Reply to: Life and death decisions of autonomous vehicles

Edmond Awad[1,2], Sohan Dsouza[1], Richard Kim[1], Jonathan Schulz[3], Joseph Henrich[3], Azim Shariff[4]*, Jean-François Bonnefon[5]*, Iyad Rahwan[1,6,7]*

[1]The Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Department of Economics, University of Exeter Business School, Exeter, EX4 4PU, UK
[3]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[4]Department of Psychology, University of British Columbia, Vancouver, Canada
[5]Toulouse School of Economics (TSM-R), CNRS, Université Toulouse Capitole, Toulouse, France
[6]Institute for Data, Systems & Society, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[7]Center for Humans & Machines, Max-Planck Institute for Human Development, Berlin, Germany

* Corresponding authors: shariffa@uci.edu; jean-francois.bonnefon@tse-fr.eu; irahwan@mit.edu

In 'The Moral Machine Experiment' [1], we argued that policymakers would benefit from being aware of citizens' preferences regarding the behaviour of autonomous vehicles in critical situations—situations in which an autonomous vehicle cannot save everyone, but can still decide to save one group of road users or another. Bigman and Gray make the important point that the way we measure these preferences can affect the results we obtain.

Actual consumer choices cannot yet be recorded. If we want the ethics of these vehicles to be decided before they hit the market, we can only collect 'stated' preferences, based on hypothetical choices. The Moral Machine Experiment used a standard method for collecting stated preferences between multidimensional outcomes: Users chose between pairs of accidents which varied along multiple dimensions, and the importance of each dimension was statistically extracted from their choices, using conjoint analysis [2]. Typical surveys can only do this for a few dimensions, because of the exponential increase in required sample size for every additional dimension. Given the unusual scale of the Moral Machine Experiment, we were able to investigate nine dimensions simultaneously.

Bigman and Gray adopted a different method. Rather than having users go through multiple pairs of nine-dimensional outcomes, they asked eight separate questions about general policy preferences, one per dimension (the human-nonhuman dimension was not used in their survey). For example:

Should self-driving cars be programmed to:
(a) Kill children and save elderly people,
(b) Kill elderly people and save children, or
(c) Treat the lives of children and elderly people equally?

Bigman and Gray report that for all but one question (saving many vs. few), the most frequent response was (c). For example, about 80% of participants said that self-driving cars should 'treat the lives of children and elderly people equally'.

These results roughly agree with the Moral Machine results on some dimensions (e.g., the weak preference for inaction), and disagree on others (e.g., the preference for saving children), but the differences between the two methods, measures, and statistical analyses make any direct comparison difficult. The two different methods may differently tap a single, stable set of preferences, or they may elicit from respondents different facets of fragmented, inconsistent preferences that have yet to be solidified. Each approach comes with its own limitations, and its own usefulness. The Moral Machine approach allows us to measure the weight of different moral priorities when pitted against each other, rather than considered in isolation; but participants cannot explicitly state that one dimension (e.g., age) should *not* be taken into account. Of course, since each scenario involved at least two moral dimensions, respondents could avoid making decisions based on dimensions they felt shouldn't be programmed into the cars. Participants who believed that the vehicle should be blind to age, for instance, could endeavor to be systematically blind to age themselves in how they responded to the scenario pairs. Had millions of participants made this choice, this would have statistically resulted in an absence of a preference for age, and it would have ranked at the bottom of the list of the nine moral dimensions we tested. It remains, though, that individuals had no opportunity to explicitly express this preference for equality.

The approach used by Bigman and Gray does offer participants the opportunity to explicitly express a preference for equality. One limitation of this approach is that measurement becomes sensitive to social desirability, experimental demands, and framing effects (which is not to say that other methods do not have this problem). For example, consider the phrasings of the three response options above, and note how the word 'kill' disappears from the third option, making it instantly more attractive at a surface level. The first two options clearly describe tradeoffs, while the third option only has positive connotations. Now imagine an opposite framing for the third option: '*the self driving car should indiscriminately kill children and elderly people.*' This is as valid a description as that used by Bigman and Gray, but it seems less attractive in this negative framing. Indeed, in their Study 2, Bigman and Gray used a framing that stands somewhere in between the positive framing used in Study 1 and the negative framing we suggested above, and this intermediate framing already had an impact on the results: for half of the questions, the frequency of the 'equality' response decreased by 16 to 27 percentage points (as can be seen from comparing their Supplementary Table 1 and Supplementary Table 2).

We should note that an unpublished portion of the Moral Machine experiment used a third method—one similar to that of Bigman and Gray, but one that avoided this loaded language confound. After making 13 decisions, users had the option to 'help us better understand [their] decisions'. Users who agreed were taken to a page where they could position one slider for each of the nine dimensions explored by the Moral Machine. For example, one slider showed a baby on the left side, an elderly person on the right side, and was labeled 'Age preference'. Users could move the slider to express how important this dimension should be—more to the left if they wanted to save younger lives, more to the right if they wanted to save older lives. Importantly, this method *did* give participants the option to "*treat the lives of children or elderly people* [or men or women, or humans or pets] *equally*"; participants could easily express such a preference by positioning the slider at the midpoint of the scale. This is, in essence, the method used by Bigman and Gray—except that it uses a continuous measure rather than a 3-point scale, and that it does not use a textual description for the midpoint of the scale.

The original position of the sliders was not systematically the middle point of the scale, but rather a rough estimation of the preference of each individual user based on their responses to the Moral Machine. Thus, users had the opportunity to move sliders if they disagreed with that estimation. More than 99% of the users who saw the slider page moved at least one slider from its original position. Figure 1A displays the final position of all sliders for these 585,531 users, thus reflecting their choices when given the option of explicitly valuing all lives equally. Figure 1B displays the final position of each slider, only for those users who actually moved it. This is a stronger test, since it restricts the data to the responses of users who actively expressed a preference.

Both figures tell a similar, three-part story. At the top of each figure, we can see that four preferences that were estimated as strong in the Moral Machine experiment (saving humans, saving more lives, saving younger lives, saving pedestrians who cross legally, see Figure 2) are confirmed as strong. For these four dimensions, the distributions of responses are clearly skewed, and the modal response is not equality. At the bottom of each figure, we can see that four preferences which were identified as weak in the Moral Machine Experiment (inaction, saving pedestrians, saving fit characters, saving women) are confirmed as weak. The modal response for these dimensions is indeed equality.

Only for one dimension do we find a clear gap between the preferences extracted from the Moral Machine and the preferences explicitly expressed by users: whereas users' scenario-based choices indicated a preference for saving high-status characters over low-status characters, their expressed preference on the sliders is to treat them equally. Here we can see the value of giving people the opportunity to express an explicit preference: While their scenario-based choices may well show an implicit bias against lower-status victims, they would likely be unhappy if this bias was actually acted upon. Of course, it's extremely unlikely policymakers would propose that autonomous vehicles should discriminate based on social status—but we can still remain vigilant for other gaps between implicit biases and explicit preferences for equality, whenever they concern characteristics that may enter policy debates.

Self-driving car fatalities are an inevitability, but the type of fatalities that ethically offend the public and derail the industry, are not. As a result, it seems important to anticipate, as accurately as we can, how the public will actually feel about the ethical decisions we program into these vehicles. Since any method used to collect these preferences will come with its own biases and limitations, the methodological diversity advocated by Bigman and Gray, and the broad involvement of psychologists more generally, will be critical to reaching that goal.

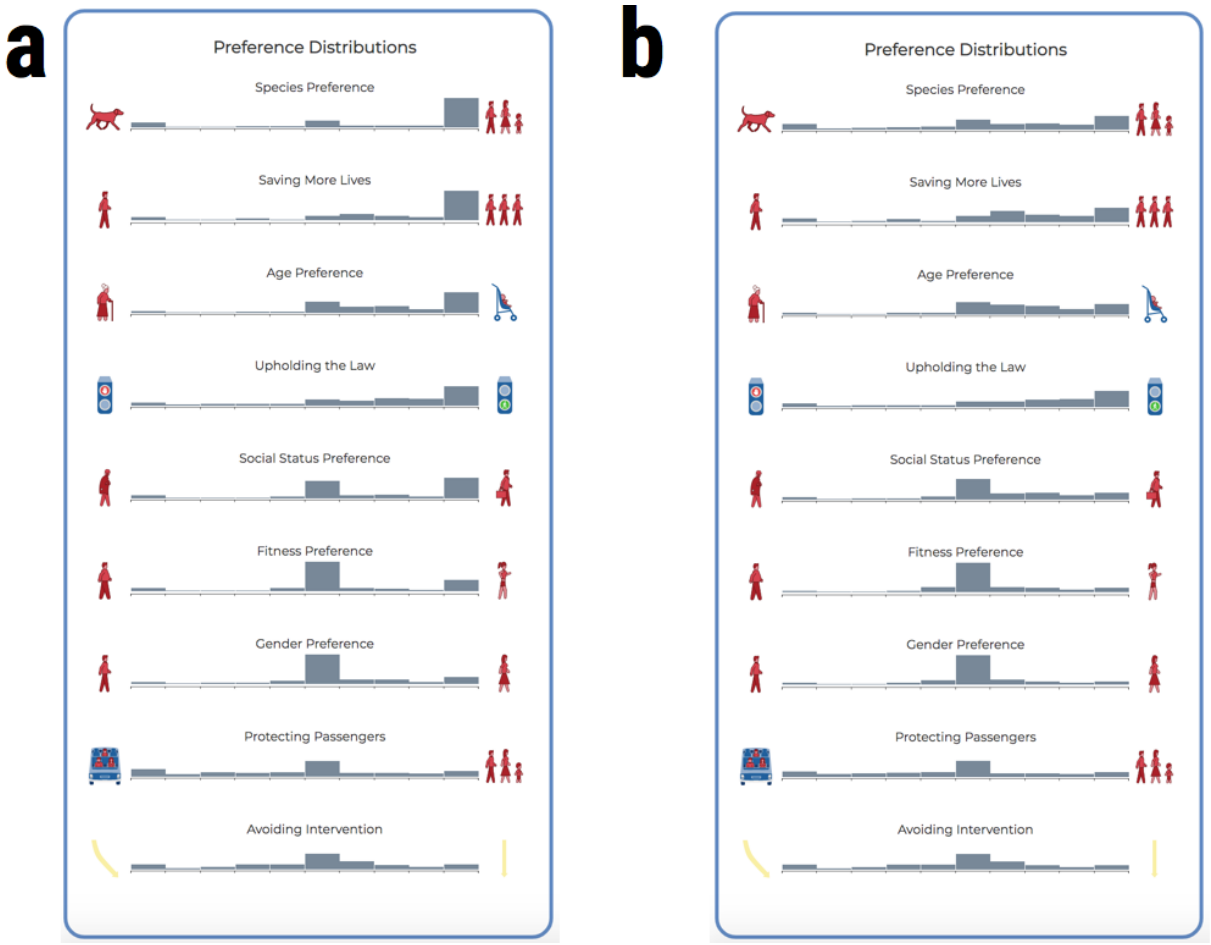

Figure 1. Distribution of explicit preferences stated by Moral Machine users. Sliders were presented with a default position determined by the responses users gave to the Moral Machine 'Judge' mode. **(a)** Preferences of users who moved at least one slider from its original position (585,531 users; > 99% of the users). **(b)** Preferences of users who changed sliders from their original position (range: 190,862 - 581,496 users). In both cases, only row 5 (Social Status Preference), shows a clear gap between the preferences extracted from the Moral Machine[1] and the preferences explicitly expressed by users.

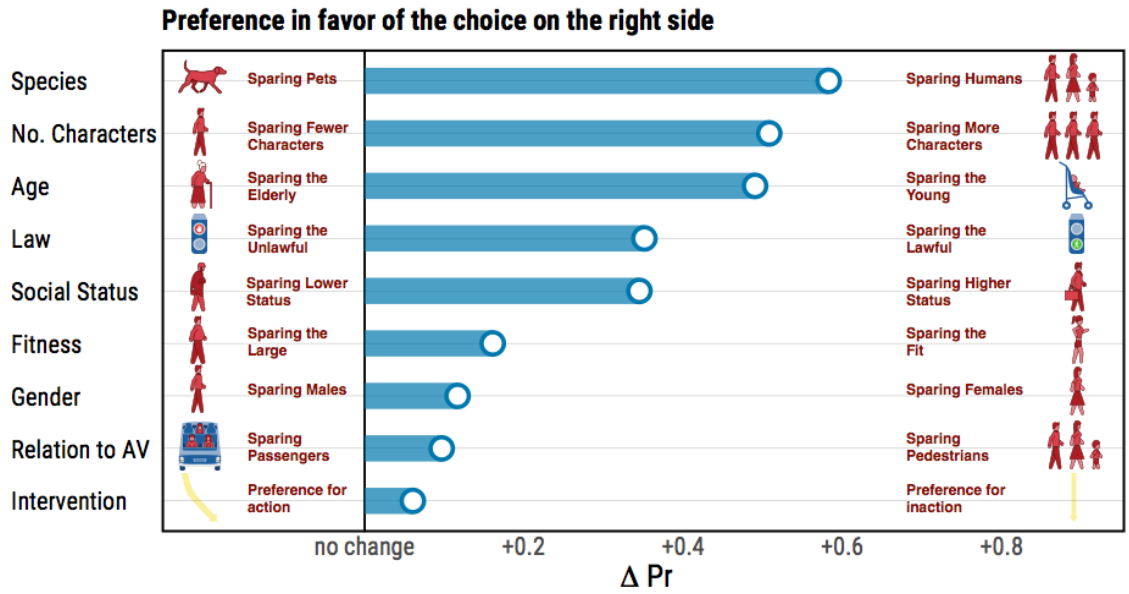

Figure 2. Preferences extracted from the conjoint analysis of the Moral Machine dataset. This figure is a simplified version of Figure 2A from *The Moral Machine Experiment.*[1] The x-axis shows the average marginal causal effect (AMCE) for each preference. In each row, Δ Pr is the difference between the probability of sparing characters possessing the attribute on the right, and the probability of sparing characters possessing the attribute on the left, aggregated over all other attributes (n = 35.2M).

ACKNOWLEDGMENTS

JFB acknowledges support from the ANR-Labex Institute for Advanced Study in Toulouse and from the ANR-3IA Artificial and Natural Intelligence Toulouse Institute.

DATA AND CODE AVAILABILITY STATEMENT

Data and code that can be used to reproduce Figs 1 and 2 is available at the following link: https://bit.ly/2VKyMhJ

ETHICAL COMPLIANCE

This study was approved by the Institute Review Board (IRB) at Massachusetts Institute of Technology (MIT). The authors complied with all relevant ethical considerations. Participants were debriefed on the purpose of the study and were given the chance to opt-out from having their data used.

REFERENCES

1. Awad, E. *et al.* The Moral Machine experiment. *Nature* **563**, 59–64 (2018).

2. Hainmueller, J., Hopkins, D. J. & Yamamoto, T. Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices Via Stated Preference Experiments. *SSRN Electronic Journal* doi:10.2139/ssrn.2231687