The Moral Machine Experiment

Leonardo Calbi, Lorenzo Cellini, Alessio Falai
Alma Mater Studiorum - University of Bologna

Outline

1. Why ethics matters for autonomous vehicles

2. The Moral Machine Experiment

3. Arguments against MME

4. Conclusions

Why ethics matters for autonomous vehicles

Ethical dilemmas in AVs

AVs will be active actors in taking life-and-death decisions in road accident scenarios [1]

- Should AVs discriminate lives on the basis of gender, race, religion, age, etc.?
- Who should be responsible for the action taken?
- Are accidents always avoidable? If not, should we implement crash-optimization techinques?
- Is it better to choose to kill someone or to let someone else die?

Self driving Uber kills woman

Unfortunately, some accidents are unavoidable

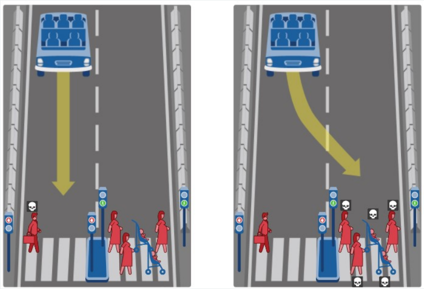
- Should the car have braked?
- Would avoiding the woman have killed the safety driver?

That's why we should prioritize analysis on such AV-related ethical dilemmas

The Moral Machine Experiment

Introduction

- MME [2] is an online experimental platform, designed to gauge social expectations about how AVs should solve moral dilemmas
- It gathered almost 40M decisions from 233 different countries in 10 languages
- What would you want an AV to do if its brakes failed?
  - Keep the lane and hit pedestrians on the road?
  - Swerve and hit pedestrians on the other lane?
  - Hit a barrier with the car?
- What would happen to the chosen characters?
  - ☠ → death
  - 🤕 → injury
  - ❓ → unknown

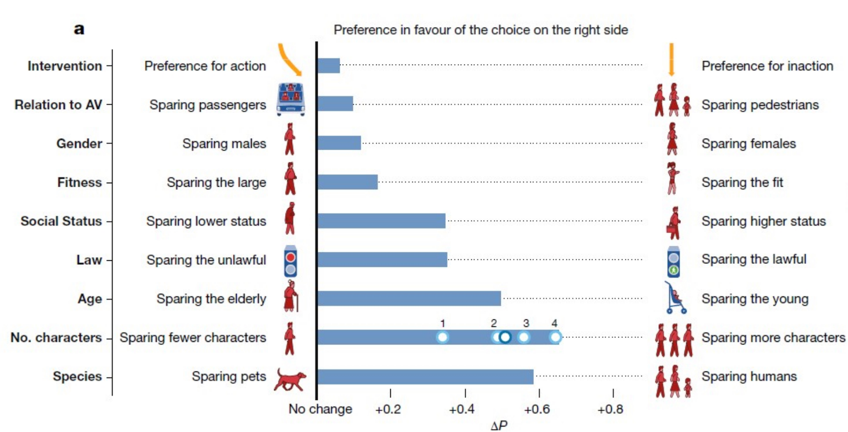


Analysis framework

The analysis was carried out on the following nine dimensions:

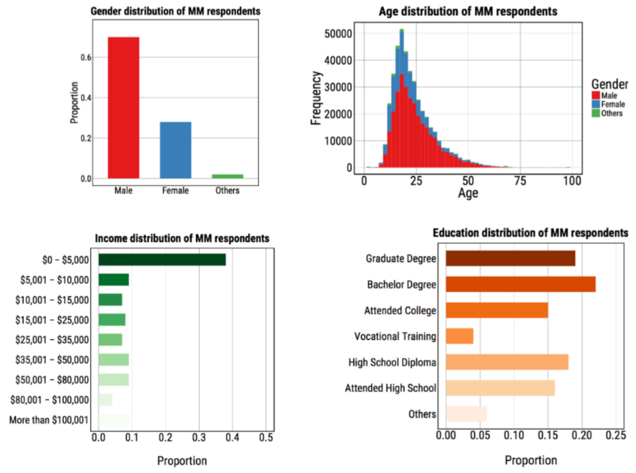
- Structural features
    1. Staying on course vs swerving
    2. Sparing passengers vs pedestrians
    3. Sparing more lives vs fewer lives
- Personal features
    4. Sparing humans vs pets
    5. Sparing men vs women
    6. Sparing the young vs the elderly
    7. Sparing pedestrians who cross legally vs jaywalking
    8. Sparing the fit vs the less fit
    9. Sparing those with higher social status vs lower social status

Relative importance



In each row, Δp is the difference between the probability of sparing characters possessing the attribute on the right, and the probability of sparing characters possessing the attribute on the left, aggregated over all other attributes

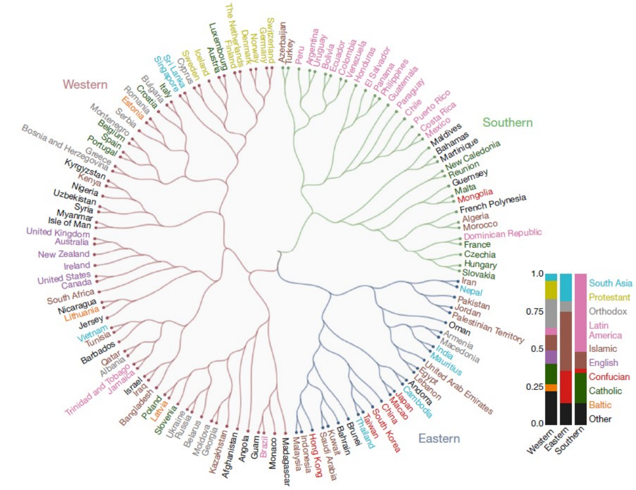
Individual variations



- Subgroup of Moral Machine users (492 921) who completed the optional demographic survey on age, education, gender, income, and political and religious views
- Including all six characteristic variables in regression-based estimators of each of the nine attributes shows that individual variations have no sizable impact on any of them (all below 0.1 ρ-value)

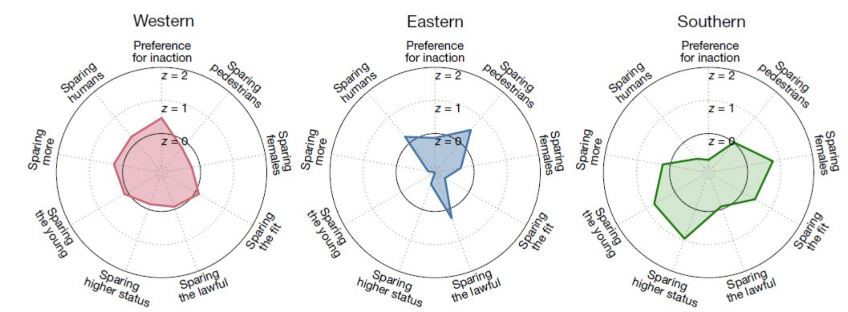
MORAL MACHINE

Hierarchical clustering of countries



- **Western cluster**: NA and EU countries (Protestant, Catholic, and Orthodox Christian cultural groups)
- **Eastern cluster**: countries such as Japan, Taiwan, Indonesia, Pakistan and Saudi Arabia (Confucianist and Islamic cultural groups)
- **Southern cluster**: consists of the Latin American countries of Central and South America, in addition to some countries that are characterized in part by French influence

Cluster preferences



- Systematic differences between individualistic and collectivistic cultures (young vs old and less vs more)
- Certain characters are preferred for demographic reasons (male vs female and rich vs poor)
- Choices related to one’s perception of law and quality of rules and institutions (pedestrian vs jaywalker)

Discussion

1. Three strong preferences: sparing human lives, more lives and young lives
2. Some preferences based on gender or social status vary considerably across countries, and appear to reflect underlying societal level preferences for egalitarianism
3. Samples are not guaranteed to be representative: this means that policymakers should not embrace MME results as the final word on societal preferences
4. Technologically unrealistic assumptions, such as the absence of uncertainty over character classification (adult vs children) and life/death outcomes

Arguments against MME

Weak points of MME

- The apparent preference for inequality in MME results is driven by the specific ”trolley-type” paradigm used by the experimenters
- MME concludes that people want AVs to make decisions about who to kill on the basis of personal features, including physical fitness, age, status and gender
  - This conclusion contradicts well-documented ethical preferences for equal treatment across demographic features and identities
  - Ignoring personal traits is also more consistent with the current technical capacities of AVs

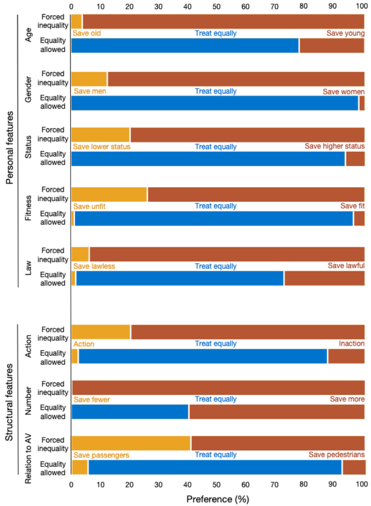
Experiments

To prove the ineffectiveness of MME, Bigman and Gray realized three different experiments [3]

1. People were randomly assigned either a forced inequality question (replication of MME in a simplified setting) or an equality-allowed condition (*”****treat*** *the lives of group A and B* ***equally****”*)
2. Similar to the first study, with a modified third option: *”AVs should decide who to save and who to* ***kill*** *without considering their personal features”*
3. Participants chose which of the two AVs should be allowed on the road: AVs based solely on structural features or both structural and personal features revealed by MME

First experiment

- Quasi-representative sample: 1174 US participants and 1178 UK participants using an online survey system
- Results from the forced inequality condition closely match the global effects of the MME
- People overwhelmingly selected the third option when it was available, revealing that they want autonomous vehicles to treat people equally
    - For example, when forced to choose between men and women, 87.7% chose to save women, but 97.9% of people actually preferred to treat both groups equally



Second experiment

- Sample: 843 US participants from an online panel
- Study 2 rules out the concern of whether participants preferred the ”treat equally” option in study 1 simply because it failed to mention killing
- Consistent with study 1, people expressed a robust preference for AVs to treat people equally by ignoring personal features
  - For example, people preferred self-driving cars to not consider gender (92.6%), fitness (88.8%) or status (84.7%)
  - The only substantial departure from study 1 was lawfulness: 53.1% of people preferred to spare law abiders over law breakers

Third experiment

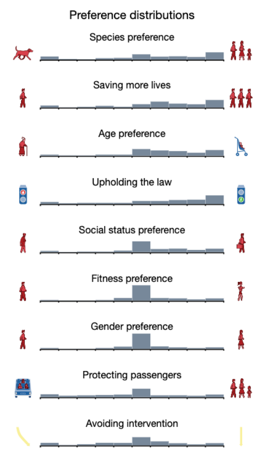
- Sample: 993 US participants from an online panel
- Personal features (law, fitness, status, gender and age) and structural features (passengers/pedestrians, fewer/more and action/inaction)
- Results show that 89.9% of participants chose the structural-features-only car, once again expressing a desire for AVs that ignore personal features in ethical dilemmas

Reply from MME's authors

- Bigman and Gray asked eight separate questions about general policy preferences (one per dimension), while MME had users go through multiple pairs of nine-dimensional outcomes
- The MME approach allows us to measure the weight of different moral priorities when pitted against each other, rather than considered in isolation, but participants cannot explicitly state that one dimension (for example, age) should not be taken into account
- The approach used by Bigman and Gray is more sensitive to social desirability, experimental demands and framing effects w.r.t. MME
- MME's authors (Awad et al.) tried an approach similar to the one by Bigman and Gray that remained unpublished [4]

MME unpublished experiment

- 585 531 MME users were taken to a page where they could position one slider for each of the nine dimensions explored by the Moral Machine: users could move the slider to express how important this dimension should be
- Values on the sliders were initialized based on user's decisions on MME questions
- Participants could easily express an equality preference by positioning the slider at the midpoint of the scale
    - This is a continuos alternative to the method used by Bigman and Gray, which relies on visual (instead of textual) cues
- A clear deviation towards equality from standard MME results is mainly observed on social status preference and over other dimesions with a lower intensity

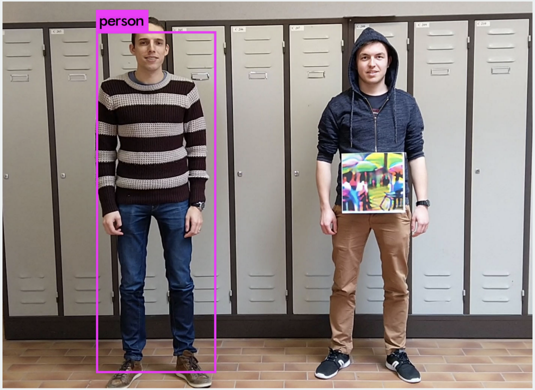


Conclusions

Personal thoughts

- MME results give us a snapshot of contemporary society: that's why we cannot rely only on them to build the basis of AVs systems
- Why should AV's producers be resposible for each and every life-and-death decision? Can't we delegate such control to the safety driver? [5]
- AVs should help us shape the future society in which we would like to live

Technical limitations



- When can a vehicle be considered fully autonomous?
- Are we ready to fully trust the underlying systems of AVs?
- Are those systems secure enough?

Thanks for the attention

References

[1] Patrick Lin. “Why Ethics Matters for Autonomous Cars”. In: *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*. Ed. by Markus Maurer et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 69–85. ISBN: 978-3-662-45854-9.

[2] Edmond Awad et al. “The Moral Machine Experiment”. In: *Nature* 563 (Nov. 2018).

[3] Yochanan Bigman and Kurt Gray. “Life and death decisions of autonomous vehicles”. In: *Nature* 579 (Mar. 2020), E1–E2.

[4] Edmond Awad et al. “Reply to: Life and death decisions of autonomous vehicles”. In: *Nature* 579 (Mar. 2020), E3–E5.

[5] Giuseppe Contissa et al. “The Ethical Knob: Ethically-customisable automated vehicles and the law”. In: *Sistemi Intelligenti* 29 (Dec. 2017), pp. 601–614.