The Thorny Challenge of Making Moral Machines: Ethical Dilemmas with Self-Driving Cars

Edmond Awad, Jean-François Bonnefon, Azim Shariff and Iyad Rahwan

Self-driving vehicles: Safe, but not 100% × Autonomous, self-driving cars are being tested and trained extensively and have already covered thousands of miles of real road driving. Incidents are remarkably rare. However, any accidents – especially if they involve fatalities – are covered broadly in media all over the world, and consumers wonder whether autonomous vehicles (AVs) are actually safe, and whether they should ever trust them. Experts agree that AVs do have the potential to benefit the world by increasing traffic efficiency, reducing pollution and eliminating up to 90 % of traffic accidents – those that are caused by driver error, tiredness, drunkenness or other human factors. Though safety is constantly improving and injuries and deaths might be significantly reduced, crashes will never be completely avoidable. And any imminent crashes will require AVs to make difficult decisions.

How to react when a crash is imminent? × Imagine, as an example, situations as depicted in Figure 1. The AV may avoid harming several pedestrians by swerving and sacrificing a passerby (A), or the AV may be faced with the choice of sacrificing its own passenger to save one (B) or more (C) pedestrians.

Although these scenarios appear unlikely, even low-probability events are bound to occur with millions of AVs on the road. Furthermore, the tradeoffs involved in these scenarios will occur in much more frequent, but less extreme scenarios: instead of choosing between certain death, the car will need to choose between slightly increasing the risk toward one group rather than toward another. AV programming must include decision rules about what to do when these situations occur. While a human driver has to spontaneously react in a

KEYWORDS
Ethics, Decision Making, AI, Autonomous Vehicles, Moral Machines

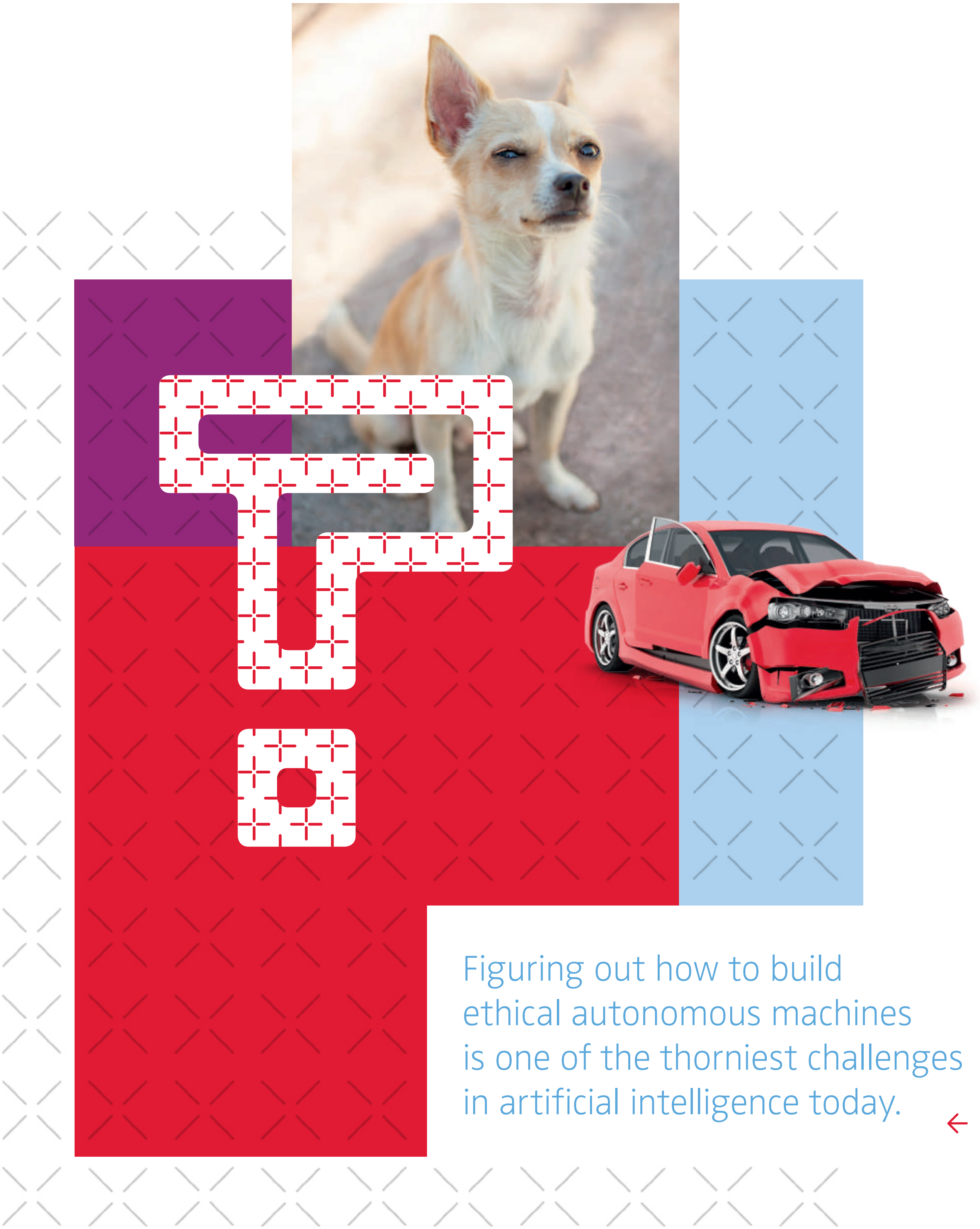
THE AUTHORS

Edmond Awad
The Media Lab, Institute for Data,
Systems and Society,
Massachusetts Institute of Technology,
Cambridge, MA, USA
awad@mit.edu

Jean-François Bonnefon
Toulouse School of Economics (TSM-R, CNRS),
Université Toulouse-1 Capitole,
Toulouse, France
jean-francois.bonnefon@tse-fr.eu

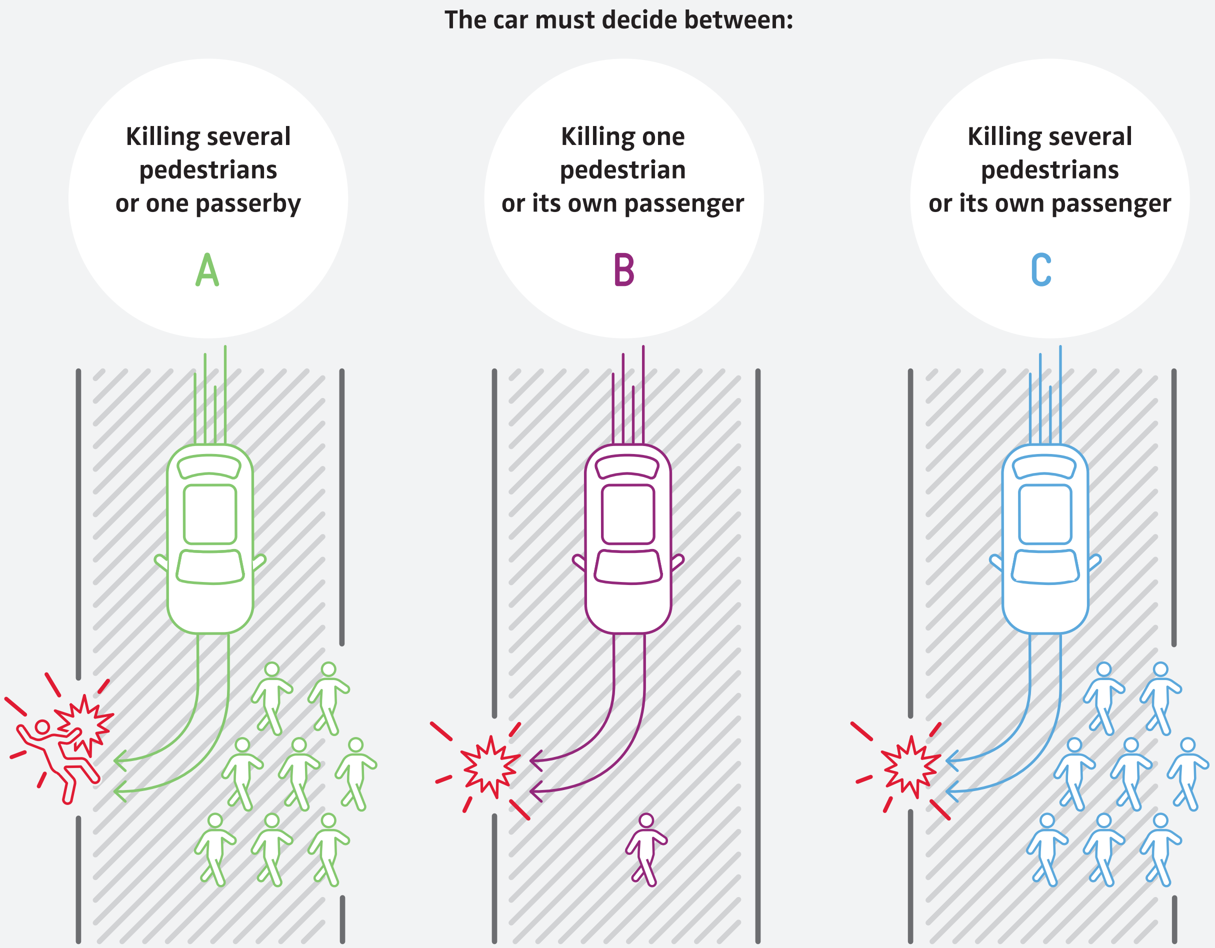
Azim Shariff
Department of Psychology,
University of British Columbia,
Vancouver, Canada
Canadashariff@psych.ubc.ca

Iyad Rahwan
The Media Lab,
Massachusetts Institute of Technology,
Cambridge, MA, USA,
Center for Humans and Machines,
Max-Planck Institute for Human Development,
Berlin, Germany
irahwan@mit.edu



Figuring out how to build ethical autonomous machines is one of the thorniest challenges in artificial intelligence today. ←

FIGURE 1 > Three traffic situations involving imminent unavoidable harm


The car must decide between:
Killing several pedestrians or one passerby
A
Killing one pedestrian or its own passenger
B
Killing several pedestrians or its own passenger
C

split second, an autonomous vehicle needs to be programmed deliberately; someone needs to define the rules, well before AVs become a global commodity. The algorithms that control AVs will need to embed moral principles guiding their decisions in situations of unavoidable harm. But what is the right moral decision in such a case? How should AI be programmed to react in such an instant? Manufacturers and regulators will need to accomplish three potentially incompatible objectives: consistency, avoiding public outrage, and not discouraging buyers. One step toward solving this problem is trying to learn how people feel about alternative decisions that self-driving vehicles' AI might have to make. This was the purpose of our Moral Machine study (see Box 1).

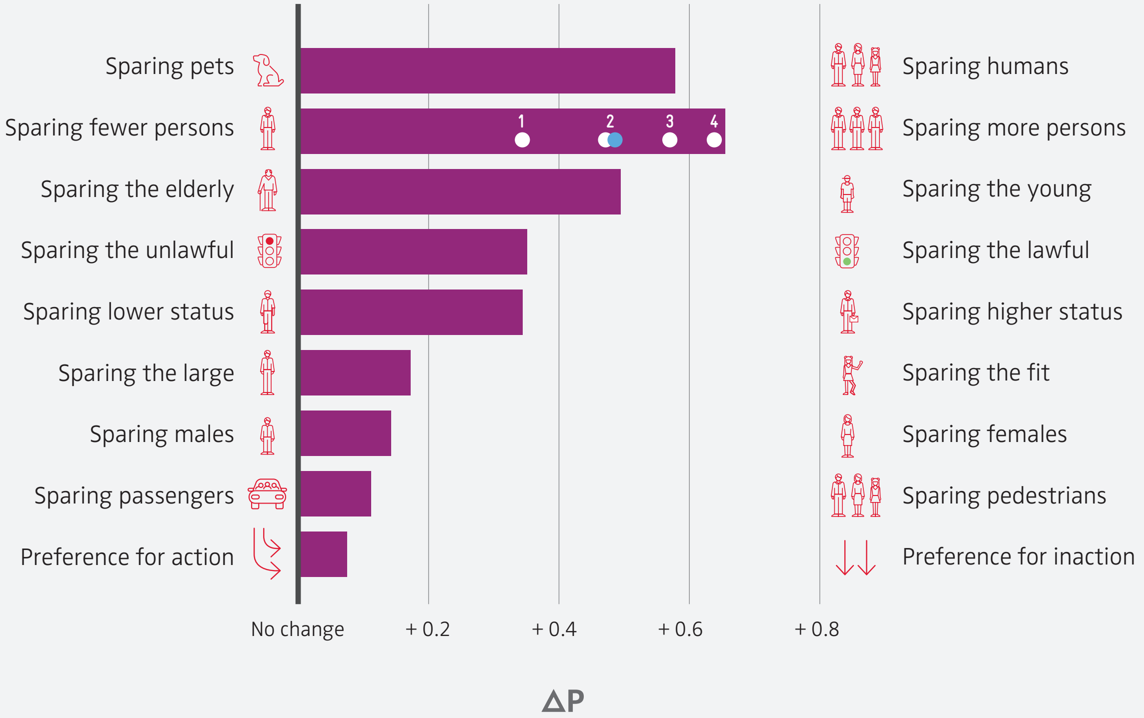
Saving passengers or pedestrians? × Another online study among U.S. residents sheds more light on the complexity of the topic of AI-driven decision-making in dangerous situations. This study explored the trade-off between saving driver and passengers versus saving pedestrians and other road users – the dilemma illustrated in Figure 1. In principle, participants approved of utilitarian AVs minimizing the number of casualties. Their moral approval increased with the number of lives that could be saved. Participants' approval of passenger sacrifice was even slightly positive when they had to imagine themselves and another person, particularly a family member, in the AV.

BOX 1

Exploring moral preferences – the moral machine experiment

With a group of MIT researchers we set out to gauge societal expectations about the ethical principles that should guide machine behavior. To address this challenge, we deployed the Moral Machine, a viral online experimental platform designed to explore the moral dilemmas faced by autonomous vehicles. This platform gathered 40 million decisions in the context of unavoidable accidents. More than two million people from 233 countries and territories participated online in our multilingual 'serious game' and revealed which harm seemed more tolerable to most people. Indeed, the most emphatic global preferences in the survey are for sparing the lives of humans over the lives of other animals; sparing the lives of many people rather than a few; and preserving the lives of the young, rather than older people (see the first three preferences in Figure 2).

FIGURE 2 > Global preferences in favour of the choice on the right side



ΔP is the difference between the probability of sparing persons possessing the attribute on the right, and the probability of sparing persons possessing the attribute on the left, aggregated over all other attributes. For the number of persons effect sizes are shown for each number of additional characters (1 to 4); the effect size for two additional persons overlaps with the mean effect of the attribute (= blue circle).

BOX 2

Cultural differences in the preference for ethical standards

While there was not much variation along the lines of demographic characteristics like age, gender, income, education, or political and religious views, the cultural background did play a role in the assessment. Some of the differences are listed below:

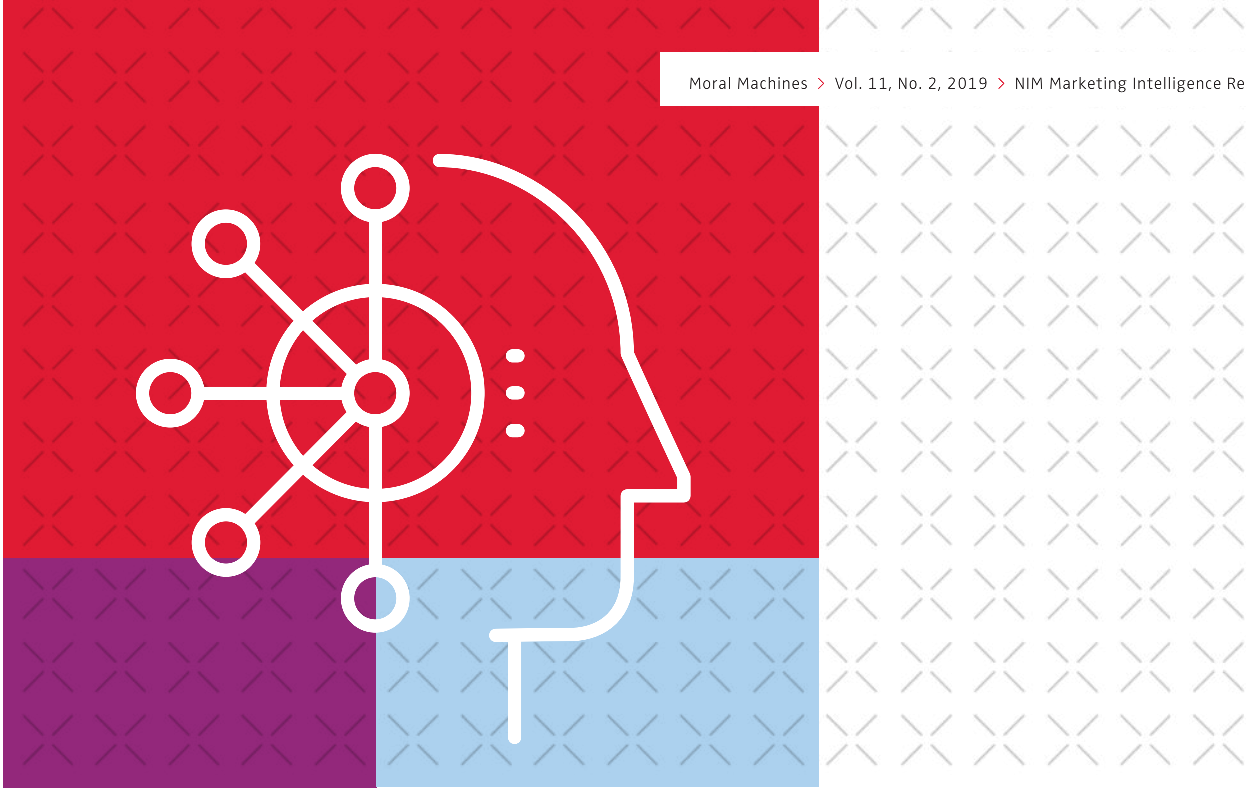
- Countries within close proximity to one another showed closer moral preferences, with three dominant clusters in the West, East, and South.
- Participants from collectivist, eastern cultures like China and Japan were less likely to spare the young over the old compared to countries in the southern cluster in which central and southern American countries dominate.
- Participants from individualistic cultures, like the UK and US, placed a stronger emphasis on sparing more lives given all the other choices – possibly, because of the greater emphasis on the value of each individual.
- Similarly, participants from poorer countries with weaker institutions turned out to be more tolerant of jaywalkers versus pedestrians who cross legally.
- Participants from countries with a high level of economic inequality showed greater gaps between the treatment of individuals with high and low social status.
- Finally, we observed some striking peculiarities, such as the strong preference for sparing women and the strong preference for sparing fit characters in the Southern cluster.

Consumers would like other people to buy AVs relying on such a utilitarian algorithm, but they would themselves prefer to ride in AVs that protect their passengers at all costs. Further, study participants disapproved of enforcing utilitarian regulations for AVs, and would be less willing to buy such AVs. Thus, the moral dilemma triggers a social dilemma that needs to be solved.

Steps towards solving the ethical dilemma of self-driving vehicles × We find ourselves in a situation that is new to the world: Vehicles are able to make decisions as far-reaching as who should live or die – without real-time human supervision. This problem is not limited to a niche market but will affect everyday transportation and all road users, no matter whether they drive, walk, or ride a bike. To be prepared to actually let AVs take over our roads, producers need to master several challenges on top of the technical ones.

- **Discuss ethics of AI on a general level** × All stakeholders should embrace the challenges of machine ethics as a unique opportunity to decide, as a community, what we believe to be right or wrong, and to make sure that machines, unlike humans, unerringly follow the agreed-upon moral preferences. We might not reach universal agreement, as indicated by the Moral Machine survey, but the fact that broad regions of the world displayed relative agreement is encouraging.

- **Work on a new social contract** × Over a century ago, cars started to become a commodity on the roads. A system of laws regulating the behavior of drivers and pedestrians (and the designs and practices of manufacturers) was introduced and has been continuously refined. Overall, this traffic system is trusted by society. In days to come, the integration of autonomous cars will



require a new social contract that provides clear guidelines on who bears responsibility for different kinds of accidents; how monitoring and enforcement will be performed; and how trust can be engendered among all stakeholders. This contract will be similarly transformational, but will probably occur over a much shorter period of time.

> **Prepare the public to build trust** × The ethical quandary of who to save in life-threatening incidents produces a social dilemma. People recognize the utilitarian approach to be the more ethical one, and as citizens, they want the cars to save a greater number of lives. But as consumers, they want self-protective cars. As a result, the adoption of either strategy brings its own risks for manufacturers: A self-protective strategy risks public outrage, whereas a utilitarian strategy may scare consumers away. To make people feel safe and trust AVs, we must encourage public discourse about the absolute reduction in risk to passengers through overall accident reduction. Otherwise, outsized media coverage of rare accidents will trigger the biased perception of risk for passengers, which might irrationally overshadow the greater safety effects.

Interesting times ahead × Figuring out how to build ethical autonomous machines is one of the thorniest challenges in artificial intelligence today. As we are about to endow millions of vehicles with decision autonomy, 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 algorithms of self-driving cars. And even if these issues are being tackled and will eventually be solved, other AV challenges such as hacking, liability, and labor displacement will remain. We face interesting times! ×

↓

FURTHER READING

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.K. and Rahwan, I. (2018): "The Moral Machine Experiment", Nature. 563. doi: 10.1038/s41586-018-0637-6.

Bonnefon, J.-F.; Shariff, A. and Rahwan, I. (2016): "The Social Dilemma of Autonomous Vehicles", Science. 352. doi: 10.1126/science.aaf2654.

Bonnefon J.-F.; Shariff A. and Rahwan, I. (2019): "The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars", Proceedings of the IEEE, Vol. 107, 502-504.

Shariff, A.; Bonnefon, J.-F. and Rahwan, I. (2017): "Psychological roadblocks to the adoption of self-driving vehicles", Nature Human Behaviour https://doi.org/10.1038/s41562-017-0202-6