Life and death decisions of autonomous vehicles

https://doi.org/10.1038/s41586-020-1987-4

Received: 2 May 2019

Accepted: 26 September 2019

Published online: 4 March 2020

Check for updates

Yochanan E. Bigman[1]✉ & Kurt Gray[1]

Arising from: E. Awad et al. *Nature* https://doi.org/10.1038/s41586-018-0637-6 (2018)

How should self-driving cars make decisions when human lives hang in the balance? The Moral Machine experiment[1] (MME) suggests that people want autonomous vehicles (AVs) to treat different human lives unequally, preferentially killing some people (for example, men, the old and the poor) over others (for example, women, the young and the rich). Our results challenge this idea, revealing that this apparent preference for inequality is driven by the specific 'trolley-type' paradigm used by the MME. Multiple studies with a revised paradigm reveal that people overwhelmingly want autonomous vehicles to treat different human lives equally in life and death situations, ignoring gender, age and status—a preference consistent with a general desire for equality[2–4].

The large-scale adoption of autonomous vehicles raises ethical challenges because autonomous vehicles may sometimes have to decide between killing one person or another[5,6]. The MME seeks to reveal people's preferences in these situations and many of these revealed preferences, such as 'save more people over fewer' and 'kill by inaction over action' are consistent with preferences documented in previous research[7,8].

However, the MME also concludes that people want autonomous vehicles to make decisions about who to kill on the basis of personal features, including physical fitness, age, status and gender (for example, saving women and killing men). This conclusion contradicts well-documented ethical preferences for equal treatment across demographic features and identities, a preference enshrined in the US Constitution, the United Nations Universal Declaration of Human Rights and in the Ethical Guideline 9 of the German Ethics Code for Automated and Connected Driving[9].

We suggest that the MME finds preferences for inequality across lives because its methodology is relatively insensitive to preferences for equality. The MME uses trolley-type dilemmas that force people to choose between killing one person (or set of people) versus killing another person (or set of people). Because this paradigm assumes inequality (for example, should we program AVs to kill men or women?), it has difficulties revealing whether people prefer equality (for example, should we program AVs to ignore gender?).

What would happen if people indicated their ethical preferences in a revised paradigm, one that allowed AVs to treat different humans equally? We explored this possibility in study 1, in which people were randomly assigned to either a 'forced inequality' or an 'equality allowed' condition. Participants were drawn from two quasi-representative samples across two Western countries (US, $N = 1{,}174$; UK, $N = 1178$).

The forced inequality condition was a simplified replication of the MME, testing whether participants thought autonomous vehicles should (1) kill group A (for example, elderly people) to save group B (for example, children) or (2) kill group B to save group A. As in the MME, we examined both personal features (for example, kill men versus women) and structural features (for example, kill many people versus few people) in driving situations. However, unlike the MME—which used composite groups that simultaneously varied both personal and structural features—we examined each of these features individually (see Supplementary Information and https://osf.io/wy8tq/?view_only=e5907f552f5e4a8a901cbdd2d4c035f6 for details and data).

As Fig. 1 shows, results from the forced inequality condition closely match the global effects of the MME. Beyond the general value of replication[10], this validates our paradigm: although we used a different sample and a simpler method, we obtained the same results as the MME.

The equality allowed condition was similar to the forced inequality condition, but with the addition of a third option, (3) treat the lives of groups A and B equally (for example, treat the lives of children and elderly people equally). As Fig.1 shows, people overwhelmingly selected this option when it was available, revealing that they want autonomous vehicles to treat people equally. For example, when forced to choose between men and women, 87.7% chose to save women, but 97.9% of people actually preferred to treat both groups equally. See Supplementary Table 1 for full results.

Admittedly, it may be difficult to program a deep sense of egalitarianism into machines, but autonomous vehicles can functionally value human lives equally by simply ignoring (or failing to detect) features such as gender, age and social class. Restricting the ethical choice set of autonomous vehicles is consistent with emerging research revealing that people prefer autonomous machines not to make important ethical decisions[11,12]. Ignoring personal features is also more consistent with the current technical capacities of AVs.

One question about our data is whether participants prefer the 'treat equally' option simply because it fails to mention killing. Study 2 ruled out this concern by replicating the equality allowed condition ($N = 843$ US participants from an online panel) with a modified third option: that autonomous vehicles should decide who to save and who to kill without considering their personal features. Consistent with study 1, people expressed a robust preference for AVs to treat people equally by ignoring personal features. For example, people preferred self-driving cars to not consider gender (92.6%), fitness (88.8%) or status (84.7%). The only substantial departure from study 1 was lawfulness: 53.1% of people preferred to spare law abiders over law breakers. See Supplementary Table 2 for full results.

Of course, AVs might sometimes have to choose between killing different sets of people, but these decisions can rely solely on structural rather than personal features. In study 3, participants ($N = 993$ US participants from an online panel) chose which of two autonomous vehicles should be allowed on the road: one that makes ethical decisions on the basis of the structural features revealed by the MME (for example,

[1]Department of Psychology and Neuroscience, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. ✉e-mail: ybigman@gmail.com

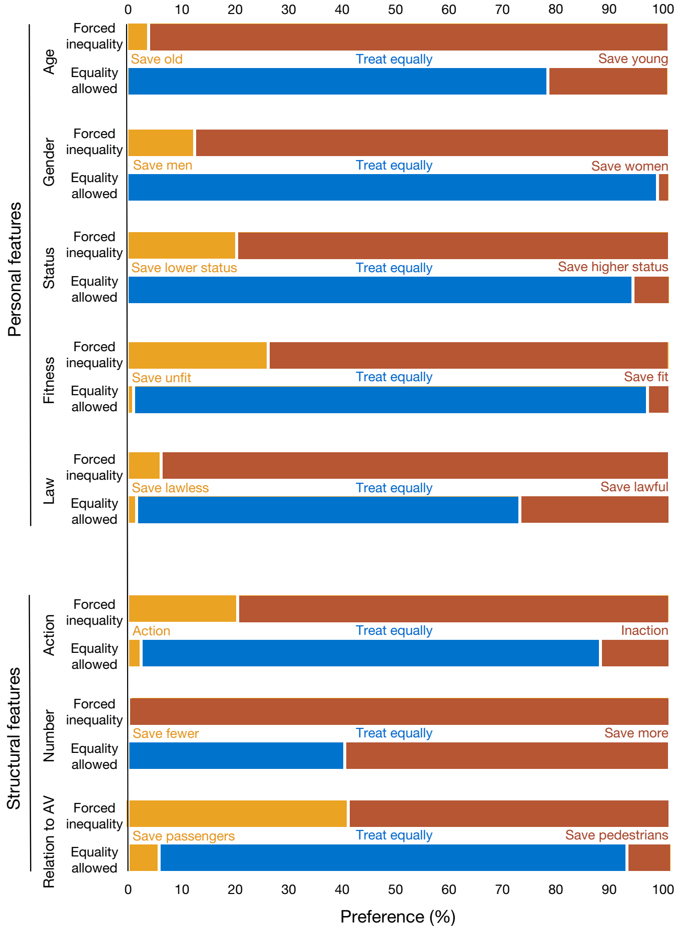

Fig. 1 | People's choices for how autonomous vehicles should be programmed to act in situations where human lives are at stake (study 1). Personal features reflect individual identity characteristics (for example, age and status) and structural features reflect characteristics of the situation. The forced inequality condition ($n = 1{,}129$) replicates the MME, which makes people choose between two options, whereas the equality allowed condition ($n = 1{,}223$) provides a third option of equal treatment. See Supplementary Fig. 1 for confidence intervals.

saving more people versus fewer, killing by inaction versus action), and another on the basis of both structural and personal features (for example, saving people based on age, gender, and status). Consistent with our predictions, 89.9% of participants chose the structural-features-only car, once again expressing a desire for AVs that ignore personal features in ethical dilemmas.

We note a number of caveats to our studies. Our samples were smaller than the millions who completed the MME. However, using quasi-representative samples in our main study (rather than a convenience sample) helps generalize the results to the populations of two large Western countries. We acknowledge that ethical preferences may vary across cultures, but our key point is that the current MME paradigm is relatively insensitive to preferences for equality, regardless of participant culture. Finally, we recognize that people often do discriminate on the basis of personal features, as sexism, classism, racism and ageism all illustrate. However, even people who implicitly act to perpetuate inequality often explicitly espouse ideas of equality[13].

To frame the MME in a broader context, consider a thought experiment about some personal features not assessed by the MME—religion, race, and disability. What might happen if the MME forced people to choose between black and white people? Aggregating people's decisions could reveal a racial bias[13], but this would not mean that people want to share the road with racist autonomous vehicles. The same logic applies to the features that were included in the MME. Do people truly want to live in a world with sexist, ageist and classist self-driving cars? This thought experiment further suggests that aggregating across forced-choice preferences may not accurately reveal how people want autonomous vehicles to be programmed to act when human lives are at stake.

Although we must be careful about interpreting the results of the MME, we emphasize its value. Every methodology has limitations, and the MME reveals both basic moral cognitive processes and global preferences for saving lives in a forced-choice paradigm. More broadly, the MME highlights the important ethical questions posed by AVs—questions that society will soon need to address.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All materials, data and code used in the studies are available at https://osf.io/wy8tq/?view_only=e5907f552f5e4a8a901cbdd2d4c035f6.

1. Awad, E. et al. The Moral Machine experiment. *Nature* **563**, 59–64 (2018).
2. Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R. & Smirnov, O. Egalitarian motives in humans. *Nature* **446**, 794–796 (2007).
3. Fehr, E., Bernhard, H. & Rockenbach, B. Egalitarianism in young children. *Nature* **454**, 1079–1083 (2008).
4. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
5. Bonnefon, J.-F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).
6. Li, J., Zhao, X., Cho, M.-J., Ju, W. & Malle, B. F. From trolley to autonomous vehicle: perceptions of responsibility and moral norms in traffic accidents with self-driving cars. *SAE Technical Paper* https://doi.org/10.4271/2016-01-0164 (2016).
7. Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R. & Hütter, M. Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *J. Pers. Soc. Psychol*. **113**, 343–376 (2017).
8. Spranca, M., Minsk, E. & Baron, J. Omission and commission in judgment and choice. *J. Exp. Soc. Psychol*. **27**, 76–105 (1991).
9. Luetge, C. The German Ethics Code for Automated and Connected Driving. *Philos. Technol*. **30**, 547–558 (2017).
10. Gertler, P., Galiani, S. & Romero, M. How to make replication the norm. *Nature* **554**, 417–419 (2018).
11. Bigman, Y. E. & Gray, K. People are averse to machines making moral decisions. *Cognition* **181**, 21–34 (2018).
12. Bigman, Y. E., Waytz, A., Alterovitz, R. & Gray, K. Holding Robots Responsible: The Elements of Machine Morality. *Trends Cogn. Sci*. **23**, 365–368 (2019).
13. Banaji, M. R. & Greenwald, A. G. *Blindspot: Hidden Biases of Good People* (Bantam Books, 2016).

Acknowledgements Y.E.B. acknowledges support from the National Science Foundation (SMA-1714298). K.G. acknowledges support from the Charles Koch Foundation and from the National Science Foundation (BCS-1823944).

Author contributions Y.E.B. and K.G. planned the research, designed the experiment, collected the data, analysed the data and wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-1987-4.
Correspondence and requests for materials should be addressed to Y.E.B.
Reprints and permissions information is available at http://www.nature.com/reprints.

nature research

Corresponding author(s): Yochanan Bigman

Last updated by author(s): YYYY-MM-DD

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Qualtrics XM
Data analysis	IBM SPSS Statistics Version 20

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data availability statement Full materials, data and code are available at https://osf.io/wy8tq/?view_only=e5907f552f5e4a8a901cbdd2d4c035f6

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

All three reported studies were quantitative.

Research sample

Study 1:
Three thousand and three people were recruited by Prolific in two nationally representative samples (on age, gender and ethnicity) one from the UK and one from the USA. After a few days of data collection the age criteria was loosened, such that older ages are still a little under-represented (e.g., for people older than 58 years old, 328 instead of 467 in the US sample and 446 instead of 463 in the UK sample).
In the UK representative sample (N = 1503), 772 were female and 731 male, 271 participants were between the ages of 18 and 27, 263 between 28 and 37, 282 between 38 and 47, 240 between 48 and 57, and 446 participants older than 58. One hundred and fifteen participants were Asian, 55 black, 31 mixed, 24 other and 1278 were white. Four of the responses were empty, such that the final sample size was 1499.
In the US representative sample (N = 1500) 769 were female and 731 male, 339 participants were between the ages of 18 and 27, 327 between 28 and 37, 258between 38 and 47, 248 between 48 and 57, and 328 participants older than 58. 96 were Asian, 197 black, 37 mixed, 30 other and 1140 were white.

Study 2:
One thousand and four people were recruited vie Amazon's Mechanical Turk (429 male, 566 female, 9 other/preferred not to disclose; Age: M = 35.00, SD = 12.22).

Study 3:
One thousand and nine people were recruited vie Amazon's Mechanical Turk (433 male, 570 female, 6 other/preferred not to disclose; Age: M = 37.26, SD = 12.80).

Sampling strategy

Study 1 used stratified sampling. Studies 2-3 used convenience samples.
Sample size: We wanted to far exceed typical power recommendations, and given that isolating the true proportion of the population is important, believed 3000 participants would keep the standard error of the mean sufficiently low for our main study (Study 2), and 1000 for the additional Studies (Studies 2 and 3).

Data collection

Data was collected on Qualtrics XM through online panels such as Prolific (Study 1) and Amazon's Mechanical Turk (Studies 2-3).

Timing

Study 1: The UK sample was collected between April 18th and April 23rd 2019. The US sample was collected between April 24th and April 30th. Data for Study 2 was collected on July 18th 2019. Data for Study 3 was collected on April 15th 2019.

Data exclusions

All exclusions were per-registered (links to the per-registration appear in the Supplemental information).

Study 1:
Participants completed three attention checks. In the first attention check they were asked what day was yesterday and what they asked for breakfast. In the second attention check participants were shown three sliders, marked X, Y and Z. They were asked to set X on 15, Y to be greater than X and evenly divisible by 10, and Z to be larger than Y. In the third attention check participants were asked if they answered questions about how a self-driving car or a human driver, and if they had an option of having people treated equally. Six hundred and forty seven participants failed at least one of the attention checks and were excluded from the analysis as planned in the pre-registration.

Study 2:
Participants completed two attention checks. In the first attention check they were asked what day was yesterday and what they asked for breakfast. In the second attention check participants were asked if they answered questions about how a self-driving car or a human driver, and if they had an option of having people treated equally. One hundred and fifty seven participants failed at least one of the attention checks and were excluded from the analysis as planned in the pre-registration.

Study 3:
Participants were asked what day was yesterday and what they asked for breakfast. Sixteen participants failed this attention check and were excluded from the analysis.

Non-participation

NA

Randomization

Randomization was done with the "randomize" function in Qualtrics.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	Participants were recruited vie online panels. While these panels perhaps do not reflect the general population perfectly, there is ample evidence that they provide good and valid results: Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. Behavior Research Methods, 48(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. Current Directions in Psychological Science, 23(3), 184–188. https://doi.org/10.1177/0963721414531598 Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. Journal of Behavioral Decision Making, 26(3), 213–224. https://doi.org/10.1002/bdm.1753
Ethics oversight	The University of North Carolina at Chapel Hill IRB

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Reply to: Life and death decisions of autonomous vehicles

https://doi.org/10.1038/s41586-020-1988-3

Published online: 4 March 2020

Edmond Awad[1,2], Sohan Dsouza[1], Richard Kim[1], Jonathan Schulz[3], Joseph Henrich[4], Azim Shariff[5]*, Jean-François Bonnefon[6]* & Iyad Rahwan[1,7,8]*

Replying to: Y. E. Bigman & K. Gray. https://doi.org/10.1038/s41586-020-1987-4 (2020)

In 'The Moral Machine experiment' (MME)[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. In the accompanying Comment[2], 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 MME used a standard method for collecting stated preferences between multidimensional outcomes: Users chose between pairs of unavoidable accidents—which varied along multiple dimensions—and the importance of each dimension was statistically extracted from their choices using conjoint analysis[3]. 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 MME, 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, they asked: should self-driving cars be programmed to (1) kill children and save elderly people, (2) kill elderly people and save children, or (3) treat the lives of children and elderly people equally?

Bigman and Gray report that for all but one question—saving many versus few—the most frequent response was (3). 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 (for example, the weak preference for inaction), and disagree on others (for example, 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 (for example, 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 should not be programmed into the cars. Participants who believed that the vehicle should be blind to age, for instance, could endeavour 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, however, 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 phrasing 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 trade-offs, whereas the third option only has positive connotations. We could suggest 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 the one 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 suggest above, and this intermediate framing appeared to have an effect on the results: for half of the questions, the frequency of the 'equality' response decreased by 16 percentage points to 27% (as can be seen by comparing their Supplementary Table 1 and Supplementary Table 2).

We should note that an unpublished portion of the MME used a third method—one similar to that of Bigman and Gray, but one that avoided this loaded language confounder. 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 labelled '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,

[1]The Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Economics, University of Exeter Business School, Exeter, UK. [3]Department of Economics, George Mason University, Fairfax, VA, USA. [4]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA. [5]Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada. [6]Toulouse School of Economics (TSM-R), , CNRS, Université Toulouse Capitole, Toulouse, France. [7]Institute for Data, Systems and Society, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany. *e-mail: shariff@psych.ubc.ca; jean-francois.bonnefon@tse-fr.eu; rahwan@mpib-berlin.mpg.de

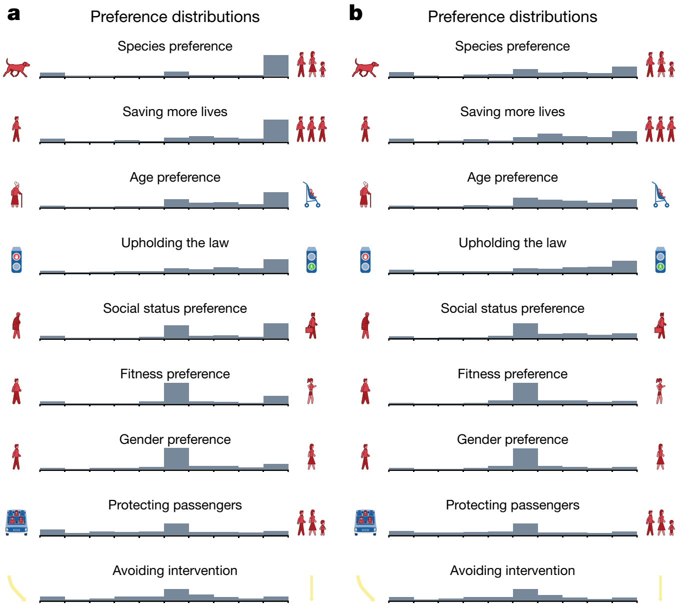

Fig. 1 | Distribution of explicit preferences stated by Moral Machine users. Sliders were presented with a default position determined by the responses users gave to the Moral Machine 'judge' mode. **a**, Preferences of users who moved at least one slider from its original position (585,531 users; >99% 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.

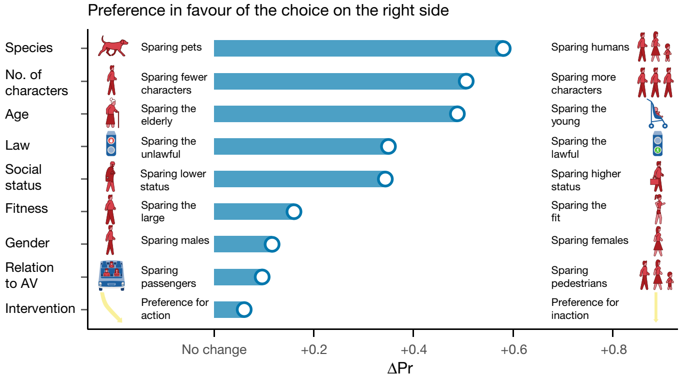

Fig. 2 | Preferences extracted from the conjoint analysis of the Moral Machine dataset. This figure is a simplified version of Fig. 2a from the MME[1]. The x axis shows the average marginal causal effect for each preference. In each row, ΔPr is the difference between the probability of sparing characters possessing the attribute on the right, and the probability of sparing characters possessing the attribute on the left, aggregated over all other attributes ($n = 35.2 \times 10^6$).

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 three-point scale and 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 the estimation. More than 99% of users who saw the slider page moved at least one slider from its original position. Figure 1a shows 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 shows 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 MME (saving humans, saving more lives, saving younger lives and saving pedestrians who cross legally; Fig. 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, four preferences that were identified as weak in the MME (inaction, saving pedestrians, saving fit characters and 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 was to treat them equally. Here we 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, the users would probably be unhappy if this bias was actually acted on. Of course, it is extremely unlikely that policymakers would propose that autonomous vehicles should discriminate on the basis of 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 have 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.

Methods

Ethical compliance

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

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

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

1. Awad, E. et al. The Moral Machine experiment. *Nature* **563**, 59–64 (2018).
2. Bigman, Y. E. & Gray, K. Life and death decisions of autonomous vehicles. *Nature* https://doi.org/10.1038/s41586-020-1987-4 (2020).
3. Hainmueller, J., Hopkins, D. J. & Yamamoto, T. Causal inference in conjoint analysis: understanding multidimensional choices via stated preference experiments. *Political Anal.* **22**, 1–30 (2014).

Acknowledgements J.-F.B. acknowledges support from the ANR-Labex Institute for Advanced Study in Toulouse, the ANR-3IA Artificial and Natural Intelligence Toulouse Institute, and the grant ANR-17-EURE-0010 Investissements d'Avenir. I.R. acknowledges funding from the Ethics & Governance of Artificial Intelligence Fund.

Author contributions I.R., A.S. and J.-F.B. planned the research. I.R., A.S., J.-F.B., E.A. and S.D. designed the experiment. E.A. and S.D. built the platform and collected the data. E.A., S.D., R.K., J.S. and A.S. analysed the data. E.A., S.D., R.K., J.S., J.H., A.S., J.-F.B. and I.R interpreted the results and wrote the paper.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-1988-3.

Correspondence and requests for materials should be addressed to A.S., J.-F.B. or I.R.

Reprints and permissions information is available at http://www.nature.com/reprints.

nature research

Corresponding author(s): Iyad Rahwan

Last updated by author(s): May 23, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was collected through the Moral Machine website (www.moralmachine.mit.edu) which was built especially for the purpose of this study.
Data analysis	Data was preprocessed, analyzed, and visualized using Python (Jupyter 3.0), R (RStudio 3.4.1), and D3.js.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study uses quantitative data collected from an online website designed as a randomized controlled multi-factorial design experiment. Part of the data is collected in a survey filled at the end.
Research sample	Research sample is Internet users who chose to visit the website and contribute to the data. Research sample has 2.3M users. The demographic information is only available for 585,565 users (aged between 15-75; 27% are females and 2% are "others").
Sampling strategy	The sample is self-selected. There was no power calculation. For Figure 2: sample size is 2.3M participants. For Figure 1: sample size for survey users is 585,565.
Data collection	Data was collected via the website http://moralmachine.mit.edu. Allocations of users to conditions was done automatically by the website, and so all researchers were blind to the experimental conditions. Data was stored in a MongoDB data base on a remote server.
Timing	Website was deployed on June 23rd, 2016. The data used in the analysis was collected continuously up until Dec 20th, 2017
Data exclusions	Similar to the MME paper, Figure 2 excluded responses for which the participant took more than 30 min. This resulted in the exclusion of 33,838 responses (out of 39.6M). Figure 1 excluded responses for which the participant did not change any of the nine sliders from their default values. This resulted in the exclusion of 34 responses (out of 585,565K).
Non-participation	Out of 2.86M completed sessions (13 scenarios), 43,979 sessions were opted out for. The number of participants dropping out (out of 2.3M) is hard to exactly know, but is comparable to the number of dropped out sessions. We did not collect the reasons to drop out.
Randomization	For Figure 1, randomization was done on the order of presentation of sliders. For Figure 2, similar to MME, users who visit the website get presented with 13 scenarios that are dawn from six main different conditions (2 scenarios from each condition + 1 fully random scenario). The six conditions vary the following aspect of characters: age, gender, fitness level, social status, number, and whether they are humans or pets. In conjunction with these six conditions, three main conditions were randomized: interventionism, relation to AV, and legality. Within each main condition, characters are sampled from a set of 20 characters (e.g. adult male, female athlete, homeless person, etc.).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	Research sample is Internet users who chose to visit the website and contribute to the data. Thus, the sample is self-selected, and it is representative of a subset of the full population.
Ethics oversight	This study was approved by the Institute Review Board (IRB) at Massachusetts Institute of Technology (MIT).

Note that full information on the approval of the study protocol must also be provided in the manuscript.