ARTICLE

https://doi.org/10.1038/s41586-018-0637-6

The Moral Machine experiment

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

With the rapid development of artificial intelligence have come concerns about how machines will make moral decisions, and the major challenge of quantifying societal expectations about the ethical principles that should guide machine behaviour. To address this challenge, we deployed the Moral Machine, an online experimental platform designed to explore the moral dilemmas faced by autonomous vehicles. This platform gathered 40 million decisions in ten languages from millions of people in 233 countries and territories. Here we describe the results of this experiment. First, we summarize global moral preferences. Second, we document individual variations in preferences, based on respondents' demographics. Third, we report cross-cultural ethical variation, and uncover three major clusters of countries. Fourth, we show that these differences correlate with modern institutions and deep cultural traits. We discuss how these preferences can contribute to developing global, socially acceptable principles for machine ethics. All data used in this article are publicly available.

We are entering an age in which machines are tasked not only to promote well-being and minimize harm, but also to distribute the well-being they create, and the harm they cannot eliminate. Distribution of well-being and harm inevitably creates tradeoffs, whose resolution falls in the moral domain[1–3]. Think of an autonomous vehicle that is about to crash, and cannot find a trajectory that would save everyone. Should it swerve onto one jaywalking teenager to spare its three elderly passengers? Even in the more common instances in which harm is not inevitable, but just possible, autonomous vehicles will need to decide how to divide up the risk of harm between the different stakeholders on the road. Car manufacturers and policymakers are currently struggling with these moral dilemmas, in large part because they cannot be solved by any simple normative ethical principles such as Asimov's laws of robotics[4].

Asimov's laws were not designed to solve the problem of universal machine ethics, and they were not even designed to let machines distribute harm between humans. They were a narrative device whose goal was to generate good stories, by showcasing how challenging it is to create moral machines with a dozen lines of code. And yet, we do not have the luxury of giving up on creating moral machines[5–8]. Autonomous vehicles will cruise our roads soon, necessitating agreement on the principles that should apply when, inevitably, life-threatening dilemmas emerge. The frequency at which these dilemmas will emerge is extremely hard to estimate, just as it is extremely hard to estimate the rate at which human drivers find themselves in comparable situations. Human drivers who die in crashes cannot report whether they were faced with a dilemma; and human drivers who survive a crash may not have realized that they were in a dilemma situation. Note, though, that ethical guidelines for autonomous vehicle choices in dilemma situations do not depend on the frequency of these situations. Regardless of how rare these cases are, we need to agree beforehand how they should be solved.

The key word here is 'we'. As emphasized by former US president Barack Obama[9], consensus in this matter is going to be important. Decisions about the ethical principles that will guide autonomous vehicles cannot be left solely to either the engineers or the ethicists. For consumers to switch from traditional human-driven cars to autonomous vehicles, and for the wider public to accept the proliferation of artificial intelligence-driven vehicles on their roads, both groups will need to understand the origins of the ethical principles that are programmed into these vehicles[10]. In other words, even if ethicists were to agree on how autonomous vehicles should solve moral dilemmas, their work would be useless if citizens were to disagree with their solution, and thus opt out of the future that autonomous vehicles promise in lieu of the status quo. Any attempt to devise artificial intelligence ethics must be at least cognizant of public morality.

Accordingly, we need to gauge social expectations about how autonomous vehicles should solve moral dilemmas. This enterprise, however, is not without challenges[11]. The first challenge comes from the high dimensionality of the problem. In a typical survey, one may test whether people prefer to spare many lives rather than few[9,12,13]; or whether people prefer to spare the young rather than the elderly[14,15]; or whether people prefer to spare pedestrians who cross legally, rather than pedestrians who jaywalk; or yet some other preference, or a simple combination of two or three of these preferences. But combining a dozen such preferences leads to millions of possible scenarios, requiring a sample size that defies any conventional method of data collection.

The second challenge makes sample size requirements even more daunting: if we are to make progress towards universal machine ethics (or at least to identify the obstacles thereto), we need a fine-grained understanding of how different individuals and countries may differ in their ethical preferences[16,17]. As a result, data must be collected worldwide, in order to assess demographic and cultural moderators of ethical preferences.

As a response to these challenges, we designed the Moral Machine, a multilingual online 'serious game' for collecting large-scale data on how citizens would want autonomous vehicles to solve moral dilemmas in the context of unavoidable accidents. The Moral Machine attracted worldwide attention, and allowed us to collect 39.61 million decisions from 233 countries, dependencies, or territories (Fig. 1a). In the main interface of the Moral Machine, users are shown unavoidable accident scenarios with two possible outcomes, depending on whether the autonomous vehicle swerves or stays on course (Fig. 1b). They then click on the outcome that they find preferable. Accident scenarios are generated by the Moral Machine following an exploration strategy that

[1]The Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA. [3]Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada. [4]Toulouse School of Economics (TSM-R), CNRS, Université Toulouse Capitole, Toulouse, France. [5]Institute for Data, Systems & Society, Massachusetts Institute of Technology, Cambridge, MA, USA. *e-mail: shariff@psych.ubc.ca; jean-francois.bonnefon@tse-fr.eu; irahwan@mit.edu

© 2018 Springer Nature Limited. All rights reserved.

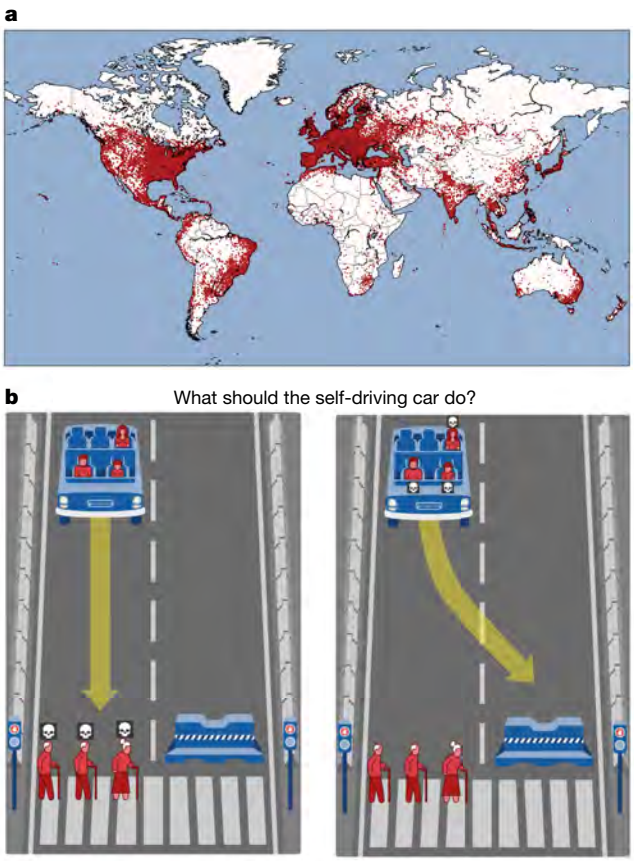


Fig. 1 | Coverage and interface. a, World map highlighting the locations of Moral Machine visitors. Each point represents a location from which at least one visitor made at least one decision ($n = 39.6$ million). The numbers of visitors or decisions from each location are not represented. **b**, Moral Machine interface. An autonomous vehicle experiences a sudden brake failure. Staying on course would result in the death of two elderly men and an elderly woman who are crossing on a 'do not cross' signal (left). Swerving would result in the death of three passengers: an adult man, an adult woman, and a boy (right).

focuses on nine factors: sparing humans (versus pets), staying on course (versus swerving), sparing passengers (versus pedestrians), sparing more lives (versus fewer lives), sparing men (versus women), sparing the young (versus the elderly), sparing pedestrians who cross legally (versus jaywalking), sparing the fit (versus the less fit), and sparing those with higher social status (versus lower social status). Additional characters were included in some scenarios (for example, criminals, pregnant women or doctors), who were not linked to any of these nine factors. These characters mostly served to make scenarios less repetitive for the users. After completing a 13-accident session, participants could complete a survey that collected, among other variables, demographic information such as gender, age, income, and education, as well as religious and political attitudes. Participants were geolocated so that their coordinates could be used in a clustering analysis that sought to identify groups of countries or territories with homogeneous vectors of moral preferences.

Here we report the findings of the Moral Machine experiment, focusing on four levels of analysis, and considering for each level of analysis how the Moral Machine results can trace our path to universal machine ethics. First, what are the relative importances of the nine preferences we explored on the platform, when data are aggregated worldwide? Second, does the intensity of each preference depend on the individual characteristics of respondents? Third, can we identify clusters of countries with homogeneous vectors of moral preferences? And fourth, do cultural and economic variations between countries predict variations in their vectors of moral preferences?

Global preferences

To test the relative importance of the nine preferences simultaneously explored by the Moral Machine, we used conjoint analysis to compute the average marginal component effect (AMCE) of each attribute (male character versus female character, passengers versus pedestrians, and so on)[18]. Figure 2a shows the unbiased estimates of nine AMCEs extracted from the Moral Machine data. In each row, the bar shows the difference between the probability of sparing characters with the attribute on the right side, and the probability of sparing the characters with the attribute on the left side, over the joint distribution of all other attributes (see Supplementary Information for computational details and assumptions, and see Extended Data Figs.1, 2 for robustness checks).

As shown in Fig. 2a, the strongest preferences are observed for sparing humans over animals, sparing more lives, and sparing young lives. Accordingly, these three preferences may be considered essential building blocks for machine ethics, or at least essential topics to be considered by policymakers. Indeed, these three preferences differ starkly in the level of controversy they are likely to raise among ethicists.

Consider, as a case in point, the ethical rules proposed in 2017 by the German Ethics Commission on Automated and Connected Driving[19]. This report represents the first and only attempt so far to provide official guidelines for the ethical choices of autonomous vehicles. As such, it provides an important context for interpreting our findings and their relevance to other countries that might attempt to follow the German example in the future. German Ethical Rule number 7 unambiguously states that in dilemma situations, the protection of human life should enjoy top priority over the protection of other animal life. This rule is in clear agreement with social expectations assessed through the Moral Machine. On the other hand, German Ethical Rule number 9 does not take a clear stance on whether and when autonomous vehicles should be programmed to sacrifice the few to spare the many, but leaves this possibility open: it is important, thus, to know that there would be strong public agreement with such programming, even if it is not mandated through regulation.

By contrast, German Ethical Rule number 9 also states that any distinction based on personal features, such as age, should be prohibited. This clearly clashes with the strong preference for sparing the young (such as children) that is assessed through the Moral Machine (see Fig. 2b for a stark illustration: the four most spared characters are the baby, the little girl, the little boy, and the pregnant woman). This does not mean that policymakers should necessarily go with public opinion and allow autonomous vehicles to preferentially spare children, or, for that matter, women over men, athletes over overweight persons, or executives over homeless persons—for all of which we see weaker but clear effects. But given the strong preference for sparing children, policymakers must be aware of a dual challenge if they decide not to give a special status to children: the challenge of explaining the rationale for such a decision, and the challenge of handling the strong backlash that will inevitably occur the day an autonomous vehicle sacrifices children in a dilemma situation.

Individual variations

We assessed individual variations by further analysing the responses of the subgroup of Moral Machine users ($n = 492{,}921$) who completed the optional demographic survey on age, education, gender, income, and political and religious views, to assess whether preferences were modulated by these six characteristics. First, when we include all six characteristic variables in regression-based estimators of each of the nine attributes, we find that individual variations have no sizable impact on any of the nine attributes (all below 0.1; see Extended Data Table 1). Of these, the most notable effects are driven by gender and religiosity of respondents. For example, male respondents are 0.06% less inclined to spare females, whereas one increase in standard deviation of religiosity of the respondent is associated with 0.09% more inclination to spare humans.

More importantly, none of the six characteristics splits its subpopulations into opposing directions of effect. On the basis of a

© 2018 Springer Nature Limited. All rights reserved.

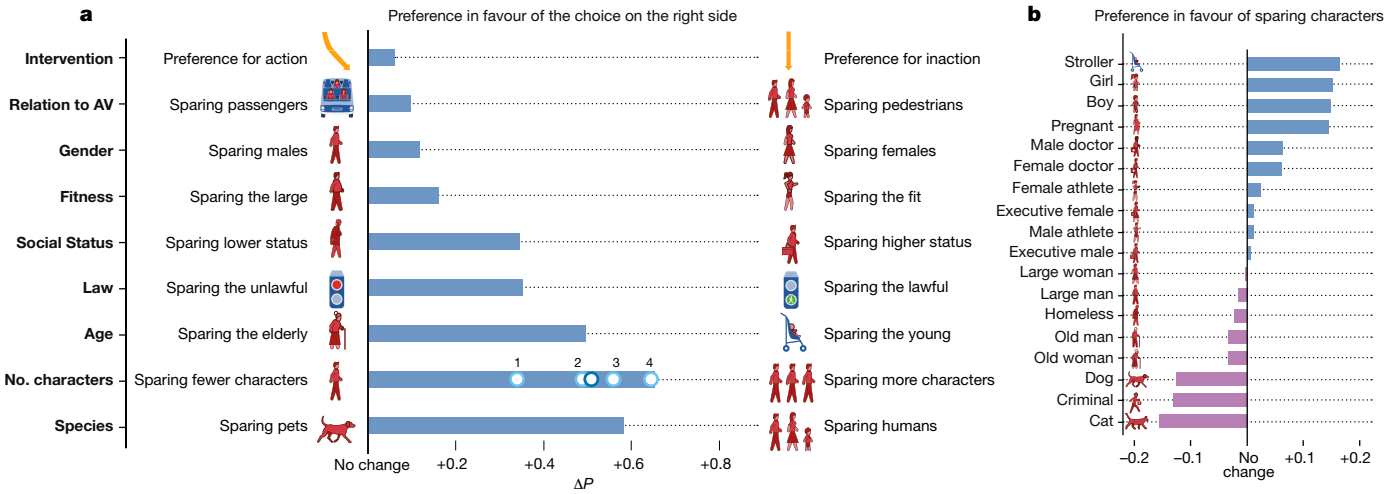


Fig. 2 | Global preferences. a, AMCE for each preference. In each row, ΔP is the difference between the probability of sparing characters possessing the attribute on the right, and the probability of sparing characters possessing the attribute on the left, aggregated over all other attributes. For example, for the attribute age, the probability of sparing young characters is 0.49 (s.e. = 0.0008) greater than the probability of sparing older characters. The 95% confidence intervals of the means are omitted owing to their insignificant width, given the sample size ($n = 35.2$ million). For the number of characters (No. characters), effect sizes are shown for each number of additional characters (1 to 4; $n_1 = 1.52$ million, $n_2 = 1.52$ million, $n_3 = 1.52$ million, $n_4 = 1.53$ million); the effect size for two additional characters overlaps with the mean effect of the attribute. AV, autonomous vehicle. **b**, Relative advantage or penalty for each character, compared to an adult man or woman. For each character, ΔP is the difference between the probability of sparing this character (when presented alone) and the probability of sparing one adult man or woman ($n = 1$ million). For example, the probability of sparing a girl is 0.15 (s.e. = 0.003) higher than the probability of sparing an adult man or woman.

unilateral dichotomization of each of the six attributes, resulting in two subpopulations for each, the difference in probability (ΔP) has a positive value for all considered subpopulations. For example, both male and female respondents indicated preference for sparing females, but the latter group showed a stronger preference (Extended Data Fig. 3). In summary, the individual variations that we observe are theoretically important, but not essential information for policymakers.

Cultural clusters

Geolocation allowed us to identify the country of residence of Moral Machine respondents, and to seek clusters of countries with homogeneous vectors of moral preferences. We selected the 130 countries with at least 100 respondents (n range 101–448,125), standardized the nine target AMCEs of each country, and conducted a hierarchical clustering on these nine scores, using Euclidean distance and Ward's minimum variance method[20]. This analysis identified three distinct 'moral clusters' of countries. These are shown in Fig. 3a, and are broadly consistent with both geographical and cultural proximity according to the Inglehart–Welzel Cultural Map 2010–2014[21].

The first cluster (which we label the Western cluster) contains North America as well as many European countries of Protestant, Catholic, and Orthodox Christian cultural groups. The internal structure within this cluster also exhibits notable face validity, with a sub-cluster containing Scandinavian countries, and a sub-cluster containing Commonwealth countries.

The second cluster (which we call the Eastern cluster) contains many far eastern countries such as Japan and Taiwan that belong to the Confucianist cultural group, and Islamic countries such as Indonesia, Pakistan and Saudi Arabia.

The third cluster (a broadly Southern cluster) consists of the Latin American countries of Central and South America, in addition to some countries that are characterized in part by French influence (for example, metropolitan France, French overseas territories, and territories that were at some point under French leadership). Latin American countries are cleanly separated in their own sub-cluster within the Southern cluster.

To rule out the potential effect of language, we found that the same clusters also emerged when the clustering analysis was restricted to participants who relied only on the pictorial representations of the dilemmas, without accessing their written descriptions (Extended Data Fig. 4).

This clustering pattern (which is fairly robust; Extended Data Fig. 5) suggests that geographical and cultural proximity may allow groups of territories to converge on shared preferences for machine ethics. Between-cluster differences, though, may pose greater problems. As shown in Fig. 3b, clusters largely differ in the weight they give to some preferences. For example, the preference to spare younger characters rather than older characters is much less pronounced for countries in the Eastern cluster, and much higher for countries in the Southern cluster. The same is true for the preference for sparing higher status characters. Similarly, countries in the Southern cluster exhibit a much weaker preference for sparing humans over pets, compared to the other two clusters. Only the (weak) preference for sparing pedestrians over passengers and the (moderate) preference for sparing the lawful over the unlawful appear to be shared to the same extent in all clusters.

Finally, we observe some striking peculiarities, such as the strong preference for sparing women and the strong preference for sparing fit characters in the Southern cluster. All the patterns of similarities and differences unveiled in Fig. 3b, though, suggest that manufacturers and policymakers should be, if not responsive, at least cognizant of moral preferences in the countries in which they design artificial intelligence systems and policies. Whereas the ethical preferences of the public should not necessarily be the primary arbiter of ethical policy, the people's willingness to buy autonomous vehicles and tolerate them on the roads will depend on the palatability of the ethical rules that are adopted.

Country-level predictors

Preferences revealed by the Moral Machine are highly correlated to cultural and economic variations between countries. These correlations provide support for the external validity of the platform, despite the self-selected nature of our sample. Although we do not attempt to pin down the ultimate reason or mechanism behind these correlations, we document them here as they point to possible deeper explanations of the cross-country differences and the clusters identified above.

As an illustration, consider the distance between the United States and other countries in terms of the moral preferences extracted from the Moral Machine ('MM distance'). Figure 4c shows a substantial

© 2018 Springer Nature Limited. All rights reserved.

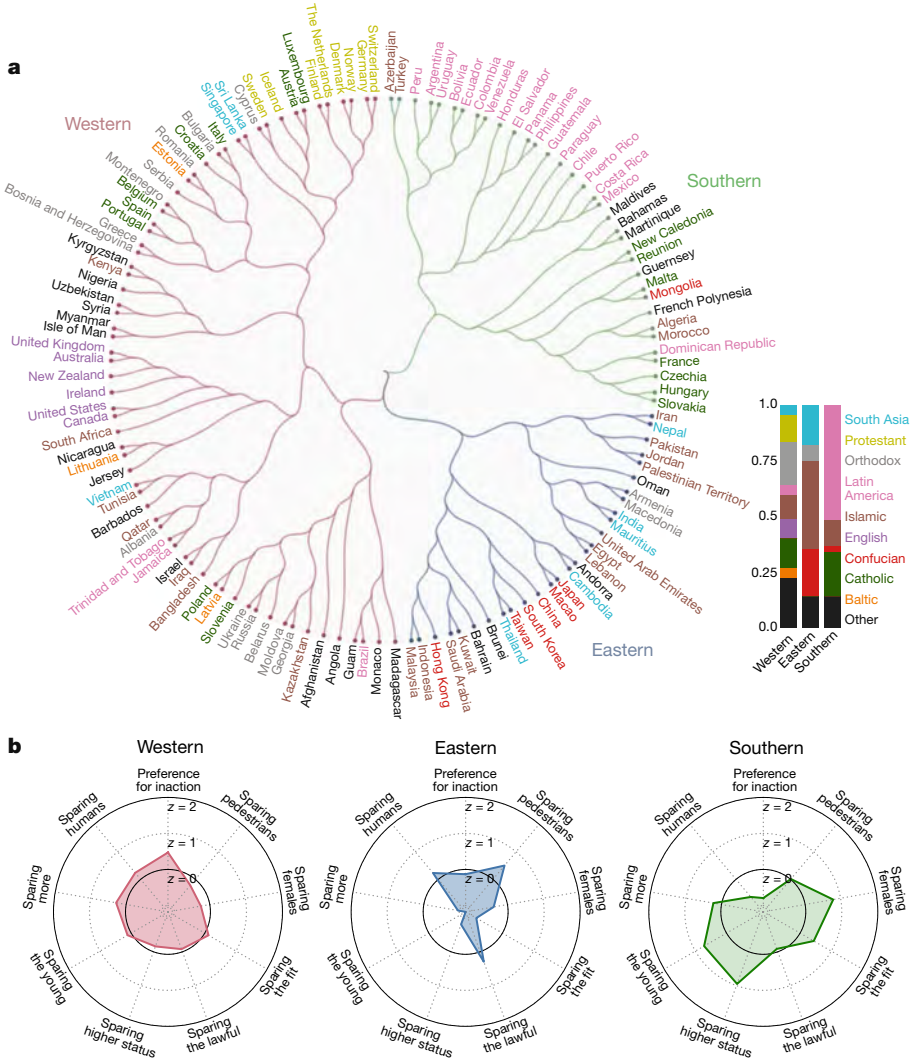


Fig. 3 | Country-level clusters. a, Hierarchical cluster of countries based on average marginal causal effect. One hundred and thirty countries with at least 100 respondents were selected (range, 101–448,125). The three colours of the dendrogram branches represent three large clusters—Western, Eastern, and Southern. Country names are coloured according to the Inglehart–Welzel Cultural Map 2010–2014[21]. Distributions across the three clusters reveal stark differences. For instance, cluster 2 (Eastern) consists mostly of countries of Islamic and Confucian cultures. By contrast, cluster 1 (Western) has large percentages of Protestant, Catholic, and Orthodox countries in Europe. **b**, Mean AMCE *z*-scores of the three major clusters. Radar plot of the mean AMCE *z*-scores of three clusters reveals a striking pattern of differences between the clusters along the nine attributes. For example, countries belonging to the Southern cluster show a strong preference for sparing females compared to countries in other clusters.

correlation ($\rho = 0.49$) between this MM distance and the cultural distance from the United States based on the World Values Survey[22]. In other words, the more culturally similar a country is to the United States, the more similarly its people play the Moral Machine.

Next, we highlight four important cultural and economic predictors of Moral Machine preferences. First, we observe systematic differences between individualistic cultures and collectivistic cultures[23]. Participants from individualistic cultures, which emphasize the distinctive value of each individual[23], show a stronger preference for sparing the greater number of characters (Fig. 4a). Furthermore, participants from collectivistic cultures, which emphasize the respect that is due to older members of the community[23], show a weaker preference for sparing younger characters (Fig. 4a, inset). Because the preference for sparing the many and the preference for sparing the young are arguably the most important for policymakers to consider, this split between individualistic and collectivistic cultures may prove an important obstacle for universal machine ethics (see Supplementary Information).

Another important (yet under-discussed) question for policymakers to consider is the importance of whether pedestrians are abiding by or violating the law. Should those who are crossing the street illegally benefit from the same protection as pedestrians who cross legally? Or should the primacy of their protection in comparison to other ethical priorities be reduced? We observe that prosperity (as indexed by GDP per capita[24]) and the quality of rules and institutions (as indexed by the Rule of Law[25]) correlate with a greater preference against pedestrians who cross illegally (Fig. 4b). In other words, participants from countries that are poorer and suffer from weaker institutions are more tolerant of pedestrians who cross illegally, presumably because of their experience of lower rule compliance and weaker punishment of rule deviation[26]. This observation limits the generalizability of the recent German ethics guideline, for example, which state that "parties involved in the generation of mobility risks must not sacrifice non-involved parties." (see Supplementary Information).

Finally, our data reveal a set of preferences in which certain characters are preferred for demographic reasons. First, we observe that higher country-level economic inequality (as indexed by the country's Gini coefficient) corresponds to how unequally characters of different social status are treated. Those from countries with less economic equality between the rich and poor also treat the rich and poor less equally in the Moral Machine. This relationship may be explained by regular encounters with inequality seeping into people's moral preferences, or perhaps because broader egalitarian norms affect both how

© 2018 Springer Nature Limited. All rights reserved.

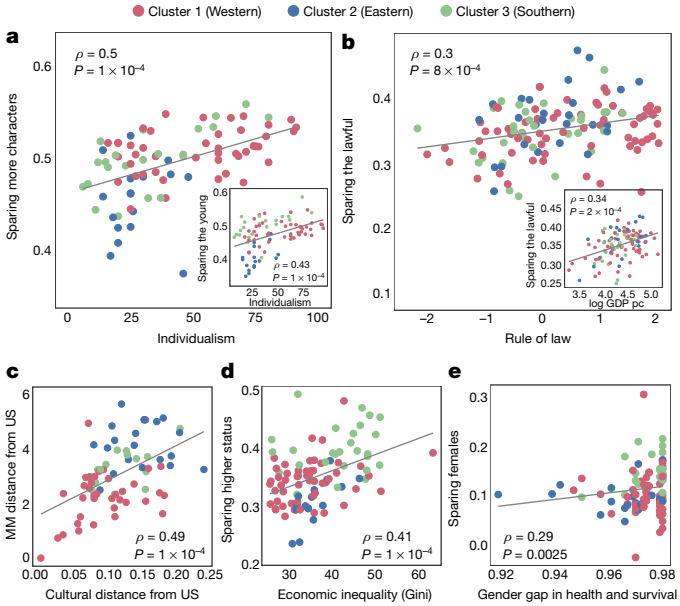


Fig. 4 | Association between Moral Machine preferences and other variables at the country level. Each panel shows Spearman's ρ and P value for the correlation test between the relevant pair of variables. **a**, Association between individualism and the preference for sparing more characters ($n = 87$), or the preference for sparing the young (inset; $n = 87$). **b**, Association between the preference for sparing the lawful and each of rule of law ($n = 122$) and log GDP per capita (pc) (inset; $n = 110$). **c**, Association between cultural distance from the United States and MM distance (distance in terms of the moral preferences extracted from the Moral Machine) from the United States ($n = 72$). **d**, Association between economic inequality (Gini coefficient) and the preference for sparing higher status ($n = 98$). **e**, Association between the gender gap in health and survival and the preference for sparing females ($n = 104$).

much inequality a country is willing to tolerate at the societal level, and how much inequality participants endorse in their Moral Machine judgments. Second, the differential treatment of male and female characters in the Moral Machine corresponded to the country-level gender gap in health and survival (a composite in which higher scores indicated higher ratios of female to male life expectancy and sex ratio at birth—a marker of female infanticide and anti-female sex-selective abortion). In nearly all countries, participants showed a preference for female characters; however, this preference was stronger in nations with better health and survival prospects for women. In other words, in places where there is less devaluation of women's lives in health and at birth, males are seen as more expendable in Moral Machine decision-making (Fig. 4e). While not aiming to pin down the causes of this variation in Extended Data Table 2, we nevertheless provide a regression analysis that demonstrates that the results hold when controlling for several potentially confounding factors.

Discussion

Never in the history of humanity have we allowed a machine to autonomously decide who should live and who should die, in a fraction of a second, without real-time supervision. We are going to cross that bridge any time now, and it will not happen in a distant theatre of military operations; it will happen in that most mundane aspect of our lives, everyday transportation. Before we allow our cars to make ethical decisions, we need to have a global conversation to express our preferences to the companies that will design moral algorithms, and to the policymakers that will regulate them.

The Moral Machine was deployed to initiate such a conversation, and millions of people weighed in from around the world. Respondents could be as parsimonious or thorough as they wished in the ethical framework they decided to follow. They could engage in a complicated weighting of all nine variables used in the Moral Machine, or adopt simple rules such as 'let the car always go onward'. Our data helped us to identify three strong preferences that can serve as building blocks for discussions of universal machine ethics, even if they are not ultimately endorsed by policymakers: the preference for sparing human lives, the preference for sparing more lives, and the preference for sparing young lives. Some preferences based on gender or social status vary considerably across countries, and appear to reflect underlying societal-level preferences for egalitarianism[27].

The Moral Machine project was atypical in many respects. It was atypical in its objectives and ambitions: no research has previously attempted to measure moral preferences using a nine-dimensional experimental design in more than 200 countries. To achieve this unusual objective, we deployed a viral online platform, hoping that we would reach out to vast numbers of participants. This allowed us to collect data from millions of people over the entire world, a feat that would be nearly impossibly hard and costly to achieve through standard academic survey methods. For example, recruiting nationally representative samples of participants in hundreds of countries would already be extremely difficult, but testing a nine-factorial design in each of these samples would verge on impossible. Our approach allowed us to bypass these difficulties, but its downside is that our sample is self-selected, and not guaranteed to exactly match the socio-demographics of each country (Extended Data Fig. 6). The fact that the cross-societal variation we observed aligns with previously established cultural clusters, as well as the fact that macro-economic variables are predictive of Moral Machine responses, are good signals about the reliability of our data, as is our post-stratification analysis (Extended Data Fig. 7 and Supplementary Information). But the fact that our samples are not guaranteed to be representative means that policymakers should not embrace our data as the final word on societal preferences—even if our sample is arguably close to the internet-connected, tech-savvy population that is interested in driverless car technology, and more likely to participate in early adoption.

Even with a sample size as large as ours, we could not do justice to all of the complexity of autonomous vehicle dilemmas. For example, we did not introduce uncertainty about the fates of the characters, and we did not introduce any uncertainty about the classification of these characters. In our scenarios, characters were recognized as adults, children, and so on with 100% certainty, and life-and-death outcomes were predicted with 100% certainty. These assumptions are technologically unrealistic, but they were necessary to keep the project tractable. Similarly, we did not manipulate the hypothetical relationship between respondents and characters (for example, relatives or spouses). Our previous work did not find a strong effect of this variable on moral preferences[12].

Indeed, we can embrace the challenges of machine ethics as a unique opportunity to decide, as a community, what we believe to be right or wrong; and to make sure that machines, unlike humans, unerringly follow these moral preferences. We might not reach universal agreement: even the strongest preferences expressed through the Moral Machine showed substantial cultural variations, and our project builds on a long tradition of investigating cultural variations in ethical judgments[28]. But the fact that broad regions of the world displayed relative agreement suggests that our journey to consensual machine ethics is not doomed from the start. Attempts at establishing broad ethical codes for intelligent machines, such as the Asilomar AI Principles[29], often recommend that machine ethics should be aligned with human values. These codes seldom recognize, though, that humans experience inner conflict, interpersonal disagreements, and cultural dissimilarities in the moral domain[30–32]. We have shown that these conflicts, disagreements, and dissimilarities, while substantial, may not be fatal.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0637-6.

© 2018 Springer Nature Limited. All rights reserved.

Received: 2 March 2018; Accepted: 25 September 2018;
Published online: 24 October 2018

1. Greene, J. *Moral Tribes: Emotion, Reason and the Gap Between Us and Them* (Atlantic Books, London, 2013).
2. Tomasello, M. *A Natural History of Human Thinking* (Harvard Univ. Press, Cambridge, 2014).
3. Cushman, F. & Young, L. The psychology of dilemmas and the philosophy of morality. *Ethical Theory Moral Pract.* **12**, 9–24 (2009).
4. Asimov, I. *I, Robot* (Doubleday, New York, 1950).
5. Bryson, J. & Winfield, A. Standardizing ethical design for artificial intelligence and autonomous systems. *Computer* **50**, 116–119 (2017).
6. Wiener, N. Some moral and technical consequences of automation. *Science* **131**, 1355–1358 (1960).
7. Wallach, W. & Allen, C. *Moral Machines: Teaching Robots Right from Wrong* (Oxford Univ. Press, Oxford, 2008).
8. Dignum, V. Responsible autonomy. In *Proc. 26th International Joint Conference on Artificial Intelligence* 4698–4704 (IJCAI, 2017).
9. Dadich, S. Barack Obama, neural nets, self-driving cars, and the future of the world. *Wired* https://www.wired.com/2016/10/president-obama-mit-joi-ito-interview/ (2016).
10. Shariff, A., Bonnefon, J.-F. & Rahwan, I. Psychological roadblocks to the adoption of self-driving vehicles. *Nat. Hum. Behav.* **1**, 694–696 (2017).
11. Conitzer, V., Brill, M. & Freeman, R. Crowdsourcing societal tradeoffs. In *Proc. 2015 International Conference on Autonomous Agents and Multiagent Systems* 1213–1217 (IFAAMAS, 2015).
12. Bonnefon, J.-F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).
13. Hauser, M., Cushman, F., Young, L., Jin, K.-X. R. & Mikhail, J. A dissociation between moral judgments and justifications. *Mind Lang.* **22**, 1–21 (2007).
14. Carlsson, F., Daruvala, D. & Jaldell, H. Preferences for lives, injuries, and age: a stated preference survey. *Accid. Anal. Prev.* **42**, 1814–1821 (2010).
15. Johansson-Stenman, O. & Martinsson, P. Are some lives more valuable? An ethical preferences approach. *J. Health Econ.* **27**, 739–752 (2008).
16. Johansson-Stenman, O., Mahmud, M. & Martinsson, P. Saving lives versus life-years in rural Bangladesh: an ethical preferences approach. *Health Econ.* **20**, 723–736 (2011).
17. Graham, J., Meindl, P., Beall, E., Johnson, K. M. & Zhang, L. Cultural differences in moral judgment and behavior, across and within societies. *Curr. Opin. Psychol.* **8**, 125–130 (2016).
18. Hainmueller, J., Hopkins, D. J. & Yamamoto, T. Causal inference in conjoint analysis: understanding multidimensional choices via stated preference experiments. *Polit. Anal.* **22**, 1–30 (2014).
19. Luetge, C. The German Ethics Code for automated and connected driving. *Philos. Technol.* **30**, 547–558 (2017).
20. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. Preprint at https://arxiv.org/abs/1109.2378 (2011).
21. Inglehart, R. & Welzel, C. *Modernization, Cultural Change, and Democracy: The Human Development Sequence* (Cambridge Univ. Press, Cambridge, 2005).
22. Muthukrishna, M. Beyond WEIRD psychology: measuring and mapping scales of cultural and psychological distance. Preprint at https://ssrn.com/abstract=3259613 (2018).
23. Hofstede, G. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations* (Sage, Thousand Oaks, 2003).
24. International Monetary Fund. *World Economic Outlook Database* https://www.imf.org/external/pubs/ft/weo/2017/01/weodata/index.aspx (2017).
25. Kaufmann, D., Kraay, A. & Mastruzzi, M. The worldwide governance indicators: methodology and analytical issues. *Hague J. Rule Law* **3**, 220–246 (2011).
26. Gächter, S. & Schulz, J. F. Intrinsic honesty and the prevalence of rule violations across societies. *Nature* **531**, 496–499 (2016).
27. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Penguin, London, 2016).
28. Henrich, J. *et al*. In search of Homo Economicus: behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.* **91**, 73–78 (2001).
29. Future of Life Institute. *Asilomar AI Principles* https://futureoflife.org/ai-principles/ (2017).
30. Haidt, J. *The Righteous Mind: Why Good People Are Divided by Politics and Religion* (Knopf Doubleday, New York, 2012).
31. Gastil, J., Braman, D., Kahan, D. & Slovic, P. The cultural orientation of mass political opinion. *PS Polit. Sci. Polit.* **44**, 711–714 (2011).
32. Nishi, A., Christakis, N. A. & Rand, D. G. Cooperation, decision time, and culture: online experiments with American and Indian participants. *PLoS ONE* **12**, e0171252 (2017).

Acknowledgements I.R., E.A., S.D., and R.K. acknowledge support from the Ethics and Governance of Artificial Intelligence Fund. J.-F.B. acknowledges support from the ANR-Labex Institute for Advanced Study in Toulouse.

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
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0637-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0637-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.S. or J.-F.B. and I.R.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

This study was approved by the Institute Review Board (IRB) at Massachusetts Institute of Technology (MIT). The authors complied with all relevant ethical considerations. No statistical methods were used to predetermine sample size. The experiments were randomized and the investigators were blinded to allocation during experiments and outcome assessment.

The Moral Machine website was designed to collect data on the moral acceptability of decisions made by autonomous vehicles in situations of unavoidable accidents, in which they must decide who is spared and who is sacrificed. The Moral Machine was deployed in June 2016. In October 2016, a feature was added that offered users the option to fill a survey about their demographics, political views, and religious beliefs. Between November 2016 and March 2017, the website was progressively translated into nine languages in addition to English (Arabic, Chinese, French, German, Japanese, Korean, Portuguese, Russian, and Spanish).

While the Moral Machine offers four different modes (see Supplementary Information), the focus of this article is on the central data-gathering feature of the website, called the Judge mode. In this mode, users are presented with a series of dilemmas in which the autonomous vehicle must decide between two different outcomes. In each dilemma, one outcome amounts to sparing a group of 1 to 5 characters (chosen from a sample of 20 characters, Fig. 2b) and killing another group of 1 to 5 characters. The other outcome reverses the fates of the two groups. The only task of the user is to choose between the two outcomes, as a response to the question "What should the self-driving car do?" Users have the option to click on a button labelled 'see description' to display a complete text description of the characters in the two groups, together with their fate in each outcome.

While users can go through as many dilemmas as they wish, dilemmas are generated in sessions of 13. Within each session, one dilemma is entirely random. The other 12 dilemmas are sampled from a space of approximately 26 million possibilities (see below). Accordingly, it is extremely improbable for a given user to see the same dilemma twice, regardless of how many dilemmas they choose to go through, or how many times they visit the Moral Machine.

Leaving aside the one entirely random dilemma, there are two dilemmas within each session that focus on each of six dimensions of moral preferences: character gender, character age, character physical fitness, character social status, character species, and character number. Furthermore, each dilemma simultaneously randomizes three additional attributes: which group of characters will be spared if the car does nothing; whether the two groups are pedestrians, or whether one group is in the car; and whether the pedestrian characters are crossing legally or illegally. This exploration strategy is supported by a dilemma generation algorithm (see Supplementary Information, which also provides extensive descriptions of statistical analyses, robustness checks, and tests of internal and external validity).

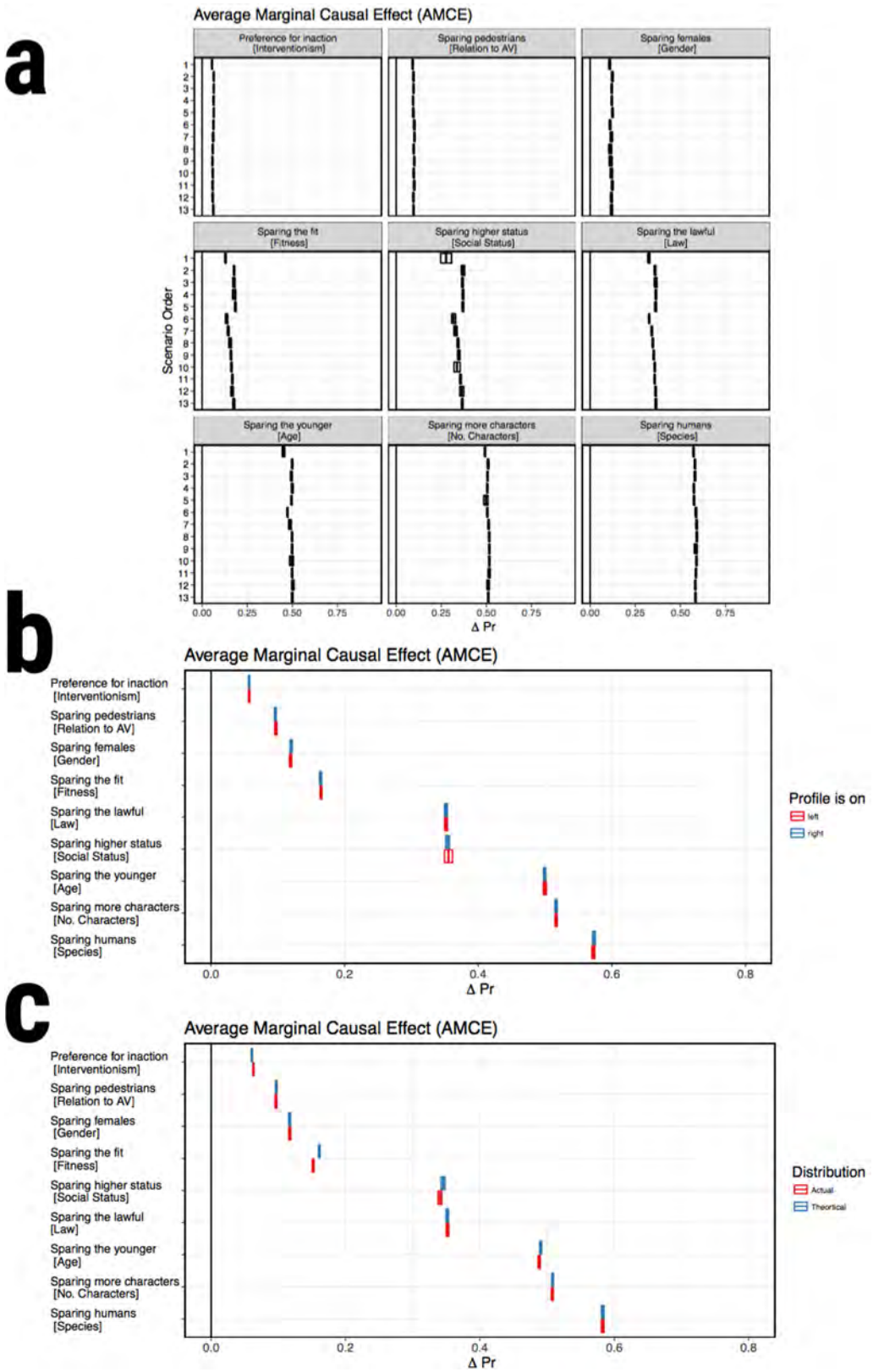
After completing a session of 13 dilemmas, users are presented with a summary of their decisions: which character they spared the most; which character they sacrificed the most; and the relative importance of the nine target moral dimensions in their decisions, compared to their importance to the average of all other users so far. Users have the option to share this summary with their social network. Either before or after they see this summary (randomized order), users are asked whether they want to "help us better understand their decisions." Users who click 'yes' are directed to a survey of their demographic, political, and religious characteristics. They also have the option to edit the summary of their decisions, to tell us about the self-perceived importance of the nine dimensions in their decisions. These self-perceptions were not analysed in this article.

The country from which users access the website is geo-localized through the IP address of their computer or mobile device. This information is used to compute a vector of moral preferences for each country. In turn, these moral vectors are used both for cultural clustering, and for country-level correlations between moral preferences and socio-economic indicators. The source and period of reference for each socio-economic indicator are detailed in the Supplementary Information.

Data availability

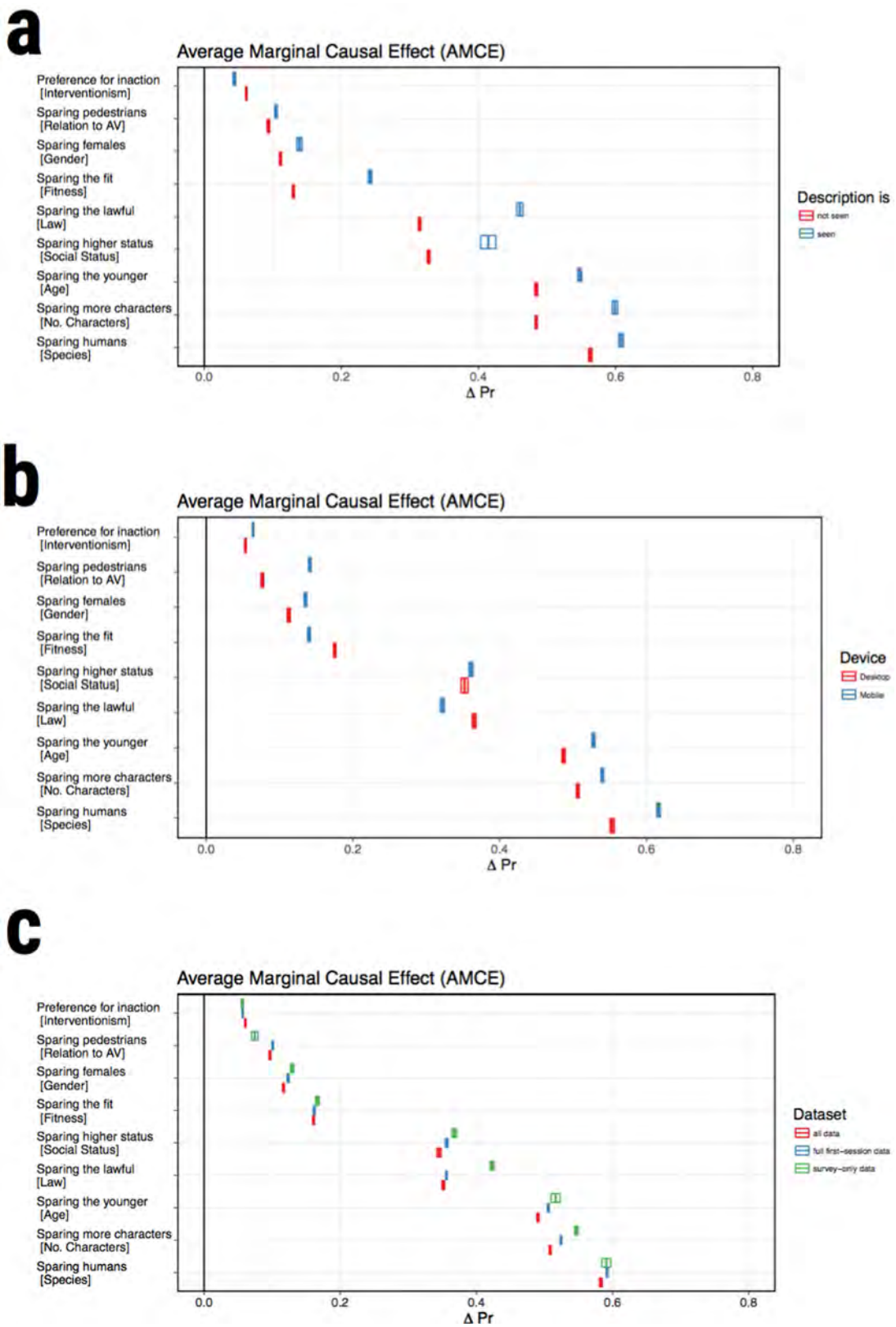
Source data and code that can be used to reproduce Figs. 2–4, Extended Data Figs. 1–7, Extended Data Tables 1, 2, Supplementary Figs. 3–21, and Supplementary Table 2 are all available at the following link: https://goo.gl/JXRrBP. The provided data, both at the individual level (anonymized IDs) and the country level, can be used beyond replication to answer follow-up research questions.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 1 | Robustness checks: internal validation of three simplifying assumptions. Calculated values correspond to values in Fig. 2a (that is, AMCE calculated using conjoint analysis). For example, 'Sparing Pedestrians [Relation to AV]' refers to the difference between the probability of sparing pedestrians, and the probability of sparing passengers (attribute name: Relation to AV), aggregated over all other attributes. Error bars represent 95% confidence intervals of the means. AV, autonomous vehicle. **a**, Validation of assumption 1 (stability and no-carryover effect): potential outcomes remain stable regardless of scenario order. **b**, Validation of assumption 2 (no profile-order effects): potential outcomes remain stable regardless of left–right positioning of choice options on the screen. **c**, Validation of assumption 3 (randomization of the profiles): potential outcomes are statistically independent of the profiles. This assumption should be satisfied by design. However, a mismatch between the design and the collected data can happen during data collection. This panel shows that using theoretical proportions (by design) and actual proportions (in collected data) of subgroups results in similar effect estimates. See Supplementary Information for more details.

© 2018 Springer Nature Limited. All rights reserved.

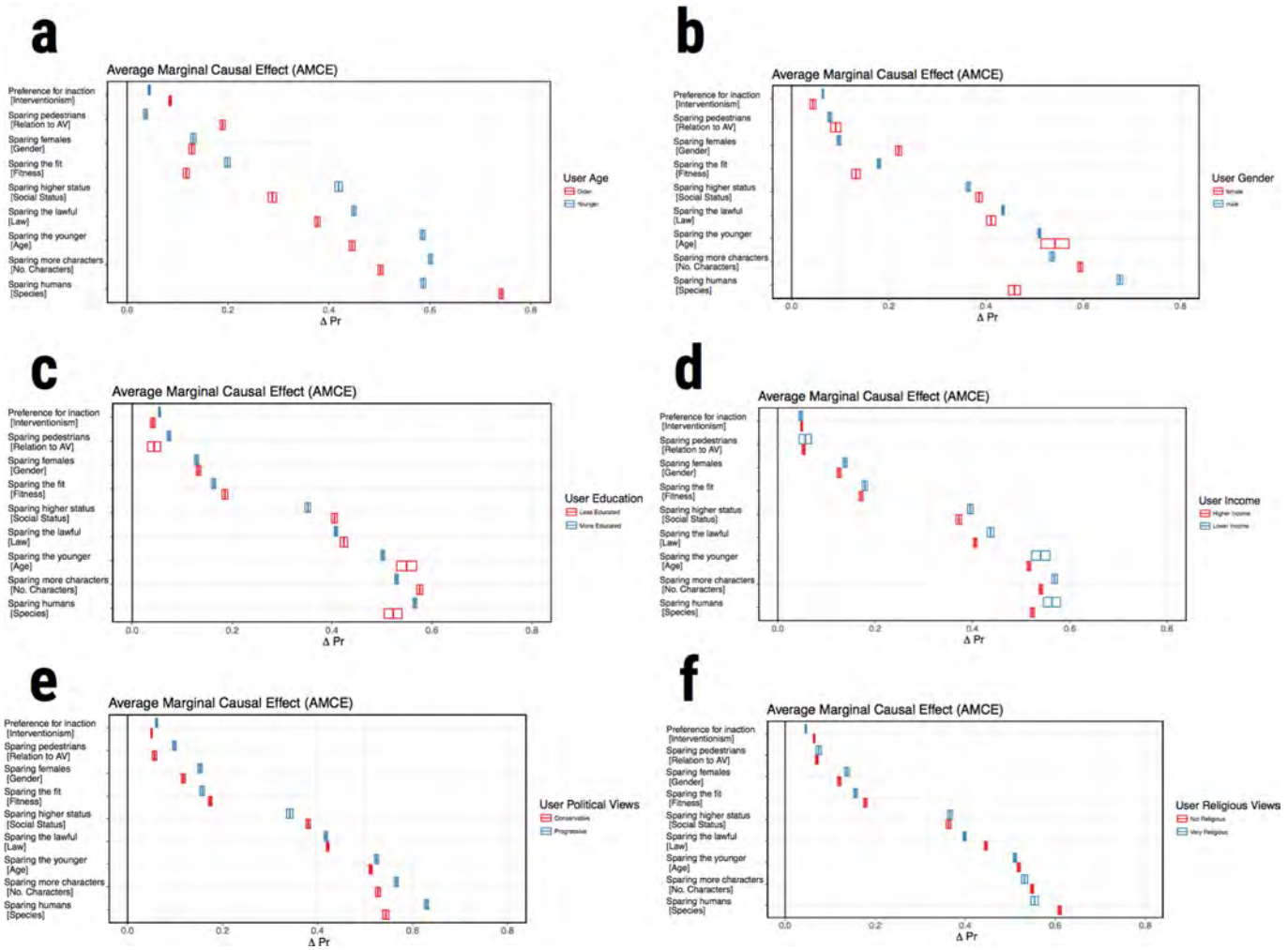


Extended Data Fig. 2 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

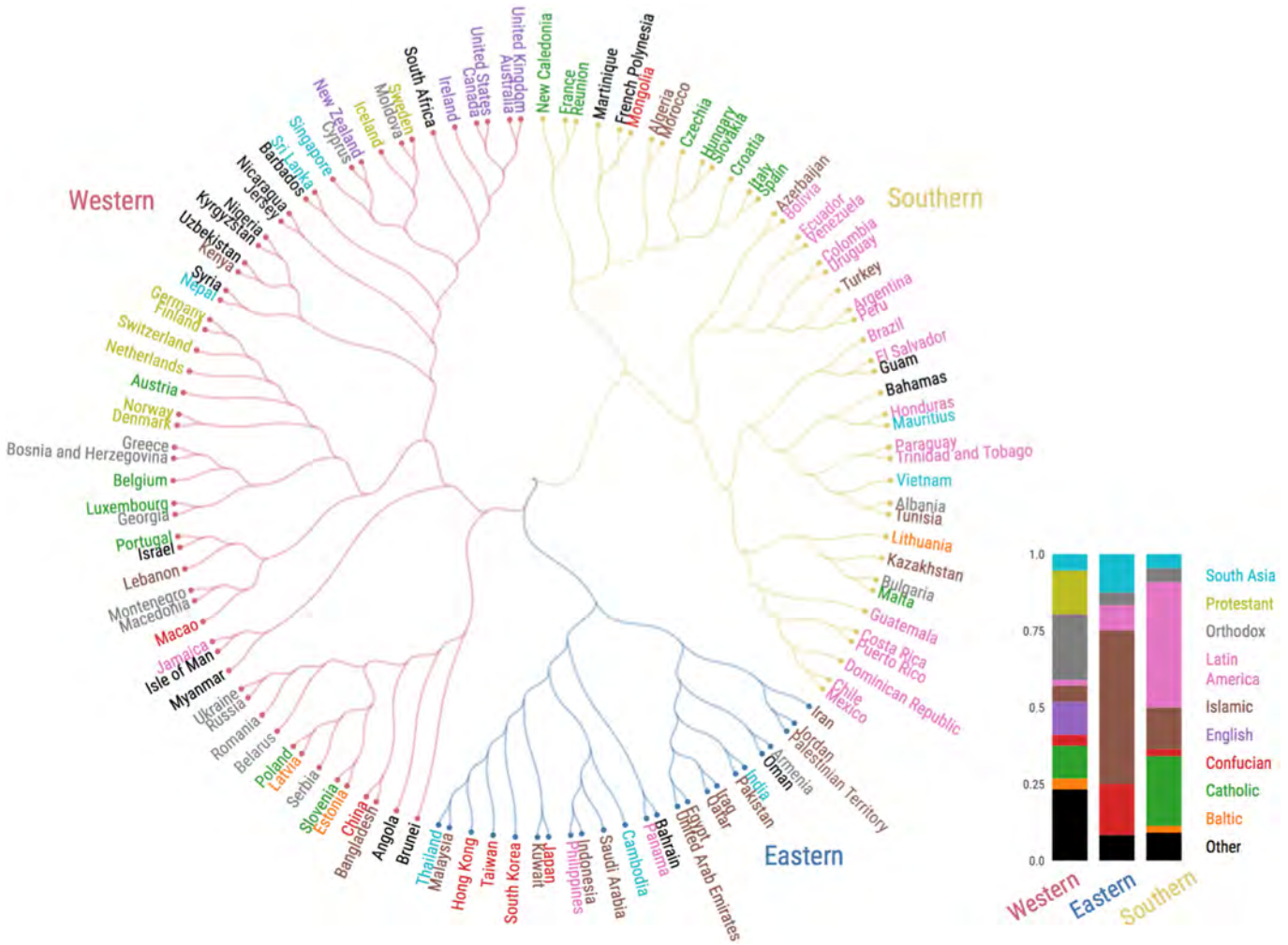
Extended Data Fig. 2 | Robustness checks: external validation of three factors. Calculated values correspond to values in Fig. 2a (AMCE calculated using conjoint analysis). For example, 'Sparing Pedestrians [Relation to AV]' refers to the difference between the probability of sparing pedestrians, and the probability of sparing passengers (attribute name: Relation to AV), aggregated over all other attributes. Error bars represent 95% confidence intervals of the means. **a**, Validation of textual description (seen versus not seen). By default, respondents see only the visual representation of a scenario. Interpretation of what type of characters they represent (for example, female doctor) may not be obvious. Optionally, respondents can read a textual description of the scenario by clicking on 'see description'. This panel shows that direction and (except in one case) order of effect estimates remain stable. The magnitude of the effects increases for respondents who read the textual descriptions, which means that the effects reported in Fig. 2a were not overestimated because of visual ambiguity. **b**, Validation of device used (desktop versus mobile). Direction and order of effect estimates remain stable regardless of whether respondents used desktop or mobile devices when completing the task. **c**, Validation of data set (all data versus full first-session data versus survey-only data). Direction and order of effect estimates remain stable regardless of whether the data used in analysis are all data, data restricted to only first completed (13-scenario) session by any user, or data restricted to completed sessions after which the demographic survey was taken. First completed session by any user is an interesting subset of the data because respondents had not seen their summary of results yet, and respondents ended up completing the session. Survey-only data are also interesting given that the conclusions about individual variations in the main paper and from Extended Data Fig. 3 and Extended Data Table 1 are drawn from this subset. See Supplementary Information for more details.

© 2018 Springer Nature Limited. All rights reserved.



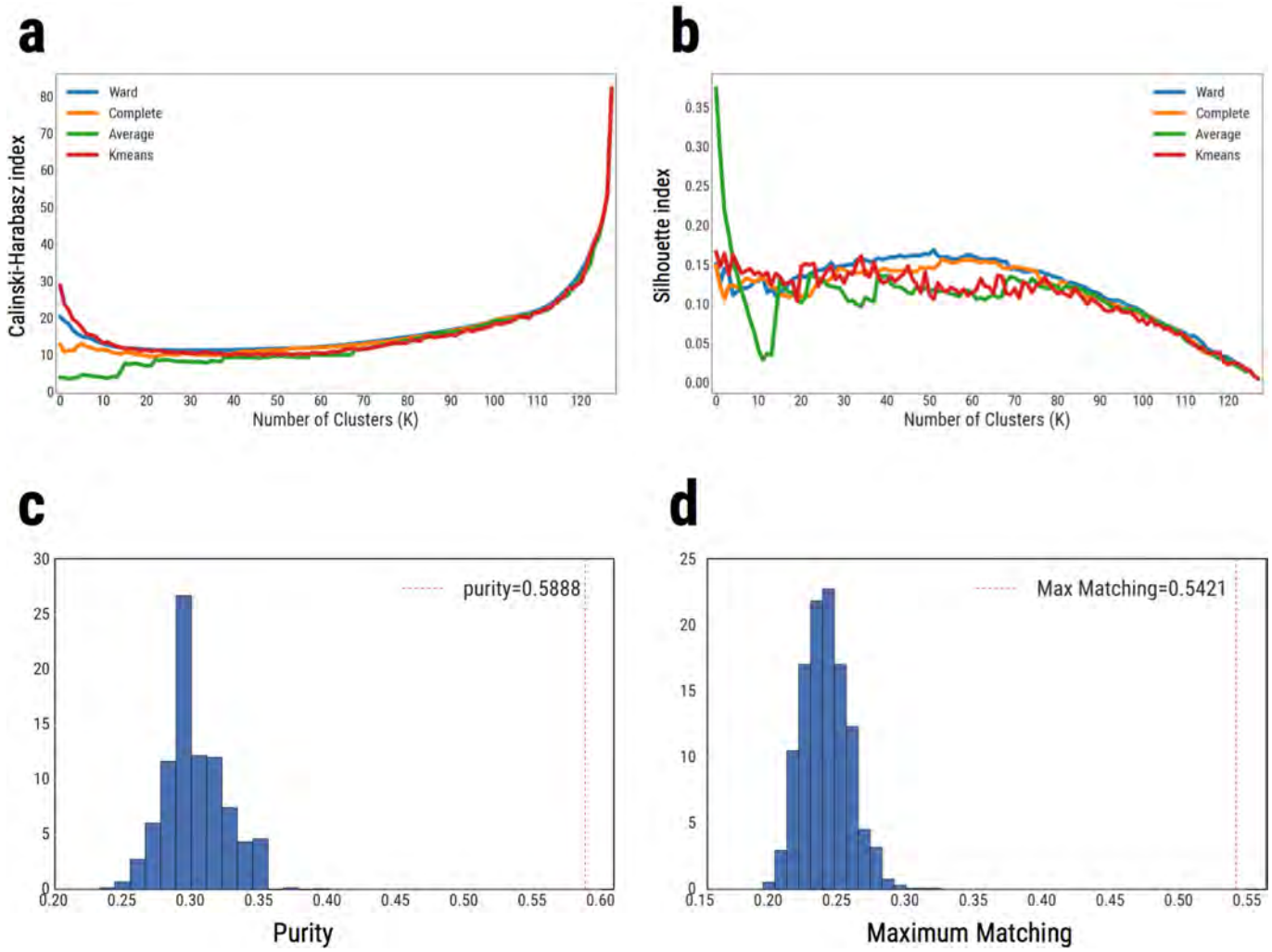
Extended Data Fig. 3 | Average marginal causal effect (AMCE) of attributes for different subpopulations. Subpopulations are characterized by respondents' age (**a**, older versus younger), gender (**b**, male versus female), education (**c**, less versus more educated), income (**d**, higher versus lower income), political views (**e**, conservative versus progressive), and religious views (**f**, not religious versus very religious). Error bars represent 95% confidence intervals of the means. Note that AMCE has a positive value for all considered subpopulations; for example, both male and female respondents indicated a preference for sparing females, but the latter group showed a stronger preference. See Supplementary Information for a detailed description of the cutoffs and the groupings of ordinal categories that were used to define each subpopulation.

© 2018 Springer Nature Limited. All rights reserved.



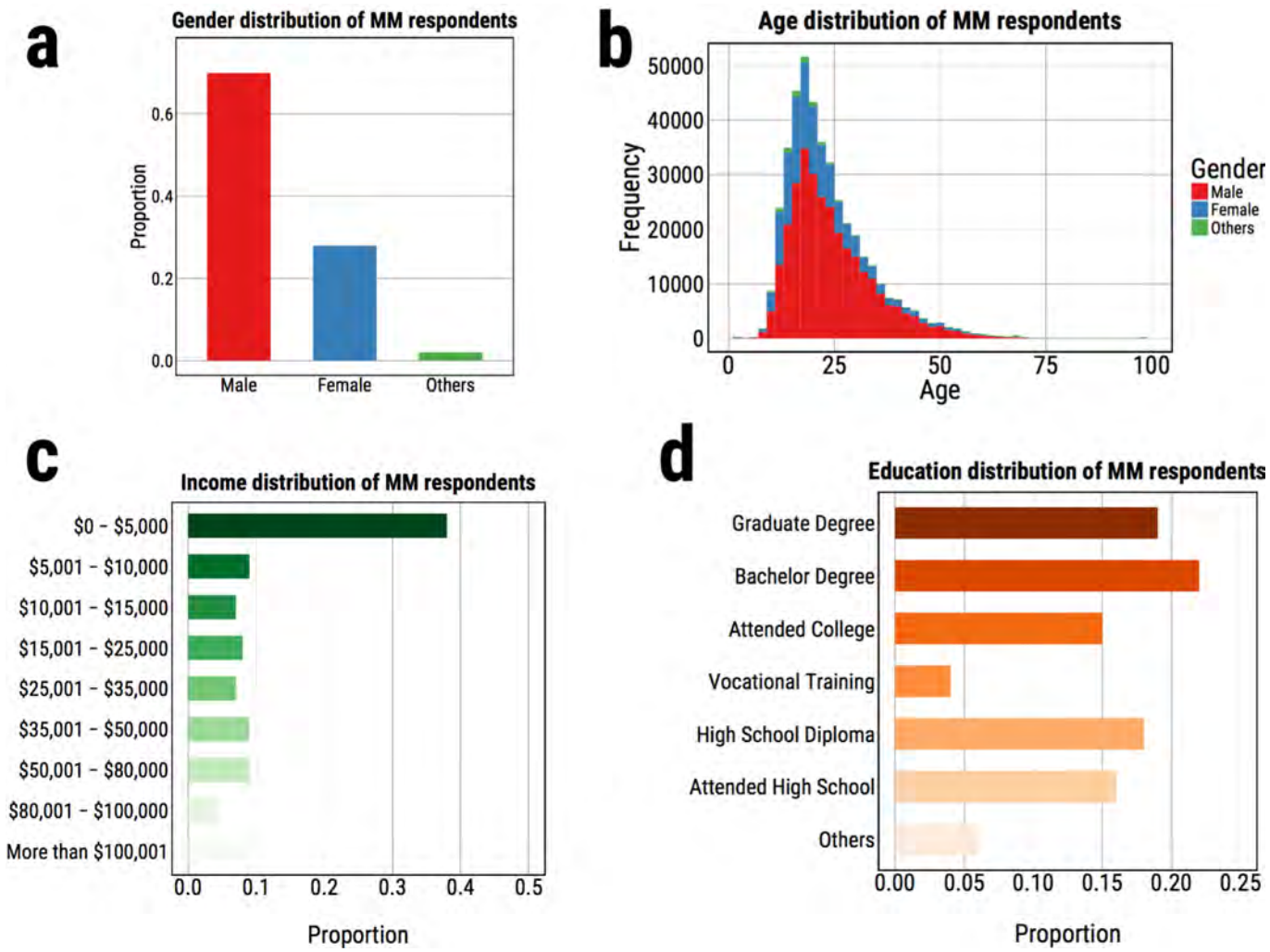
Extended Data Fig. 4 | Hierarchical cluster of countries based on country-level effect sizes calculated after filtering out responses for which the linguistic description was seen, thus neutralizing any potential effect of language. The three colours of the dendrogram branches represent three large clusters: Western, Eastern, and Southern. The names of the countries are coloured according to the Inglehart–Welzel Cultural Map 2010–2014[21]. See Supplementary Information for more details. The dendrogram is essentially similar to that shown in Fig. 3a.

© 2018 Springer Nature Limited. All rights reserved.



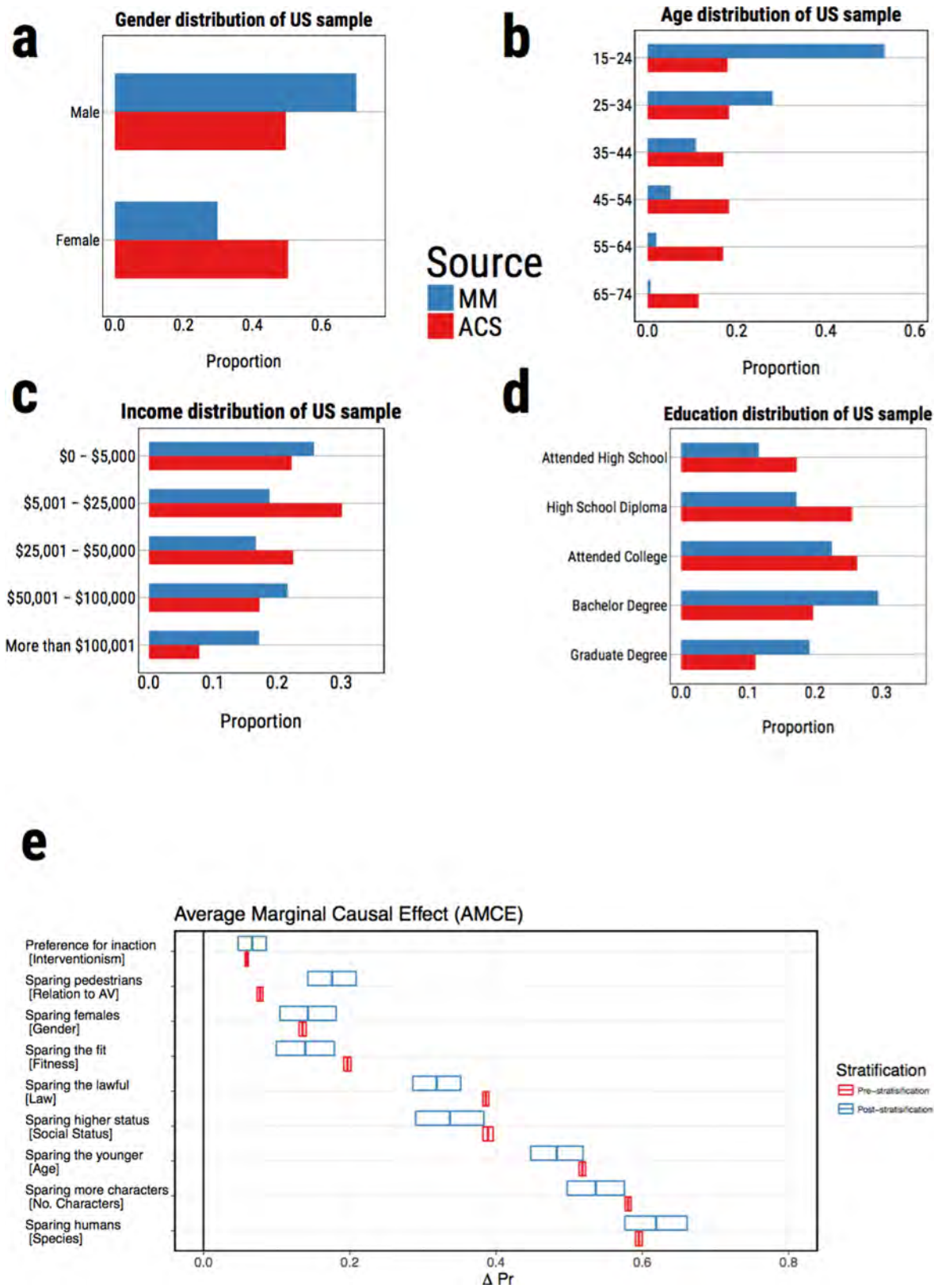
Extended Data Fig. 5 | Validation of hierarchical cluster of countries. **a**, **b**, We use two internal metrics of validation of three linkage criteria of calculating hierarchical clustering (Ward, Complete and Average) in addition to the *K*-means algorithm: **a**, Calinski–Harabasz index; **b**, silhouette index. The *x* axis indicates the number of clusters. For both internal metrics, a higher index value indicates a 'better' fit of partition to the data. **c**, **d**, We use two external metrics of validation of the used hierarchical clustering algorithm (Ward) versus those of random clustering assignment: **c**, purity; **d**, maximum matching. The histogram shows the distributions of purity and maximum matching values derived from randomly assigning countries to nine clusters. The red dotted lines indicate purity and maximum matching values computed from the clustering output of the hierarchical clustering algorithm using ACME values. See Supplementary Information for more details.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Demographic distributions of sample of population that completed the survey on Moral Machine (MM) website. Distributions are based on gender (**a**), age (**b**), income (**c**), and education attributes (**d**). Most users on Moral Machine are male, went through college, and are in their 20s or 30s. While this indicates that the users of Moral Machine are not a representative sample of the whole population, it is important to note that this sample at least covers broad demographics. See Supplementary Information for more details.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Demographic distributions of US sample of population that completed the survey on Moral Machine website versus US sample of population in American Community Survey (ACS) data set. a–d, Only gender (**a**), age (**b**), income (**c**), and education (**d**) attributes are available for both data sets. The MM US sample has an over-representation of males and younger individuals compared to the ACS US sample. **e**, A comparison of effect sizes as calculated for US respondents who took the survey on MM with the use of post-stratification to match the corresponding proportions for the ACS sample. Except for 'Relation to AV' (the second smallest effect), the direction and order of all effects are unaffected. See Supplementary Information for more details.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Regression table showing the individual variations for each of the nine attributes

	Demographics								
	Preference for Inaction	Sparing Pedestrians	Sparing the Lawful	Sparing Females	Sparing the Fit	Sparing Higher Status	Sparing the Young	Sparing More Characters	Sparing Humans
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Male	-0.015***	-0.022***	0.020***	-0.061***	0.024***	-0.009***	-0.018***	-0.024***	0.085***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
Age	0.001*	0.037***	-0.014***	0.008***	-0.019***	-0.022***	-0.020***	-0.011***	0.019***
	(0.0004)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Income	-0.003***	-0.008***	-0.010***	-0.008***	0.004***	-0.002	-0.004***	-0.003***	-0.007***
	(0.0004)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Is college educated	-0.010***	0.001	0.016***	-0.001	-0.008***	-0.012***	-0.016***	-0.009***	0.037***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Political views (conservative to progressive)	0.001	0.011***	-0.002*	0.014***	-0.007***	-0.012***	0.004***	0.009***	0.011***
	(0.0003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Religiousity	0.038***	0.064***	-0.083***	0.054***	-0.059***	-0.003	-0.016*	0.010	0.091***
	(0.003)	(0.005)	(0.006)	(0.007)	(0.007)	(0.009)	(0.006)	(0.006)	(0.005)
Constant	0.503***	0.565***	0.696***	0.585***	0.545***	0.680***	0.751***	0.772***	0.743***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
Structural Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6,477,161	2,542,020	1,547,713	1,100,816	993,252	356,165	1,064,506	1,168,238	1,105,292

Dependent variables are recorded as to whether the preferred option was chosen (for example, whether the respondent spared females). Continuous predictor variables are all standardized. All models include structural covariates (remaining attributes of a scenario). Coefficients are estimated using a regression-based estimator with cluster-robust standard errors. *$P < 0.01$, **$P < 0.001$, ***$P < 0.0001$. See Supplementary Information for more details.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | Country-level OLS regressions showing the relationships between key ethical preferences and various social, political and economic measures

	Sparing...			
	More Characters	The Lawful	Higher Status	Females
Individualism	**0.88***** **(5.69)**	-0.52*** (-2.96)	0.02 (0.11)	-0.07 (-0.38)
Rule of Law	-0.37** (-2.60)	**0.53***** **(3.29)**	-0.25 (-1.56)	0.24 (1.50)
Economic Inequality	0.23* (1.86)	-0.30** (-2.05)	**0.32**** **(2.28)**	0.46*** (3.23)
Female Health/Survival	0.12 (1.15)	0.06	0.24* (1.96)	**0.07** **(0.53)**
N	56	56	56	56
R^2	0.65	0.48	0.52	0.48

Pairwise exclusion was used for missing data. Predicted relationships are shown in bold. *$P < 0.10$, **$P < 0.05$, ***$P < 0.01$. See Supplementary Information for more details.

© 2018 Springer Nature Limited. All rights reserved.