Autonomous vehicles: the trolley problem and its consequences

Nina Papenfuß

25.02.2020

Structure

1.Introduction

- 1.1. Philosophical background
- 1.2. Relevance for AVs

2. The Moral Machine Experiment

- 2.1. Moral preferences
- 2.2. Individual variations based on demographics
- 2.3. Moral clusters
- 2.4. Cultural correlation
- 2.5. Consequences for a universal ethics

3. Conclusion & Outlook

Introduction

Thought experiment:
Simplify complex issues through hypothetical scenarios
“isolate and test desired variables” (Lin [7])

Introduction

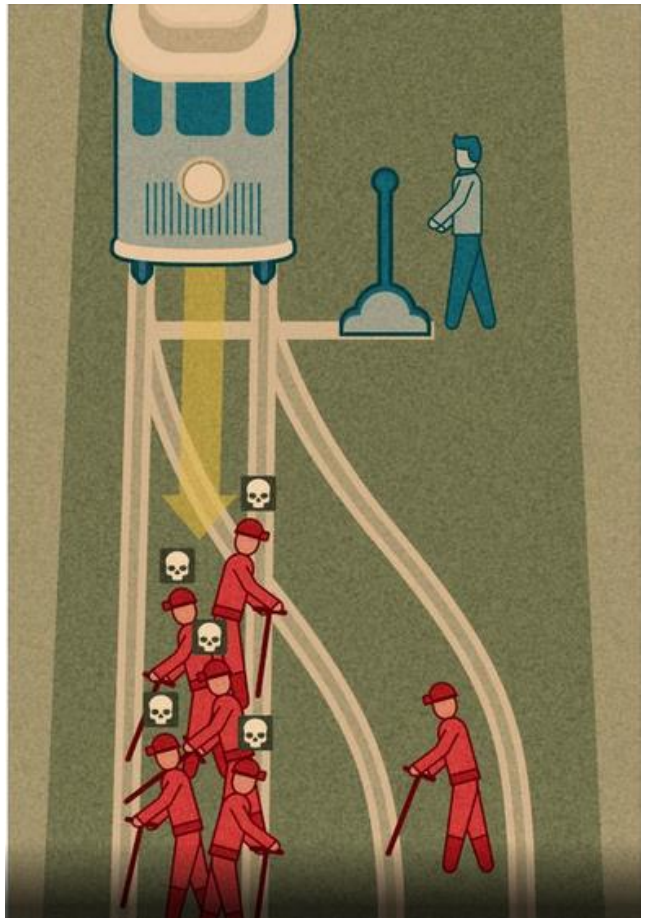
Thought experiment:
Simplify complex issues through hypothetical scenarios
“isolate and test desired variables” (Lin [7])

Moral dilemma:

„the agent is required to do each of two (or more) actions; the agent can do each of the actions; but the agent cannot do both (or all) of the actions. The agent thus **seems condemned to moral failure**; no matter what she does, she will do something wrong (or fail to do something that she ought to do) (…) So in addition to the features mentioned above, in order to have a *genuine* moral dilemma it must also be true that neither of the conflicting requirements is overridden“ (SEP [A])

Philosophical Background

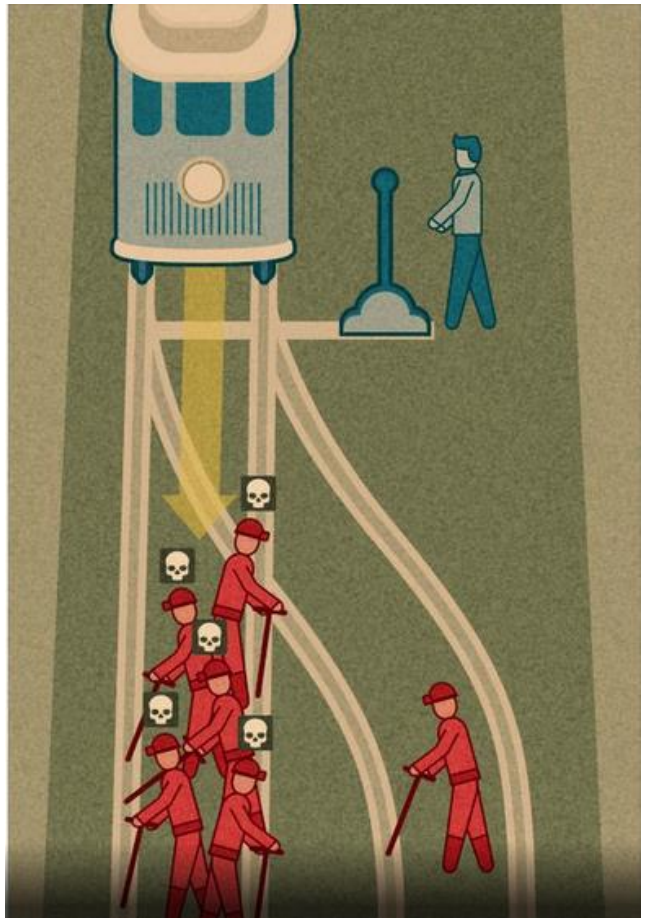
Philippa Foot 1967:
A trolley is out of control and about to roll over five persons on the track. Through pulling the switch the trolley could be diverted onto a different track. Unfortunately there is another person. Is it allowed to pull the switch and therefore accept the death of one person in order to save the lives of five persons? [B]



[C]

Philosophical Background

Philippa Foot 1967:
A trolley is out of control and about to roll over five persons on the track. Through pulling the switch the trolley could be diverted onto a different track. Unfortunately there is another person. Is it allowed to pull the switch and therefore accept the death of one person in order to save the lives of five persons? [B]

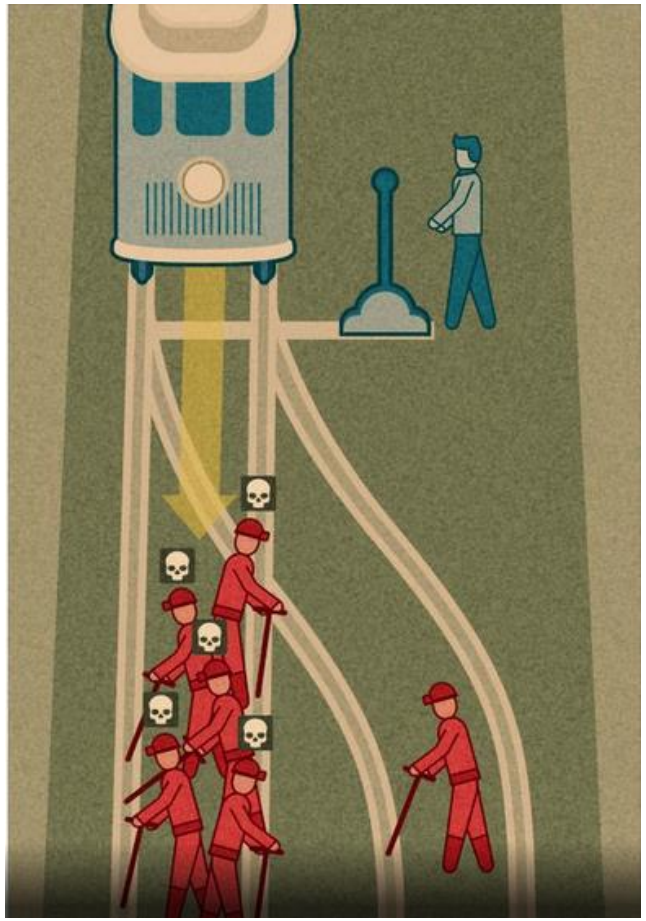


[C]

Cosequentialists: Switch

Philosophical Background

Philippa Foot 1967:
A trolley is out of control and about to roll over five persons on the track. Through pulling the switch the trolley could be diverted onto a different track. Unfortunately there is another person. Is it allowed to pull the switch and therefore accept the death of one person in order to save the lives of five persons? [B]



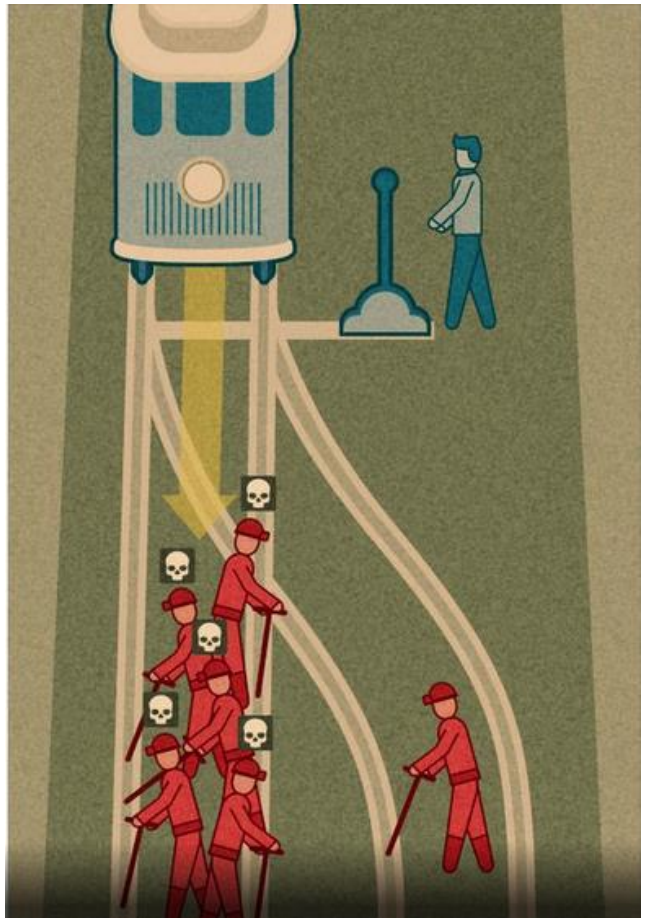
[C]

Cosequentialists: Switch

Double Effect: Switch

Philosophical Background

Philippa Foot 1967:
A trolley is out of control and about to roll over five persons on the track. Through pulling the switch the trolley could be diverted onto a different track. Unfortunately there is another person. Is it allowed to pull the switch and therefore accept the death of one person in order to save the lives of five persons? [B]



[C]

Cosequentialists: Switch

Double Effect: Switch

Deontic Logic: notSwitch

Philosophical Background

Philippa Foot 1967

A trolley is out of contro
to roll over five perso
Through pulling th
trolley could b
different track
is another person
pull the switch and t
the death of one person
save the lives of five person

Somewhat analogous for AV

[C]

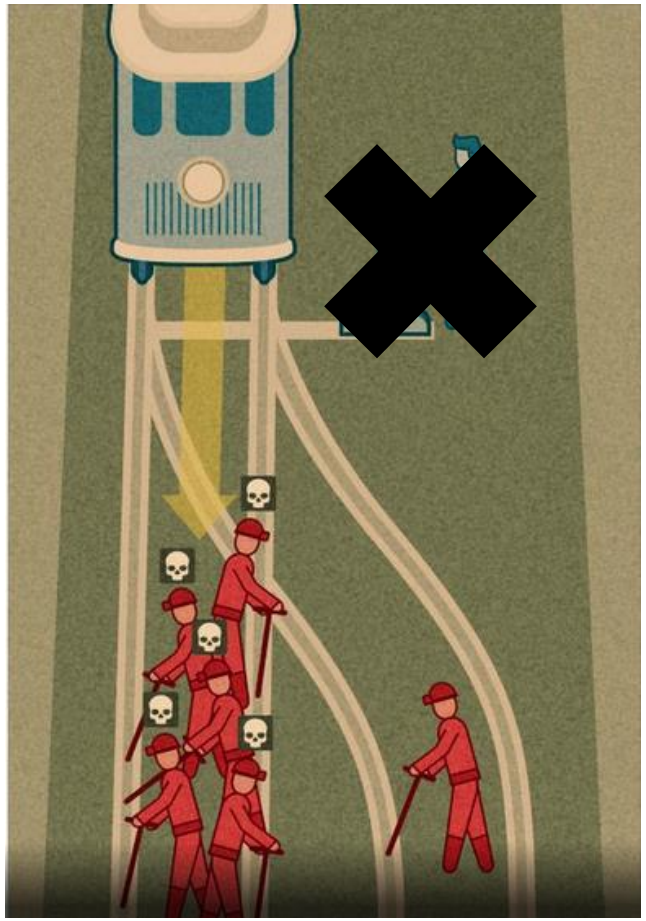
Cosequentialists: Switch

Double Effect: Switch

Deontic Logic: notSwitch

Relevance for AVs

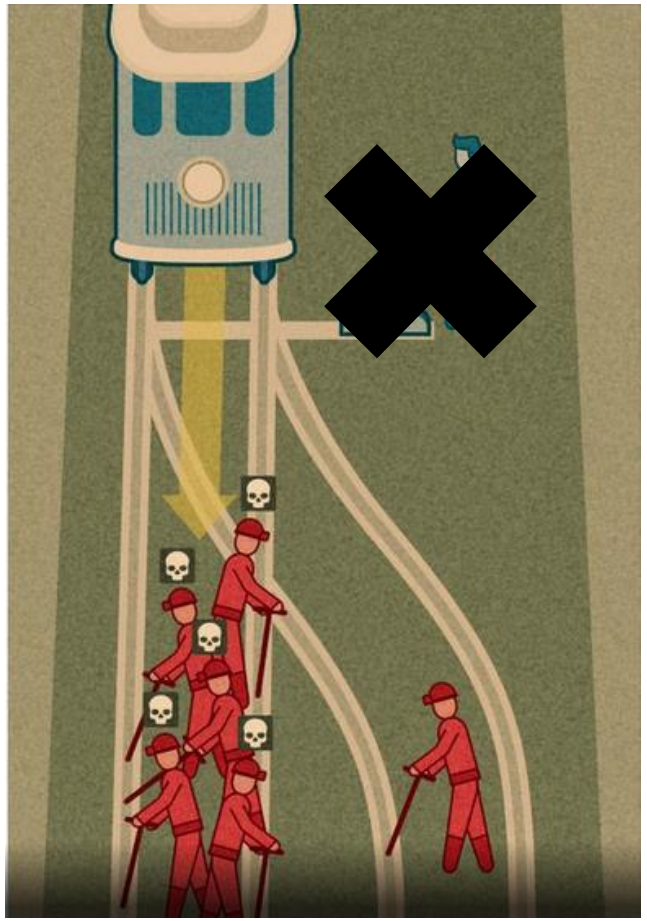
According to an AV's calculation, it is about to roll over five persons if it stays on its lane. By swerving the AV would hit another person. Is the AV allowed to swerve and therefore bring by the death of one person in order to save the lives of five persons?



[C]

Relevance for AVs

According to an AV's calculation, it is about to roll over five persons if it stays on its lane. By swerving the AV would hit another person. Is the AV allowed to swerve and therefore bring by the death of one person in order to save the lives of five persons?

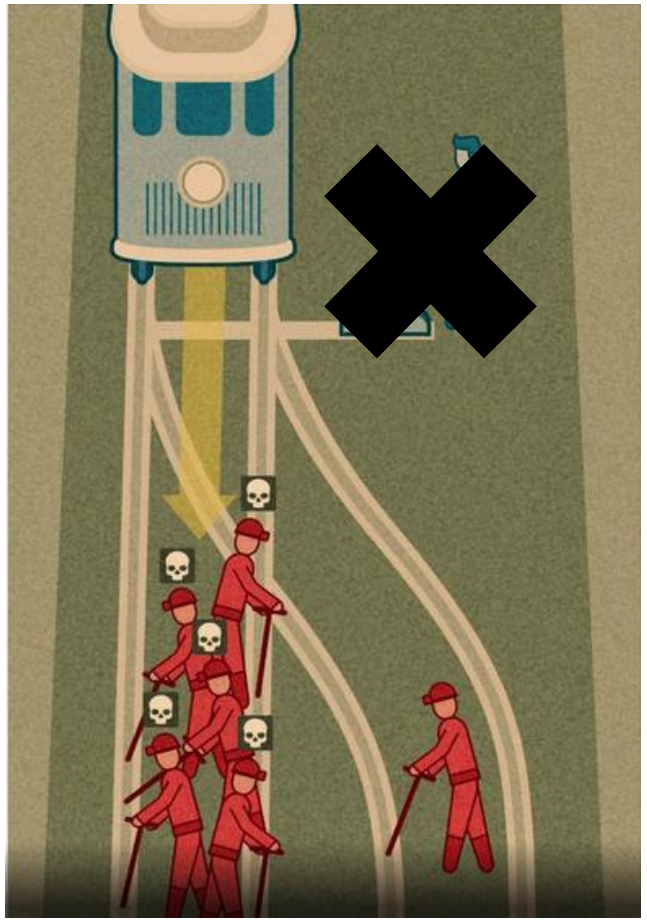


[C]

Cosequentialists: Switch

Relevance for AVs

According to an AV's calculation, it is about to roll over five persons if it stays on its lane. By swerving the AV would hit another person. Is the AV allowed to swerve and therefore bring by the death of one person in order to save the lives of five persons?



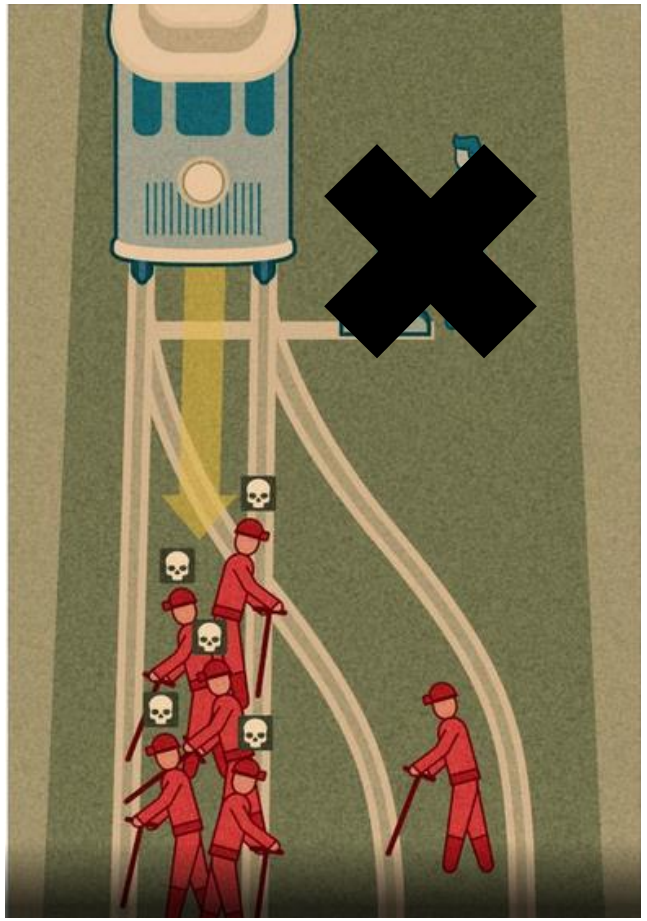
[C]

Cosequentialists: Switch

Double Effect possible?

Relevance for AVs

According to an AV's calculation, it is about to roll over five persons if it stays on its lane. By swerving the AV would hit another person. Is the AV allowed to swerve and therefore bring by the death of one person in order to save the lives of five persons?



[C]

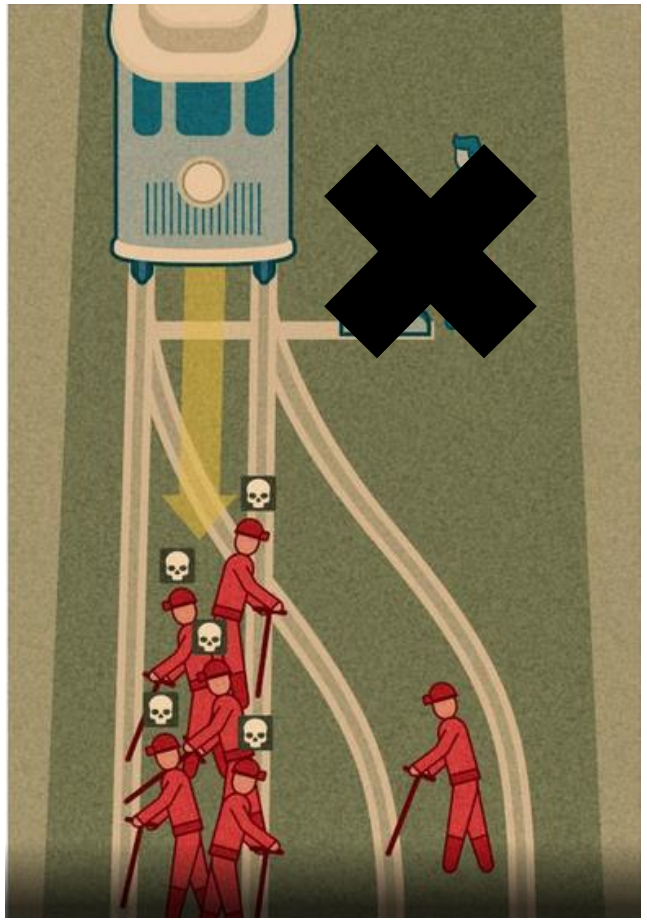
Cosequentialists: Switch

Double Effect possible?

Deontic Logic: notSwitch

Relevance for AVs

According to an AV's calculation, it is about to roll over five persons if it stays on its lane. By swerving the AV would hit another person. Is the AV allowed to swerve and therefore bring by the death of one person in order to save the lives of five persons?



[C]

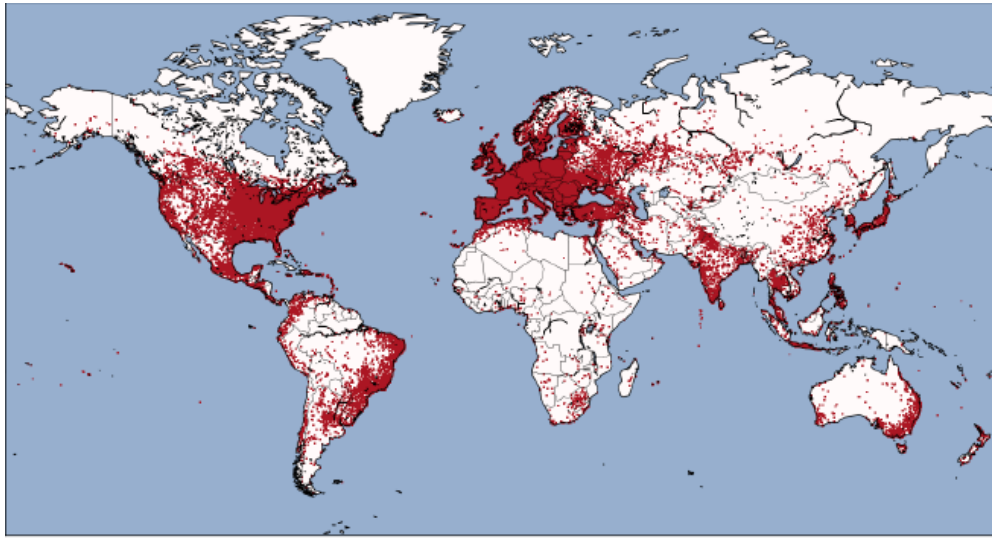
Cosequentialists: Switch

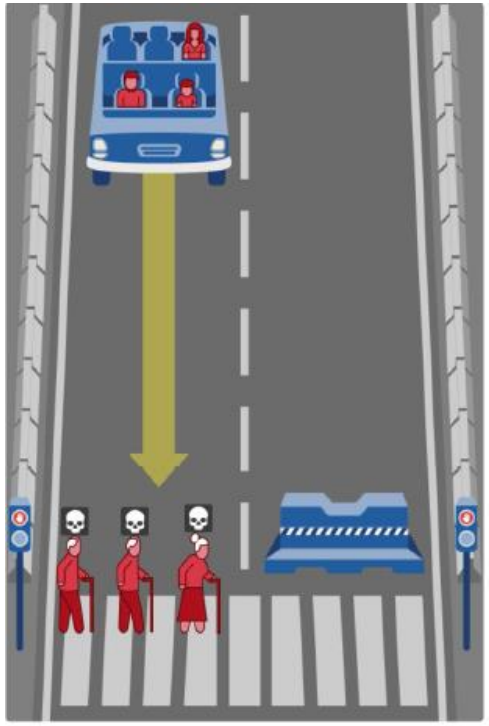
Double Effect possible?

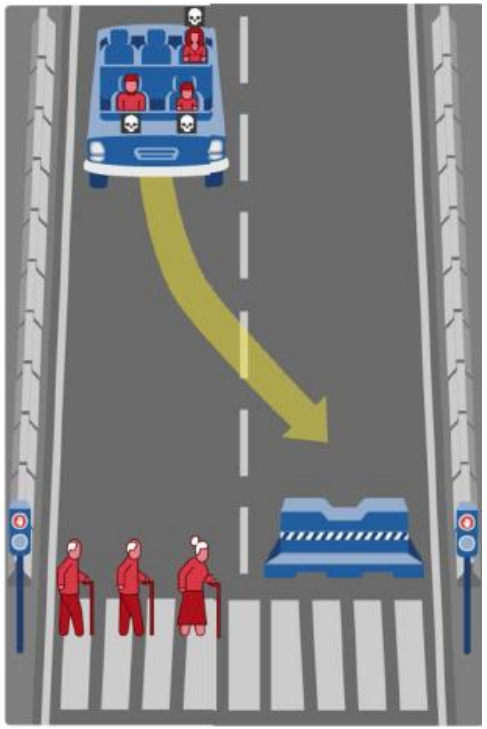
Deontic Logic: notSwitch

Flip a coin?

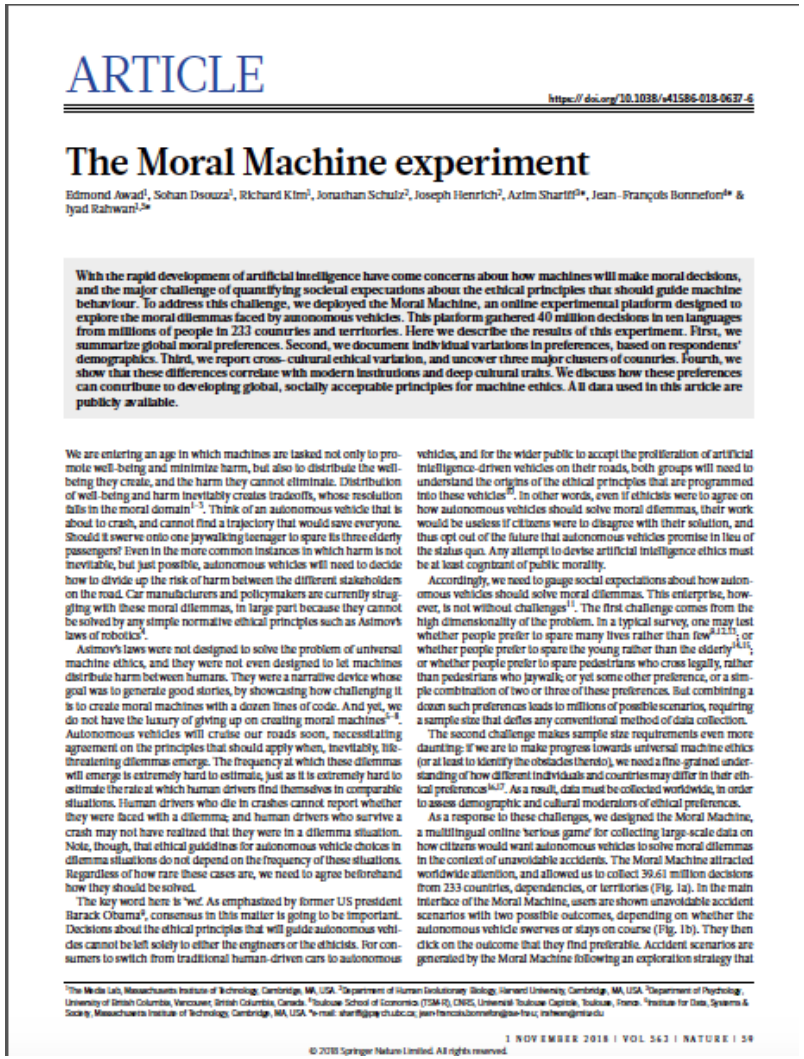
The Moral Machine Experiment







Website: http://moralmachine.mit.edu
2.3 Million self-selected participants
All information of this slide: [1]



ARTICLE

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

The Moral Machine experiment

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

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

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

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

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

Accordingly, we need to gauge social expectations about how autonomous vehicles should solve moral dilemmas. This enterprise, however, is not without challenges[11]. The first challenge comes from the high dimensionality of the problem. In a typical survey, one may test whether people prefer to spare many lives rather than few[8,12,13]; or whether people prefer to spare the young rather than the elderly[14,15]; or whether people prefer to spare pedestrians who cross legally, rather than pedestrians who jaywalk; or yet some other preference, or a simple combination of two or three of these preferences. But combining a dozen such preferences leads to millions of possible scenarios, requiring a sample size that defies any conventional method of data collection.

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

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

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

Awad, Dsouza, Kim, Schulz, Henrich, Shariff, Bonnefon, Rahwan 2018 [1]

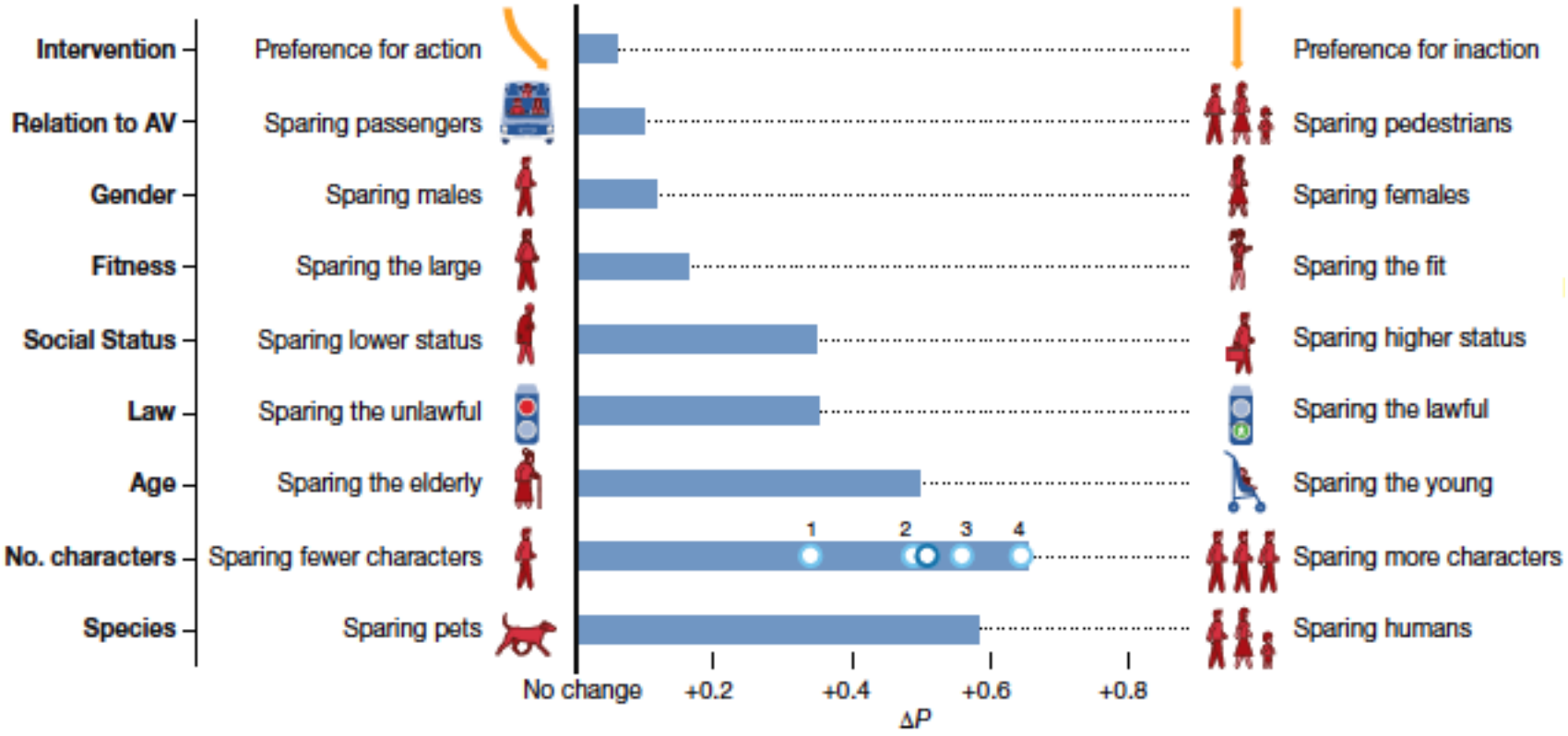
1. Moral preferences

2. Individual variations based on demographics

3. Cultural cluster

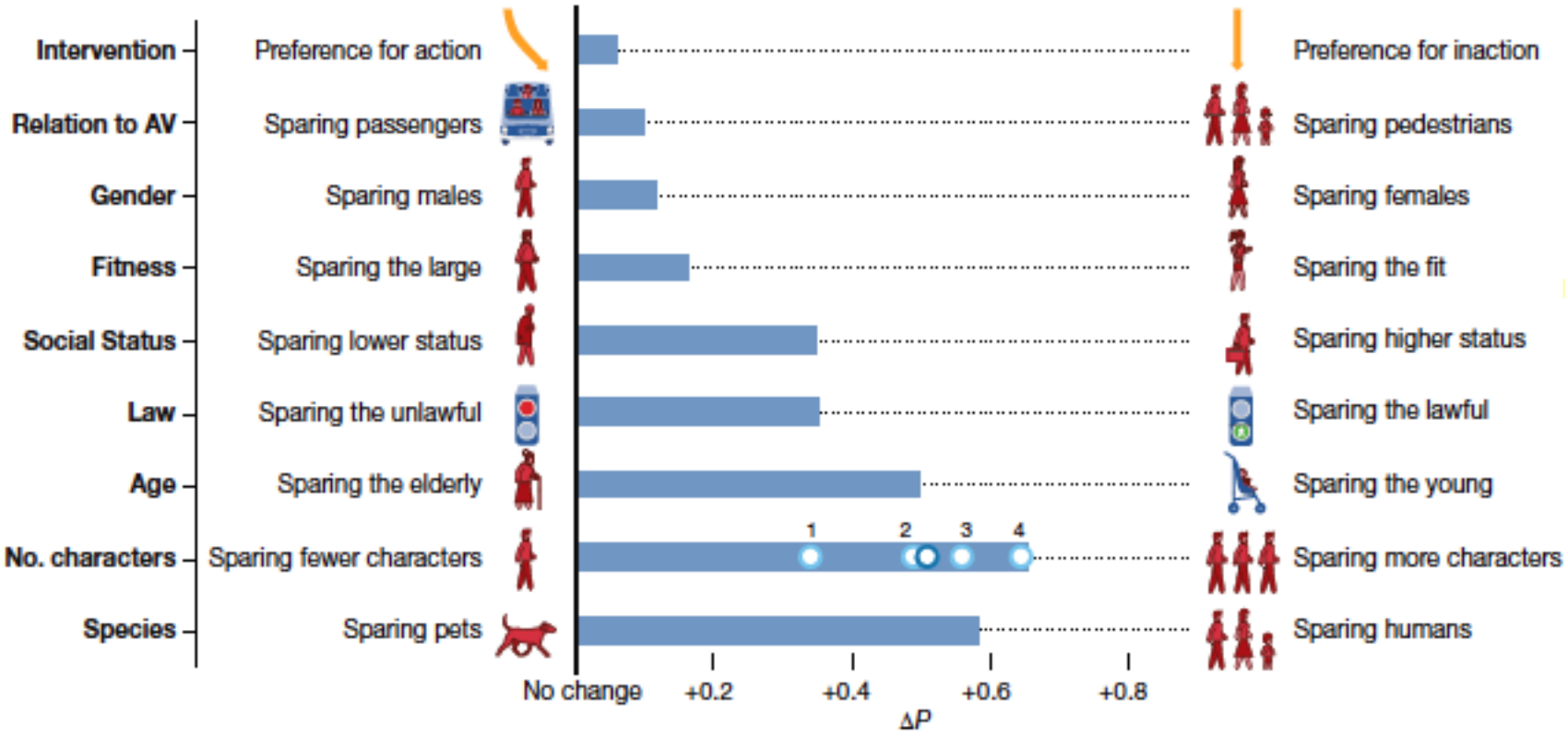
4. Cultural correlation

Moral preferences

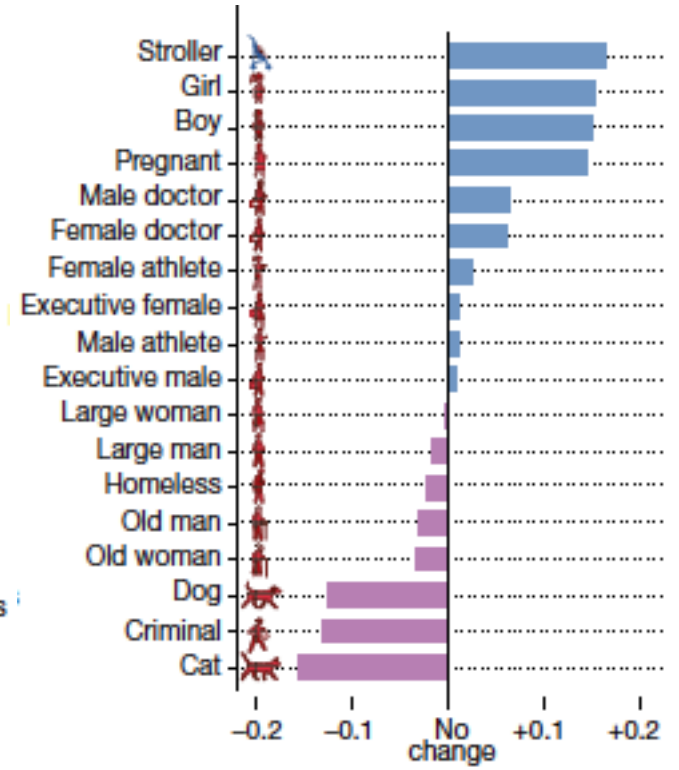


Average Minimal Component Effect (AMCE) [1]

Moral preferences



Average Minimal Component Effect (AMCE) [1]



Relative advantage/penalty [1]

Individual variations based on demographics

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

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

Regression table showing the individual variations for each of the nine attributes [1]

Individual variations based on demographics

Demographics

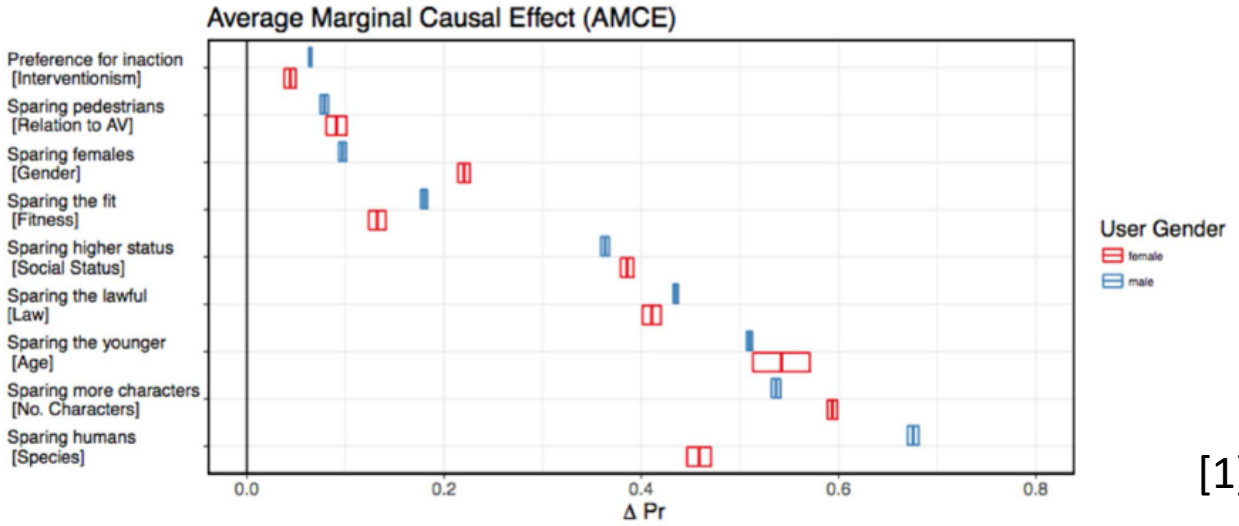
	Preference for Inaction	Sparing Pedestrians	Sparing the Lawful	Sparing Females	Sparing the Fit	Sparing Higher Status	Sparing the Young	Sparing More Characters	Sparing Humans
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Male	-0.015***	-0.022***	0.020***	-0.061***	0.024***	-0.009***	-0.018***	-0.024***	0.085***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

No sizable impact

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

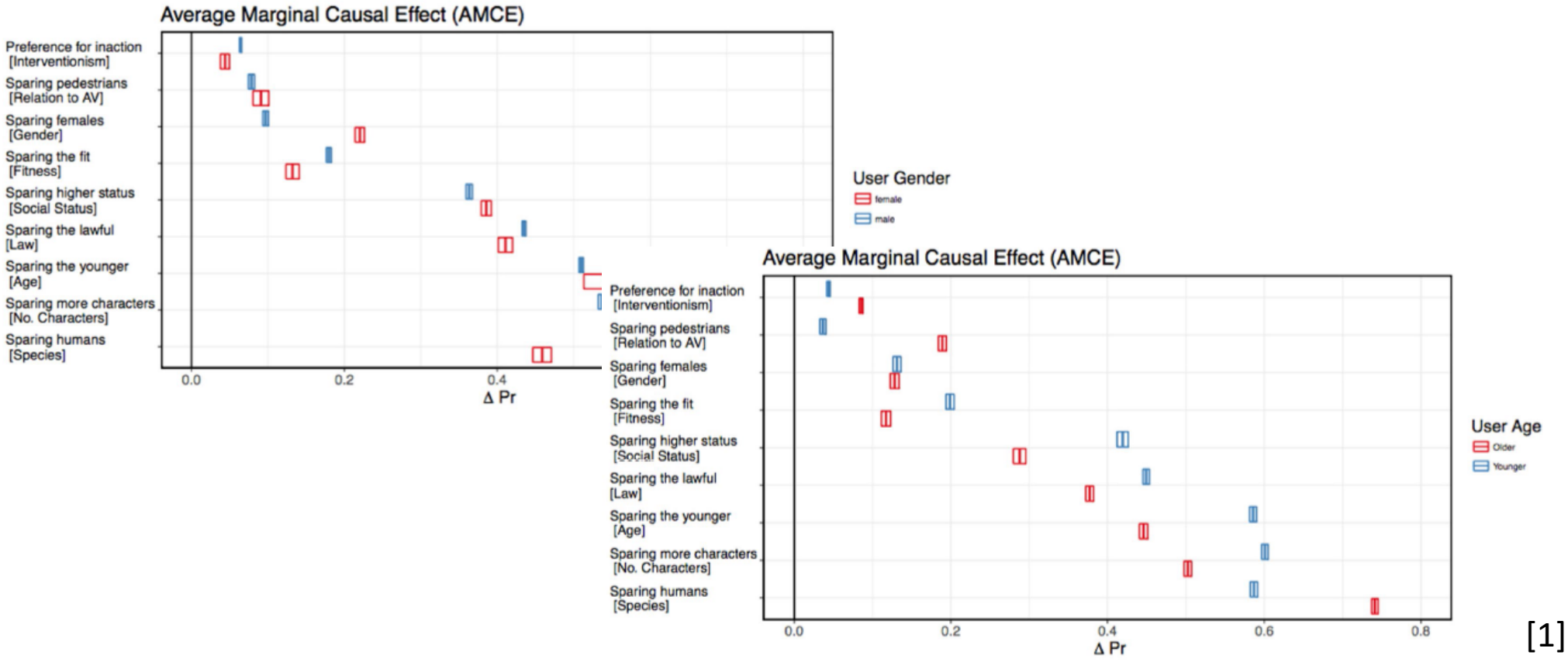
Regression table showing the individual variations for each of the nine attributes [1]

Individual variations based on demographics



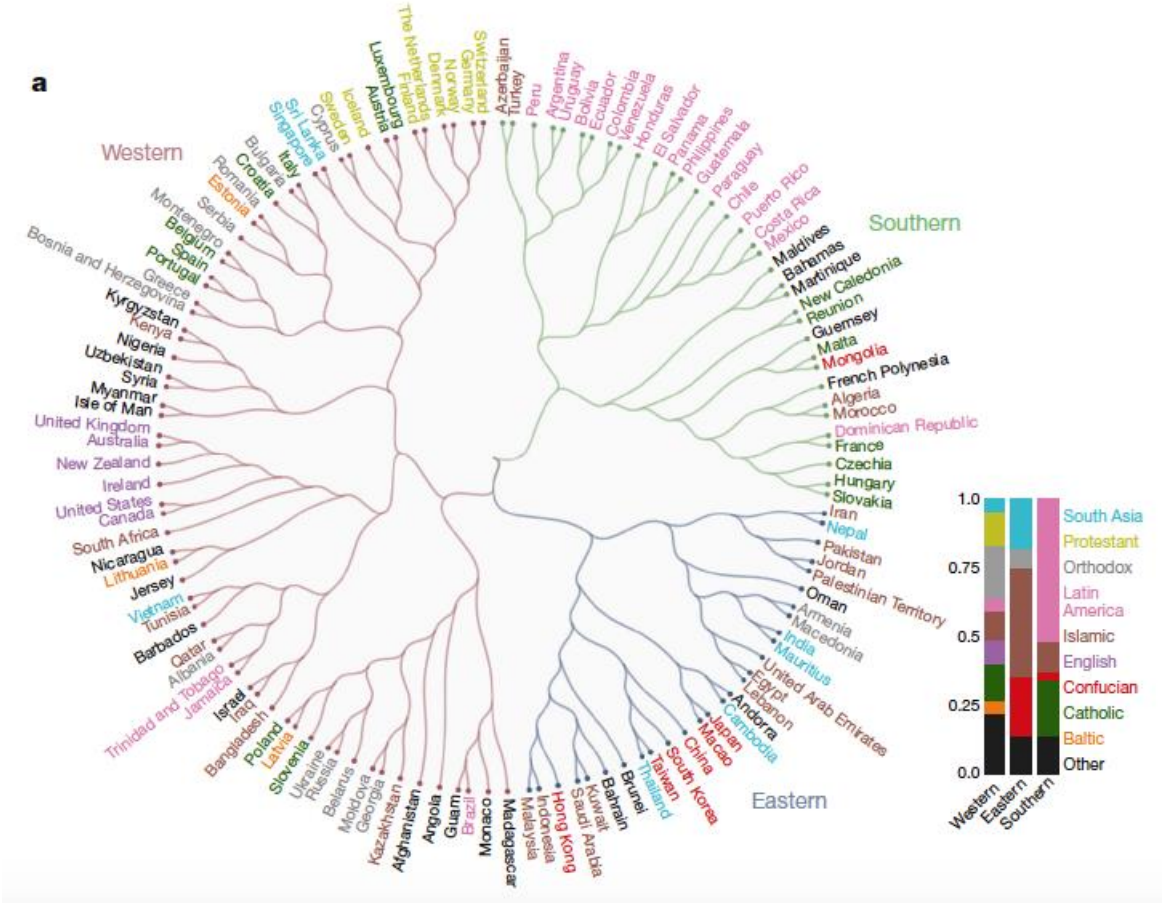
[1]

Individual variations based on demographics



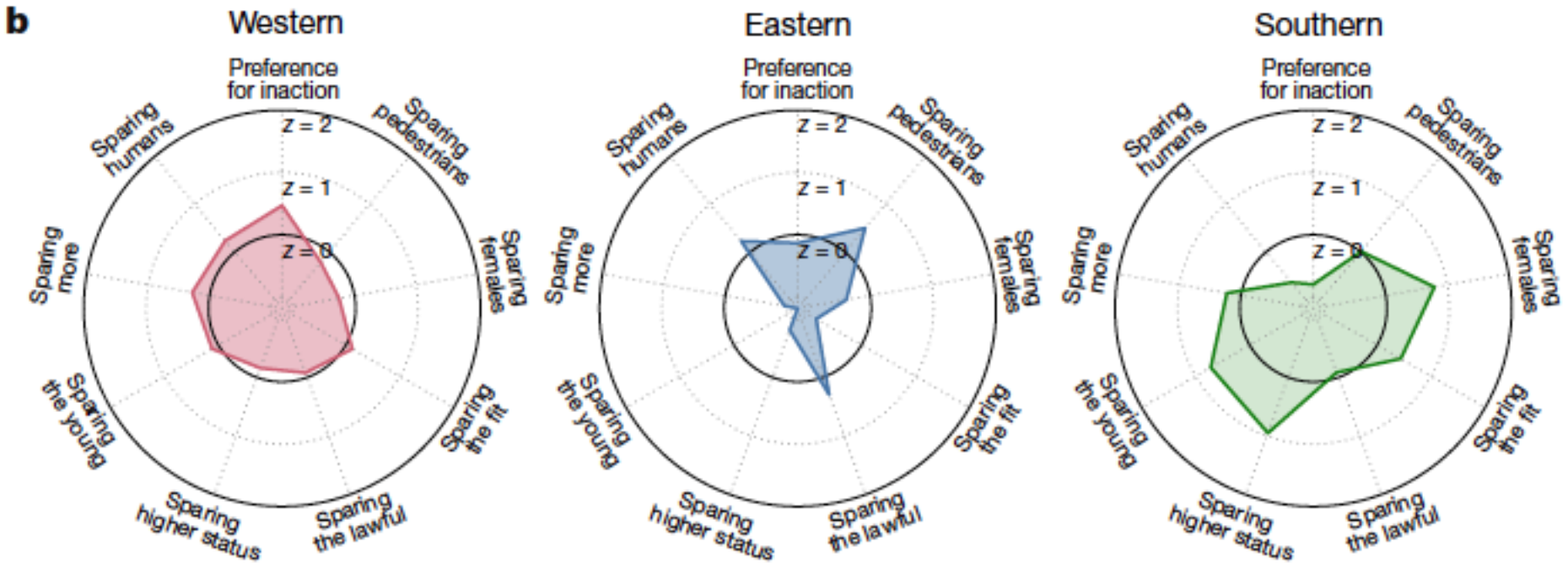
[1]

Cultural clusters



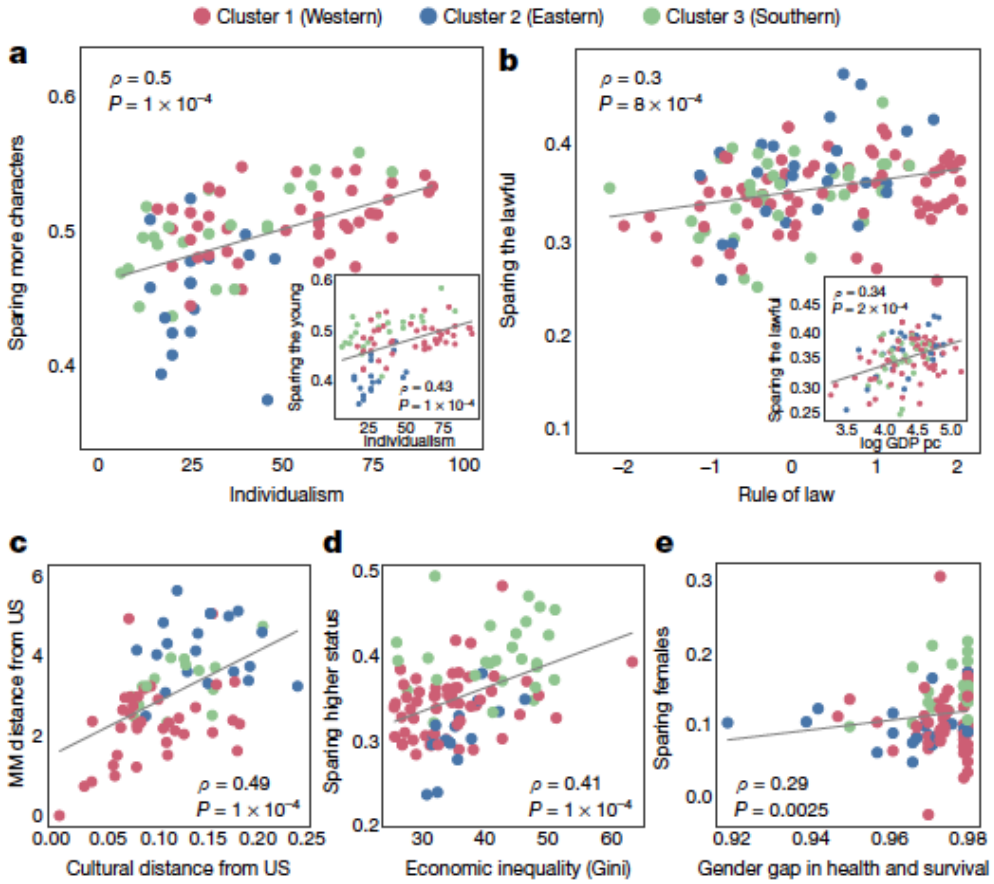
Hierarchical cluster of countries based on AMCE [1]

Cross-cluster differences



Mean AMCE z-scores [1]

Moral and cultural correlation



Association between Moral Machine preferences and other variables at the country level (Spearman's ρ and P values) [1]

Consequences for a universal ethics

1. Human > Pet

Consequences for a universal ethics

1. Human > Pet

2. Sparing more people

Consequences for a universal ethics

1. Human > Pet

2. Sparing more people

3. Young > Old

Consequences for a universal ethics

1. Human > Pet

2. Sparing more people

3. Young > Old

Ethik-Kommission 2017:

Der „Schutz menschlichen Lebens in einer Rechtsgüterabwägung [besitzt] höchste Priorität. Die Programmierung ist deshalb im Rahmen des technisch Machbaren so anzulegen, im Konflikt Tier- oder Sachschäden in Kauf zu nehmen, wenn dadurch Personenschäden vermeidbar sind.“ [D]

Consequences for a universal ethics

1. Human > Pet

2. Sparing more people

3. Young > Old

Ethik-Kommission 2017:

„(…)Eine Aufrechnung von Opfern ist untersagt. Eine allgemeine Programmierung auf eine Minderung der Zahl von Personenschäden kann vertretbar sein. Die an der Erzeugung von Mobilitätsrisiken Beteiligten dürfen Unbeteiligte nicht opfern.“ [D]

Consequences for a universal ethics

1. Human > Pet

2. Sparing more people

3. Young > Old

Ethik-Kommission 2017:

„Bei unausweichlicher Unfallsituation ist jede Qualifizierung nach persönlichen Merkmalen (Alter, Geschlecht, körperliche oder geistige Konstitution) strikt untersagt.“ [D]

Conclusion

1. Ethical principles for AVs – especially concerning moral dilemmas - need to be discussed collectively and democratically (no correct/moral answer, only way to guarantee acceptance).
2. Empirical research about moral preferences should only be one tool and should be used with caution (bias, positivism).
3. The debate needs to be embedded in a debate on who is prioritized on the streets (statistical dilemma [3]) and include even more factors (e.g. ecological).
4. At the moment there is a tilt in the debate (See: Greene et al. [5], Jobin et al.[6])

Outlook

Awad, Dsouza, Shariff, Rahwan, Bonnefon: Universals and variations in moral decisions made 42 countries by 70,000 participants. 01/2020.[2]

Bibliography

[1] Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon J.-F., Rahwan, I.: *The Moral Machine Experiment.* In: Nature 563, pp. 59-64 (2018).

[2] Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon J.-F., Rahwan, I.: *Universals and variations in moral decisions made 42 countries by 70,000 participants*. In: PNAS 117, pp. 2332-2337 (2020).

[3] Bonnefon, J.-F., Shariff, A., Rahwan, I.: *The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars*. In: Proceedings of the IEEE 107, pp. 502-504 (2019).

[4] Crawford, K., Calo, R.: *There is a blind spot in AI research*. In: Nature 538, pp.311-313 (2016).

[5] Greene, D., Hoffmann, A. L., Stark, L.: *Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning*. In: Proc. 52nd Hawaii International Conference on System Sciences, pp. 2122-2131 (2019).

[6] Jobin, A., Ienca, M., Vayena, E.: *The global landscape of AI ethics guidelines*. In: Nature Machine Intelligence 1, pp. 389-399 (2019).

[7] Lin, Patrick: „Why Ethics matters“. In: *Autonomes Fahren*, Maurer, Gerdes et al, 2015.

[8] Thomson, J. J.: *The trolley problem*. In: Yale Law Journal 94, pp. 1395-1415 (1985).

[9] Winfield, A.F., Michael, K., Pitt, J., Evers, V.: *Machine Ethics: The Design and Gouvernance of Ethical AI and Autonomous Systems*. In: Proceedings of the IEEE 107, pp. 509-517 (2019).

Online-Quellen

- [A] https://plato.stanford.edu/entries/moral-dilemmas/ (SEP)
- [B] https://plato.stanford.edu/entries/philippa-foot/
- [C] https://www.spiegel.de/wissenschaft/mensch/trolley-problem-wuerden-sie-einen-menschen-opfern-um-fuenf-andere-zu-retten-a-f7714fe4-a8c4-440c-989e-e7de9f669d04#
- [D] https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.html