nature
SUPPLEMENTARY INFORMATION
https://doi.org/10.1038/s41586-018-0637-6
In the format provided by the authors and unedited.

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]*

[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

The Moral Machine Experiment

Supplementary Information

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, University of British Columbia, Vancouver, 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 02139, USA*

* *Corresponding authors: shariff@psych.ubc.ca; jean-francois.bonnefon@tse-fr.eu; irahwan@mit.edu*

1 Overview of the Moral Machine

The Moral Machine website offers four different modes, Judge, Design, Browse, and Classic. The Judge mode is the central data-gathering feature of the site, whose results we report in the main paper, and whose design is described in this Supporting Information document. The Design mode provides users with the possibility to create their own scenarios; the Browse mode allows to explore these user-generated scenarios; and the Classic mode features three classic trolley dilemmas

(Switch, Bridge, Loop) using the same visual presentation as the scenarios in the Judge mode, but different icons and an olde time sepia color scheme. Data of the Design, Browse and Classic modes are not discussed in the main article, and we accordingly focus on the Judge mode in this Supplementary Information.

The Judge mode is illustrated in Fig. S1. In this mode, users are presented with a series of moral dilemmas, with a simple point-and-click (or, in the case, of the mobile version, toggle-and-commit) method to choose which outcome of the two possible for a given scenario was deemed by the user to be most acceptable. A Moral Machine session comprises 13 scenarios, after which the user is presented with a summary of their choices along with how they compare to other users. Users are also asked to complete an optional survey (see below). Users can go through as many sessions as they wish, or leave the site mid-session.

The website was initially available in English only, but was later translated into nine additional languages: *Arabic*, *Chinese*, *French*, *German*, *Japanese*, *Portuguese*, *Korean*, *Spanish*, and *Russian*. Translation was performed through a process of forward-translation and back-translation by two bilingual native speakers of each of the nine languages. Multilingualism helped to understand cultural specificities of non-English-speaking countries, both by reaching more representative samples of the (monolingual) non-English-speaking inhabitants of these countries, and by collecting more accurate judgments by the (bilingual) non-native English-speaking inhabitants of these countries [?].

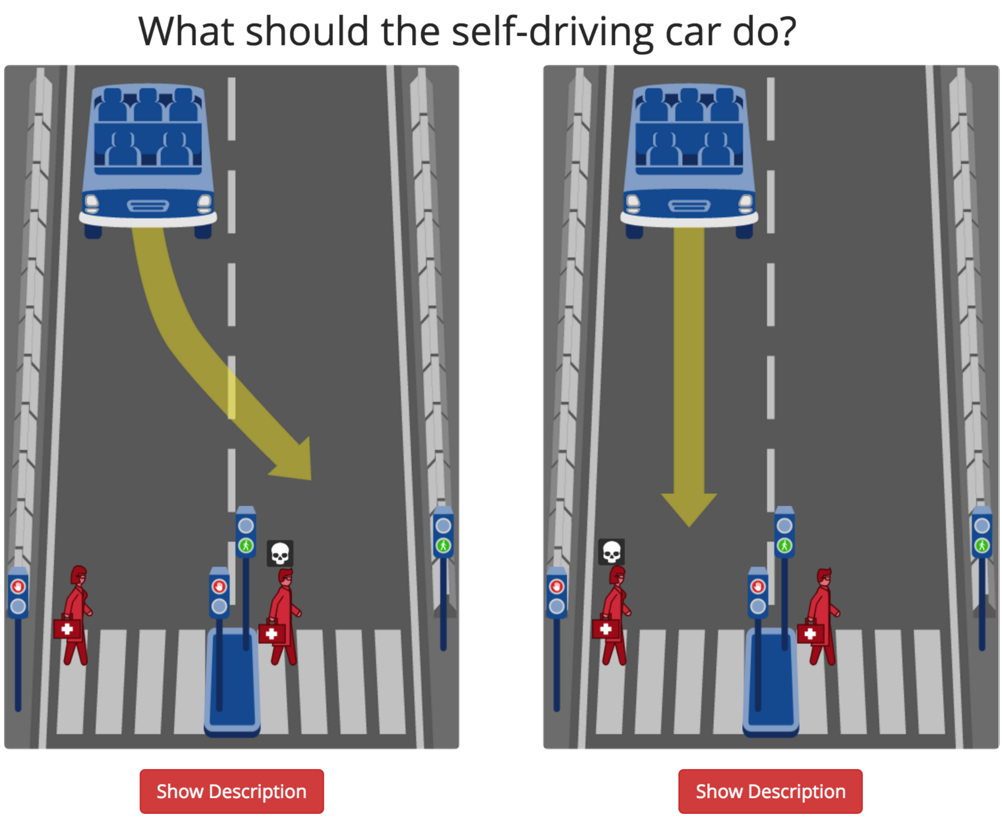


Figure S1: *Moral Machine* interface – Judge Interface. An example of a dilemma scenario: an AV experiences a sudden brake failure. Staying on course would result in the death of a female doctor, crossing on a "do not cross" signal (right). Swerving would result in the death of a male doctor, crossing on a "go ahead" signal (left).

2 Scenario Generation

Each scenario features characters from the following set: $C = \{$*Man*, *Woman*, *Pregnant Woman*, *Baby in Stroller*, *Elderly Man*, *Elderly Woman*, *Boy*, *Girl*, *Homeless Person*, *Large Woman*, *Large Man*, *Criminal*, *Male Executive*, *Female Executive*, *Female Athlete*, *Male Athlete*, *Female Doctor*, *Male Doctor*, *Dog*, *Cat*$\}$.

The scenarios are generated using randomization under constraints, so that scenarios explore the following dimensions:

1. **Species.** This dimension tests the extent to which users are willing to save or sacrifice pets vs. humans. We consider two sets of characters: 1) pets: $S_1 = \{Dog, Cat\}$, and 2) humans: $S_2 = C \setminus S_1$. The number of characters on each side[1] (same number on both sides) z is sampled from the set of positive integers $\{1, 2, \ldots, 5\}$. Then, z pairs of characters are sampled (unordered sampling with replacement) from the Cartesian product of the two sets $S_1 \times S_2$ (e.g. $(\texttt{Dog}, \texttt{Female Doctor})$). The first entries of the ordered pairs (i.e. pets) go to one side, while the second entries of the ordered pairs (i.e. humans) go to the other side. Accordingly,[2] the number of distinct scenarios for this dimension is $\sum_{i=1}^{5} \left[\binom{x_1+i-1}{i}\binom{x_2+i-1}{i}\right]$, where $x_1 = |S_1| = 2$, and $x_2 = |S_2| = 18$. Hence, the number of distinct scenarios for this dimension is $N_{Species} = 193,038$.

2. **Social Value.**[3] This dimension tests the extent to which users are willing to save/sacrifice characters of higher social value (e.g. a *Pregnant Woman,* or a *Male Executive*) when put against characters of lower social value (e.g. a *Criminal*). We consider three sets of characters, corresponding to three levels: 1) characters of low social value: $L_1 = \{$*Homeless*

[1] We use the term *side* to refer to one of the two options that the cars will choose to save/kill. Depending on the *relationship to vehicle* dimension (mentioned later), the *side* can refer to inside the car, or on the zebra crossing ahead or on the other lane.

[2] Note that in all cases we do unordered sampling with replacement. Hence, the formula $\binom{n+k-1}{k}$.

[3] Note here that "social value" refers to the *perceived* social value i.e. the widespread perception of the characters. We do not endorse the valuation of any humans above others, and we do not suggest that AVs should discriminate on the basis of any of the classifications presented in Moral Machine.

Person, *Criminal*$\}$, 2) characters of neutral social value: $L_2 = \{$*Man*, *Woman*$\}$, and 3) characters of high social value: $L_3 = \{$*Pregnant Woman*, *Male Executive*, *Female Executive*, *Female Doctor*, *Male Doctor*$\}$. In the main article, we restrict the analysis of this dimension to *Homeless Person* vs *Male and Female Executives*, for a cleaner focus on social status. The number of characters on each side (same number on both sides) z is sampled from the set of positive integers $\{1, 2, \ldots, 5\}$. Then, z pairs of characters are sampled (unordered sampling with replacement) from the following set: $(L_1 \bigtimes L_2) \cup (L_1 \bigtimes L_3) \cup (L_2 \bigtimes L_3)$. The first entries of the ordered pairs (i.e. lower-level characters) go to one side, while the second entries of the ordered pairs (i.e. higher-level characters) go to the other side. For example, (`Criminal`, `Man`), (`Woman`, `Male Doctor`), and (`Homeless Person`, `Female Executive`) are all possible sampled pairs where the first entries are strictly of lower value than the second entries. Given this, the number of distinct scenarios of this dimension is

$$\sum_{i=1}^{5}\sum_{j=0}^{i}\left[\binom{x_1+j-1}{j}\binom{x_2+i-j-1}{i-j}\binom{x_2+x_3+j-1}{j}\binom{x_3+i-j-1}{i-j}\right]$$

where $x_1 = |L_1| = 2$, $x_2 = |L_2| = 2$, and $x_3 = |L_3| = 5$. Hence, the number of distinct scenarios of this dimension is $N_{SocialV} = 58,547$.

3. **Gender.** This dimension tests the extent to which users are willing to save/sacrifice female characters when put against male characters. We consider two sets of characters: 1) female characters: $G_1 = \{$*Woman*, *Elderly Woman*, *Girl*, *Large Woman*, *Female Executive*, *Female Athlete*, *Female Doctor*$\}$, 2) male characters: $G_2 = \{m \mid m = g(f), f \in G_1\}$, where g is a bijection that maps each female character to its corresponding male character (e.g. $g($*Female Athlete*$) =$ *Male Athlete*). To generate a scenario of this dimension, the number of characters on each side (same number on both sides) z is sampled from the set of positive

integers $\{1, 2, \ldots, 5\}$. Then, z pairs of characters are sampled (unordered sampling with replacement) from: $\{(f, m) \mid f \in G_1, m = g(f)\}$. The first entries of the ordered pairs (i.e. female characters) go to one side, while the second entries of the ordered pairs (i.e. male characters) go to the other side. Given this, the number of distinct scenarios of this dimension is $\binom{x+4}{5} - 1$, where $x = |G_1| + 1 = 8$. Hence, the number of distinct scenarios of this dimension is $N_{Gender} = 791$.

4. **Age.** This dimension tests the extent to which users are willing to save/sacrifice characters of younger age when put against characters of older age. We consider three sets of characters, corresponding to three levels: 1) characters of young age: $A_1 = \{$*Boy*, *Girl*$\}$, 2) neutral adult characters: $A_2 = \{$*Man*, *Woman*$\}$, and 3) elderly characters: $A_3 = \{$*Elderly Man*, *Elderly Woman*$\}$. Consider the following two gender-preserving bijections $a_1 : A_1 \to A_2$, and $a_2 : A_2 \to A_3$ (e.g. $a_1($*Boy*$) =$ *Man*, and $a_2($*Woman*$) =$ *Elderly Woman*). To generate a scenario of this dimension, the number of characters on each side (same number on both sides) z is sampled from the set of positive integers $\{1, 2, \ldots, 5\}$. Then, z pairs of characters are sampled (unordered sampling with replacement) from the following set:

$$\{(y, n) \mid y \in A_1, n = a_1(y)\} \cup$$

$$\{(n, d) \mid n \in A_2, d = a_2(n)\} \cup$$

$$\{(y, d) \mid y \in A_1, d = a_2 \circ a_1(y)\}$$

The first entries of the ordered pairs (i.e. younger characters) go to one side, while the second entries of the ordered pairs (i.e. older characters) go to the other side. Given this, the number of distinct scenarios of this dimension is $\binom{x+4}{5} - 1$, where $x = |A_1| + |A_2| + |A_1| + 1 = 2+2+2+1 = 7$. Hence, the number of distinct scenarios of this dimension is $N_{Age} = 461$.

5. **Fitness.** This dimension tests the extent to which users are willing to save/sacrifice characters of higher physical fitness when put against characters of lower physical fitness. We consider three sets of characters, corresponding to three levels: 1) characters of low fitness: $F_1 = \{$*Large Man*, *Large Woman*$\}$, 2) characters of neutral fitness: $F_2 = \{$*Man*, *Woman*$\}$, and 3) characters of high fitness: $F_3 = \{$*Male Athlete*, *Female Athlete*$\}$. Consider the following two gender-preserving bijections $f_1 : F_1 \to F_2$, and $f_2 : F_2 \to F_3$ (e.g. $f_1($*Large Man*$) =$ *Man*, and $f_2($*Woman*$) =$ *Female Athlete*). To generate a scenario of this dimension, the number of characters on each side (same number on both sides) z is sampled from the set of positive integers $\{1, 2, \ldots, 5\}$. Then, z pairs of characters are sampled (unordered sampling with replacement) from the following set:

$$\{(l, n) \mid l \in F_1, n = f_1(l)\} \cup$$

$$\{(n, f) \mid n \in F_2, f = f_2(n)\} \cup$$

$$\{(l, f) \mid l \in F_1, f = f_2 \circ f_1(l)\}$$

The first entries of the ordered pairs (i.e. characters of lower fitness) go to one side, while the second entries of the ordered pairs (i.e. characters of higher fitness) go to the other side. Given this, the number of distinct scenarios of this dimension is $\binom{x+4}{5} - 1$, where $x = |F_1| + |F_2| + |F_1| + 1 = 2+2+2+1 = 7$. Hence, the number of distinct scenarios of this dimension is $N_{Fitness} = 461$.

6. **Utilitarianism.** This dimension tests the extent to which users are willing to save/sacrifice a group of characters when put against the same group of characters in *addition* to a positive number of characters. To generate a scenario of this dimension, the number of characters on each side (same number on both sides) z is sampled from the set of positive integers $\{1, 2, \ldots, 4\}$. Then, z pairs of characters are sampled (unordered sampling with replacement) from the following set: $\{(c, c) \mid c \in C\}$, where C is the set of all characters, defined above. This will create two sides with identical groups of characters. Then, the number of *additional* characters u is sampled from the set of positive integers $\{1, \ldots, 5 - z\}$. Then, the u *additional* characters are sampled (unordered sampling with replacement) from C. All the *additional* characters go to the same side. Given this,[4] the number of distinct scenarios of this dimension is $\binom{x+4}{5} - \binom{x_0+4}{5}$, where $x_0 = |C| + 1 = 21$, and $x = x_0 + |C| = 41$. Hence, the number of distinct scenarios of this dimension is $N_{Utilitarian} = 1, 168, 629$.

Given that the six dimensions above are mutually exclusive in terms of the generated scenarios, the overall number of distinct scenarios of the six dimensions equal to the sum of the numbers above i.e. $N = 1, 421, 927$. Each user is presented with two randomly sampled scenarios of each of the above dimensions, in addition to one completely random scenario (that can

[4]To see how this calculation is done, consider the following set

$$X = \{(c, c) \mid c \in C\} \cup \{(c, \_) \mid c \in C\} \cup \{(\_, \_)\}$$

where "_" refers to no character in that entry. Now drawing unordered 5 samples with replacement can be done in $\binom{|X|+4}{5}$ ways. However, this includes undesirable cases e.g. drawing $(\_, .)$ five times, where "." is a character or "_". Thus, the subtracted term.

have any number of characters on each side and in any combination of characters). These together make the 13 scenarios per session. The order of the 13 scenarios is also counterbalanced over sessions. Using a similar calculation as before, the number of distinct random scenarios is $\left[\binom{x+4}{5} - 1\right]^2$, where $x = |C| + 1 = 21$. Hence, the number of distinct completely random scenarios is $N_{Random} = 14,102,512,516$. These, of course, include scenarios from the six dimensions above.

In addition to the above six dimensions, the following three dimensions are randomly sampled in conjunction with every scenario of the six dimensions above:

1. **Interventionism.** This dimension tests the extent to which the omission bias (i.e. the favorability of omission/inaction over the commission/action). In every scenario, the car has to make a decision as to stay (omission) or to swerve (commission). To model this dimension, each of the generated scenarios would have one side as the omission, and the other as the commission, or vice versa. This multiplies the number of scenarios by two. To see why, consider a gender-dimension scenario. It can have two possibilities when Interventionism is added: females sacrificed on omission vs. males sacrificed on commission, and vice versa.

2. **Relationship to vehicle.** This dimension tests the preference to save the passengers over the pedestrians and to what degree it differs from the case of saving pedestrians over other group of pedestrians. Each scenario presents a tradeoff of either between passengers and pedestrians, or between pedestrians and other groups of pedestrians. A large concrete barrier serves as a visual indicator of the case where the passengers may be sacrificed. Pedestrians are ren-

dered over a zebra crossing, which is split by an island in case of a pedestrian vs pedestrian scenario. Pedestrians can be crossing either ahead of the car (for the case of passengers vs. pedestrians), on the other lane (also for the case of passengers vs. pedestrians), or on both lanes (for the case of pedestrians vs. pedestrians). To model this dimension, each of the generated scenarios would have both sides on zebra crossings; one side inside the car, and the other on the zebra crossing; or vice versa. This multiplies the number of scenarios by three. To see why, consider again the gender-dimension scenario. It can have in conjunction with this dimension the following possibilities: female passengers vs. male pedestrians, female pedestrians vs. male passengers, and female pedestrians vs. male pedestrians.

3. **Concern for law.** This dimension tests the effect of adding legal complications in the form of pedestrian crossing signals. Scenarios can have no crossing signals (no legal complications), crossing signals on either side of the crossing, that all have the same light color, red or green (for the case passengers vs. pedestrians), or crossing signals on either side of each lane's crossing, if split by an island, where the light color of one side is different from the light color of the other side e.g. green vs. red (for the case of pedestrians vs. pedestrians). In the last case, the crossing signal on the main lane can be green (i.e. legal crossing), in which case, the crossing signal on the other lane is red (illegal crossing), or vice versa. In the case of matching green/red light crossing signals, the two signals are either both green (legal) or red (illegal). To model this dimension, each of the generated scenarios would have no legal complication, one side as legal, or the same side as illegal (the other side will be a function of this side). This multiplies the number of scenarios by three. To see why, consider again

the gender-dimension scenario. It can have in conjunction with this dimension the following possibilities: female pedestrians with no legal considerations, female pedestrians crossing legally, and female pedestrians crossing illegally. The other side would always feature male pedestrians/passengers with their legal considerations determined as a function of the legal considerations of the female pedestrians.

The above three extra dimension can be factored independently from each other. Hence, they all together multiply the number of distinct scenarios by 18. Thus, the overall number of distinct scenarios of the nine dimensions (i.e. excluding the completely random scenarios) is $M = 18 \times N = 25,594,686$ (or approximately $26M$).

The stay/swerve outcomes are rendered on the fly by overlaying vector graphic stylized icons of the characters and dynamic objects on a static image background depicting the respective outcome course, and the left/right position of each outcome is switched randomly, so as to avoid any bias from handedness. A short delay featuring an animated visual distraction is forced between choice commitment and the rendering of the next scenario, so as to allow the user to mentally clear and shift.

The damage level to each character is depicted using either a skull icon (death), an equal-armed cross icon (injury), or a question mark icon (unknown). For simplicity, scenarios generated in the *Judge* interface have the possibility of death only. The other two levels (injury and unknown) are only used in the *Design* interface.

Apart from the instructions available on the main page, a brief description of each outcome may also be viewed by clicking a button below the depiction of each outcome, describing the circumstances of the vehicle (autopilot with sudden brake failure), its course in that outcome, and any pedestrian crossing signal(s) involved, as well as a list of the impacted characters and the damage to them that will result in that outcome.

After the user has completed assessing all 13 scenarios, they are presented with a summary of their decisions, a sample of which can be seen in Fig. S2.

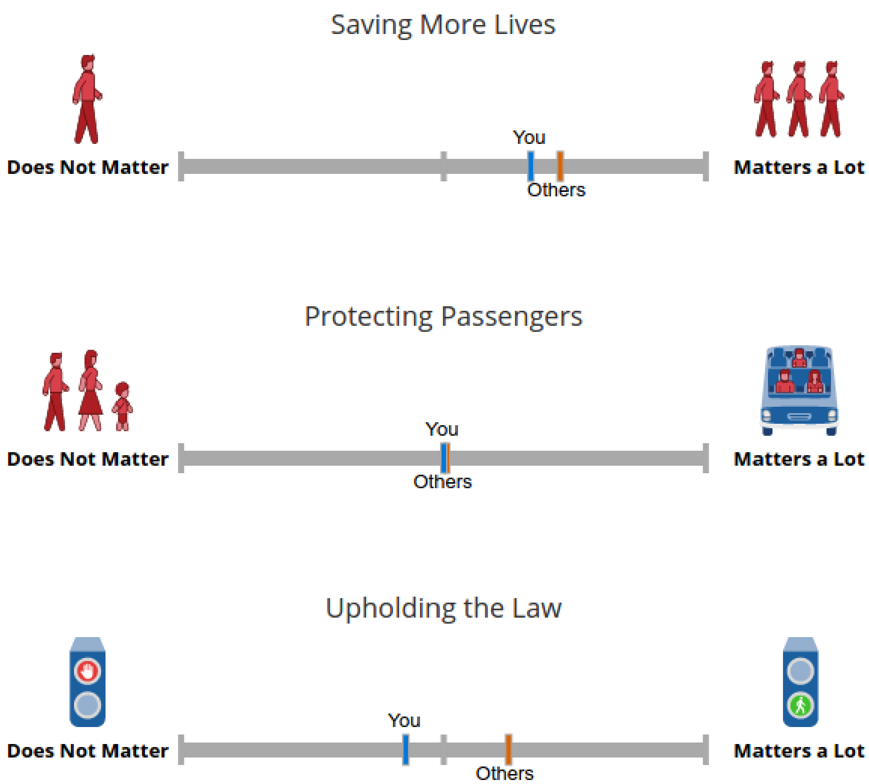


Figure S2: A sample of the summary results shown to users upon finishing all 13 scenarios. Only three sliders out of the overall nine sliders are shown.

Demographic Survey Four months after the initial deployment, an extension of the user result interface was added to collect demographic information and feedback on the user's perception of their own moral priorities along each dimension. This survey helps us understand the type of users visiting our website, and to assess the effects of demographic characteristics, political views or religious beliefs on moral preferences [1]. The survey contains demographic questions about age, gender, income, education, religious views, and political views. Further, it asks users to provide their stated preferences over the nine dimensions using sliders. Additionally, the survey contains four questions that concern the attitude towards machine intelligence. However, we do not use respondents' answers to these questions here or in the main manuscript.

Whether the option to do the survey appeared before or after the user saw their *Results* page was counterbalanced between users. In addition, the survey questions were presented in four blocks. Each block contains one group of questions: (a) the stated preference sliders, (b) the demographic questions (age, gender, income, and education), (c) the political and religious view questions, and (d) the "attitude towards machine intelligence" questions. The order of the blocks and the order of questions within each block was also counterbalanced between users.

3 Estimating Causal Effects

We employ conjoint analysis to identify causal effects of multiple treatment components (i.e. factors) simultaneously. In a conjoint design, respondents are asked to choose from, or rate profiles that represent multiple attributes. For example, respondents may be asked to perform multiple tasks of voting for one out of three candidates after being presented with the age, gender, education, income, etc. of each, and the components of competing profiles are manipulated independently. Conjoint analysis has been introduced in political science by Green and Rao [2]. Similar tools were independently introduced by sociologists under names like “vignettes” or “factorial surveys” [3,4]. We follow the framework proposed by Hainmueller et al. [5], in which they proposed the potential outcomes framework of causal inference [6,7] to analyze the causal properties of conjoint analysis. They proposed two variations of outcome variable: choice-based variable (choosing one of multiple profiles), and rating-based variable (rating each profile). Their framework allows for non-parametric identification of causal effects under few testable assumptions that do not include any modeling assumptions. Following, we review notation, assumptions, causal quantities of interest, identification strategies, and estimation strategies, before we present how this can be applied to our data. Then, we present our results, and conclude this part with robustness checks.

Notation. Consider a random sample of N respondents, indexed by $i \in \{1, \ldots, N\}$, from a population $\mathcal{P}$. Each respondent i is presented with K tasks, indexed by $k \in \{1, \ldots, K\}$. In each task k, respondent i chooses from (or rate each of) J profiles, indexed by $j \in \{1, \ldots, J\}$. Each profile j in task k is characterised by L attributes, indexed by $l \in \{1, \ldots, L\}$. Each attribute l is assumed to have D_l levels, indexed by $d_l \in \{t_{l1}, \ldots, t_{lD_l}\}$ (continuous attributes are discretised).

For example, in the case of Moral Machine, the number of respondents is $N = 2.3\ M$. Each respondent is presented with $K = 13$ tasks.[5] Each task consists of $J = 2$ profiles (left vs.

[5]Each complete session consists of 13 scenarios. A respondent may choose to complete more/less than a one complete session. For that, a clean version of the analysis considers only the first complete session per respondent.

right in the desktop interface or default vs. non-default in the mobile interface). Each profile is characterised by $L = 4$ attributes (Interventionism, Relation to AV, legality, and characters type). Each of the four attributes has $D_l = 2$ levels.[6] For the characters type, we will repeat the analysis for each of the six dimensions (gender, age, fitness, social value/status, species, and utilitarianism), and each of these attributes has two levels.

Let $\mathcal{C}$ be the Cartesian product of all possible levels of attributes: $\mathcal{C} = \bigtimes_{l=1}^{L}\{t_{l1}, \ldots, t_{lD_l}\}$. We use an L-dimensional vector $T_{ijk} = [T_{ijk1}, \ldots, T_{ijkL}]^\top \in \mathcal{T} \subseteq \mathcal{C}$ to denote a treatment that is presented to respondent i as the jth profile in her kth task, where T_{ijkl} is the lth attribute of the profile, and $\mathcal{T}$ is the domain of all profiles of interest. An example of a treatment is [`Omission, Passengers, No Legality, Males`], which represents a profile in which an AV with all-male passengers would hit a barrier killing all passengers, if left without intervention. We also use $\overline{\mathbf{T}}_i = (1, \ldots, \mathbf{T}_{iK})$ to denote the set of all JK profiles presented to respondent i, where $\mathbf{T}_{ik} = [T_{i1k}, \ldots, T_{iJk}]^\top$ is the set of all attribute values for all J profiles in the task k presented to respondent i. An example of a realization of $\mathbf{T}_{ik}$ is:

$$\bar{\mathbf{t}} = \begin{bmatrix} \texttt{Omission} & \texttt{Passengers} & \texttt{No Legality} & \texttt{Males} \\ \texttt{Commission} & \texttt{Pedestrians} & \texttt{No Legality} & \texttt{Females} \end{bmatrix}$$

which represents a task with two profiles. The task features an AV with all-male passengers

This makes the number of respondents N less than 2.3 M.

[6] The attribute legality has $D_l = 3$ levels, but when we analyze the effect of this attribute, we only consider the two levels of legal crossing and illegal crossing.

that would hit a barrier killing all passengers, if left without intervention. On the other hand, if the AV swerves, it will, instead, kill a group of all-female pedestrians crossing on a road with no crossing signals.

Given $\bar{\mathbf{t}}$, a realization of $\mathbf{T}_{ik}$ (or a sequence of profile attributes), let

$$Y_{ik}(\bar{\mathbf{t}}) = [Y_{i1k}(\bar{\mathbf{t}}), \ldots, Y_{iJk}(\bar{\mathbf{t}})]^\top$$

be the J-dimensional vector of potential outcomes for respondent i when presented with task k, where $Y_{ijk}(\bar{\mathbf{t}})$ is the potential outcome for profile j. Using the choice tasks, we have:

$$\forall i \; \forall k \; \forall \bar{\mathbf{t}} \; : \; \sum_{j=1}^{J} Y_{ijk}(\bar{\mathbf{t}}) = 1$$

Assumptions. The conjoint analysis performed here relies on the following three simplifying assumptions (taken from [5]). The first assumption is that the potential outcomes remain stable regardless of the task order k. This means that a respondent's response to a treatment is the same whether she answers the task before or after answering other tasks.

Assumption 1 (Stability and No Carryover Effects). Potential outcomes always take the same value as long as all the profiles in the same choice task have the same set of attributes. Formally:

$$\forall i \; \forall j \; \forall k, k' \; \forall \overline{\mathbf{T}}_i, \overline{\mathbf{T}'}_i : [\mathbf{T}_{ik} = \mathbf{T}'_{ik'} \Rightarrow Y_{ijk}(\overline{\mathbf{T}}_i) = Y_{ijk'}(\overline{\mathbf{T}'}_i)]$$

where $\overline{\mathbf{T}}_i = (1, \ldots, \mathbf{T}_{iK})$ and $\overline{\mathbf{T}'}_i = (1, \ldots, \mathbf{T}'_{iK})$.

The second assumption is that the potential outcomes remain stable regardless of the profile order j within a task. This means that a respondent's response to a task is the same no matter how the profiles are ordered within this task.

Assumption 2 (No Profile-Order Effects). Potential outcomes always take the same value regardless of the order of the profiles in the same choice task. Formally:

$$\forall i\ \forall j, j'\ \forall k\ \forall \mathbf{T}_{ik}, \mathbf{T}'_{ik} : [(T_{ijk} = T'_{ij'k} \wedge T_{ij'k} = T'_{ijk}) \Rightarrow Y_{ij}(\mathbf{T}_{ik}) = Y_{ij'}(\mathbf{T}'_{ik})]$$

where $\mathbf{T}_{ik} = [1, \ldots, \mathbf{T}_{iJk}]^\top$ and $\mathbf{T}'_{ik} = [1, \ldots, \mathbf{T}'_{iJk}]^\top$.

The third assumption is that the potential outcomes are statistically independent of the profile. This means that the attributes of each profile are randomly generated. This holds if attributes were randomly assigned to each profile. This assumption has a second part in which every combination of attribute values for which potential outcomes are defined has a non-zero probability.

Assumption 3 (Randomization of the Profiles). Potential outcomes are statistically independent of the profiles. Further, all the possible attribute combinations for which potential outcomes are defined have non-zero probability. Formally:

$$\forall i\ \forall j\ \forall k\ \forall l\ \forall \mathbf{t} : [(Y_i(\mathbf{t}) \perp\!\!\!\perp T_{ijkl})\ \wedge\ (0 < p(\mathbf{t}) \equiv p(\mathbf{T}_{ik} = \mathbf{t}) < 1)]$$

where independence is the pairwise independence between each element of $Y_i(\mathbf{t})$ and T_{ijkl}.

Our design ensures the satisfaction of the first part of the third assumption and it allows us to partially test for the first two assumptions. Given our design restrictions (to be mentioned later), the second part of assumption 3 is only satisfied for combinations of interest.

Causal quantities of interest – identification and estimation strategies The main quantity of interest is the *average marginal component effect* (AMCE), and it represents the marginal effect of an attribute l averaged over the joint distribution of other attributes. Under assumptions 1, 2, and 3 above, this quantity is given by:

$$\begin{aligned}\hat{\bar{\pi}}_l(t_1, t_0, p(\mathbf{t})) = \sum_{(t,\mathbf{t})\in\widetilde{\mathcal{T}}} &\Big\{\mathbb{E}\big[Y_{ijk}|T_{ijkl} = t_1, T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t}\big] \\ &- \mathbb{E}\big[Y_{ijk}|T_{ijkl} = t_0, T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t}\big]\Big\} \\ &\times p\Big(T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t}|(T_{ijk[-l]}, \mathbf{T}_{i[-j]k}) \in \widetilde{\mathcal{T}}\Big)\end{aligned}$$

where $T_{ijk[-l]}$ is the $(L-1)$-vector of other components (for choice task k, profile j, faced by respondent i), $\mathbf{T}_{i[-j]k}$ is the $[(J-1)\times L]$-matrix of other profiles (for choice task k faced by respondent i), and $\widetilde{\mathcal{T}}$ is the intersection of the support of $p(T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t}|T_{ijkl} = t_1)$ and $p(T_{ijk[-l]} = t, \mathbf{T}_{i[-j]k} = \mathbf{t}|T_{ijkl} = t_0)$.

Consider the following assumption:

Assumption 4 (Conditionally Independent Randomization). An attribute component of a treatment can take any value after conditioning on some of the other attribute values. Formally:

$$\forall i \; \forall j \; \forall k \; \forall l : [T_{ijkl} \perp\!\!\!\perp \{T^S_{ijk}, T_{i[-j]k}\} | T^R_{ijk}]$$

where T^R_{ijk} is an L^R-dimensional subvector of $T_{ijk[-l]}$, and T^S_{ijk} is the relative complement of T^R_{ijk} w.r.t. $T_{ijk[-l]}$.

Under assumptions 1,2,3, and 4, AMCE can be non-parametrically estimated using the following unbiased subclassification estimator:

$$\hat{\hat{\bar{\pi}}}_l(t_1, t_0, p(\mathbf{t})) = \sum_{t^R \in \mathcal{T}^R} \left\{ \frac{\sum_{i=1}^N \sum_{j=1}^J \sum_{k=1}^K Y_{ijk} \mathbb{1}\{T_{ijkl} = t_1, T^R_{ijk} = t^R\}}{n_{1t^R}} - \frac{\sum_{i=1}^N \sum_{j=1}^J \sum_{k=1}^K Y_{ijk} \mathbb{1}\{T_{ijkl} = t_0, T^R_{ijk} = t^R\}}{n_{0t^R}} \right\} \times \mathbb{P}\Big(T^R_{ijk} = t^R\Big)$$

where n_{dt^R} is the number of profiles for which $T_{ijkl} = t_d$ and $T^R_{ijk} = t^R$, and $\mathcal{T}^R$ is the intersection of the supports of $p(T^R_{ijk} = t^R | T_{ijkl} = t_1)$ and $p(T^R_{ijk} = t^R | T_{ijkl} = t_0)$. A proof is provided in [5].

Given the correspondence between subclassification and linear regression, one can use linear regression to compute the above estimator. As a result, the use of linear regression would allow for non-parametric (unbiased) estimation, even though the outcome variable is binary.

A special case of Assumption 4 is when T^R_{ijk} has a length of zero, which corresponds to a completely independent randomization. In this case, the above equation reduces to:

$$\hat{\bar{\hat{\pi}}}_l(t_1, t_0, p(\mathbf{t})) = \sum_{t^R \in \mathcal{T}^R} \left\{ \frac{\sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{k=1}^{K} Y_{ijk} \mathbb{1}\{T_{ijkl} = t_1\}}{n_1} - \frac{\sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{k=1}^{K} Y_{ijk} \mathbb{1}\{T_{ijkl} = t_0\}}{n_0} \right\}$$

where n_d is the number of profiles for which $T_{ijkl} = t_d$.

For interaction, we do two types of interactions (both were suggested in [5]): 1) interaction between different components (formalized as the *average component interaction effect* (ACIE)), and 2) interaction between components and respondent's background information (formalized as conditional AMCE). Both can be estimated by subclassification and applying linear regression estimators (while including interaction terms).

For variance estimation, as proposed in [5], we calculate within-respondent cluster-robust standard errors to account for the fact that the observed choice outcomes within each task are strongly negatively correlated for each respondent (a respondent will choose one outcome over the other, especially in the 2-profile outcomes we consider), and to account for the fact that tasks within each respondents are expected to be positively correlated (given respondent's unobserved characteristics).

Empirical design restrictions. It would be very simple, for the analysis, to consider all possible combinations of attributes when creating profiles, and to test for all possible combinations of pairs of profiles. However, this would result in unrealistic scenarios. As such, two types of restrictions were applied: one at the level of profiles, and another at the level of profile pairing. An example of the former is a restriction that disallows some combination of levels of the two attributes *relation to AV* and *legality*. Specifically, profiles that feature *passengers* cannot have legal considerations (*legal crossing* or *illegal crossing*). This restriction stems from the requirement of a realistic scenario in which passengers make no legality-related decisions. This restriction does not exist for *pedestrians* who can make (il)legal decisions by crossing on a green or a red signal. Therefore, comparing pooled profiles of pedestrians to pooled profiles of passengers would result in a biased estimate of effect despite the random generation of the two factors, *legality* and *relation to AV*. Fortunately, the fourth assumption above allows for conditioning on the *no legality* level in order to produce an unbiased estimate for the effect of *relation to AV* (that is pedestrians and passengers are compared only when there is no legality involved). Similarly, the same assumption allows us to condition on the *pedestrian* level in order to produce an unbiased estimate for the effect of *legality* (that is legal crossing and illegal crossing are compared only for pedestrians).

An example of the second type of restrictions (at the profile pairing level) is one that disallows two profiles that both feature *omission* to be presented in the same task. Not enforcing this restriction would result in a dilemma in which an AV has two decisions to make, these two decisions would result in different outcomes, yet both decisions are no action decisions (i.e. *omission*). One can easily see the value of enforcing such restriction. However, other similar restrictions that were enforced at the level of profiles pairing are probably less obvious. For example, an argument can be provided against pairing profiles that both require commission (though, it is less obvious); omission level profiles are needed in every task, because it should be always possible for the AV to do nothing. Given this, a restriction was put at the level of profile pairing, that is a profile with omission decision is always paired with a profile with commission decision, and vice versa. Similar pairing-level restrictions were also applied to other attributes. For example, while profiles with pedestrians can be paired with profiles with either pedestrians or passengers, two profiles with passengers cannot be pit against each other. The reason is simply because tasks that weigh passengers against another group of passengers would require adding another AV, a more complicated task that we would like to postpone to future work for its game-theoretic consideration. Note that when calculating the effect of *relation to AV* only tasks that pair pedestrians profiles with passengers profiles are considered; that is scenarios that weigh pedestrians to another group of pedestrians are excluded from analysis for this attribute. This level of restriction was also enforced beyond the first three attributes, and applied to the character types attribute; up until recently, males were always weighed against females, elderly always against young, fit always against large, high social value always against low social value, humans always against pets, and more characters always against fewer characters (utilitarian). Some of these restrictions are indeed crucial and removing them would result in unrealistic tasks such as the case for utilitarian scenarios; profiles with more characters can only be paired with profiles with fewer characters. In other character types, these restrictions are not needed but are still justifiable; for example in the case of age, one can take profiles with young character to mean younger characters instead. In this case, profiles with younger characters can only be paired with profiles with older characters (and the same for profiles with older characters). Similar justification can be made for fitness, and social value character types. This restriction however is not justified for the gender and species character types. One can imagine realistic scenarios in which all-male characters are weighed against all-male characters, or all-pets against all-pets. This type of restriction does not bias the estimator either, but it limits the interpretation of the estimated effect to scenarios that follow these restrictions. Recently, a small proportion of scenarios that weigh a group of characters against a similar group of characters, sampled from five out of the six character types above (utilitarianism is excluded) were added. However, they were not included in the main results to avoid inconsistent interpretations. One can expect that including these scenarios, and re-weighting to give them an equal representation in data would result in halving the effect sizes of the five character type attributes; increasing the effect sizes of interventionism, relation to AV, and legality; and having no effect on utilitarianism attribute. This would make for an unfair comparison between the nine attributes.

One final consideration is the use of non-uniform distributions for some of the treatment levels. An example of this is in *legality* attribute. For a simple analysis, one would assign each of the three levels: *no legality*, *legal crossing*, and *illegal crossing* with equal probability. However, we chose to present respondents with scenarios that have no legal considerations (*no legality*) more often than with scenarios that involve legal considerations. As such using the estimator above would provide a biased estimate of the marginal effect of other attributes. To see why, note how given that two attributes can "interact", and probability of levels for the first attribute are not uniform, then the marginal effect of the second attribute will be biased, or will have a different interpretation that is conditional on the used distributions (which do not reflect distributions in reality, and thus are not defensible). However, this can be fixed by weighting observation by the inverse probability of the corresponding levels of legality (i.e. inverse probability of treatment). Weighting observations here is done using the distribution chosen by design (theoretical probability of each subgroup), and it simply ensures that the regression function would calculate weighted means instead of pooled means. Alternatively, one can avoid weighting by using a modified version of the estimator above that would calculate the mean of each subgroup (each combination of attributes) and then would take the mean of means for each attribute level.

Results. We consider "forced choice" outcome in which respondents choose one of the two profiles they are shown in each task. Each profile describes a potential decision that can be made by an AV, and it would result in sacrificing a group of characters in order to spare another group of characters. Thus, each task describes a dilemma faced by an AV in which it has to make one of two decisions (two profiles). Respondents choose the decision (a profile described by four attributes) that they prefer for the AV to make over the other decision. The less a profile is chosen, the more the decision of sparing the characters in it is preferred by respondents.

The data is re-shaped so that each observation is a profile, and the outcome is a binary variable representing whether characters in this profile were spared by the respondent (spared here means the respondent chose for the AV to sacrifice the characters in the opposite profile).

Under the assumptions above, AMCEs were non-parametrically estimated using a simple linear regression of the binary choice variable of sparing on the dummy variable of the attribute with clustered standard errors. For *relation to AV* attribute, only scenarios that have pedestrians vs. passengers, and that have no traffic lights (no legal complications) were considered. For *law* attribute, only scenarios that have legal crossing vs. illegal crossing, and that have pedestrians vs. pedestrians were considered. For Social Status, scenarios that include at least one of male/female doctor, pregnant, or criminal were excluded, in order to have a cleaner interpretation (though including these scenarios yields a similar result and effect size). For all other attributes, all scenarios were considered, but observations were re-weighted, as mentioned above, using inverse probability of treatment for each subgroup. Coefficients of the treatment variables represent the value of interest, the AMCE for that treatment (attribute).

Figure S3 shows the AMCEs for the nine attributes. These AMCEs can be also meaningfully compared to each other. Coefficients and their standard errors are shown for each attribute level as compared to the baseline level of each attribute. The baseline levels of each attribute were chosen as to make all coefficient signals on the same side (all are positive). For each attribute, the AMCE (x-axis) represents the increase in probability of sparing a group of characters when the attribute value changes from a baseline value (values at the left e.g. sparing passengers) to the other value (values at the right e.g. sparing pedestrians). In other words, for each attribute/row, ΔP is the difference between the probability of sparing characters described by the attribute level on the righthand side and the probability of sparing characters described by the attribute level on the lefthand side (aggregated over other attributes). For example, in the case of *age* attribute, the chosen probability of sparing a group of young characters is 0.49 (SE = 0.0008) greater than the chosen probability of sparing a group of elderly characters, when an AV is to choose between sparing a group of elderly characters and a group of young characters. The same goes for the other attributes. In the case of *intervention* attribute, the probability of choosing inaction (to keep the AV on its track) is 0.06 (SE = 0.0004) higher than the preference for action (to swerve the AV off its track). As noted earlier, these interpretations are restricted to the cases where two different groups of characters are on each side of the dilemma.

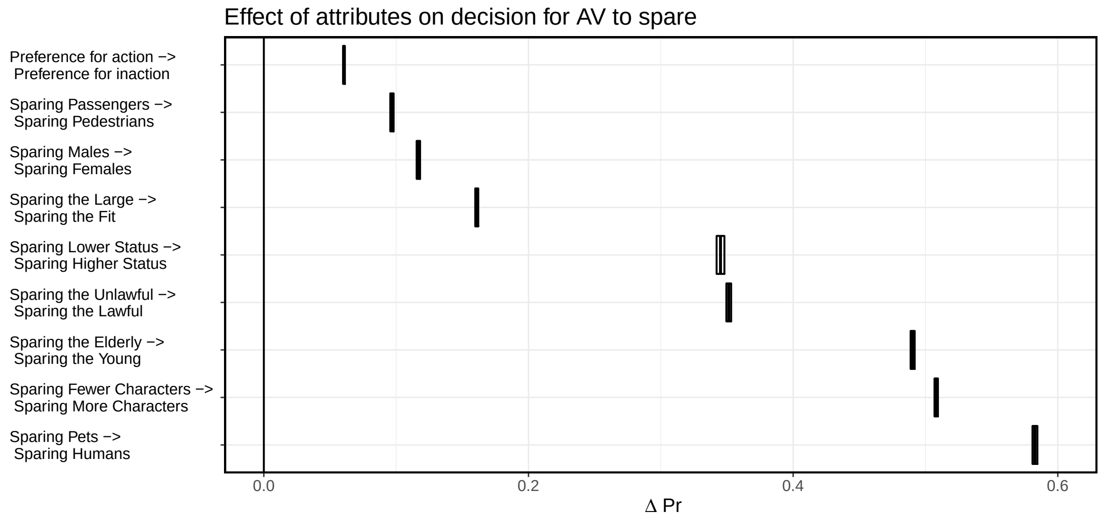
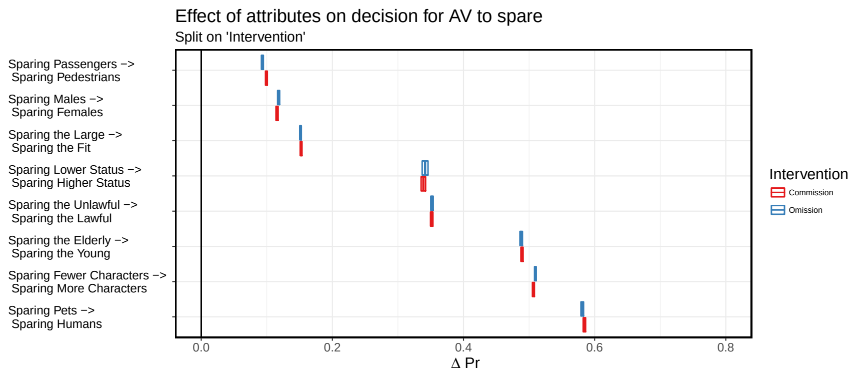


Figure S3: Average marginal causal effect (AMCE) of attributes in Moral Machine. Each row represents the difference between the probability of saving characters possessing the attribute on the right, and the probability of saving characters possessing the attribute on the left, aggregated over all other attributes. Estimates are the coefficients of a simple linear regression of the binary choice variable on the dummy variable of the attribute with clustered standard errors, and boxes show the 95% CI of the mean. Fig.2 (a) in the main manuscript is a more visually appealing version of this figure.

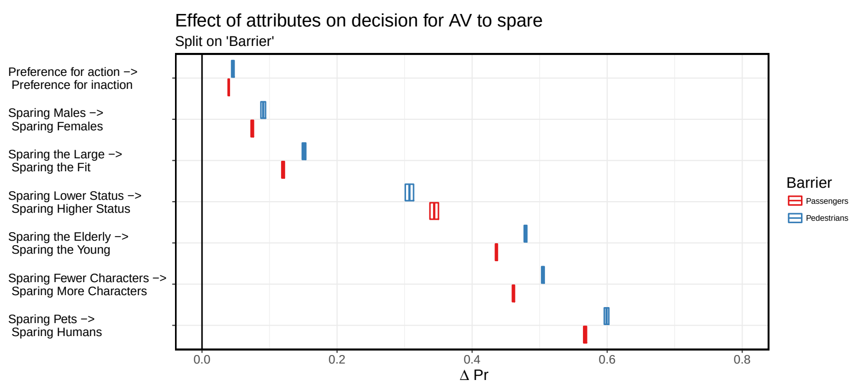
Next, we investigate interaction effects between attributes. This would be useful to analyze whether the causal effect of some attributes vary depending on the value of another attribute. For example, the causal effect of character type attributes may be different for passengers than for pedestrians. For this example, one would condition on whether characters are passengers or pedestrians, and calculate AMCEs for each of the six character type attributes, as shown in Figure S4 (b).

One can alternatively condition on interventionism and legality (as in Figure S4 (a), (c)).

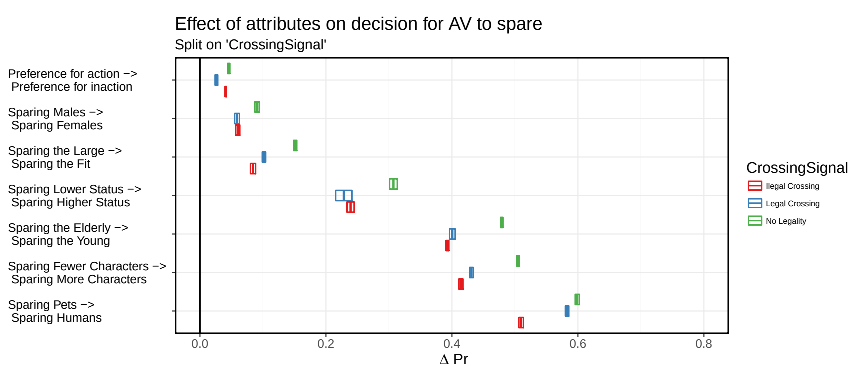
Another interesting type of interaction is between respondents' background and attributes. This would be useful to analyze whether the causal effect of some attributes vary depending on respondents' background characteristics. For example, the causal effect of gender or age attributes may very depending on the respondents' gender or age. For these examples, one would condition on whether respondents are males/females or on whether respondents are young/elderly, as shown in Extended Data Fig. 4 (a), (b). For example, $\Delta P|\texttt{female respondents}$ vs. $\Delta P|\texttt{male respondents}$. In the case of gender, the two subpopulations are males and females. As for the other respondents' background variables, a cutoff or a grouping of categories was used to define each subpopulation as the following: for age (15-75 years old), political views, and religious views, upper quartiles (Older, Progressive, Very Religious) and lower quartiles (Younger, Conservative, Not Religious) were considered. For income, ordinal categories corresponding to annual income that is below $\$10K$ were grouped together (Lower Income) and ordinal categories corresponding to annual income that is more than $\$80K$ were grouped together (Higher Income). For education, Vocational Training, High School degree or lower categories were grouped together (Less Educated); while only Graduate Degree category was considered as its own group (More Educated). Indicated values outside the above-mentioned values were discarded from this analysis. Note how in each subfigure in Extended Data Fig. 4, the AMCE of each of the nine attributes has a positive value for all subpopulations e.g. both males and females indicated preference for Sparing Females, but the latter group showed stronger preference.



(a) Split on Interventionism



(b) Split on Relation to AV



(c) Split on Legality

Figure S4: Average marginal causal effect (AMCE) of attributes in Moral Machine conditioned on (a) interventionism, (b) relation to AV, and (c) legality.

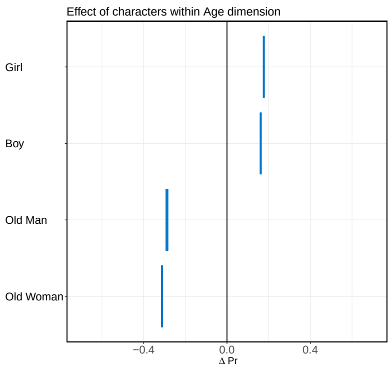
There are three points to make about this kind of analysis. First, there is the danger of contamination effects whereby whichever task was completed first–the ethical scenarios or the background characteristics survey–may have affected the results of the task completed subsequently. In Moral Machine, for reasons related to respondent fatigue and boredom (very crucial aspect of such a gamified survey), the survey always came after respondents completed a set of 13 scenarios. While we believe the risk of contamination is low, it is more likely in the case for political and religious views than for age, gender, income or education, so one must be careful when interpreting the interactions of the political and religious views. Second, the data used for this type of interaction is limited to respondents who took the optional survey, who might not be representative of the other survey non-takers. However, we show below that the AMCEs of these two groups (survey takers and survey non-takers) are not substantially different.

Third, by conditioning on each demographic attribute unilaterally, we neglect the fact that these attributes are correlated in our data e.g. our female respondents are more likely to be young. This could potentially mean that differences based on gender are driven by differences based on age. To address this, we include all six characteristic variables in regression-based estimators of each of the nine attributes. The dependent variable for each attribute is re-coded as to whether the respondent chose the preferred option or not. For example, in the case of Fitness the dependent variable becomes whether the respondent spared the fit or not (binary), while the treatment variable of "fitness" is removed from the regression. Furthermore, to account for the non-uniform treatments of Legality and Relation to AV (which was accounted for by re-weighting in the main effect), the treatment variables corresponding to Legality and Relation to AV (Structural Covariates) were included. Moreover, "Income" variable is turned from bracketed categorical variable into a continuous variable by choosing the midpoint of each bracket, while using $\$150K$ for the top bracket ($> \$100K$; calculated as the mean of an assumed Pareto Type I distribution with $\alpha = 3$, $x = 100K$). Then, it is standardized along with the other continuous predictor variables: Age, Political views, and Religiosity, while Education variable is re-coded into a binary variable: "Is college educated". To account for correlation of responses within a respondent, cluster-robust standard errors were computed. Finally, to account for multiple comparisons, more conservative significance cutoff threshold were used at: 0.01, 0.001 and 0.0001 (see Extended Data Fig. 3).

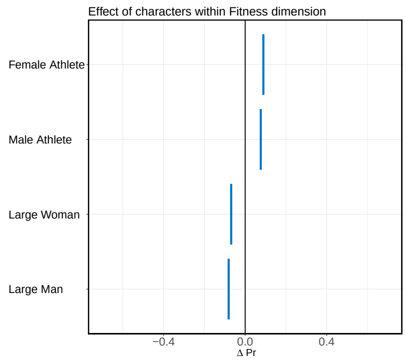
The scale of the website allowed for randomization over other elements, which helped strengthen the external validity of the results. Aside from the above-mentioned nine attributes, other attributes were also varied. These were the number of characters in the profile, the difference in number of characters between the paired profiles, and the characters themselves. This allows us to study the effect of these attributes, using the same tools. However, some restrictions were also enforced for these variables. As for the first two, their effect can be only studied within the "number of characters" dimension and the completely random scenarios. This is due to fixing the number of characters among the paired profiles in other character type attributes. Moreover, the characters that are used in the dilemmas spanned a diverse possible groups of characters. Out of 20 different characters, a subset of relevant characters were considered within each characters type attribute (as explained earlier in the previous subsection).

In order to identify causal effects of characters, we employed a similar approach to the above.

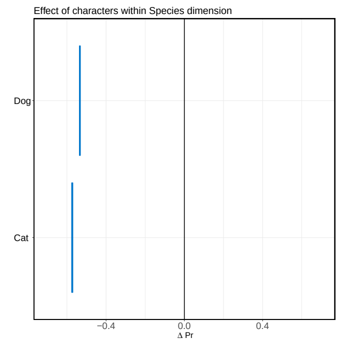
Consider the context-independent effect of character, which is shown in S5 (f). First, only scenarios that are generated "randomly" (i.e. other than the six dimensions), and that feature one character on one of the two sides are used for this calculation. Then, effects were non-parametrically estimated using a simple linear regression of the binary choice variable of sparing on the dummy variable of the character with clustered standard errors (while re-weighting, as before, using inverse probability of subgroups, that was chosen by design). The baseline case for each character is an adult man or an adult woman (i.e. dummy variable of a character is 1 for the character, and 0 for adult man/woman). So, Figure S5 (f) shows the effect of replacing an adult man/woman by each of the other characters, ordered from the most positive to the most negative. For each character/row, ΔP is the difference between probability of sparing this character (when it is alone) and the probability of sparing one adult man/woman (aggregated over attributes Interventionism, Law, and Relation to AV). For example, the probability of sparing a girl is 0.15 (SE = 0.003) higher than the probability of sparing an adult man/woman, while the probability of sparing a cat is 0.16 (SE = 0.003) lower than the probability of sparing an adult man/woman. Other figures are generated similarly after conditioning on an attribute (e.g. age, social status). For example, Figure S5 (a) shows that within the age attribute, replacing one neutral character (man or woman) with an elderly man decrease the probability of choice for sparing this character by 0.03 ($SE = 0.003$). For this example, these effects are conditional on the attribute (or context) of age. One can also show the change in probability as a result of replacing two and three (neutral characters).



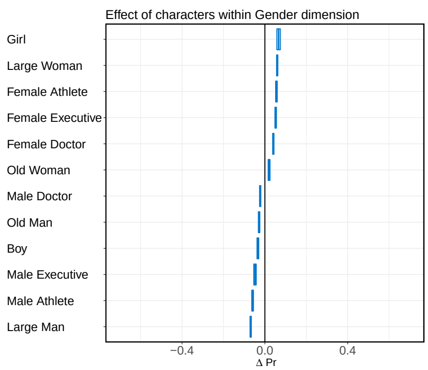
(a) Within Age



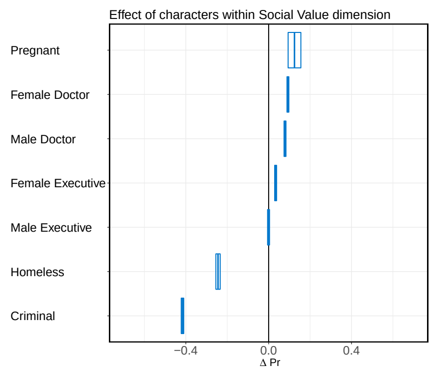
(b) Within Fitness



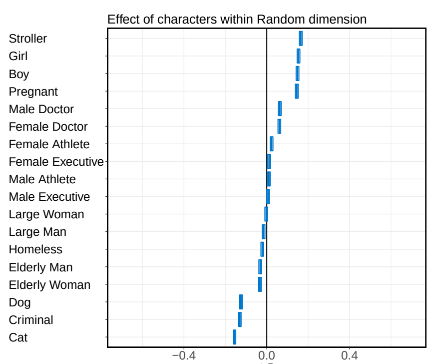
(c) Within Species



(d) Within Gender



(e) Within Social value



(f) Within Random

Figure S5: Characters effect within dimensions: (a) age, (b) fitness, (c) species, (d) gender, and (e) social value. (f) context-independent characters effect on the probability of choice for sparing this character. Fig.2 (b) in the main manuscript is a more visually appealing version of figure (f).

Robustness Checks. We now turn to checking the validity of the found results. We start with internal validity. For that we need to check that the above-mentioned assumptions hold. For assumption 1 (no carryover effect), this means that a respondent's choice of a profile over another in a task is the same regardless whether the respondent had seen other tasks before or not. This can be achieved by estimating the AMCEs for each attribute in 13 sub-samples: 1st task subsample, ..., 13th task subsample. We are able to perform this because the order of the different tasks is randomized (so each dimension and each combination of attributes can appear in any order). Extended Data Fig. 1 (a) shows the AMCEs for each of the nine attributes (ten levels) across the 13 task orders. One can see how the AMCEs estimates are very similar for each task order. Even when two AMCEs of an attribute are different for different task orders, the difference is small and both are in the same direction.

We now turn to assumption 2 (no profile order effect). This means that a respondent's choice of a profile over another is the same regardless of the order of the two profiles within the task. Similarly, this can be achieved by estimating the AMCEs for each attribute in two sub-samples: left-hand (default in mobile) profiles and right-hand (non-default in mobile) profiles. We are able to perform this because the order of the two profiles is randomized within a task (so each dimension and each combination of attributes can appear on the left or on the right of the screen). Extended Data Fig. 1 (b) shows the AMCEs for each of the nine attributes across the two orders. One can see how the AMCEs estimates are very similar for each profile order. Even when two AMCEs of an attribute are different for different profile orders, the difference is small and both are in the same direction.

As for assumption 3 (randomization of profiles). This means that attributes of each profile are randomly generated. This should be satisfied by design. However, one can still check for that, just to confirm that nothing wrong had happened during the randomization process. First, we note that the actual representation of subgroups in our data almost matches the designed (true) proportions of these subgroups. To check the extent of such mismatch, we compare the effects calculated by weighting using actual proportions to the effects calculated by weighting using theoretical (designed) proportions. Extended Data Fig. 1 (c) shows the AMCEs for each of the nine attributes across the two weighting schemes. One can see how the AMCEs estimates are very similar for each distribution (designed vs. actual).

We now turn to external validation. There are various factors to check. Two important factors relate to the interface effect. First, whether respondents recognized the characters and were able to differentiate between them can influence their judgment. It is possible that some characters are hard to recognize. A useful feature of the website, that would help testing this is the "show description" button. When this button is clicked, a description of the scenario and the involved characters is provided. This description is not provided by default to avoid cognitive load. Thus, similar to the check for profile order, we split data into two subsamples: 1) description is seen, and 2) description is not seen. Extended Data Fig. 2 (a) shows the AMCEs for each of the nine attributes across these two subsamples. Indeed, when respondents saw the description of the scenario, attributes had larger effect. This means that the reported effects are possibly underestimated.

The second important factor is the platform used. Mainly, two platforms were used by respondents: desktop and mobile (the third one is the tablet by a smaller fraction). These two platforms have slightly different interfaces, which can result in different assessment. For example, because of the small screen of the mobile, only one profile can be seen at a time, unlike the case with desktop in which respondents can see both profiles together (which makes it easier for comparison). Additionally, characters are smaller on mobile which might make it harder to recognize them. To check for any platform effect, we split data into two subsamples: 1) data collected from desktop, and 2) data collected from mobile. Extended Data Fig. 2 (b) shows the AMCEs for each of the nine attributes across these two subsamples (platforms). One can see some differences for some AMCEs. However, for every attribute, the difference between the two platforms is not very large and both values are on the same direction.

Another check we perform in regard to external validity is the difference between three different data sets. As explained earlier, we use for analysis the full data set. However, a case could be made against using all the collected data. After all, the full data set includes incomplete sessions, and it includes multiple sessions per same respondent. Despite accounting for within-user correlations, results could be biased to users who chose to take more sessions. Further, the inclusion of repeated sessions per user can also have the problem of adapting respondents, who upon seeing a summary of their results decided to make different decisions. For all these reasons, we constructed another data set that includes the first completed session by each user. This second data set includes equal representation of each user, and it does not include post-summary responses (i.e. responses to scenarios presented after seeing one's summary page), which deals with the above issues. However, it introduces another issue, that it excludes a possibly different group of respondents.

The third data set we consider is the responses in a sessions after which a survey was filled by the corresponding respondent. Recall that when analyzing the interaction between respondent's background attributes and profile attributes, we had to limit this analysis to a subset of the full data; a subset that includes only the observations for which respondents chose to take the survey. One can see how this subset of data could result in different results from the full data set. After all, the respondents contributing to this data are the ones who voluntarily chose to take the optional survey at the end. Those respondents can be fundamentally different from the remaining participants. Extended Data Fig. 2 (c) shows the AMCEs for each of the nine attributes across the three data sets. One can see some differences between the three data sets, but the direction and the magnitudes of the effects are in agreement.

Another point that can be made is about the representativeness of respondents. After all, respondents in our case are the ones who chose to visit the website and take the test. It is important first to note what kind of demographics these respondents represent. Extended Data Fig. 8 shows that most users are male, went through college, and are between their 20s and 30s. While this indicates that the users of *Moral Machine* are not equally representative of all demographics, it is important to note that this sample at least covers broad demographics. The unequal representation, however, could be taken to mean that it represents the population that uses the Internet, which includes, to be more specific, the tech-savvy users and AI/AV enthusiasts. These are the individuals that are the most interested in the technology of the AVs, and are thus the most likely to have formed an opinion about this technology, and most likely to adopt this technology in the future. Furthermore, compared to data collected from lab-based experiments, online experiments, and field experiments for research conducted in psychology, cognitive science, and behavioral economics, this respondent sample falls on the less biased side of the spectrum [8].

In addition, the problem of disproportionate representation of demographics can be partly dealt with in different ways, if one is willing to make further assumptions. For example, given that 70% of participants are males, the reported effect sizes are skewed towards male preferences, which differ from female preferences in some aspects (as shown in Extended Data Fig. 4 (b)). One way to deal with this follows from the above argument to treat the sample of respondents as a representative sample of the “population of interest” (e.g. tech-savy, AI/AV enthusiasts). This entails conceding that this population is comprised of 28% females, the truth of which is hard to validate. Another way to deal with it, is to re-weight the observations so that female respondents have 50% representation in the data. This can be done easily by re-weighting using inverse probability of each group of respondents (males vs. females). In fact the result of doing so can be inferred from Extended Data Fig. 4 (b). The effects would be pulled in the direction of female preferences half way. One can simply look at that Extended Data Fig. 4 to know how the effects would become if we try to balance the representation of high-income vs. low-income, highly-educated vs. less-educated, etc. However, one has to note that re-weighting is not perfect either, and it relies on the assumption that, for example, females who chose to take the Moral Machine test are very similar to females who did not choose to do so. This, of course, might not be the case. The same holds for other demographic attributes. Furthermore, it is very likely that our sample is not representative with respect to the combination of the observed demographic variables (e.g. disproportionate representation of young non-religious males). In the next subsection, we provide a post-stratification analysis for the respondents in the US only (who took our survey). However, given the problem mentioned above about post-stratification, and the difficulty of obtaining data and performing same analysis for all countries, we opt for treating the respondents sample as representative of the "population of interest" in the main paper.

A similar but perhaps more interesting case could be made about the country of respondents. The reported results are conditional on the proportion of respondents visiting from each country. As such, some countries (e.g. US, Russia, Canada, Germany, France) have more representation than others. The same choices as above hold here; either accept these percentages as representative of the "population of interest", or re-weight relying on the above assumption of representative samples from each country (e.g., Indians who took the test are similar to Indians who did not). One might object here that not only is this a strong assumption, but also that re-weighing in this regard would give small countries like Luxembourg an equal weight to large populous countries. Another possibility is to re-weight so that, instead of countries having equal representation in the data, that countries have weights proportional to their actual population proportion. This would give higher weight to countries with high population (e.g. China, India, Russia, US, etc.), which some would find problematic for other reasons. So, each of the three possibilities we just presented is both justifiable and open to criticism.

Post-Stratification. To analyze the extent of performing post-stratification, we focus on US respondents only. From Moral Machine data we extract data for respondents from US who took the survey. Discarded from this data are respondents who indicated age outside (15-75) years old, and respondents who indicated “vocational training” or “others” as their educational level. Income levels are shrinked from the previous nine levels, and Age levels are created using equal-width discretization. For external “representative” dataset, we use population-level data from US Census Bureau’s 2012–2016 American Community Survey (ACS) 5-year Public Use Microdata Sample (PUMS), accessed through American FactFinder [9]. This dataset has $15M$ records. Extended Data Fig. 9 (a)-(d) shows the proportion of the two US population samples (MM vs. ACS) across levels of each of the four attributes: Age (6 levels), Gender (2 levels), Income (5 levels), and Education (5 levels). One can see that MM US-sample has an over-representation from Male population and from young population, as compared to the ACS US-sample.

After that, we calculate the number of people in each cell of Age (6 levels) x Gender (2 levels) x Income (5 levels) x Education (5 levels) = 300 levels. The ACS dataset has data that cover all 300 levels. However, MM Survey in US data has only data that cover 280 cells. The remaining 20 cells cover less than 4% of the ACS data (those cells were discarded). The levels above are chosen minimally to cover as many cells as possible.

In order to perform post-stratification, we first calculate the effect size of the nine attributes for the above MM dataset. Then, we follow the following procedure : 1) calculate the percentage of respondents in each of the 280 cells in MM data and in ACS data, 2) divide the numbers for ACS by the numbers for MM, 3) use the answers to re-weight instances in the regression function.

Extended Data Fig. 9 (e) shows the comparison between pre- and post-stratification. It shows that, except for Sparing Pedestrians, re-weighting does not make much difference for effect sizes. In fact, for the attributes that have the strongest effects (the last five rows in Extended Data Fig. 9 (e)), there is no change in order and there are only small changes in estimates. This analysis is of course not sufficient to establish what the effect would be had we had representative samples. First, as mentioned before, post-stratification is not perfect because it assumes that the participating subjects in a cell are similar to non-participating subjects from that cell, which is a strong assumption. Second, for cross-country results, one would need to repeat the same analysis for each of the remaining 129 countries.

4 Identifying Cross-Country Variations

In this section, we describe our approach to identifying cross-cultural differences and similarities among countries in ethical preferences observed in Moral Machine.

In addition to the response data, Moral Machine captures approximate geo-location information of the respondents through the IP addresses of the computers and mobile devices that the respondents used to access the website. Using the geo-location information, we identified the country of residence of the respondents at the time when the respondent engaged in judgement mode of Moral Machine. With the knowledge of country of residence, we divided the judgments based on the respondent's country of residence, which gave us the information to study the cultural differences in preferences among countries represented in Moral Machine.

In order to maintain consistency and high fidelity of the AMCE values, we excluded judgments from countries that had fewer than 100 respondents; as a result, we narrowed down the analysis to 130 countries. Using the responses collected from each country, we computed the nine ACME values for each country using the methodology described in Sec. 3, and in order to compare the distributions of the effect sizes in a comparable scale, we converted the nine ACME values into nine z-scores ($\frac{x-\mu}{\sigma}$) for each country.

Figure S6 shows a matrix plot for pairwise correlation and scatter plots of the nine attributes at the level of countries (130 countries). Remarkably, the nine attributes show only few pairwise notable correlations: (More, Young), (Young, High), (High, Inaction), and (High, Fit).

Hierarchical/Agglomerative Clustering While there is panoply of clustering algorithms in machine learning literature such as K-means, Gaussian Mixture Models, and DBSCAN [10], most clustering algorithms do not provide results that enable us to analyze structural patterns within a cluster. A methodology that uncovers structural patterns within clusters is a general family of clustering algorithms called *hierarchical clustering* algorithms. Hierarchical clustering algorithms build nested clusters by merging or splitting the clusters successively in multiple iteration and represent the nested structure as a dendrogram tree. The root of the tree encompasses all of the data, and the leaves of the tree represents a single data point as it's own cluster. Agglomerative algorithm performs hierarchical clustering via a bottom up approach wherein each data point starts as its own cluster. The algorithm builds nested clusters by merging or splitting the clusters successively in multiple iterations and represents the nested structure as a dendrogram tree. The root of the tree holds all data points, and the individual leaf of the tree represents single data point as its own cluster.

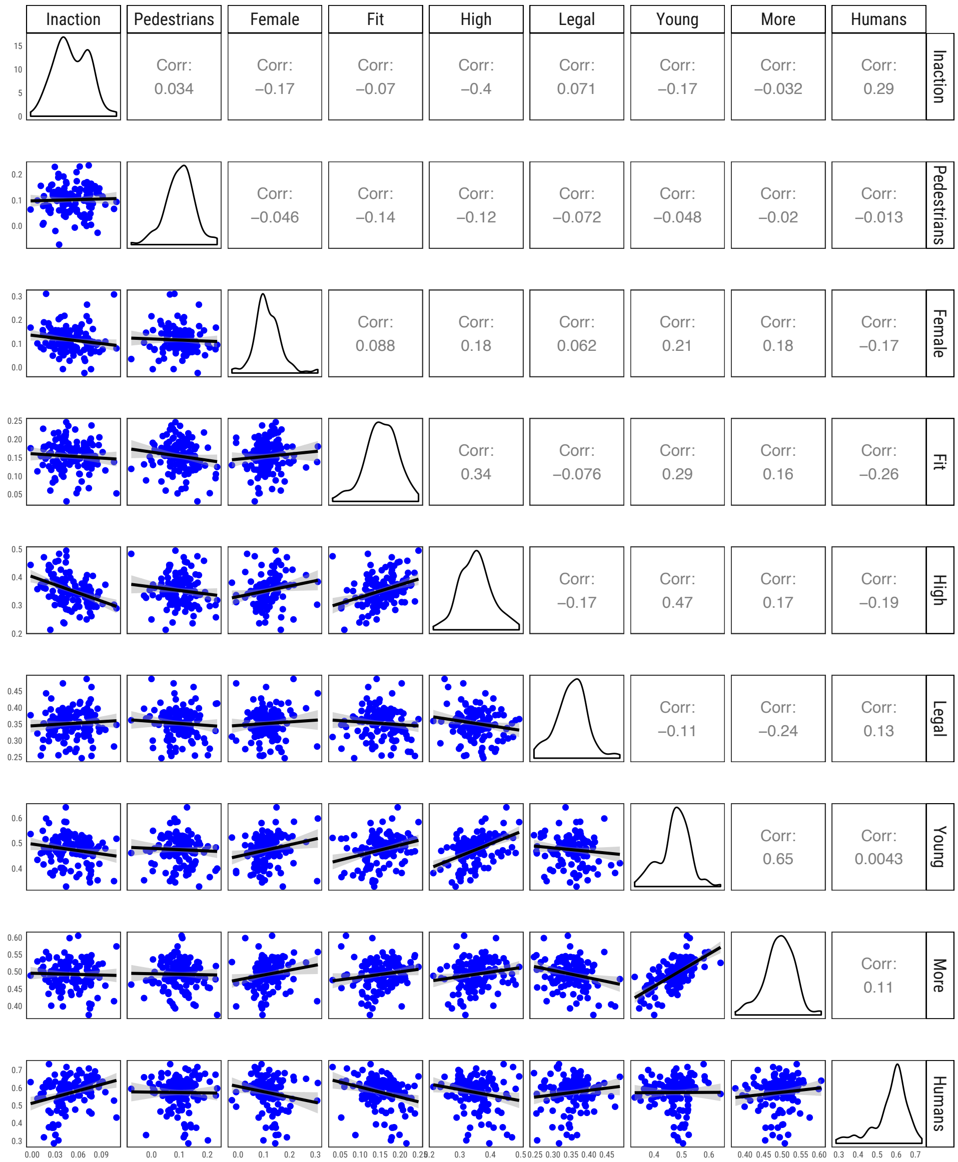


Figure S6: **Pairwise correlations of the nine attributes at the level of country.** One can see that attributes have few pairwise correlations, with the exception of some interesting ones like the one between sparing more and sparing the young.

Adopting notations from [11], we define a data set X as a set of N data points represented as vectors of F dimensions. A clustering of X is a set of disjoint clusters that partitions X into K

groups $C = \{c_1, c_2, ..., c_K\}$ where $\cup_{c_k \in C} c_k = X$ and $c_l \cap c_k = \emptyset, \forall k \neq l$.

In the first iteration, $\binom{N}{2}$ distance values between base N clusters are computed using a distance metrics. We used Euclidean distance for our analysis. After the distance values have been evaluated, a linkage criteria is used to determine two clusters c_i and c_j to combine to form a new cluster c_k. There are many linkage criterions and depending on the choice of linkage criterion, agglomerative hierarchical clustering algorithm can yield different dendrogram structures and clustering outcomes. We explored the following three popular linkage criterions:

- Ward variance minimization (Ward)

$$d(c_k, c_l) = \sqrt{\frac{|c_l| + |c_i|}{N} d(c_l, c_i)^2 + \frac{|c_l| + |c_j|}{N} d(c_l, c_j)^2 - \frac{|c_l|}{T} d(c_i, c_j)^2} \tag{1}$$

  where c_k is a newly formed clustered consisting of clusters c_i and c_j and c_l is an unused cluster.

- Complete or Maximum linkage (Complete)

$$d(c_k, c_l) = \max_{c_i \in c_k} \max_{c_j \in c_l} \left\{ d(c_i, c_j) \right\} \tag{2}$$

- Average linkage (Average)

$$d(c_k, c_l) = \sum_{c_i \in c_k} \sum_{c_j \in c_l} \frac{d(c_i, c_j)}{|c_k| * |c_l|} \tag{3}$$

Validation. In machine learning literature, a substantial volume of research exists in theoretical guarantees and empirical evaluations of supervised machine learning algorithms. In contrast, research in validation metrics of unsupervised machine learning algorithms, including clustering algorithms, is a relatively novel and open-area of research. Nevertheless, machine learning researchers have introduced several metrics to evaluate clustering algorithms, and the vast majority of the metrics can be classified into two categories: internal and external metrics. For both internal metrics, higher index value indicates "better" fit of partition to the data.

Internal Validation. Internal metrics are values derived from the data itself to measure the goodness of clusters by computing *compactness* and *separation* of the clusters. Here, we use the internal metrics to compare the three distance measures of clusters and select the best viable distance metric to study cross-cultural differences in AMCE values. Of the numerous metrics in literature, we utilized two well-established internal metrics:

- Calinski-Harabasz Index [12] as defined by

$$CH(C) = \frac{N-K}{K-1} \frac{\sum_{c_k \in C} |c_k| d(\bar{c}_k, \bar{X})}{\sum_{c_k \in C} \sum_{x_i \in c_k} ||x_i, \bar{c}_k||_2)} \tag{4}$$

  which measures cohesion based on the distances from the points in a cluster to its centroid $\bar{c}_k = \frac{1}{|c_k|} \sum_{x_i \in c_k} x_i$. The separation is measured as the distance from the centroids to the global centroid $\bar{X} = \frac{1}{N} \sum_{x_i \in X} x_i$.

- Silhouette Index [13] as defined by

$$Sil(C) = \frac{1}{N} \sum_{c_k \in C} \sum_{x_i \in c_k} \frac{b(x_i, c_k) - a(x_i, c_k)}{\max\big(a(x_i, c_k), b(x_i, c_k)\big)} \tag{5}$$

  where

$$a(x_i, c_k) = \frac{1}{|c_k|} \sum_{x_j \in c_k} ||x_i - x_j||_2 \tag{6}$$

$$b(x_i, c_k) = \min_{c_l \in C c_k} \left\{ \frac{1}{|c_l|} \sum_{x_j \in c_l} ||x_i - x_j||_2 \right\} \tag{7}$$

which measures cohesion as the distance between all points in the same cluster and separation as the distance to the nearest neighbor.

We tested the performances of the three linkage criterions by computing the two internal metrics from the outputs of the algorithm across increasing number of clusters $|C|$. By increasing the number of clusters and testing the algorithms fit independently, we measured consistency in the performances of the algorithms as the algorithm explore deeper levels of the dendrogram hierarchy. We also included K-means algorithm as a benchmark in this experiment.

As measured by Calinski-Harabasz Index, the Ward variance minimization criterion outperforms Complete and Average methods where $|C| < 20$; albeit, it performs relatively poor compared to the K-means. However, as the cluster size increases, all four algorithms converge toward the same index value (Extended Data Fig. 6 (a)).

As measured by Silouette Index, Average criterion outperforms other algorithms where $|C| < 10$; however, as the cluster size increase, Ward variance minimization criterion outperforms other methods until the cluster number is large enough that all four methods converges to the same value (Extended Data Fig. 6 (b)).

This validation process using the two internal metrics reveals that Ward variance minimization criterion yields partitions that are relatively stronger fit on the country level AMCE values. Henceforth, we applied Ward variance minimization criterion in our hierarchical clustering analysis in the following section.

External Validation. In contrast to internal metrics, external metrics utilize information not available in the data source such as clustering output performed by human experts or broadly agreed on clustering output as labels (i.e. ground truth) to measure *goodness of fit* of clustering algorithm. Here, we define the number of clusters in the hierarchical clustering algorithm to nine clusters, and compared the distribution of countries (see Table S1) in the nine clusters against nine cultural groups of Inglehart-Welzel (IW) cultural map, which are (1) South Asia, (2) Protestant, (3) Orthodox, (4) Latin America, (5) Islamic, (6) English, (7) Confucian, (8) Catholic, and (9) Baltic [14].

We use two external metrics, *Purity* and *Maximum Matching* as measures of goodness of fit. Purity quantifies the extent that cluster i contains points (i.e. countries) only from one partition in the ground truth clusters, and it is computed as

$$purity = \frac{1}{N} \sum_{i=1}^{K} \max_{j \in (1,\ldots,K)} n_{ij} \tag{8}$$

where K is the number of clusters/cultural groups, N is the number of countries, and n_{ij} is the total number of countries that are both in cluster i and in cultural group j.

Purity permits multiple clusters to correspond to one partition in the ground truth. Ideally, one would like to evaluate the quality of a 1-to-1 correspondence between clusters and cultural groups. For that, *maximum matching* computes the proportion of mutual countries in 1-to-1 cluster-group pairs.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Switzerland	Italy	United Kingdom	Poland	Cambodia
Germany	Bulgaria	Austria	Latvia	Japan
Norway	Croatia	New Zealand	Slovenia	Macao
Denmark	Romania	Ireland	Ukraine	China
Netherlands	Estonia	United States	Russia	South Korea
Finland	Serbia	Canada	Belarus	Taiwan
Luxembourg	Montenegro	South Africa	Moldova	Thailand
Austria	Belgium	Lithuania	Georgia	Kuwait
Iceland	Spain	Vietnam	Kazakhstan	Saudi Arabia
Sweden	Portugal	Tunisia	Brazil	Hong Kong
Cyprus	Greece	Qatar		Indonesia
Sri Lanka	Bosnia & Herzegovina	Albania		Malaysia
Singapore		Trinidad & Tabago		
		Jamaica		
		Iraq		
		Bangladesh		

Cluster 6	Cluster 7	Cluster 8	Cluster 9	
Iran	New Caledonia	Azerbaijan	Kenya	
Nepal	Reunion	Turkey		
Pakistan	Malta	Peru		
Jordan	Mongolia	Argentina		
Palestinian Territory	Algeria	Uruguay		
Armenia	Morocco	Bolivia		
Macedonia	Dominican Republic	Ecuador		
India	France	Columbia		
Mauritius	Czech Republic	Venezuela		
United Arab Emirates	Hungary	Honduras		
Egypt	Slovakia	El Salvador		
Lebanon		Panama		
		Philippines		
		Guatemala		
		Paraguay		
		Chile		
		Puerto Rico		
		Costa Rica		
		Mexico		

Table S1: **Partition of countries into nine clusters found via hierarchical clustering algorithm.** We leave out countries such as Nigeria and Israel that are not classified in the Inglehart-Welzel (IW) cultural map [14].

In order to evaluate the quality of the clusters we found in Table S1 and how closely they can be matched to Inglehart-Welzel (IW) cultural groups, we consider the set of our clusters and the cultural groups as the vertices of a graph G. Here, *maximum matching* refers to the quality of our clusters, and it is the value of the maximum weighted bipartite matching, where an edge between a cluster and a group is weighted by the number of shared countries between them.

Formally, let $\mathcal{M}$ be the set of all possible perfect matchings of G. A matching $M \in \mathcal{M}$ is a set of edges. Let $n(e_{ij})$ be a function that computes the weight of each edge e_{ij} between cluster i and cultural group j. *Maximum matching* is given by:

$$maxmatching = \frac{1}{N} \max_{M \in \mathcal{M}} \sum_{e_{ij} \in M} n(e_{ij}) \tag{9}$$

Application of hierarchical clustering algorithm to AMCE values yields purity value of 0.5888 and maximum matching value of 0.5421. In addition, in order to evaluate hierarchical clustering algorithm, we tested the algorithm's purity and maximum matching metrics against those of random clustering assignments. We generated 1000 randomly assigned clusters and computed purity and maximum matching values. Extended Data Fig. 6 shows the distribution of (c) purity and (d) maximum matching values of the randomly assigned clusters. The red dotted lines mark the purity and max-matching values of the outcome of hierarchical clustering algorithm. The striking differences in the values from hierarchical clustering algorithm against those from random clustering assignment suggests that hierarchical clustering algorithm yields robust clustering outcomes.

Furthermore, they indicate that hierarchical clustering algorithm using AMCE values in Moral Machine yields a clustering pattern consistent with broadly accepted understanding of global cultural groups.

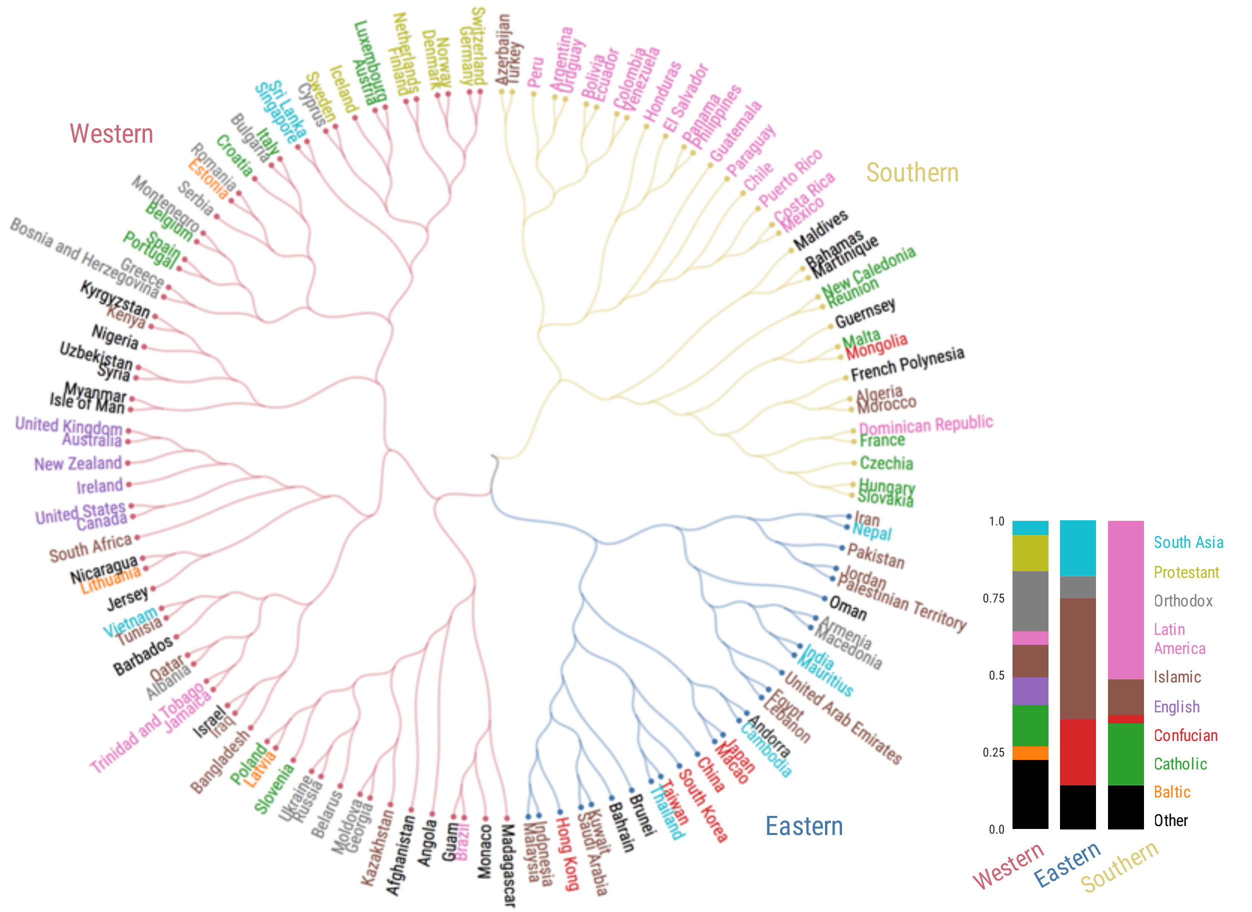


Figure S7: **Hierarchical Cluster of Countries based on average marginal causal effect.** One hundred thirty countries with at least 100 user responses are selected. Three colors of the dendrogram branches represent three large clusters – Western, Eastern, and Southern. Names of the countries are colored according to Inglehart-Welzel cultural map [14]. This is Fig.3 (a) from the main manuscript, copied here for convenience.

Results The structure of the dendrogram (See Figure S7) reveals patterns that are consistent with our existing views about cultural and geographical groupings among countries around the world. At the highest level, the dendrogram has three major clusters, which we label Western, Eastern, and Southern. We use the labels based on the overall pattern of countries represented in each cluster. For instance, the Western cluster contains many European and North American countries whereas the Eastern cluster contains the countries in the Middle-East and the Far East. The Southern cluster consists of many countries in South and Central 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. These patterns in the clusters suggests that the judgment data in Moral Machine has captured moral preferences consistent in geographic and cultural clusters of countries around the world.

Aggregating the AMCE values of the countries into three large clusters show striking differences in the distribution of preferences between the clusters along the nine dimensions (See Figure S8). For instance, countries in the Western cluster show stronger preference for inaction in the Intervention dimension compared to those of the other clusters whereas the Southern cluster exhibits stronger preference for sparing female characters compared to the other cluster. On the other hand, all three clusters share similar distributions on the dimension about relation to AV suggesting that there are certain moral dimensions that most cultures concur.

To further analyze the relationships between countries based on the moral preferences re-

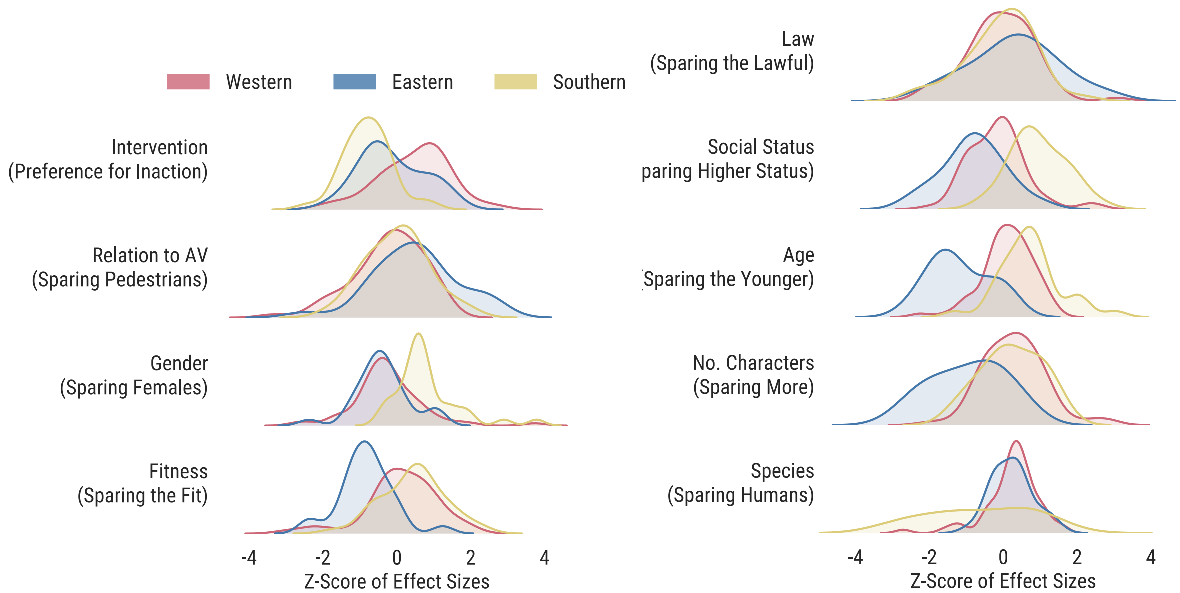
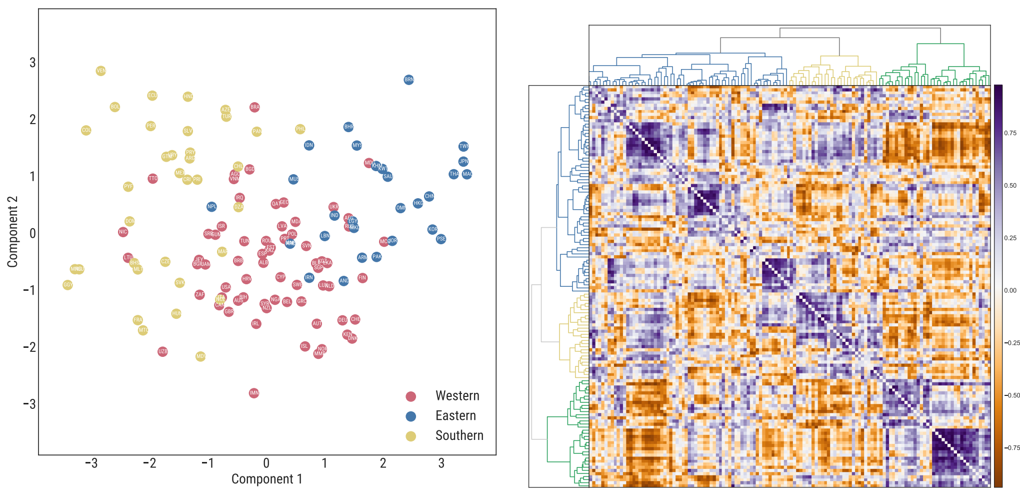


Figure S8: **Density plots of the AMCE z-scores show differences in distributions of the effect sizes between the three large clusters.** Higher z-scores indicate stronger preference for the default choice; for instance, the high z-score distribution in the Southern cluster along the Gender dimension reveals that the countries in this cluster have stronger preference for sparing female characters.

vealed in Moral Machine, we conducted dimensionality reduction using Principle Component Analysis (PCA) on the original nine AMCE values (Figure S9(a)). We transformed the data that consisted of the original nine AMCE values into two dimensional values. Distribution of the countries along the two new basis reveals that the three large clusters divide the countries into three cluster that are consistent with variance maximizing dimensions.

Pearson correlation values in AMCE z-scores between countries (Figure S9(b)) show consistent pattern that matches the structural patterns in the dendrogram. Countries that belong to the



(a) (b)

Figure S9: **Principle component analysis and Pearson correlation matrix.** (a) Scatter plot of countries along the first two principle components after performing dimensionality reduction using Principle Component Analysis. (b) Pearson correlation matrix of the nine AMCE values and dendrogram from hierarchical clustering.

same cluster show generally higher degree of correlation validating the outcome of the hierarchical clustering algorithm.

Neutralizing language effect The reader might note that the clusters might have formed as a result of the language used (recall that the website is available in ten languages). An example for this with respect to the French language can be seen in Figure 3 (a) from the main manuscript. Some countries in the Southern cluster are identified as former French colonies. There can be various cultural or otherwise reasons for these countries having closer responses to each other. Nevertheless, one might suspect that the use of the French version of the website had a special effect on responses from those countries. The French language seems to be the only clear example for this case. For example, Spain is in a different cluster from the Southern cluster which includes many Spanish-speaking countries. One needs to note here that the scenarios on Moral Machine are heavily pictorial and the only use of language is in the optional textual description which is not shown by default to respondents. In fact, when we exclude responses for which the description was seen by respondents (that is, consider only scenarios where respondents depended only on pictures to make their decisions) we find that the same clusters persist with some minimal changes, especially in regard to the French-related countries (see Extended Data Fig. 5).

5 Explaining Cross-Country Variations

Societies vary widely along many dimensions: the proper functioning of large-scale institutions like democracy, the consistency of the rule of law, corruption, and GDP per capita. One of the important dimensions is one that relates to cultural differences captured by Hofstede's Individualism-collectivism dimension. In this section, we establish that cross-societal variation in these dimensions are highly correlated of Moral Machine decisions. In other words, inter-societal differences in Moral Machine behavior vary systematically with underlying societal characteristics, as opposed to being independent. While we document a systematic variation, we do not attempt to pin down the ultimate reasons for the inter-societal variation. A growing body of literature suggests, however, that they are deeply rooted in societies longer-term history (see [15] for an overview).

The systematic variation highlights that societies attribute different weights to moral dilemmas arising through technological innovations. This has important implications when designing and implementing regulations. On the one hand, objectives and implementing procedures are highly contingent on societal values and norms. On the other hand, even if there was such a thing as a universally morally right behavior based on a normative theory, from which formal regulations could be inspired, they would be only effective and followed if they are perceived to be legitimate and palatable to the local population. The variation we document suggests that universal regulations face an uphill battle and that regulations will need to take societal differences into account. We document cultural differences to the degree that individuals value rules and rule-following.

Individuals in the Moral Machine make choices in trade-offs between sparing rule-followers

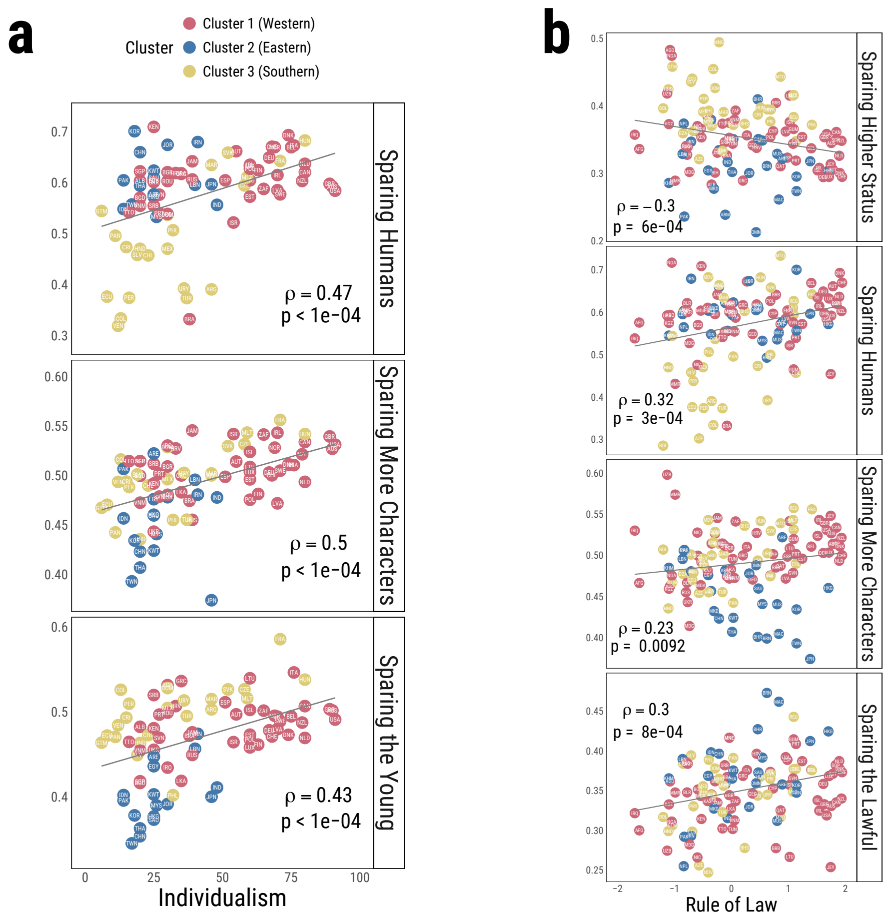


Figure S10: (a) Individualism and Moral Machine Behavior. Top panel reveals the association between individualism and the preference for inaction, middle panel the association with the probability to spare more (vs fewer) people, and bottom panel the probability to spare the young. (b) Rule of Law and Moral Machine Behavior. Top panel reveals the association between societies rule of law and the probability to spare the higher status, second panel the association with the probability to spare humans, third panel the association with the probability to spare more characters, and the last panel the association with the probability to spare rule-followers. Colors follow clusters colors from Figure 3 in the manuscript.

vs rule-flouters. A higher propensity to spare rule-followers reflects a greater respect (or perceived legitimacy) of formal rules. Consequently, we expect participants from societies that score higher on the governance indicator rule of law–which measures perceptions of the quality of ones society's body of legal rules and institutions–exhibit a higher propensity to spare rule-followers. On the one hand, individuals growing up in societies with well-functioning institutions learn to trust institutions and to internalize the norm so that deviations do not pay off for them. On the other hand, weak institutions can be an expression of lower societal inclination to follow formal rules [16]. Figure S10 (b) indeed reveals a positive association: a one-standard deviation increase in the rule of law (based on the world banks governance indicator) is associated with an 3.4 percentage point increase in sparing rule-followers over rule-flouters. Given the well-established importance of well-functioning institutions for economic prosperity, log GDP per capita measure is likewise highly predictive of sparing rule-followers (see Figure 4 (b) in the main text).

Systematic variation also exists regarding other choices. Consistent with the ideal that emphasizes equality before the law and human rights, individuals from high rule of law countries are also more likely to spare more characters, more likely to favor humans over non-humans, and less likely to favor higher-status (over lower-status) characters. Higher societal-level inequality (as measured by the Gini-coefficient) is likewise associated with a higher propensity of sparing higher-status (over lower-status) individuals. This may be due to the fact that people with higher status – and more likely to be early adopters of driverless cars – are overrepresented in our sample, especially in countries with high income inequality. This points to a potential moral hazard in these countries, whereby driverless cars would protect their wealthy owners at the expense of others (see

Extended Data Fig. 7 for the simple and linear regression models including our key cross-national predictor variables.)

The difference between "individualistic" cultures (which emphasize personal freedom and achievement) and "collectivistic" cultures (which stress embeddedness into a larger group) has emerged as a key and frequently discussed distinction in cross-cultural research [17]. The distinction has been used to explain differences in institutional quality and economic prosperity [18] [19], and is likely deeply rooted in societies kin-network structures (Schulz, Bahrami-Rad, Beauchamp, and Henrich in prep). Tight kin-network structures are not only negatively associated with individualism but can also explain modern day institutional failure [20]. Figure S10 (a) demonstrates that individualism is highly predictive of Moral Machine choices. Consistent with the emphasis on the individual, the probability of sparing more (vs fewer) people increases in individualistic countries, and the value assigned to human (vs non-human) lives is likewise higher. Meanwhile, the often very hierarchical structure of collectivist societies, with its emphasis on conformity and obedience towards older relatives, is reflected in the relatively lower likelihood of sparing younger people in the Moral Machine decisions made by those in more collectivist countries.

Overall, this demonstrates that societal indicators are predictive of choices on Moral Machine. The cluster analysis has already demonstrated that the range of participating societies exhibit clustering in these choices. Now, we go a step further by investigating cultural transmission. This is an important factor in explaining cultural similarities among societies. Culture is transmitted vertically from parent to child, as well as horizontally across populations. Work by Spolaore and Wacziarg [21] and Muthukrishna et al. [22] has demonstrated that genetic relatedness at the societal level is significantly associated with cultural similarities. The idea is that (i) populations that are more closely genetically related shared a longer common ancestry and thus will have had less time to diverge from each other on culturally transmitted traits and (ii) genetically more closely related populations have a higher propensity to adopt novel norms and values from each other. Genetic relatedness (or distance) measures differences in gene distributions across populations. It is based on neutral genes that change randomly over time and do not impact behavior. The measure thus reflects common ancestry (or the time elapsed when different populations were separated). Ancestrally closer populations are genetically more similar, share a longer common history and thus face fewer barriers to transmission of culture.

To investigate whether ancestrally closer societies also behave similarly in the Moral Machine we correlated genetic distance to the US (based on [23]) to Moral Machine distance to the US. Moral Machine behavioral distance (see Figure S11 (a)) is computed using Euclidean distance of the nine attribute values of each country from those of the US. Figure S11 (b) reveals a high correlation. To rule out the possibility that correlation is trivially driven by geographic proximity (ancestral closer population residing closer together), we ran a simple OLS regression analysis controlling for geodesic distance, and demonstrate that the relation holds (Table 1, Column 1). This suggests that common ancestry, which is determined in the distant past when human's started to emigrate from Africa, and Moral Machine behavior are closely linked. Societies that are genetically more related also behave in a similar way.

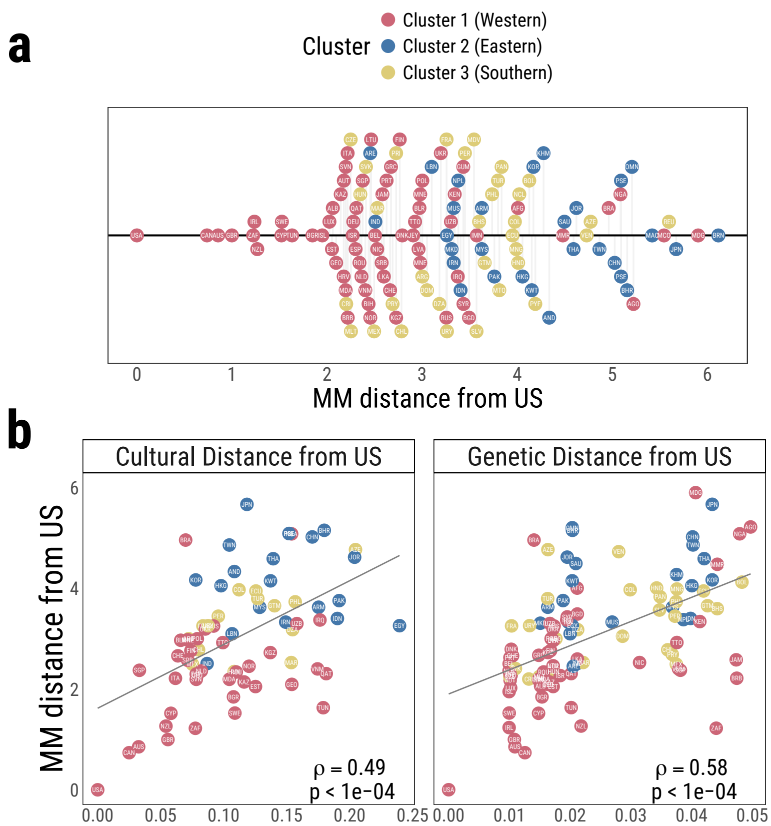


Figure S11: Moral Machine scale of AI Ethics distance. (a) Distance between each country and US. Vertical positioning (height and signal) has no indication beyond visual clarity. (b) Association between MM Distance and cultural distance (left panel, based on the WVS) and genetic distance (right panel). Each distance is measured to US. Colors follow clusters' colors from Figure 3 in the manuscript.

Using a measure of cultural distance based on Muthukrishna et al. [22] likewise reveals a highly significant association. The cultural distance indicator of Muthukrishna et al. is based on a synthesize of a large set of world value survey questions. Resting on this large set of questions it draws a comprehensive picture of cultural distance. The association between the Moral Machine distance and cultural distance again demonstrates that cross-societal differences in Moral Machine behavior are systematic (and likewise not trivially explained by geodesic distance, Table S2, Column 2).

	MM distance from US	
	(1)	(2)
Genetic distance	26.79**	
	(10.991)	
Cultural distance		9.32***
		(2.611)
Geodesic distance	-0.00	-0.00
	(0.000)	(0.000)
Constant	2.82***	2.05***
	(0.282)	(0.382)
N	96	65
R^2	0.063	0.183

Table S2: **Country-level OLS regressions of Moral Machine distance from the US on Genetic distance from the US (Column 1) and cultural distance from the US (Column 2).** The regression controls for geodesic distance (in km) from the US. Robust standard errors are reported in parentheses. Asterisks refer to the following significance levels: $* \, p < 0.10$, $** \; p < 0.05$, $* * * \, p < 0.01$

Indicators:

- Rule of law: This indicator is one of the World Bank's governance indicators (see [24]). Rule of law captures perceptions of the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, property rights, the police, and the courts, as well as the likelihood of crime and violence.

- Individualism: This indicator is based on Hofstede [25] (retrieved from http://geert-hofstede.com/, accessed 28.10.2015). According to Hofstede, individualism (vs collectivism) captures the following underlying concept: The high side of this dimension, called individualism, can be defined as a preference for a loosely-knit social framework in which individuals are expected to take care of only themselves and their immediate families. Its opposite, collectivism, represents a preference for a tightly-knit framework in society in which individuals can expect their relatives or members of a particular in-group to look after them in exchange for unquestioning loyalty. A society's position on this dimension is reflected in whether people's self-image is defined in terms of "I" or "we." The indicator is based on 30 questions and largely rests on IBM employees around the world.

- Genetic distance: This indicator is based on [23]. This measure captures how distant human societies are in terms of the frequency of neutral genes among them. As such, [21] describe it as a molecular clock that characterizes the degree of relatedness between human populations in terms of the number of generations that separate them from a common ancestor population. Genetic distance is measured by a fixation index (F_{ST}). If two populations have identical allele frequencies at a given locus, F_{ST} is zero. If two populations are completely different F_{ST} takes the value one. The underlying data on F_{ST} genetic distance is based on [26], who compiled data on 267 populations. [23] match populations to countries, using ethnic composition data by country from [27]. The indicator is weighted by the frequency of different populations residing in a country. It thus represents the expected genetic distance between two randomly selected individuals, one from each country.

- Cultural distance: This indicator is based on [22]. They apply the method from population genetics and calculate a fixation index (F_{ST}) for cultural distance based on answers to the World Value Survey. Among several technical advantages (e.g. it does not assume that traits fall along a single dimension and it can handle binary, continuous and nominal traits) it rests on a comprehensive set of WVS based questions.

- Economic inequality (Gini coefficient). This indicator is based on the UN development report of the year 2015 or latest [28]. The Gini coefficient measures the deviation of the distribution of income among individuals or households within a country from a perfectly equal distribution. A value of 0 represents absolute equality, a value of 100 absolute inequality.

- Gender gap in health and survival: This is a subindex of the World Economic Forum's Global Gender Gap Index [29]. It measures disparities between the healthy lives of men and women focusing on two dimensions: first, the sex ratio at birth, which provides an indication of the number of girls that are ”missing” due to sex-selective abortions or infanticide. The second dimension is the life expectancy difference between women and men, up to a maximum score of parity.

- Log Gross Domestic Product (GDP) per capita. The GDP per capita data is purchasing power parity-adjusted GDP per capita in constant 2017 international dollars from the World Economic Outlook Database by International Monetary Fund [30].

References

1. Hohenberger, C., Spörrle, M. & Welpe, I. M. How and why do men and women differ in their willingness to use automated cars? the influence of emotions across different age groups. *Transportation Research Part A: Policy and Practice* **94**, 374–385 (2016).

2. Green, P. E. & Rao, V. R. Conjoint measurement for quantifying judgmental data. *Journal of Marketing research* 355–363 (1971).

3. Jasso, G. & Rossi, P. H. Distributive justice and earned income. *American Sociological Review* 639–651 (1977).

4. Wallander, L. 25 years of factorial surveys in sociology: A review. *Social Science Research* **38**, 505–520 (2009).

5. Hainmueller, J., Hopkins, D. J. & Yamamoto, T. Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis* **22**, 1–30 (2014).

6. Splawa-Neyman, J., Dabrowska, D. M., Speed, T. P. *et al.* On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science* **5**, 465–472 (1990).

7. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66**, 688 (1974).

8. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behavioral and brain sciences* **33**, 61–83 (2010).

9. US Census Bureau, 2012–2016 American Community Survey (ACS) 5-year Public Use Microdata Sample (PUMS). Accessed via FactFinder: `https://www.census.gov/programs-surveys/acs/data/pums.html`. Accessed: 2018-06-06.

10. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (1996). `10.1.1.71.1980`.

11. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M. & Perona, I. An extensive comparative study of cluster validity indices. *Pattern Recognition* **46**, 243–256 (2013).

12. Calinski, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods* (1974).

13. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* (1987). `z0024`.

14. Inglehart, R. & Welzel, C. *Modernization, cultural change, and democracy: The human development sequence* (Cambridge University Press, 2005).

15. Spolaore, E. & Wacziarg, R. How Deep Are the Roots of Economic Development ? *Journal of Economic Literature* **51**, 325–369 (2013).

16. Gächter, S. & Schulz, J. Intrinsic honesty and the prevalence of rule violations across societies. *Nature* **531** (2016).

17. Hofstede, G. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations across Nations.* (SAGE Publications, 2003).

18. Gorodnichenko, Y. & Roland, G. Individualism, innovation, and long-run growth. *Proceedings of the National Academy of Sciences* **108**, 21316–21319 (2011).

19. Gorodnichenko, Y. & Roland, G. Culture, Institutions and Democratization (2016).

20. Schulz, J. F. The Churches' Bans on Consanguineous Marriages, Kin-Networks and Democracy (2017). URL `http://www.ssrn.com/abstract=2877828`.

21. Spolaore, E. & Wacziarg, R. Ancestry, Language and Culture. In Ginsburgh, V. & Weber, S. (eds.) *The Palgrave Handbook of Economics of Languages*, chap. 6 (Palgrave Macmillan, London, 2016).

22. Muthukrishna, M. *et al.* A WEIRD scale of cultural distance. *(submitted)* .

23. Spolaore, E. & Wacziarg, R. Ancestry and Development: New Evidence (2017).

24. 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).

25. Hofstede, G. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations* (Sage publications, 2003).

26. Pemberton, T. J., DeGiorgio, M. & Rosenberg, N. A. Population structures in a comprehensive genomic data set on human microsatellite variation. *G3-Genes/Genomes/Genetics* **3**, 903–919 (2013).

27. Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S. & Wacziarg, R. Fractionalization. *Journal of Economic Growth* **8**, 155–194 (2003).

28. World Bank. The world bank GINI report. Tech. Rep. (2017).

29. World Economic Forum. The global gender gap report. Tech. Rep. (2017).

30. International Monetary Fund. World economic outlook database. `http://www.imf.org/external/pubs/ft/weo/2017/01/weodata/index.aspx` (2017).