Psychological roadblocks to the adoption of self-driving vehicles

Azim Shariff*, Jean-François Bonnefon*, and Iyad Rahwan*

Correspondence: azim.shariff@uci.edu; jean-francois.bonnefon@tse-fr.eu; irahwan@mit.edu

Standfirst
Self-driving cars offer a bright future, but only if the public can overcome the psychological challenges that stand in the way of widespread adoption. We discuss three—ethical dilemmas, overreactions to accidents, and the opacity of the cars' decision-making algorithms—and propose steps towards addressing them.

The widespread adoption of autonomous vehicles (AVs) promises to make us happier, safer, and more efficient. Manufacturers are speeding past the remaining technical challenges to the cars' readiness. But the biggest roadblocks standing in the path of the mass adoption may be psychological, not technological; 78% of Americans report fearing riding in an AV, with only 19% indicating they would trust the car[1].

Trust—the comfort in making oneself vulnerable to another entity in the pursuit of some benefit—has long been recognized as critical to the adoption of automation, and becomes even more important as both the complexity of automation and the vulnerability of the users increase[2]. For AVs, which will need to navigate our complex urban environment with the power of life and death, trust will determine how widely they are adopted by consumers, and how tolerated they are by everyone else. Achieving the bright future promised by AVs will require overcoming the psychological barriers to trust. Here we diagnose three factors underlying this resistance and offer a plan of action (see Table).

The Dilemmas of Autonomous Ethics

The necessity for AVs to make ethical decisions leads to a series of dilemmas for their designers, regulators, and the public at large[3]. These begin with the need for an AV to decide how it will operate in situations where its actions could decrease the risk of harming its own passengers by *increasing* the risk to a potentially larger number of non-passengers (e.g. pedestrians, other drivers). While these decisions will most often involve probabilistic tradeoffs in small-risk manoeuvres, at its extreme the decision could involve an AV determining whether to harm its passenger to spare the lives of two or more pedestrians, or vice-versa (see Figure).

<Figure about here>

In handling these situations, the cars may operate as utilitarians, minimizing total risk to people regardless of who they are, or as self-protective, placing extra weight on the safety of their own passengers. Human drivers make such decisions instinctively in a split-second, and thus cannot be expected to abide by whatever ethical principle they formulated in the comfort of their armchair. But AV manufacturers have the luxury of moral deliberation and thus the responsibility of deliberation.

The existence of this ethical dilemma in turn produces a social dilemma. People are inconsistent about what principles they want AVs to follow, recognizing the utilitarian approach to be the most ethical, and as citizens, wanting cars to save the greater number. But as consumers, they want self-protective cars[3]. As a result, adopting either strategy brings its own risks for manufacturers—a self-protective strategy risks public outrage, whereas a utilitarian strategy may scare consumers away.

Both the ethical and social dilemmas will need to be addressed to earn the trust of the public. And because it seems unlikely that regulators will adopt the strictest self-protective solution—in which AVs would never harm their passengers, however small the danger to passengers, and large the risk to others—we will have to grapple with consumers' fear that their car might someday decide to harm them.

To overcome that fear, we need to make people feel both safe and virtuous about owning an AV. To make people feel safe, we must understand how to most effectively convey the *absolute* reduction in risk to passengers due to overall accident reduction, so that it is not irrationally overshadowed by a potentially small increase in *relative* risk that passengers face in relation to other road users.

Communication about the overall safety benefits of AVs could be further leveraged to appeal to potential consumers' concerns about self-image and reputation. Virtue signalling is a powerful motivation for buying ethical products—but only when the ethicality is conspicuous[4]. Allowing the altruistic benefits of AVs to reflect on the consumer can change the conversation about AV ethics and prove itself to be a marketing asset. The most relevant example of successful virtue consumerism is that of the Toyota Prius, a hybrid-electric automobile whose distinctive shape has allowed owners to signal their environmental commitment. However, whereas "green" marketing can backfire for those politically unaligned with the environmental movement[5], the package of virtues connected with AVs—safety, but also reductions in traffic and parking congestion—contain uncontroversial values that allow consumers to advertise themselves as safe, smart, and prosocial.

<Table about here>

Risk Heuristics and Algorithm Aversion

When the first traffic fatality involving Tesla's Autopilot occurred in May 2016, it was covered by every major news organization—a feat unmatched by any of the other 40,200 US traffic fatalities that year. We can expect an even larger reaction the first time an AV kills a pedestrian, or kills a child, or two AVs crash into each other. Outsized media coverage of crashes involving AVs may feed and amplify people's fears by tapping into the availability heuristic (risks are subjectively higher when they come to mind easily) and affective heuristic (risks are perceived to be higher when they evoke a vivid emotional reaction). As with airplane crashes, the more disproportionate—and disproportionately sensational—the coverage that AV accidents receive, the more exaggerated people will perceive the risk and dangers of these cars in comparison to those of traditional human-driven ones. Worse, for AVs these reactions may be compounded by *algorithm aversion*[6], the tendency for people to more rapidly lose faith in an erring decision-making algorithm than in humans making comparable errors.

These reactions could derail the adoption of AVs through numerous paths; it could directly deter consumers, it could provoke politicians to enact suffocating restrictions, or it could create outsized liability issues—fuelled by court and jury overreactions—that compromise the financial feasibility of AVs. Each path could slow or even stall widespread adoption.

Countering these powerful psychological effects may prove especially difficult. Nevertheless, there are opportunities. AV spokespeople should prepare the public for the inevitability of accidents—not overpromising infallibility, but still emphasizing AVs' safety advantages over human drivers. One barrier that prevents people from adopting (superior) algorithms over human judgment is overconfidence in one's own performance[7]—something famously prevalent in driving. Manufacturers should also be open about algorithmic improvements. AVs are better portrayed as being perfected, not as perfect.

Politicians and regulators can also play a role in managing overreaction. Though human themselves, and ultimately answerable to the public, legislators should resist capitulating to the public's fears of low-probability risks[8]. Instead they should educate the public about the actual risks and, if moved to act, do so in a calculated way, perhaps by offering the public "fear placebos" [8]—high-visibility, low-cost gestures that do the most to assuage the publics' fears without undermining the real benefits that AVs might bring.

Asymmetric Information and the Theory of the Machine Mind

The dubious reputation of the CIA is sometimes blamed on the asymmetry between the secrecy of their successes and the broad awareness of their failures. AVs will face a similar challenge. Passengers will be acutely aware of the cars' rare failures—leading to the issues described above—but may be blissfully unaware of all the cars' small successes and optimizations.

This asymmetry of information is part of a larger psychological barrier to the trust in AVs: the opacity to the decision-making occurring under the hood. If trust is characterized by the willingness to yield vulnerability to another entity, it is critical that people can comfortably predict and understand the behaviour of the other entity. Indeed, the European Union General Data Protection Regulation recently established the citizen's "right to [...] obtain an explanation of the decision reached […] and to challenge the decision" made by algorithms[9].

However, full transparency may be neither possible nor optimal. AV intelligence is driven in part by machine learning, in which computers learn increasingly sophisticated patterns without being explicitly taught. This leaves underlying decision-making processes opaque even to the programmer (let alone the passenger). But even if a detailed account of the computer's decisions were available, it would only offer the end-user an incomprehensible deluge of information. The trend in many "lower stakes" computer interfaces (e.g. web-browsers) has thus been in the opposite direction—hiding the complex decision-making of the machine in order to present a simple, minimalistic user experience. For AVs, whereas some transparency can improve trust, too much transparency into the explanations for the car's actions can overwhelm the passenger, increasing anxiety[10].

Thus, what is most important for generating trust and comfort is not full transparency but communication of the right amount *and kind* of information to allow people to develop mental models (an abstract representation of the entity's perceptions and decision rules) of the cars[5]—a sort of theory of the machine mind. There is already a robust literature investigating what information is most crucial to communicate, however most of this research has been conducted on AI in industrial, residential, or software interface settings. Not all of it will be perfectly transferable to AVs, so researchers need to investigate what information best fosters predictability, trust and comfort in this new and specific setting. Moreover, AVs will need to communicate not just with their passengers, but with pedestrians, fellow drivers, and the other stakeholders on the road. Currently, people decipher the intentions of other drivers through explicit signals (blinkers, horns, gestures) and through assumptions based on the mental models formed of drivers (why is she slowing down here? Why is he positioning himself like that?). Everyone on the road will need to adjust their human models to those of AVs, and the more research delineating what information people find crucial and comforting, the more seamless and less panicky this transition will be.

A new social contract

Automobiles began their transformational integration into our lives over a century ago. In this time, a system of laws regulating the behaviour of drivers and pedestrians, and the designs and practices of manufacturers, has been introduced and continuously refined. Today, the technologies that mediate these regulations, and the norms, fines and other punishments that enforce them, maintain just enough trust in the traffic system to keep it tolerable. Tomorrow, the integration of autonomous cars will be similarly transformational, but will occur over a much shorter timescale. In that time, we will need a new social contract that provides clear guidelines about who is responsible for different kinds of accidents, how monitoring and enforcement will be performed, and how trust among all stakeholders can be engendered. Many challenges remain—hacking, liability, and labour displacement issues, most significantly— but this social contract will be bound as much by psychological realities as by technological and legal ones. We have identified several here, but more work remains. We believe it is morally imperative for behavioural scientists of all disciplines to weigh in on this contract. Every day the adoption of autonomous cars is delayed is another day that people will continue to lose their lives to the non-autonomous human drivers of yesterday.

References

1. American Automobile Association. “Americans Feel Unsafe Sharing the Road with Fully AVs” Retrieved July 17, 2017 from http://newsroom.aaa.com/2017/03/americans-feel-unsafe-sharing-road-fully-self-driving-cars/
2. Wortham, R. H., & Theodorou, A. *Connection Science*, **29**, 242-248 (2017).
3. Bonnefon, J. F., Shariff, A., & Rahwan, I. *Science*, **352**, 1573-1576 (2016).
4. Griskevicius, V., Tybur, J. M., & Van den Bergh, B. *Journal of Personality and Social Psychology*, **98**, 392 (2010).
5. Gromet, D. M., Kunreuther, H., & Larrick, R. P. *Proceedings of the National Academy of Sciences*, **110**, 9314-9319 (2013).
6. Dietvorst, B. J., Simmons, J. P., & Massey, C. *Journal of Experimental Psychology: General*, **144**, 114 (2015).
7. Sieck, W. R., & Arkes, H. R. *Journal of Behavioral Decision Making*, **18**, 29-53 (2005).
8. Sunstein, C. R., & Zeckhauser, R. *Environmental and Resource Economics*, **48**, 435-449 (2011).
9. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April. Official Journal of the European Union (2016).
10. Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. *International Journal on Interactive Design and Manufacturing,* **9**, 269-275 (2014).

Competing interests

The authors declare no competing interests.

Table: A summary of the psychological challenges to AVs, and suggest actions for overcoming them.

Psychological Challenge	**Suggested Actions**
The Dilemmas of Autonomous Ethics People are torn between how they want AVs to ethically behave; they morally believe the vehicles should operate under utilitarian principles, but prefer to buy vehicles that prioritize their own lives as passengers. The idea of a car sacrificing its passengers deters people from purchasing a AV.	• Shift the discussion from the relative risk of injury to the absolute risk. • Appeal to consumers' desire for virtue signaling.
Risk Heuristics and Algorithmic Aversion The novelty and nature of AVs will result in outsized reactions in the face of inevitable accidents. Such overreactions risk slowing or stalling the adoption of AVs.	• Prepare the public for the inevitability of accidents. • Openly communicate algorithmic improvement. • Manage public overreaction with "fear placebos" and information about actual risks levels.
Asymmetric Information and the Theory of the Machine Mind A lack of transparency into the underlying decision-making processes can make it difficult for people to predict the AVs' behavior, diminishing trust.	• Research the type of information required to form trustable mental models of AVs.

Figure: A schematic example of the ethical tradeoffs AVs will need to make between the lives of passengers and pedestrians[3] taken from the Moral Machine web site (http://moralmachine.mit.edu), which we launched to collect large-scale data from the public. So far, we have collected over 30 million decisions from over 3 million people.