. †Corresponding author. Email: irahwan@mit.edu

road) relative to the traffic situation that participants had just considered. Participants who failed this attention check (typically 10% of the sample) were discarded from subsequent analyses.

Detailed statistical results for all studies are provided in the supplementary materials (tables S1 to S8). Overall, participants strongly agreed that it would be more moral for AVs to sacrifice their own passengers when this sacrifice would save a greater number of lives overall.

In study one (n = 182 participants), 76% of participants thought that it would be more moral for AVs to sacrifice one passenger rather than kill 10 pedestrians [with a 95% confidence interval (CI) of 69 to 82]. These same participants were later asked to rate which was the most moral way to program AVs, on a scale from 0 (protect the passenger at all costs) to 100 (minimize the number of casualties). They overwhelmingly expressed a moral preference for utilitarian AVs programmed to minimize the number of casualties (median = 85) (Fig. 2A). However, participants were less certain that AVs would be programmed in a utilitarian manner (67% thought so, with a median rating of 70). Thus, participants were not worried about AVs being too utilitarian, as often portrayed in science-fiction works. If anything, they imagined future AVs as being less utilitarian than they should be.

In study two (n = 451 participants), participants were presented with dilemmas that varied the number of pedestrians' lives that could be saved, from 1 to 100. Participants did not think that AVs should sacrifice their passenger when only one pedestrian could be saved (with an average approval rate of 23%), but their moral approval increased with the number of lives that could be saved ($P < 0.001$), up to approval rates consistent with the 76% observed in study one (Fig. 2B).

Participants' approval of passenger sacrifice was even robust to treatments in which they had to imagine themselves and another person, particularly a family member, in the AV (study three, n = 259 participants). Imagining that a family member was in the AV negatively affected the morality of the sacrifice, as compared with imagining oneself alone in the AV ($P = 0.003$). But even in that strongly aversive situation, the morality of the sacrifice was still rated above the midpoint of the scale, with a 95% CI of 54 to 66 (Fig. 3A).

Still, study three presents the first hint of a social dilemma. On a scale of 1 to 100, respondents were asked to indicate how likely they would be to buy an AV programmed to minimize casualties (which would, in these circumstances, sacrifice them and their co-rider family member), as well as how likely they would be to buy an AV programmed to prioritize protecting its passengers, even if it meant killing 10 or 20 pedestrians. Although the reported likelihood of buying an AV was low even for the self-protective option (median = 50), respondents indicated a significantly lower likelihood ($P < 0.001$) of buying the AV when they imagined the situation in which they and their family member would be sacrificed for the greater good (median = 19). In other words, even though participants still agreed that utilitarian AVs were the most moral, they preferred the self-protective model for themselves.

Study four (n = 267 participants) offers another demonstration of this phenomenon. Participants were given 100 points to allocate between different types of algorithms, to indicate (i) how moral the algorithms were, (ii) how comfortable participants were for other AVs to be programmed in a given manner, and (iii) how likely participants would be to buy an AV programmed in a given manner. For one of the algorithms, the AV would always swerve when it was about to run over people on the road. Figure 3B shows the points allocated to the AV equipped with this algorithm, in three situations: (i) when it swerved into a pedestrian to save 10 people, (ii) when it killed its own passenger to save 10 people, and (iii) when it swerved into a pedestrian to save just one other pedestrian. The algorithm that swerved into one to save 10 always received many points, and the algorithm that swerved into one to save one always received few points. The algorithm that would kill its passenger to save 10 presented a hybrid profile. Like the high-valued algorithm, it received high marks for morality (median budget share = 50) and was considered a good algorithm for other people to have (median budget share = 50). But in terms of purchase intention,

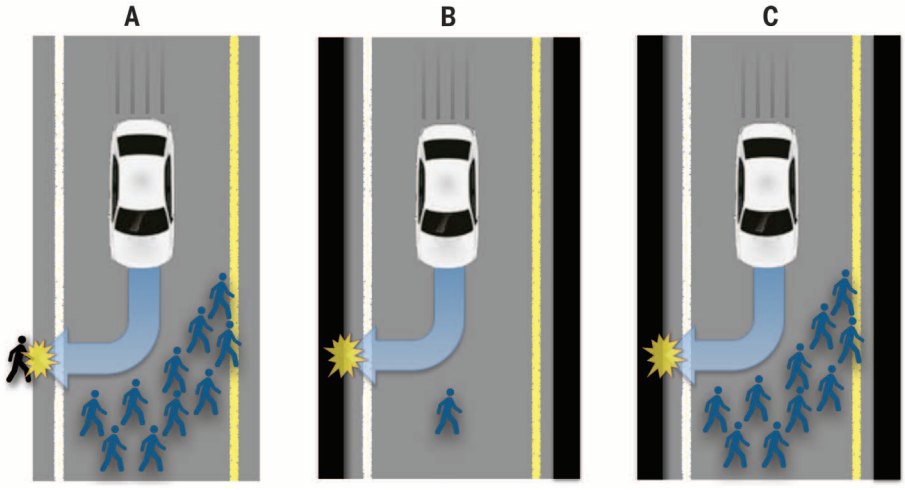
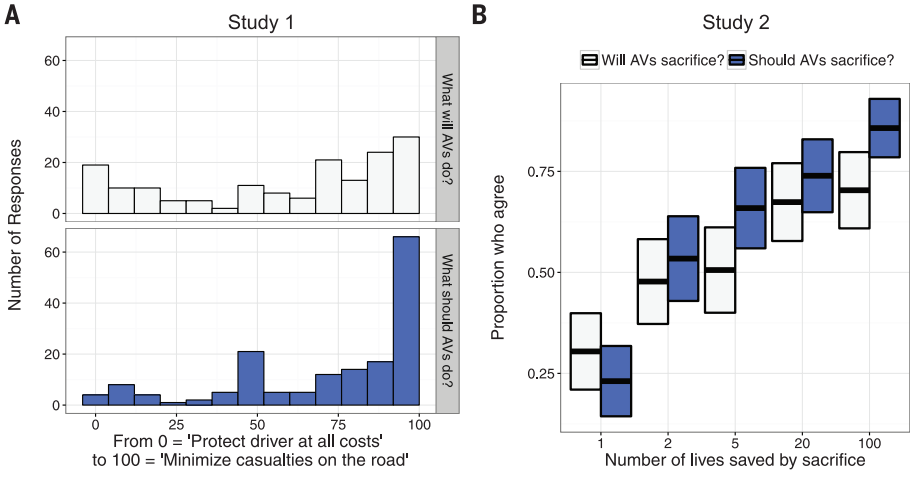


Fig. 1. Three traffic situations involving imminent unavoidable harm. The car must decide between (**A**) killing several pedestrians or one passerby, (**B**) killing one pedestrian or its own passenger, and (**C**) killing several pedestrians or its own passenger.

Fig. 2. Considering the greater good versus the life of the passenger. (**A** and **B**) In studies one and two, when asked which would be the most moral way to program AVs, participants expressed a preference for AVs programmed to kill their passengers for the greater good. This preference was strong, provided that at least five lives could be saved [(A) shows detailed results for 10 lives]. On average, participants were more confident that AVs should pursue the greater good than whether AVs would actually be programmed to do so. In (B), boxes show the 95% CI of the mean.



it received significantly fewer points than the high-valued algorithm ($P < 0.001$) and was, in fact, closer to the low-valued algorithms (median budget share = 33). Once more, it appears that people praise utilitarian, self-sacrificing AVs and welcome them on the road, without actually wanting to buy one for themselves.

This is the classic signature of a social dilemma, in which everyone has a temptation to free-ride instead of adopting the behavior that would lead to the best global outcome. One typical solution in this case is for regulators to enforce the behavior leading to the best global outcome. Indeed, there are many similar societal examples involving trade-off of harm by people and governments (*15*–*17*). For example, some citizens object to regulations that require children to be immunized before starting school. In this case, the parental decision-makers choose to minimize the perceived risk of harm to their child while increasing the risk to others. Likewise, recognition of the threats of environmental degradation have prompted government regulations aimed at curtailing harmful behaviors for the greater good. But would people approve of government regulations imposing utilitarian algorithms in AVs, and would they be more likely to buy AVs under such regulations?

In study five ($n = 376$ participants), we asked participants about their attitudes toward legally enforcing utilitarian sacrifices. Participants considered scenarios in which either a human driver or a control algorithm had an opportunity to self-sacrifice to save 1 or 10 pedestrians (Fig. 3C). As usual, the perceived morality of the sacrifice was high and about the same whether the sacrifice was performed by a human or by an algorithm (median = 70). When we inquired whether participants would agree to see such moral sacrifices legally enforced, their agreement was higher for algorithms than for human drivers ($P < 0.002$), but the average agreement still remained below the midpoint of the 0 to 100 scale in each scenario. Agreement was highest in the scenario in which algorithms saved 10 lives, with a 95% CI of 33 to 46.

Finally, in study six ($n = 393$ participants), we asked participants specifically about their likelihood of purchasing the AVs whose algorithms had been regulated by the government. Participants were presented with scenarios in which they were riding alone, with an unspecified family member, or with their child. As in the previous studies, the scenarios depicted a situation in which the algorithm that controlled the AV could sacrifice its passengers to minimize casualties on the road. Participants indicated whether it was the duty of the government to enforce regulations that would minimize the casualties in such circumstances, whether they would consider the purchase of an AV under such regulations, and whether they would consider purchasing an AV under no such regulations. As shown in Fig. 3D, people were reluctant to accept governmental regulation of utilitarian AVs. Even in the most favorable condition, when participants imagined only themselves being sacrificed to save 10 pedestrians, the 95% CI for whether people thought it was appropriate for the government to regulate this sacrifice was only 36 to 48. Finally, participants were much less likely to consider purchasing an AV with such regulation than without ($P < 0.001$). The median expressed likelihood of purchasing an unregulated AV was 59, compared with 21 for purchasing a regulated AV. This is a huge gap from a statistical perspective, but it must be understood as reflecting the state of public sentiment at the very beginning of a new public issue and is thus not guaranteed to persist.

Three groups may be able to decide how AVs handle ethical dilemmas: the consumers who buy the AVs; the manufacturers that program the AVs; and the government, which may regulate the kind of programming manufacturers can offer and consumers can select. Although manufacturers may engage in advertising and lobbying to influence consumer preferences and government regulations, a critical collective problem consists of deciding whether governments should regulate the moral algorithms that manufacturers offer to consumers.

Our findings suggest that regulation for AVs may be necessary but also counterproductive. 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

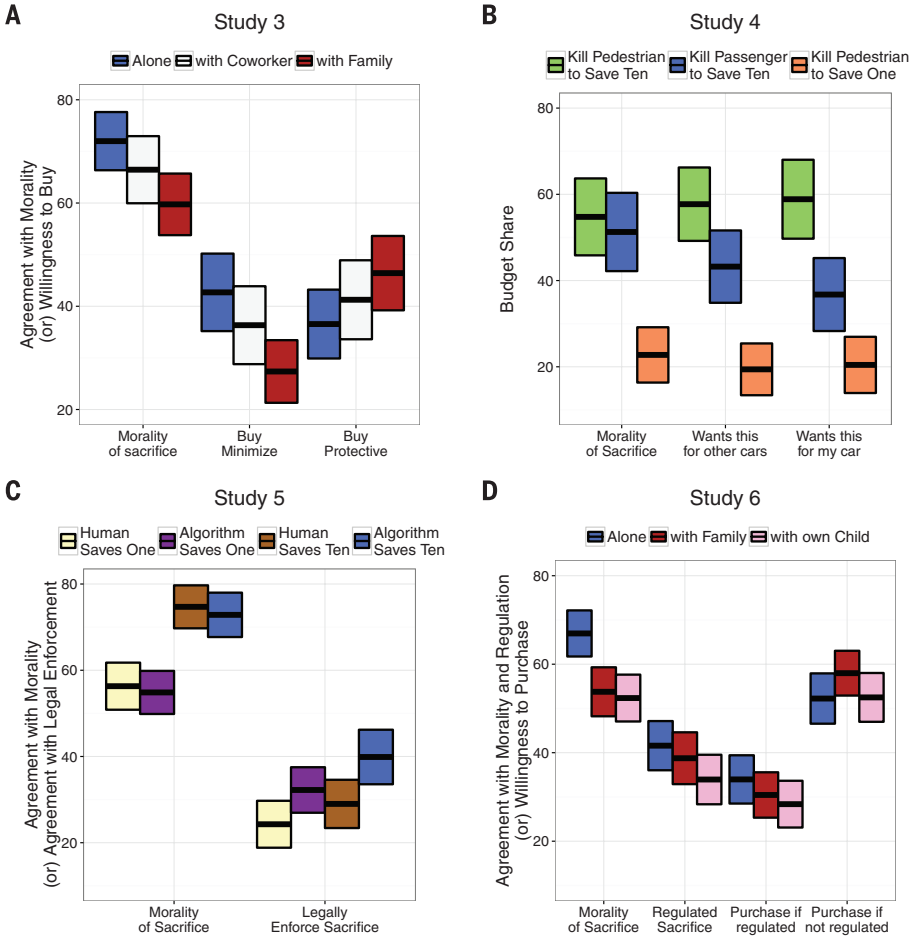


Fig. 3. Toward regulation and purchase (studies three to six). (**A** to **D**) Boxes show the 95% CI of the mean. In all studies, participants expressed a moral preference for AVs sacrificing their passengers to save a greater number of pedestrians. This moral preference was robust for situations in which participants imagined themselves in the AV in the company of a co-worker, a family member, or their own child. However, participants did not express a comparable preference for buying utilitarian AVs, especially when they thought of family members riding in the car [(A) and (B)]. Additionally, participants disapproved of regulations enforcing utilitarian algorithms for AVs and indicated that they would be less likely to purchase an AV under such regulations [(C) and (D)].

lives saved by making AVs utilitarian may be outnumbered by the deaths caused by delaying the adoption of AVs altogether. Thus, car-makers and regulators alike should be considering solutions to these obstacles.

Moral algorithms for AVs will need to tackle more intricate decisions than those considered in our surveys. For example, our scenarios did not feature any uncertainty about decision outcomes, but a collective discussion about moral algorithms will need to encompass the concepts of expected risk, expected value, and blame assignment. Is it acceptable for an AV to avoid a motorcycle by swerving into a wall, considering that the probability of survival is greater for the passenger of the AV than for the rider of the motorcycle? Should AVs account for the ages of passengers and pedestrians (*20*)? If a manufacturer offers different versions of its moral algorithm, and a buyer knowingly chose one of them, is the buyer to blame for the harmful consequences of the algorithm's decisions? Such liability considerations will need to accompany existing discussions of regulation (*21*), and we hope that psychological studies inspired by our own will be able to inform this discussion.

Figuring out how to build ethical autonomous machines is one of the thorniest challenges in artificial intelligence today (*22*). As we are about to endow millions of vehicles with autonomy, a serious consideration of algorithmic morality has never been more urgent. Our data-driven approach highlights how the field of experimental ethics can provide key insights into the moral, cultural, and legal standards that people expect from autonomous driving algorithms. For the time being, there seems to be no easy way to design algorithms that would reconcile moral values and personal self-interest—let alone account for different cultures with various moral attitudes regarding life-life trade-offs (*23*)—but public opinion and social pressure may very well shift as this conversation progresses.

REFERENCES AND NOTES

1. B. Montemerlo *et al.*, *J. Field Robot.* **25**, 569–597 (2008).
2. C. Urmson *et al.*, *J. Field Robot.* **25**, 425–466 (2008).
3. M. M. Waldrop, *Nature* **518**, 20–23 (2015).
4. B. van Arem, C. J. van Driel, R. Visser, *IEEE Trans. Intell. Transp. Syst.* **7**, 429–436 (2006).
5. K. Spieser *et al.*, in *Road Vehicle Automation*, G. Meyer, S. Beiker, Eds. (Lecture Notes in Mobility Series, Springer, 2014), pp. 229–245.
6. P. Gao, R. Hensley, A. Zielke, "A roadmap to the future for the auto industry," *McKinsey Quarterly* (October 2014); www.mckinsey.com/industries/automotive-and-assembly/our-insights/a-road-map-to-the-future-for-the-auto-industry.
7. N. J. Goodall, in *Road Vehicle Automation*, G. Meyer, S. Beiker, Eds. (Lecture Notes in Mobility Series, Springer, 2014), pp. 93–102.
8. K. Gray, A. Waytz, L. Young, *Psychol. Inq.* **23**, 206–215 (2012).
9. J. Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion* (Pantheon Books, 2012).
10. W. Wallach, C. Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford University Press, 2008).
11. F. Rosen, *Classical Utilitarianism from Hume to Mill* (Routledge, 2005).
12. J. D. Greene, *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them* (Atlantic Books, 2014).
13. S. Côté, P. K. Piff, R. Willer, *J. Pers. Soc. Psychol.* **104**, 490–503 (2013).
14. J. A. C. Everett, D. A. Pizarro, M. J. Crockett, *J. Exp. Psychol. Gen.* **145**, 772–787 (2016).
15. N. E. Kass, *Am. J. Public Health* **91**, 1776–1782 (2001).
16. C. R. Sunstein, A. Vermeule, *Stanford Law Rev.* **58**, 703–750 (2005).
17. T. Dietz, E. Ostrom, P. C. Stern, *Science* **302**, 1907–1912 (2003).
18. R. M. Dawes, *Annu. Rev. Psychol.* **31**, 169–193 (1980).
19. P. A. M. Van Lange, J. Joireman, C. D. Parks, E. Van Dijk, *Organ. Behav. Hum. Decis. Process.* **120**, 125–141 (2013).
20. E. A. Posner, C. R. Sunstein, *Univ. Chic. Law Rev.* **72**, 537–598 (2005).
21. D. C. Vladeck, *Wash. Law Rev.* **89**, 117–150 (2014).
22. B. Deng, *Nature* **523**, 24–26 (2015).
23. N. Gold, A. M. Colman, B. D. Pulford, *Judgm. Decis. Mak.* **9**, 65–76 (2014).

ACKNOWLEDGMENTS

J.-F.B. gratefully acknowledges support through the Agence Nationale de la Recherche–Laboratoires d'Excellence Institute for Advanced Study in Toulouse. This research was supported by internal funds from the University of Oregon to A.S. I.R. is grateful for financial support from R. Hoffman. Data files have been uploaded as supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/352/6293/1573/suppl/DC1
Materials and Methods
Supplementary Text
Fig. S1
Tables S1 to S8
Data Files S1 to S6

15 January 2016; accepted 21 April 2016
10.1126/science.aaf2654

Science

The social dilemma of autonomous vehicles

Jean-François Bonnefon, Azim Shariff and Iyad Rahwan

Science **352** (6293), 1573-1576.
DOI: 10.1126/science.aaf2654

Codes of conduct in autonomous vehicles

When it becomes possible to program decision-making based on moral principles into machines, will self-interest or the public good predominate? In a series of surveys, Bonnefon *et al.* found that even though participants approve of autonomous vehicles that might sacrifice passengers to save others, respondents would prefer not to ride in such vehicles (see the Perspective by Greene). Respondents would also not approve regulations mandating self-sacrifice, and such regulations would make them less willing to buy an autonomous vehicle.

Science, this issue p. 1573; see also p. 1514