The Moral Machine experiment

Citation: Awad, E. et al. "The Moral Machine experiment." Nature 563, 7729 (October 2018): 59–64 © 2018 Springer Nature Limited

Published Version: http://dx.doi.org/10.1038/s41586-018-0637-6

Publisher: Springer Science and Business Media LLC

Permanent Link: https://hdl.handle.net/1721.1/125065

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

The Moral Machine Experiment: 40 Million Decisions and the Path to Universal Machine Ethics

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]*

[1]The Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[3]Department of Psychology & Social Behavior, University of California, Irvine, CA 92697, USA
[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 02139, USA

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

Abstract

With the rapid development of Artificial Intelligence technology come widespread concerns about how machines will behave in morally charged situations. Addressing these concerns raises the major challenge of quantifying societal expectations about the ethical principles that should guide machine behavior. To address this challenge, we deployed the Moral Machine, an Internet-based experimental platform that is designed to explore the multi-dimensional moral dilemmas faced by autonomous vehicles. This platform enabled us to gather 40 million decisions in ten languages from over 2.3 million people in 233 countries and territories, and thus to assess the paths and obstacles to machine ethics. First, we summarize global moral preferences, some of them strong (e.g., sparing more lives or younger individuals), and some of them weak (e.g., sparing women, sparing pedestrians). Second, we document individual variations of these preferences, based on respondents' demographics (e.g., age, gender, income), and find that demographics do not change the direction of any preference. Third, we report cross-cultural ethical variation, and uncover three major clusters of countries exhibiting substantial differences along key moral preferences about autonomous vehicles. Fourth, we show that these differences correlate with modern institutions, but also with deep cultural traits. We discuss how these three layers of preferences can contribute to developing global, harmonious, and socially acceptable principles for machine ethics. All data used in this article can be accessed and downloaded at https://goo.gl/JXRrBP.

We are entering an age in which machines are not only tasked to promote well-being and minimize harm, but also to distribute the well-being they create, and the harm they cannot eliminate. Distributing well-being and harm inevitably creates tradeoffs, whose resolution falls in the moral domain[1,2,3]. Think of an autonomous vehicle (AV) 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? Should it swerve away

from a group of children and kill a single adult passenger instead? Even in the more common instances in which harm is not inevitable, but just possible, AVs will need to decide how to divvy 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 like 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 to give up on creating moral machines[5,6,7,8]. AVs will cruise our roads soon, necessitating agreement on the principles that should apply when, inevitably, life-threatening dilemmas emerge. The frequency at which these dilemma 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 AV choices in dilemma situations do not depend on the frequency of these situations. Whether these cases are rare, very rare, or extremely rare, we need to agree beforehand on how they should be solved.

The keyword here is “we”. As emphasized by former U.S. president Barack Obama[9], consensus in this matter is going to be important. Decisions about the ethical principles that will guide AVs cannot be left to solely to either the engineers or the ethicists. For consumers to switch from traditional human-driven cars to AVs, and for the wider public to accept the proliferation of AI-driven vehicles on their roads, both groups will need to understand the origins of the ethical principles programmed into these vehicles[10]. In other words, even if ethicists were to agree on how AVs 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 AVs promise in lieu of the status quo. Any attempt to devise AI ethics must be at least cognizant of public morality.

Accordingly, we need to gauge social expectations about the way AVs 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 of 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 toward universal machine ethics (or at least identify the obstacles thereto), we need a fine-grained understanding of how different individuals and different 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 the way citizens would want AVs 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 in 233 countries, dependencies, or territories (Fig.1 (a)). In the main interface of the Moral Machine, users are shown unavoidable accident scenarios with two possible outcomes, depending on whether the AV swerves or stays on course (Fig.1 (b)). They then click on the outcome that they find preferable. Accidents scenarios are generated by the Moral Machine following an exploration strategy that focuses on nine factors: sparing humans (vs. pets), staying on course (vs. swerving), sparing passengers (vs. pedestrians), sparing more lives (vs. fewer lives), sparing men (vs. women), sparing the young (vs. the elderly), sparing pedestrians who cross legally (vs. jaywalk), sparing the fit (vs. the less fit), and sparing those with higher social status (vs. lower social status). Additional characters were included in some scenarios (e.g., criminals, pregnant women, 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 can complete a survey that collects, among other variables, demographic information such as gender, age, income, and education, as well as religious and political attitudes. Participants are geolocated so that their coordinates can be used in a clustering analysis that seeks to identifies groups of countries or territories with homogeneous vectors of moral preferences.

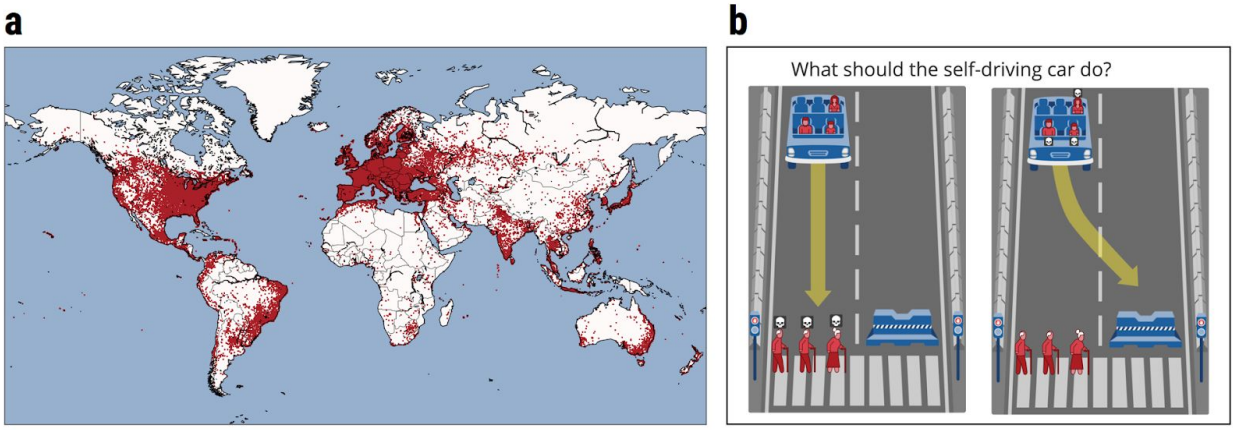

Figure 1. (a) World map highlighting the locations of Moral Machine visitors. *Each point represents a location from which at least one visitor made at least one decision. The number of visitors or decisions from each location are not represented.* ***(b) Moral Machine interface.*** *An AV experiences a sudden brake failure. Staying on course would result in the death of two elderly men and an elderly woman, 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).*

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 individual characteristics of respondents? Third, can we identify clusters of countries with homogeneous vectors of moral preferences? Fourth, do cultural and economic variations between countries predict variations in their vectors of moral preferences?

RESULTS

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 vs. female character, passengers vs. pedestrians, etc.)[18]. Fig.2 (a) 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 SI for computational details and assumptions).

As shown in Fig.2 (a), 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 starkly differ 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 AVs. As such, it provides an important context for interpreting our findings and their relevance to other countries which would attempt to follow the German example in the future. German Ethical Rule #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 #9 does not take a clear stance on whether and when AVs 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.

In contrast, German Ethical Rule #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 AVs to preferentially spare children, or for that matter, women over men, overweight persons over athletes, or executives over homeless persons--all of which we see weaker but clear effects for. 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 AV sacrifices children in a dilemma situation.

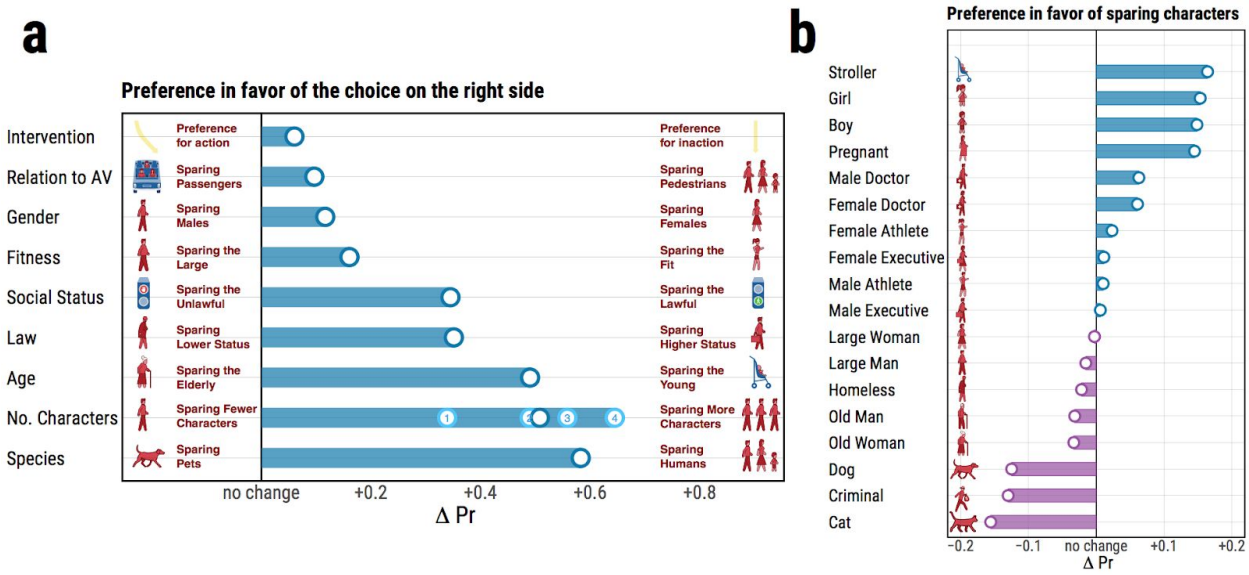

Figure 2. (a) Average marginal causal effect (AMCE) for each preference. *In each row, ΔPr is the difference between the probability of sparing characters possessing the attribute on the right, and the probability of sparing characters possessing the attribute on the left, aggregated over all other attributes. For example (age) the probability of sparing young characters is 0.49 (SE = 0.0008) greater than the probability of sparing older characters. The 95% CIs of the means are omitted due to their insignificant width, given the sample size. For the number of characters (No. characters), effect sizes are shown for each number of additional characters (1 to 4); the effect size for 2 additional characters overlaps with the mean effect of the attribute.* ***(b) Relative advantage or penalty for each character, compared to an adult man or woman.*** *For each character, ΔPr is the difference the between probability of sparing this character (when presented alone) and the probability of sparing one adult man or woman. For example, the probability of sparing a girl is 0.15 (SE = 0.003) higher than the probability of sparing an adult man/woman.*

Policymakers cannot solely rely on global trends when assessing social expectations, though. For machine ethics to be truly universal, preferences should be shared by all people, in all countries. If these conditions are not met, then policymakers should be aware of individual and national variations in the preferences extracted from the Moral Machine.

INDIVIDUAL VARIATIONS

We assessed individual variations by further analyzing the responses of the subgroup of Moral Machine users (N = 492,921) who filled 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 Figure S6 in SI). Of these, the most notable impacts are driven by gender and religiosity of respondents. For example, male respondents are 0.06 percentage point less inclined to spare females and 0.09 percentage point more inclined to spare humans compared to female respondents, while one increase in standard deviation of religiosity of respondent is associated with 0.09 more inclination to spare humans, 0.08 less inclination to spare the lawful, 0.06 more inclination to spare pedestrians, and 0.06 less inclination to spare the fit.

More importantly, none of the six characteristics splits its subpopulations into opposing directions of effect. Based on a unilateral dichotomization of each of the six attributes, resulting in two subpopulations per each, $\Delta \Pr$ has a positive value for all considered subpopulations e.g. both male and female respondents indicated preference for Sparing Females, but the latter group showed stronger preference (see Figure S5 in SI). In sum, the individual variations 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 exhibiting homogeneous vectors of moral preferences. We selected the 130 countries with at least 100 respondents (N range = [101 - 448,125]), standardized the 9 target AMCEs of each country, and conducted a hierarchical clustering on these 9 scores, using Euclidean distance and ward variance minimization algorithm[20]. This analysis identified three distinct "moral clusters" of countries. These are shown in Fig.3 (a), 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 Protestant / Scandinavian countries, and a sub-cluster containing Commonwealth / English-speaking countries.

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

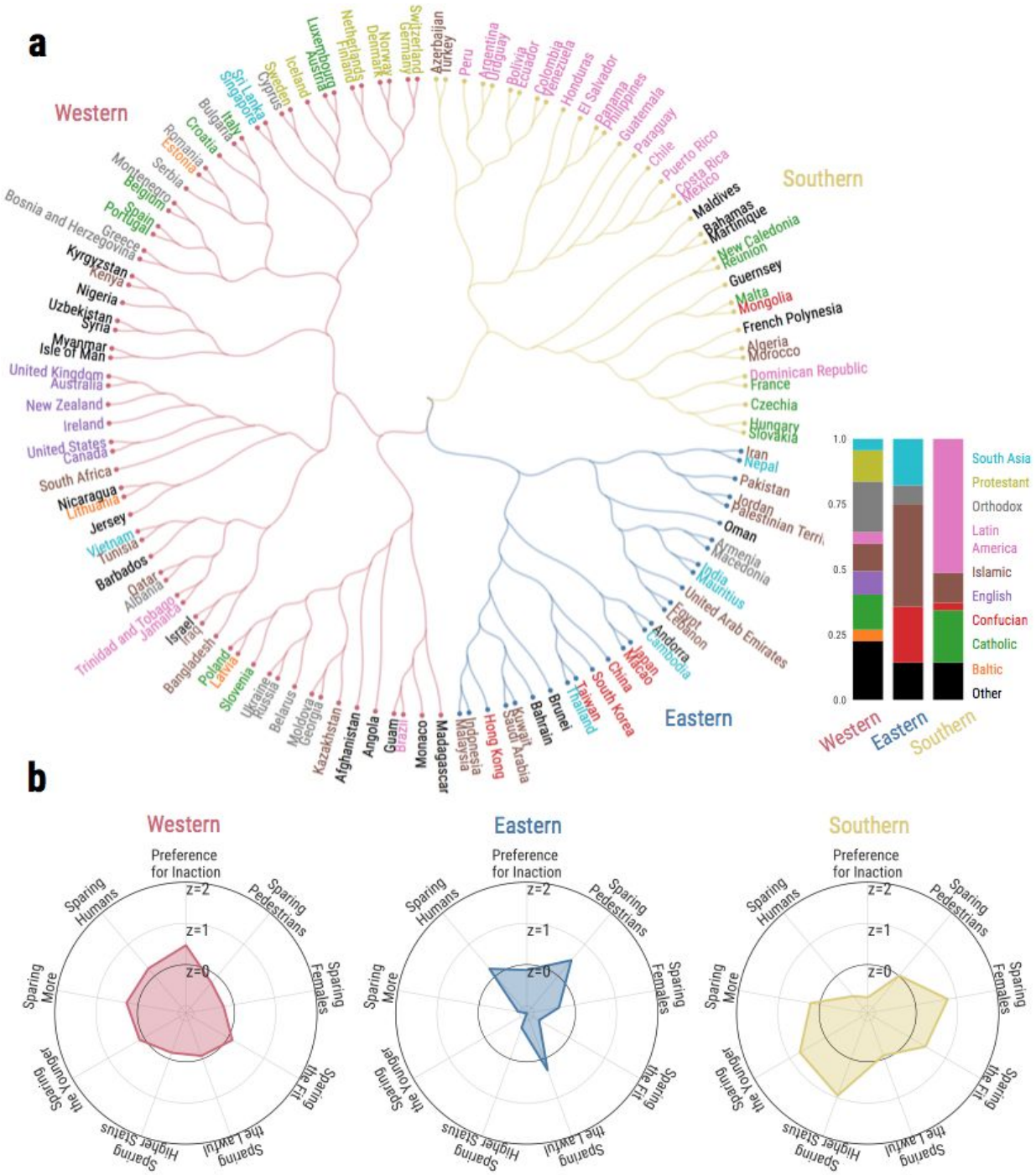

Figure 3. (a) Hierarchical Cluster of Countries based on average marginal causal effect. *One hundred thirty countries with at least 100 respondents are selected. Three colors of the dendrogram branches represent three large clusters -- Western, Eastern, and Southern. Names of the countries are colored according to Inglehart-Welzel Cultural Map 2010-2014[21]. Distributions across three clusters reveal stark differences. For instance, cluster 2 (Eastern) mostly consists of countries of Islamic and Confucian cultures. In contrast, cluster 1 (Western) has large percentages of Protestant, Catholic, and Orthodox countries of Europe.* ***(b) Mean AMCE z-scores of the three major clusters.*** *Radar plot of the mean AMCE z-scores of three clusters reveals striking pattern of differences between the clusters along the nine attributes. For example, countries belonging to the Southern cluster shows strong preference for the female gender compared to those of other clusters.*

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 e.g., 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 emerge when the clustering analysis is restricted to participants who only relied on the pictorial representations of the dilemmas, without accessing their written descriptions (see SI for more details). Furthermore, it is noteworthy that Spain appears in the Western cluster, adjacent to Portugal and other European countries, rather than in the Spanish-speaking Latin American sub-cluster.

This clustering pattern (which is fairly robust, see SI for details) 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.3 (b), 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 about 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, like 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.3 (b), though, suggests that manufacturers and policymakers should be, if not responsive, at least cognizant of moral preferences in the countries in which they design AI 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 AVs 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. While we do not attempt to pin down the ultimate reason or mechanism behind these correlations, we document them here as they point at possible deeper explanations of the cross-country differences and the clusters identified in the previous section.

As an illustration, consider the distance between the US and other countries in terms of the moral preferences extracted from the Moral Machine (MM distance). Figure 4c shows a substantial correlation ($\rho = 0.49$) between this MM distance and the cultural distance from the US based on the World Values Survey[22]. In other words, the more culturally similar a country is to the US, the more similarly its people play the Moral Machine.

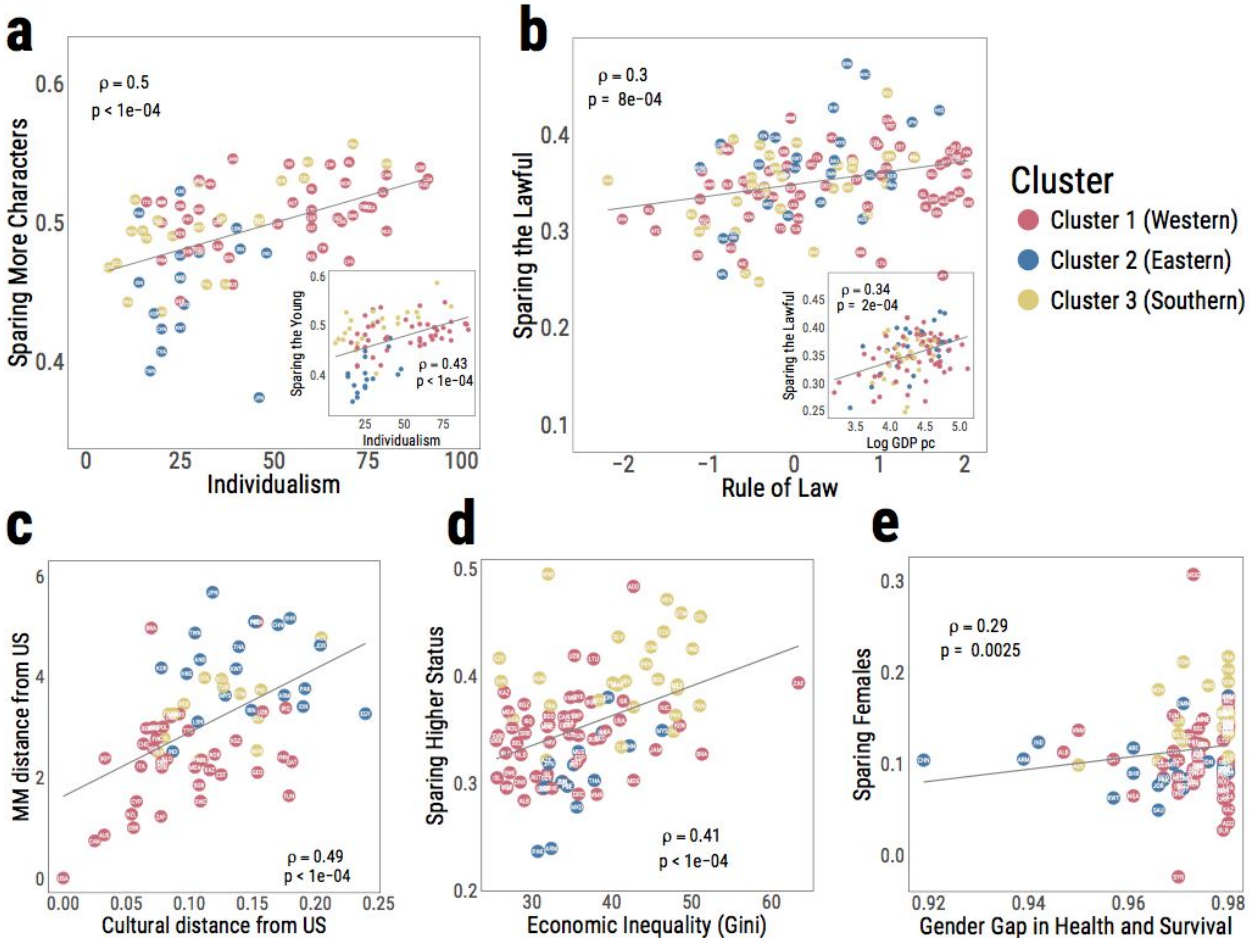

Figure 4. Association between Moral Machine preferences and other variables at the country level. *Each panel shows Spearman's ρ and p-value for the correlation test between the relevant pair of variables. (a) Association between individualism and each of the preference for sparing more characters, and the preference for sparing the young (inset). (b) Association between the preference for sparing the lawful and each of rule of law and log of GDP per capita (inset). (c) Association between cultural distance from US and MM distance (distance in terms of the moral preferences extracted from the Moral Machine) from US. (d) Association between economic inequality (Gini coefficient) and the preference for sparing higher status. (e) Association between the gender gap in health and survival and the preference for sparing females.*

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 (Figure 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 (Figure 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.

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 somewhat 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 (Figure 4b and inset). In other words, participants from countries which 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."

Finally, our data revealed 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 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 of a devaluation of women's lives in health and at birth, males are seen as more expendable in Moral Machine decision-making (Figure 4e). While not aiming to pin down the causes of these variation in table S3 in the SI, 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, outside of real-time supervision. We are going to cross that bridge any time now, and it will not happen in a distant theater 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 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 (e.g., the preference for

sparing pedestrians, or the preference for the car not to change direction) seem too weak to be fully relevant for policy debates. Other judgments 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 ever attempted to measure moral preferences using a 9-dimensional experimental design, in more than 200 countries. To achieve this unusual objective, we employed the unusual method of deploying 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 9-factorial design in each of these samples would verge into the impossible. Our approach allowed 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. 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; just as the post-stratification analysis we report in the SI. 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. Our hope is indeed that the Moral Machine experiment will provide the impetus for governments to organize inclusive public consultations.

Even with a sample size as large as ours, we could not do justice to all the complexity of AV 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, etc. 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 relation between respondents and characters (e.g. relatives, spouses). Our previous work did not find a strong impact of this variable on moral preferences[12], but this previous work only reached out to American respondents. Finally, what the Moral Machine captured is the state of moral preferences before AVs are deployed on the road, and before they start facing moral dilemmas. Preferences may well change after AVs start making their ethical decisions, and people die as a result, moving the discussion from abstract to identified victims, with all the psychological baggage that comes with such a change[28]. But to understand these changes, we will need the baseline that is provided by the Moral Machine.

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[29]. 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, like the *Asilomar AI Principles*[30],

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[31,32,33]. Here we showed that these conflicts, disagreements, and dissimilarities, while substantial, may not be fatal.

METHODS

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 SI), 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 AV 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, see Figure 2b) and to kill 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 labeled '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 time 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 whose details are presented in the SI, 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 if 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 are not analyzed 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 is detailed in the SI.

AUTHOR CONTRIBUTIONS

IR, AS and JFB planned the research. IR, AS, JFB, EA and SD designed the experiment. EA and SD built the platform and collected the data. EA, SD, RK, JS, and AS analyzed the data. All authors interpreted the results and wrote the paper.

ACKNOWLEDGMENTS

IR, EA, SD, and RK acknowledge support from the Ethics and Governance of Artificial Intelligence Fund. JFB acknowledges support from the ANR-Labex Institute for Advanced Study in Toulouse.

DATA AND CODE AVAILABILITY STATEMENT

Source data and code that can be used to reproduce Fig 2, Fig 3 and Fig 4; Supplementary Figures S3-S21; and Supplementary Table S2 are all available in the following link: https://goo.gl/JXRrBP

ETHICAL COMPLIANCE

This study was approved by the Institute Review Board (IRB) at Massachusetts Institute of Technology (MIT). The authors complied with all relevant ethical considerations.

REFERENCES

1. Greene, J. *Moral Tribes: Emotion, Reason and the Gap Between Us and Them.* (Atlantic Books, 2013).

2. Tomasello, M. *A Natural History of Human Thinking.* (Harvard University Press, 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, 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 University Press, 2008).
8. Dignum, V. Responsible Autonomy. in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* 4698–4704 (International Joint Conferences on Artificial Intelligence Organization, 2017).
9. Dadich, S. Barack Obama, Neural Nets, Self-Driving Cars, and the Future of the World. *Wired* (2016).
10. Shariff, A., Bonnefon, J.-F. & Rahwan, I. Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour* **1,** 694–696 (2017).
11. Conitzer, V., Brill, M. & Freeman, R. Crowdsourcing societal tradeoffs. in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems* 1213–1217 (International Foundation for Autonomous Agents and Multiagent Systems, 2015).
12. Bonnefon, J.-F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352,** (2016).
13. Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R. & Mikhail, J. A dissociation between moral judgments and justifications. *Mind and Language* **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. *Current Opinion in Psychology* **8,** 125–130 (2016).
18. Hainmueller, J., Hopkins, D. J. & Yamamoto, T. Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices Via Stated Preference Experiments. *SSRN Electronic Journal* (2013). doi:10.2139/ssrn.2231687
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. *arXiv [stat.ML]* (2011).
21. Inglehart, R. & Welzel, C. *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. (Cambridge University Press, 2005).
22. Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C., Gedranovich, A., McInerney, J., & Thue, B. A WEIRD Scale of Cultural Distance. *(submitted)*
23. Hofstede, G. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*. (SAGE Publications, 2003).
24. International Monetary Fund. World Economic Outlook Database. (2017).
25. Kaufmann, D., Kraay, A. & Mastruzzi, M. The Worldwide Governance Indicators: Methodology and Analytical Issues. *Hague Journal on the Rule of 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 UK, 2016).

28. Lee, S. & Feeley, T. H. The identifiable victim effect: a meta-analytic review. *Social Influence* **11,** 199–215 (2016).

29. Henrich, J. *et al.* In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *Am. Econ. Rev.* **91,** 73–78 (2001).

30. Asilomar AI Principles. *Future of Life Institute* Available at: https://futureoflife.org/ai-principles/. (Accessed: 5th January 2017)

31. Haidt, J. *The Righteous Mind: Why Good People Are Divided by Politics and Religion.* (Knopf Doubleday Publishing Group, 2012).

32. Gastil, J., Braman, D., Kahan, D. & Slovic, P. The Cultural Orientation of Mass Political Opinion. *PS Polit. Sci. Polit.* **44,** 711–714 (2011).

33. 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).