Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?

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

[1]enter for Research in Management, Toulouse School of Economics, Toulouse, 31000, France
[2]Department of Psychology, 1277 University of Oregon, Eugene, OR 97403-1227, USA
[3]Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA

October 13, 2015

Abstract

The wide adoption of self-driving, Autonomous Vehicles (AVs) promises to dramatically reduce the number of traffic accidents. Some accidents, though, will be inevitable, because some situations will require AVs to choose the lesser of two evils. For example, running over a pedestrian on the road or a passer-by on the side; or choosing whether to run over a group of pedestrians or to sacrifice the passenger by driving into a wall. It is a formidable challenge to define the algorithms that will guide AVs confronted with such moral dilemmas. In particular, these moral algorithms will need to accomplish three potentially incompatible objectives: being consistent, not causing public outrage, and not discouraging buyers. We argue to achieve these objectives, manufacturers and regulators will need psychologists to apply the methods of experimental ethics to situations involving AVs and unavoidable harm. To illustrate our claim, we report three surveys showing that laypersons are relatively comfortable with utilitarian AVs, programmed to minimize the death toll in case of unavoidable harm. We give special attention to whether an AV should save lives by sacrificing its owner, and provide insights into (i) the perceived morality of this self-sacrifice, (ii) the willingness to see this self-sacrifice being legally enforced, (iii) the expectations that AVs will be programmed to self-sacrifice, and (iv) the willingness to buy self-sacrificing AVs.

1 Introduction

In 2007, a sequence of technical advances enabled six teams to complete the DARPA Urban Challenge, the first benchmark test for autonomous driving in realistic urban environments [Montemerlo, B. et al, 2008, Urmson, C. et al, 2008]. Since then, major research labs spearheaded efforts to build and test Autonomous Vehicles (AVs). The Google Car, for example, has already succeeded in covering thousands of miles of real-road driving [Waldrop, 2015]. AVs promise world-changing benefits by increasing traffic efficiency [Van Arem et al., 2006], reducing pollution [Spieser, K. et al, 2014], and eliminating up to 90% of traffic accidents [Gao et al., 2014].

Not all accidents will be avoided, though, and accidents involving driverless cars will create the need for new kinds of regulation—especially in cases where harm cannot be entirely avoided.

Fig. 1 illustrates three situations in which harm is unavoidable. In one case, an AV may avoid harming several pedestrians by swerving and sacrificing a passer-by. In the other two cases, the AV is faced with the choice of sacrificing its own passenger, who can be its owner, in order to save one or more pedestrians. The behavior of AVs in these situations of unavoidable harm raises non-trivial questions, both legal and moral, which have commanded public interest but have not yet been empirically investigated. Indeed, while we could easily trace multiple public discussions of these situations, we have not identified a single scientific article dealing with the choices an AV must make in situations of unavoidable harm.

On the legal side, since the control algorithm ultimately makes the decision to hit a pedestrian, passer-by, or wall, it is not obvious whether the passenger should be held legally accountable for this outcome. Some argue that liability must shift from the driver to the manufacturer, because failure to anticipate decisions like those in Fig. 1 may amount to negligence in design under Product Liability law [Villasenor, 2014]. In this case, car manufacturers may become the de facto insurers against accidents, or may rely on the expertise of existing motor insurers [Jain et al., 2015]. Today, facing regulation that lags behind technology [UK Department for Transport, 2015], manufacturers of semi-autonomous vehicles are considering ad hoc workarounds. For example, to minimize manufacturer liability associated with their new 'automated overtaking' feature, Tesla Motors will actually require the driver to initiate the feature, thus ensuring legal responsibility for the maneuver's consequences falls with the driver [Ramsey, 2015].

In this context, defining the algorithms that will guide AVs in situations of unavoidable harm is a formidable challenge, given that the decision to run over pedestrians, to swerve into a passer-by, or to self-destruct is a moral one [Marcus, 2012]. Indeed, in situations where harm cannot entirely be avoided, it must be distributed, and the distribution of harm is universally considered to fall within the moral domain [Haidt, 2012]. Accordingly, the control algorithms of AVs will need to embed moral principles guiding their decisions in situations of unavoidable harm [Wallach and Allen, 2008]. In this respect, manufacturers and regulators will need to accomplish three potentially incompatible objectives: being reasonably consistent, not causing public outrage, and not discouraging buyers.

Not discouraging buyers is a commercial necessity—but it is also in itself a moral imperative, given the social and safety benefits AVs provide over conventional cars. Meanwhile, avoiding public outrage, that is, adopting moral algorithms that align with human moral attitudes, is key to fostering public comfort with allowing the broad use of AVs in the first place. However, to pursue these two objectives simultaneously 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 car should swerve. This would fit a utilitarian moral doctrine [Rosen, 2005], according to which the moral course of action is to minimize the death toll. But consider then the case displayed in Fig. 1c. The utilitarian course of action, in that situation, would be for the car to swerve and kill its owner—but a driverless car programmed

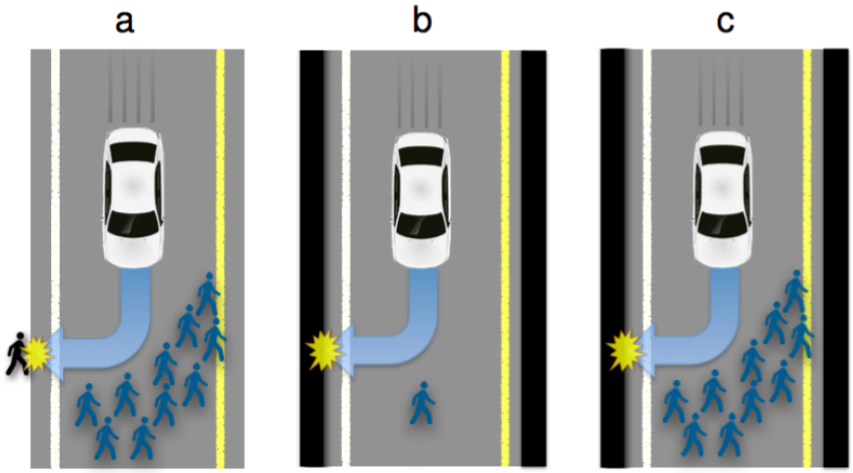

Figure 1: Three traffic situations involving imminent unavoidable harm. (a) The car can stay on course and kill several pedestrians, or swerve and kill one passer-by. (b) The car can stay on course and kill one pedestrian, or swerve and kill its passenger. (c) The car can stay on course and kill several pedestrians, or swerve and kill its passenger.

to follow this course of action might discourage buyers, who may consider that 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.

We suggest that a way out of this conundrum is to adopt a data-driven approach, inspired by experimental ethics [Greene, 2014a], to identify moral algorithms that people are willing to accept as citizens and to be subjected to as car owners. Indeed, situations of unavoidable harms, as illustrated in Fig. 1, bear a striking resemblance with the flagship dilemmas of experimental ethics—that is, the so called 'trolley problems'. In typical trolley problems, an agent has the opportunity to save several persons by sacrificing the life of a single individual. This decision is commonly labeled as utilitarian, since it amounts to accepting the infliction of harm in the name of the greater good.

Trolley problems inspired a huge number of experiments, aimed at identifying both the antecedents and the cognitive processing leading to a utilitarian decision [Greene, 2014c, Greene, 2014b]. The results of these experiments do not apply well to situations involving AVs, though. For example, some pressing questions about AVs do not make sense in typical trolley problems that involve a human decision-maker: Would people purchase AVs programmed to be utilitarian? What are their expectations about the way car manufacturers will program AVs? Other questions could have been addressed in a human context, but were not because they were

only made salient in the context of AVs. For example, would people accept legal enforcement of utilitarian decisions, and more so for AVs than for humans? And finally, perhaps the most striking problem created by AVs, at least in the public opinion, is whether a car may decide to sacrifice its passenger in order to save several lives on the road. How do people balance their motivations for self-preservation as drivers with the protection of pedestrians and other drivers, when considering whether to own an AV, when thinking about how AV will or should be programmed, and when pondering which programming might be legally enforced?

We believe that regulators and manufacturers will soon be in pressing need of answers to these questions, and that answers are most likely to come from surveys employing the protocols of experimental ethics. To be clear, we do not mean that these thorny ethical questions can be solved by polling the public and enforcing the majority opinion. Survey data will inform the construction and regulation of moral algorithms for AVs, not dictate them. The importance of survey data must not be underestimated, though, since algorithms that go against the moral expectations of citizens (or against the preferences or consumers) are likely to impair the smooth adoption of AVs.

Having argued that autonomous vehicles need experimental ethics, we offer to illustrate this approach in the rest of this article, by reporting a series of three surveys built upon the protocols commonly used to study trolley problems, which we adapted to the new context of traffic accidents involving AVs. In this first series of surveys, we focus on the question of whether respondents would approve of utilitarian AVs, programmed to minimize the death toll of an accident, even when it requires to sacrifice the car's own passenger.

2 Materials and Methods

We conducted three online surveys in June 2015. All studies were programmed on Qualtrics survey software and recruited participants from the Mechanical Turk platform, for a compensation of 25 cents. In all studies, participants provided basic demographic information, such as age, sex, and religiosity. At the end of all studies, three 7-point scales measured the overall enthusiasm that participants felt for AVs: How excited they felt about a future in which autonomous self-driving cars are an everyday part of our motoring experience, how fearful they were of this future (reverse coded), and how likely they were to buy an autonomous vehicle, should they become commercially available by the time they would next purchase a new car. Cronbach alpha for these three items was .82 in Study 1, .78 in Study 2, and .81 in Study 3. Regression analyses (shown in the Supplementary Information) showed that enthusiasm was significantly greater for male participants, in all studies; and that enthusiasm was significantly greater for younger and less religious participants in two studies, although these two effects were smaller than that of sex. Given these results, all subsequent analyses controlled for sex, age and religiosity.

2.1 Study 1

Participants (241 men and 161 women, median age = 30) read a vignette in which one or more pedestrians could be saved if a car were to swerve into a barrier, killing its passenger. Participants were randomly assigned to one group of a 2×2×2 between-participant design, which manipulated the number of pedestrians which could be saved (one vs ten), whether the driver or the self-driving car would make the decision

to swerve, and whether participants were instructed to imagine themselves in the car, or to imagine an anonymous character. The experiment was programmed so that the sex of this anonymous character always matched that of the participant (this information being collected early in the experiment). In addition to the vignette, participants were given a pictorial representation of the SWERVE and STAY options. See the Supplementary Information for detailed examples of the vignettes and their pictorial representation.

2.2 Study 2

Participants (144 men and 166 women, median age = 31) were randomly assigned to one of four versions of a vignette, in which either 1 or 10 pedestrians could be saved if an autonomous, self-driving car were to swerve into either a barrier (killing its passenger), or another pedestrian (killing that pedestrian). Participants were told that three algorithms could be implemented in the car for such a contingency: ALWAYS STAY, ALWAYS SWERVE, and RANDOM. Additionally, participants were given a pictorial representation of the SWERVE and STAY options. Participants then answered three questions, introduced in random order for each participant: How would you rate the relative morality of these three algorithms? How would you rate your relative willingness to buy a car with one of these three algorithms? How would you rate the relative comfort with having each of the car with one of these three algorithms on the roads in your part of the world? To provide their responses, participants assigned points to each of the three algorithms, taken from an overall budget of 100 points. See the Supplementary Information for detailed examples of the vignettes, their pictorial representation, and a screen capture of the budget splitting procedure.

2.3 Study 3

Participants (97 men and 104 women, median age = 30) read a vignette in which ten pedestrians could be saved if a car were to swerve into a barrier, killing its passenger. Participants were randomly assigned to one of two groups, manipulating whether participants were instructed to imagine themselves in the car, or to imagine an anonymous character. The experiment was programmed so that the sex of this anonymous character always matched that of the participant (this information being collected early in the experiment). Thus, the experimental vignettes were similar to that used in Study 1 (restricted to situations involving ten pedestrians). Participants were asked whether the moral course of action was to swerve or stay on course (dichotomous response variable), and whether they expected future driverless cars to be programmed to swerve or to stay on course in such a situation (dichotomous response variable). Next, participants were asked to indicate whether it would be more appropriate to "protect the driver at all costs" or to "maximize the number of lives saved," on a continuous slider scale anchored at these two expressions.

3 Results

Results suggest that participants were generally comfortable with utilitarian AVs, programmed to minimize an accident's death toll. This is especially true of situations that do not involve the sacrifice of the AV's owner. For example, Fig. 2b (right) displays the responses of participants in Study 2, who expressed their willingness to buy

(and their comfort with others buying) different types of AVs. An analysis of covariance showed that participants preferred AVs programmed to swerve into a passer-by when it could save 10 (but not one) pedestrians, $F(1, 141) = 41.3$, $p < .001$, whether they were thinking of buying the car themselves or having other people drive the car.

Critically important, however, are preferences and expectations about whether an AV should be programmed to save lives by sacrificing its owner. Our results provide information on (i) the morality of this self-sacrifice, (ii) the willingness to see this self-sacrifice being legally enforced, (iii) the expectations that AVs will be programmed to self-sacrifice, and (iv) the willingness to buy self-sacrificing cars.

First, participants in Study 1 approved of AVs making utilitarian self-sacrifices to the same extent as they approved this decision for human drivers (Fig. 2a, left). This was consistent with the results of an analysis of covariance where morality was the dependent outcome, and the predictors were the number of lives that could be saved; whether the agent making the sacrifice decision was human or machine; whether participants pictured themselves or another person in the car; and all the 2-way and 3-way interaction terms. (As in all analyses and as explained in the Materials and Methods section, age, sex, and religiosity were entered as covariates). No effect was detected as significant but that of the number of lives which could be saved, $F(1, 367) = 48.9$, $p < .001$.

Second, participants in Study 1 were more amenable to legally enforcing self-sacrifice for AVs than for human drivers (Fig. 2a, right). When the dependent variable in the analysis of covariance was the the legal enforceability of self-sacrifice, the analysis detected a significant effect of the number of lives that could be saved, $F(1, 367) = 7.3$, $p < .01$; and an effect of whether the decision-maker was human or machine, $F(1, 367) = 9.2$, $p < .01$. Participants were more willing to consider legal enforcement of the sacrifice when it saved a greater number of lives, and when the decision was made by a self-driving car, rather than by a human driver.

Third, whereas 3/4 of participants in Study 3 believed that an AV should self-sacrifice to save ten lives, less than 2/3 (a significantly lower proportion, McNemar's $\chi^2 = 6.7$, $p < .01$) believed that AVs would actually be programmed that way in the future. Similarly, whereas these participants believed that AVs should, generally speaking, place the greater good over the safety of the passenger, they were significantly less confident that manufacturers would programmed AVs with this goal, $F(1, 179) = 18.9$, $p < .001$. Thus, rather than being concerned about AVs being too utilitarian (as is often portrayed in science fiction), participants were generally wary that AVs would be programmed to protect their passengers at all costs.

Fourth, whereas participants in Study 2 generally supported others buying AVs programmed for utilitarian self-sacrifice, they were less willing to buy such AVs themselves (even when the sacrifice would save ten pedestrians). We ran an analysis of covariance of the budget share allocated to the AV programmed to self-sacrifice, in which the predictors were the number of lives that could be saved and whether participants were asked about their own willingness to buy such car, or their comfort with seeing other people drive such a car. Participants allocated a larger budget share to the self-sacrificing AV when it could save 10 pedestrians, $F(1, 132) = 7.6$, $p < .01$, and when they were asked about other people traveling in the AV, rather than

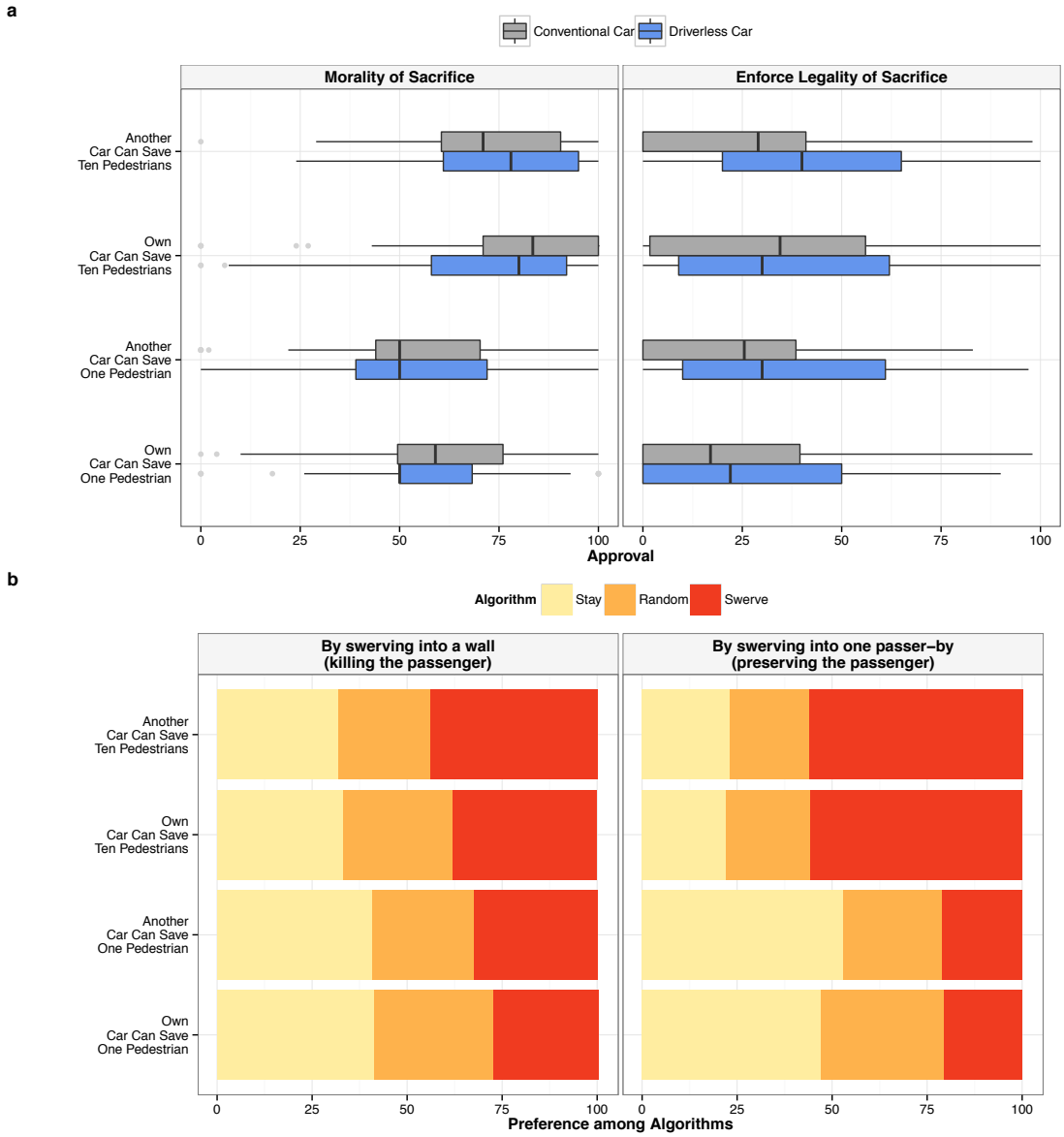

Figure 2: Selected experimental results. (a) People approve of autonomous vehicles (AVs) sacrificing their passenger to save pedestrians, and show greater willingness to see this action legally enforced when the decision is made by an AV than by a human driver. (b) People are willing to buy AVs programmed to swerve into a passer-by or into a wall in order to save pedestrians, although their utilitarianism is qualified by a self-preserving bias.

themselves, $F(1, 135) = 4.9$,$p = .03$. When considering which AV they would buy, participants were split between AVs that would self-sacrifice, AVs that would not, and AVs that would randomly pick a course of action.

4 Discussion

Three surveys suggested that respondents might be prepared for autonomous vehicles programmed to make utilitarian moral decisions in situations of unavoidable harm. This was even true, to some extent, of situations in which the AV could sacrifice its owner in order to save the lives of other individuals on the road. Respondents praised the moral value of such a sacrifice the same, whether human or machine made the decision. Although they were generally unwilling to see self-sacrifices enforced by law, they were more prepared for such legal enforcement if it applied to AVs, than if it applied to humans. Several reasons may underlie this effect: unlike humans, computers can be expected to dispassionately make utilitarian calculations in an instant; computers, unlike humans, can be expected to unerringly comply with the law, rendering moot the thorny issue of punishing non-compliers; and finally, a law requiring people to kill themselves would raise considerable ethical challenges.

Even in the absence of legal enforcement, most respondents agreed that AVs should be programmed for utilitarian self-sacrifice, and to pursue the greater good rather than protect their own passenger. However, they were not as confident that AVs would be programmed that way in reality—and for a good reason: They actually wished others to cruise in utilitarian AVs, more than they wanted to buy utilitarian AVs themselves. What we observe here is the classic signature of a social dilemma: People mostly agree on what should be done for the greater good of everyone, but it is in everybody's self-interest not to do it themselves. This is both a challenge and an opportunity for manufacturers or regulatory agencies wishing to push for utilitarian AVs: even though self-interest may initially work against such AVs, social norms may soon be formed that strongly favor their adoption.

We emphasize that our results provide but a first foray into the thorny issues raised by moral algorithms for AVs. For example, our scenarios did not feature any uncertainty about decision outcomes, but future work will have to test scenarios that introduce 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 car, than for the rider of the motorcycle? Should different decisions be made when children are on board, since they both have a longer time ahead of them than adults, and had less agency in being in the car in the first place? 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?

Figuring out how to build ethical autonomous machines is one of the thorniest challenges in artificial intelligence today [Deng, 2015]. As we are about to endow millions of vehicles with autonomy, taking algorithmic morality seriously has never been more urgent. Our data-driven approach highlights how the field of experimental ethics can give us key insights into the moral and legal standards that people expect from autonomous driving algorithms. When it comes to split-second moral judgments, people may very well expect more from machines than they do

from each other.

References

[Deng, 2015] Deng, B. (2015). Machine ethics: The robot's dilemma. *Nature*, 523:24–26.

[Gao et al., 2014] Gao, P., Hensley, R., and Zielke, A. (2014). A road map to the future for the auto industry.

[Greene, 2014a] Greene, J. D. (2014a). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics*, 124:695–726.

[Greene, 2014b] Greene, J. D. (2014b). The cognitive neuroscience of moral judgment and decision making. In Gazzaniga, M. S., editor, *The cognitive neurosciences V*, pages 1013–1023. MIT Press.

[Greene, 2014c] Greene, J. D. (2014c). *Moral tribes: Emotion, reason and the gap between us and them*. Atlantic Books.

[Haidt, 2012] Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Pantheon Books.

[Jain et al., 2015] Jain, N., O'Reilly, J., and Silk, N. (2015). Driverless cars: Insurers cannot be asleep at the wheel.

[Marcus, 2012] Marcus, G. (2012). Moral machines.

[Montemerlo, B. et al, 2008] Montemerlo, B. et al (2008). Junior: The stanford entry in the urban challenge. *J. Field Robotics*, 25:569–597.

[Ramsey, 2015] Ramsey, M. (2015). Who's responsible when a driverless car crashes? tesla's got an idea.

[Rosen, 2005] Rosen, F. (2005). *Classical utilitarianism from Hume to Mill*. Routledge.

[Spieser, K. et al, 2014] Spieser, K. et al (2014). Toward a systematic approach to the design and evaluation of automated mobility-on-demand systems: A case study in Singapore. In Meyer, G. and Beiker, S., editors, *Road Vehicle Automation*, pages 229–245. Springer.

[UK Department for Transport, 2015] UK Department for Transport (2015). The pathway to driverless cars.

[Urmson, C. et al, 2008] Urmson, C. et al (2008). Autonomous driving in urban environments: Boss and the urban challenge. *J. Field Robotics*, 25:425–266.

[Van Arem et al., 2006] Van Arem, B., Van Driel, C. J., and Visser, R. (2006). The impact of cooperative adaptive cruise control on traffic-flow characteristics. *IEEE Transactions on Intelligent Transportation Systems*, 7:429–436.

[Villasenor, 2014] Villasenor, J. (2014). Products liability and driverless cars: Issues and guiding principles for legislation.

[Waldrop, 2015] Waldrop, M. M. (2015). Autonomous vehicles: No drivers required. *Nature*, 518:20–23.

[Wallach and Allen, 2008] Wallach, W. and Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

A Experimental Methods

The article is based on three online surveys conducted in June 2015. All studies were programmed on Qualtrics and recruited participants from the Mechanical Turk platform, for a compensation of 25 cents. Study 1 involved 402 participants (161 women, median age = 30); Study 2 involved a different sample of 310 participants (166 women, median age = 31); and Study 3 involved yet a different sample of 201 participants (104 women, median age = 30). We did not exclude any participant from the analyses, but because some participants failed to answer some questions, degrees of freedom show small variations between analyses. Participants provided basic demographic information, such as age, sex, ethnicity, political ideology, and religiosity (the latter being measured on a continuous slider from *not at all religious* to *extremely religious*). At the end of all studies, three questions measured the overall enthusiasm that participants felt for driverless cars: How excited they felt about a future in which autonomous self-driving cars are an everyday part of our motoring experience (7-point scale from *not at all* to *very much*), how fearful they were of this future (same scale, reverse coded), and how likely they were to buy an autonomous vehicle, should they become commercially available by the time they would next purchase a new car (7-point scale from *not at all likely* to *extremely likely*).

A.1 Study 1

Participants read a vignette in which one or more pedestrians could be saved if a car were to swerve into a barrier, killing its passenger. Participants were randomly assigned to one group of a 2×2×2 between-participant design, which manipulated the number of pedestrians which could be saved (one vs ten), whether the driver or the self-driving car would make the decision to swerve, and whether participants were instructed to imagine themselves in the car, or to imagine an anonymous character. The experiment was programmed so that the sex of this anonymous character always matched that of the participant (this information being collected early in the experiment). Here is an example of the vignette (one pedestrian, participant in the car, driver makes the decision):

> You are the sole passenger driving at the speed limit down a main road. Suddenly, a pedestrian appears ahead, in the direct path of the car. You have the option to: SWERVE off to the side of road, where you will impact a barrier, killing you but leaving the pedestrian unharmed, or STAY on your current path, where you will kill the pedestrian, but you will be unharmed.

And here is the mirror example, as it would appear to a male participant (i.e., the anonymous individual in the car is male):

> A man is the sole passenger in an autonomous self-driving vehicle traveling at the speed limit down a main road. Suddenly, 10 people appear ahead, in the direct path of the car. The car could be programmed to: SWERVE off to the side of road, where it will impact a barrier, killing the passenger but leaving the ten pedestrians unharmed, or STAY on its current path, where it will kill the 10 pedestrians, but the passenger will be unharmed.

In addition to the vignette, participants were given a pictorial representation of the SWERVE and STAY options. Figure 3 (top) displays the images that were shown to participants, in the

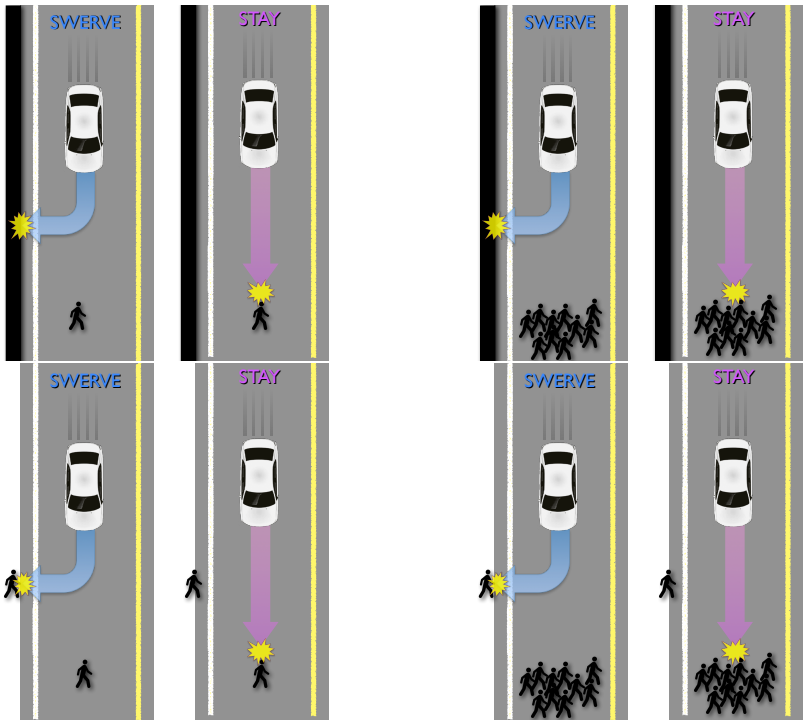

Figure 3: Pictorial representation of the SWERVE and STAY options in Study 1 (top) and Study 2 (all).

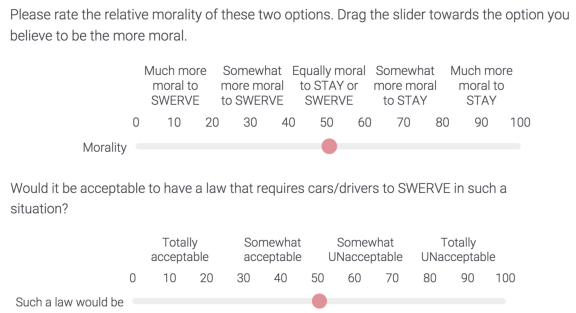

Figure 4: Main measures of Study 1 (moral approval and legal enforceability of sacrifice).

conditions where one or ten pedestrians could be saved.

In Study 1, we recorded the perceived morality of the SWERVE option (i.e., the utilitarian sacrifice) as well as it legal enforceability, that is, the extent to which participants found it acceptable to have a law that would require cars/drivers to SWERVE. The sliders that were used for these two measures are shown in Figure 4. Responses were reverse coded so that higher scores would reflect higher moral approval, and higher legal enforceability of the SWERVE option.

A.2 Study 2

Participants were randomly assigned to one of four versions of a vignette, in which either 1 or 10 pedestrians could be saved if an autonomous, self-driving car were to swerve into either a barrier (killing its passenger), or another pedestrian (killing that pedestrian). Participants were told that three algorithms could be implemented in the car for such a contingency: ALWAYS STAY, ALWAYS SWERVE, and RANDOM. Here is an example of the vignette (one pedestrian, swerve into barrier):

> You are the sole passenger riding in an autonomous, self-driving car that is traveling at high speed down a main road. Suddenly, one pedestrian appears ahead, in the direct path of the car. In preparing the car for such an eventuality, the computer engineers can program the car for three options:
>
> **always stay.** In such a situation, the car would continue on its path, kill the pedestrian on the main road, but you as the passenger will be unharmed.
>
> **always swerve.** In such a situation, the car would swerve quickly, diverting the car onto the side road where it will kill you as the passenger, but the pedestrian on the main road will be unharmed.
>
> **random.** In such a situation, the car would be programmed to randomly choose to either STAY or SWERVE.

And here is the mirror example (ten pedestrian, swerve into another pedestrian):

> You are the sole passenger riding in an autonomous, self-driving car that is traveling at high speed down a main road. Suddenly, ten pedestrians appears ahead, in the direct path of the car. The car could however swerve to the right where it would kill one pedestrian on the side of the road. In preparing the car for such an eventuality, the computer engineers can program the car for three options:
>
> **always stay.** In such a situation, the car would continue on its path, kill the ten pedestrians on the main road, but the pedestrian on the side of the road would be unharmed.
>
> **always swerve.** In such a situation, the car would swerve quickly, diverting the car onto the side road where it will kill the one pedestrian on the side

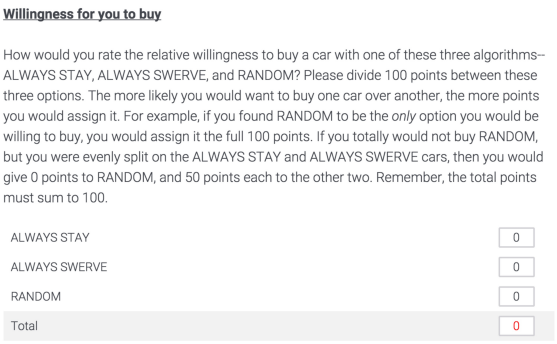

Figure 5: Illustration of the budget-splitting procedure used in Study 2, here applied to the willingness to buy self-driving cars programmed to ALWAYS STAY, ALWAYS SWERVE, or RANDOM.

of the road, but the pedestrian on the main road will be unharmed.

random. In such a situation, the car would be programmed to randomly choose to either STAY or SWERVE.

Additionally, participants were given a pictorial representation of the SWERVE and STAY options (see Figure 3). Participants then answered three questions, introduced in random order for each participant: How would you rate the relative morality of these three algorithms? How would you rate your relative willingness to buy a car with one of these three algorithms? How would you rate the relative comfort with having each of the car with one of these three algorithms on the roads in your part of the world? To provide their responses, participants assigned points to each of the three algorithms, taken from an overall budget of 100 points (Figure 5).

A.3 Study 3

Participants read a vignette in which ten pedestrians could be saved if a car were to swerve into a barrier, killing its passenger. Participants were randomly assigned to one of two groups, manipulating whether participants were instructed to imagine themselves in the car, or to imagine an anonymous character. The experiment was programmed so that the sex of this anonymous character always matched that of the participant (this information being collected early in the experiment). Thus, the experimental vignettes were similar to that used in Study 1 (restricted to situations involving ten pedestrians).

Participants were asked whether the moral course of action was to swerve or stay on course (dichotomous response variable), and whether they expected future driverless cars to be programmed to swerve or to stay on course in such a situation (dichotomous response variable). Next, participants were asked to indicate whether it would be more appropriate to *protect the driver at all costs* or to *maximize the number of lives saved,* on a continuous slider scale anchored at these two expressions.

B Experimental Results

In all studies, we computed an index of enthusiasm about driverless cars by averaging the excitement participants felt about a future with driverless cars, their fearfulness about this future (reverse coded), and the likelihood that they might buy a driverless car next, if such cars were available. Cronbach alpha for these three items was .82 in Study 1, .78 in Study 2, and .81 in Study 3. Regression analyses showed that enthusiasm was significantly greater for male participants, in all studies; and that enthusiasm was

significantly greater for younger and less religious participants in two studies (see Table 1), although these two effects were much smaller than that of sex. Given these results, all subsequent analyses controlled for sex, age and religiosity.

B.1 Study 1

To analyze participants' rating of the morality of the sacrifice (SWERVE), we ran an analysis of covariance where morality was the dependent outcome. The predictors were the number of lives that could be saved; whether the agent making the sacrifice decision was human or machine; whether participants pictured themselves or another person in the car; and all the 2-way and 3-way interaction terms. As in all analyses, age, sex, and religiosity were entered as covariates. No effect was detected as significant but that of the number of lives which could be saved, $F(1, 367) = 48.9$, $p < .001$.

We ran a similar analysis in which the morality of the sacrifice was replaced as the dependent outcome with the legal enforceability of SWERVE. The analysis detected a significant effect of the number of lives that could be saved, $F(1, 367) = 7.3$, $p < .01$; and an effect of whether the decision-maker was human or machine, $F(1, 367) = 9.2$, $p < .01$. Participants were more willing to consider legal enforcement of the sacrifice when the decision was made by a self-driving car, rather than by a human driver.

This was mostly true, though, of situations in which they did not think of themselves in the car, as shown by analyzing separately the data provided by participants picturing another person in the car, and the data provided by participants picturing themselves in the car. In the first dataset, whether the decision maker was human or machine had a large impact, $F(1, 181) = 12.0$, $p < .001$; in the second dataset, this predictor had a nonsignificant impact, $F < 1$, $p = .37$.

B.2 Study 2

Our primary interest in Study 2 was participants' willingness to buy (or to have others buy) cars programmed with an ALWAYS SWERVE algorithm. We first analyzed the data provided by participants who faced a similar situation than in Study 1, that is, a dilemma between crashing the car into a barrier (killing its passenger) or staying on course (killing 1 or 10 passengers). We ran an analysis of covariance of the budget share allocated to the ALWAYS SWERVE car, in which the predictors were the number of lives that could be saved and whether participants were asked about their own willingness to buy such car, or their comfort with seeing other people drive such a car. Participants allocated a larger budget share to ALWAYS SWERVE when it could save 10 pedestrians, $F(1, 132) = 7.6$, $p < .01$, and when they were asked about other people driving the car, rather than buying the car themselves, $F(1, 135) = 4.9$,$p = .03$.

In the new situation in which the dilemma was between between swerving into a pedestrian, or staying on course and killing 1 or 10 pedestrians, participants' responses followed a purely utilitarian logic. The analysis of covariance showed that they allocated a significantly larger budget share to ALWAYS SWERVE when it could save 10 pedestrians, $F(1, 141) = 41.3$, $p < .001$, whether they were thinking of buying the car themselves or having other people drive the car.

	Study 1 (N = 402)		Study 2 (N = 302)		Study 3 (N = 201)	
	β	t	β	t	β	t
Sex (Women)	-0.96	-5.59^{***}	-0.88	-4.74^{***}	-0.71	-2.77^{**}
Age	-0.02	-2.79^{**}	-0.02	-2.57^{*}	-0.02	-1.60
Religiosity	-0.01	-2.45^{*}	-0.01	-2.15^{*}	-0.01	-0.99

Table 1: Impact of demographic variables on general enthusiasm for self-driving cars.

B.3 Study 3

Logistic regressions for the two dichotomous response variables (including the control variables sex, age and religiosity) did not detect any significant effect of the experimental condition (imagine yourself vs imagine another individual in the car). Though this may appear inconsistent with the results from Study 2, recall that here we are asking about moral appropriateness whereas there we were talking about willingness to buy.

Overall, 75% of participants deemed more moral to swerve, but only 65% predicted that driverless cars would be programmed to swerve. The difference between these two proportions was statistically significant, McNemar's $\chi^2(1) = 6.7$, $p < .01$.

On a scale from -50 (protect the driver at all costs) to $+50$ (maximize the number of lives saved), the average response was $+24$ (SD = 30) to the question of what was morally appropriate, but it was only $+8$ (SD =35) to the question of how future driverless cars would be programmed. The difference between these two responses was significant, as shown by a repeated-measure analysis of covariance which included the experimental condition as a between-group predictor, and sex, age and religiosity as control variables, $F(1, 179) = 18.9$, $p < .001$. The only other significant effect was that of the age covariate, $F(1, 176) = 7.8$, $p < .01$.