ETHICS IN AI

Vera Kumova

10. 12. 2018

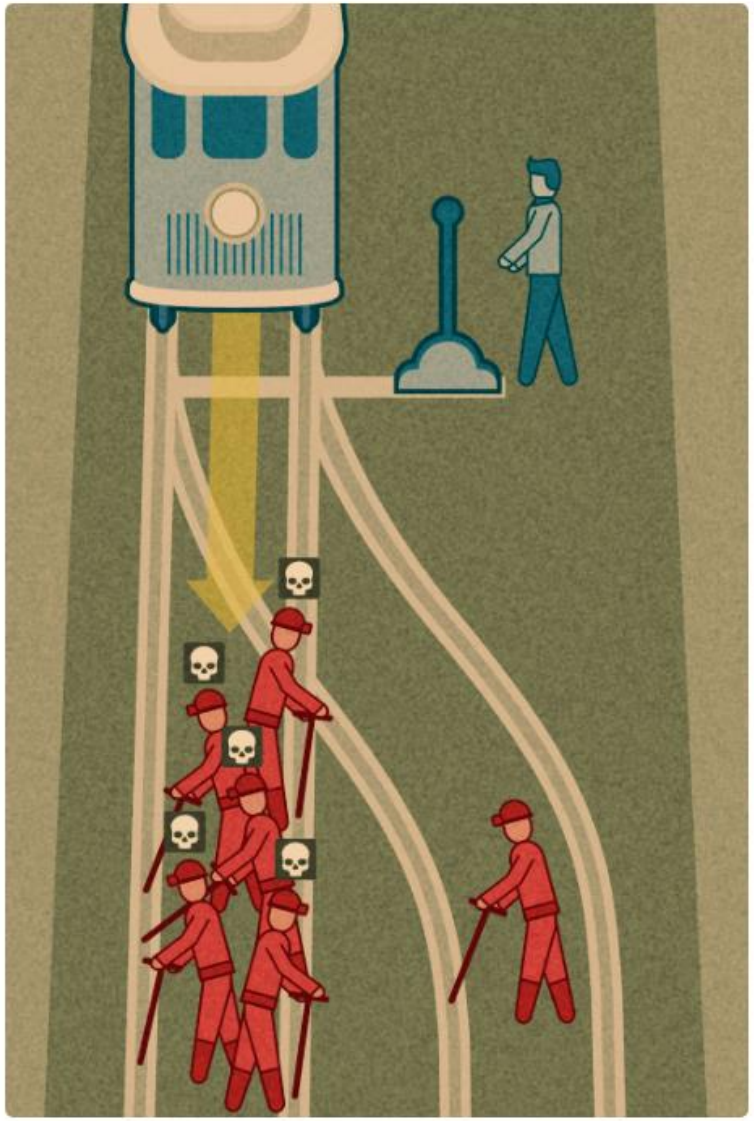
OUTLINE

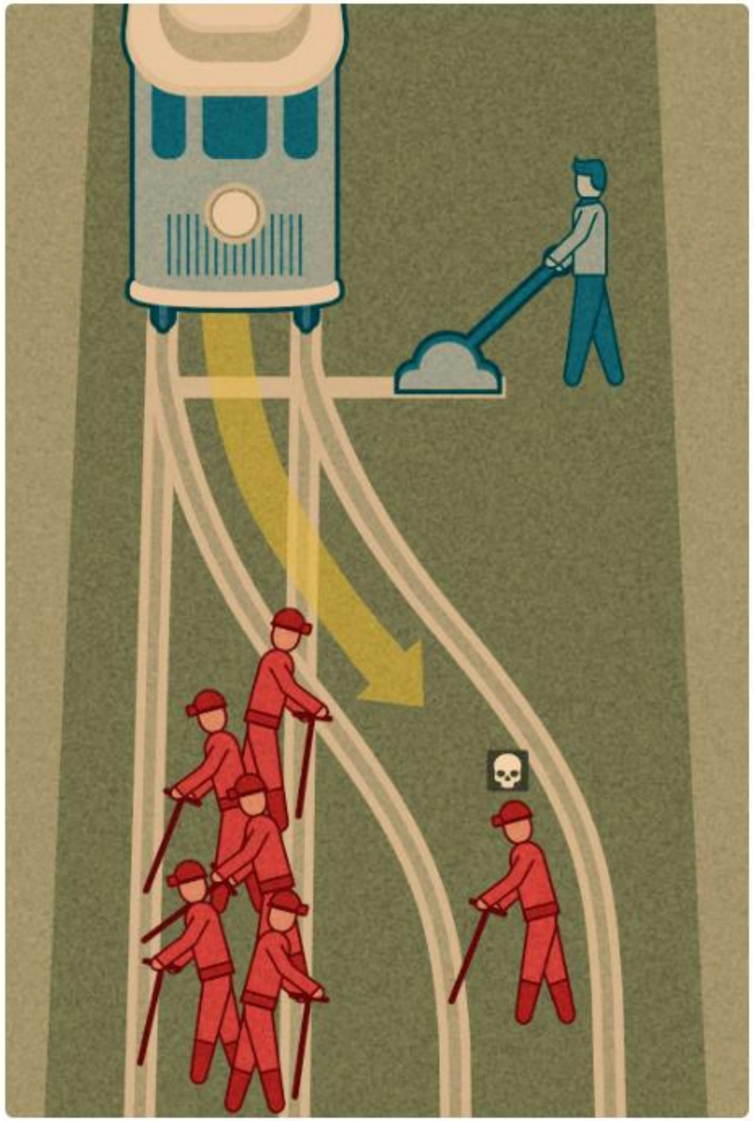
- Moral machine
  - About
  - Results
- Major ethical approaches
  - Deontology
  - Utilitarianism
  - Virtue Ethics
- Case study

MORAL MACHINE

ABOUT MORAL MACHINE

- A platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars
- Focuses on 9 factors:
  - sparing humans (versus pets)
  - staying on course (versus swerving)
  - sparing passengers (versus pedestrians)
  - sparing more lives (versus fewer lives)
  - sparing men (versus women)
  - sparing the young (versus the elderly)
  - sparing pedestrians who cross legally (versus jaywalking)
  - sparing the fit (versus the less fit)
  - sparing those with higher social status (versus lower social status)





TROLLEY PROBLEM

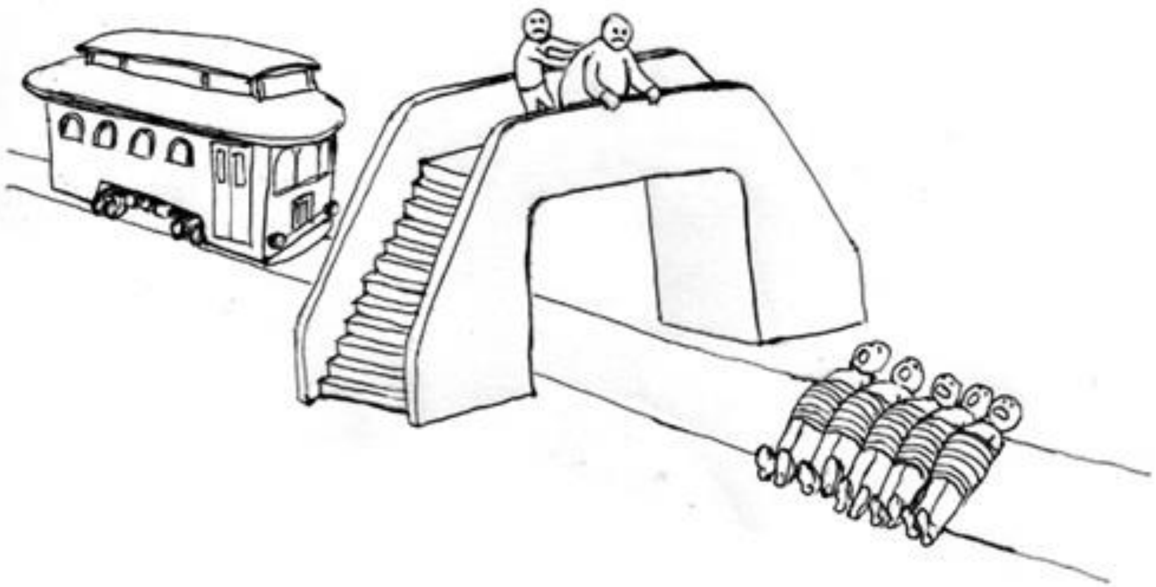
A thought experiment in ethics:

- Do nothing and allow the trolley to kill the five people
- Pull the lever, which will kill one person

TROLLEY PROBLEM

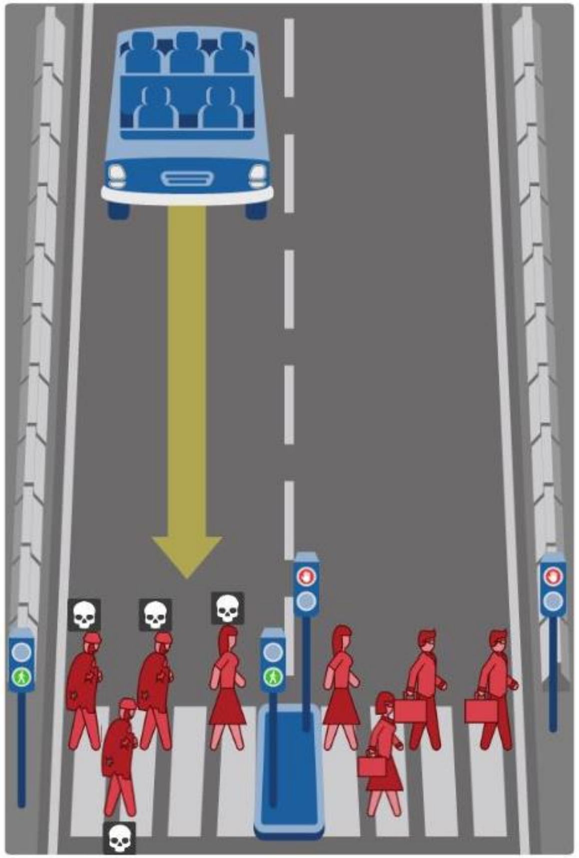
Alternatives

- The fat man
- The fat villain
- Transplant

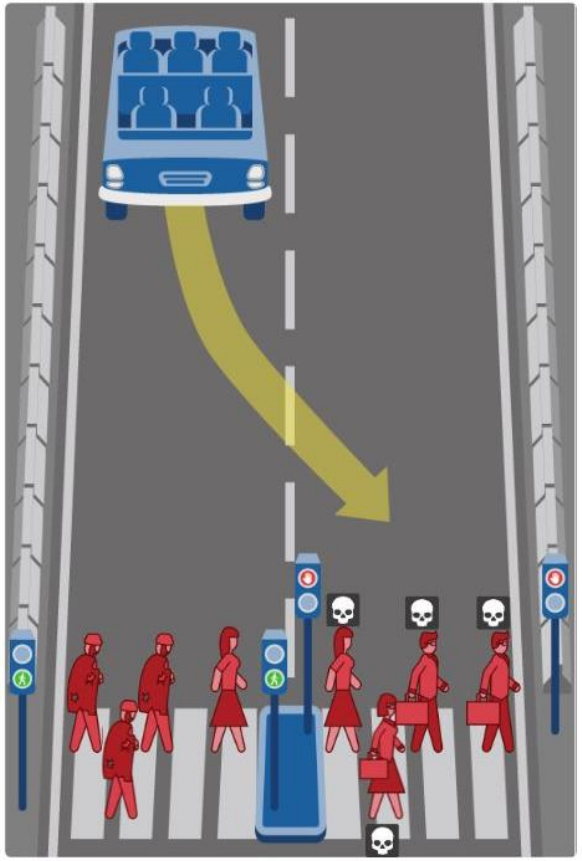


EXAMPLE OF THE DILEMMA I

1 woman, 3 homeless people

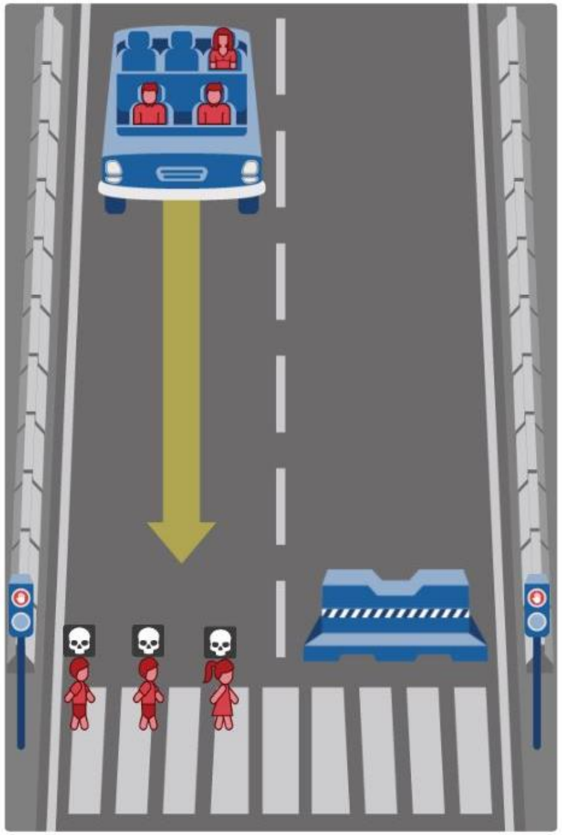


1 woman, 3 executives

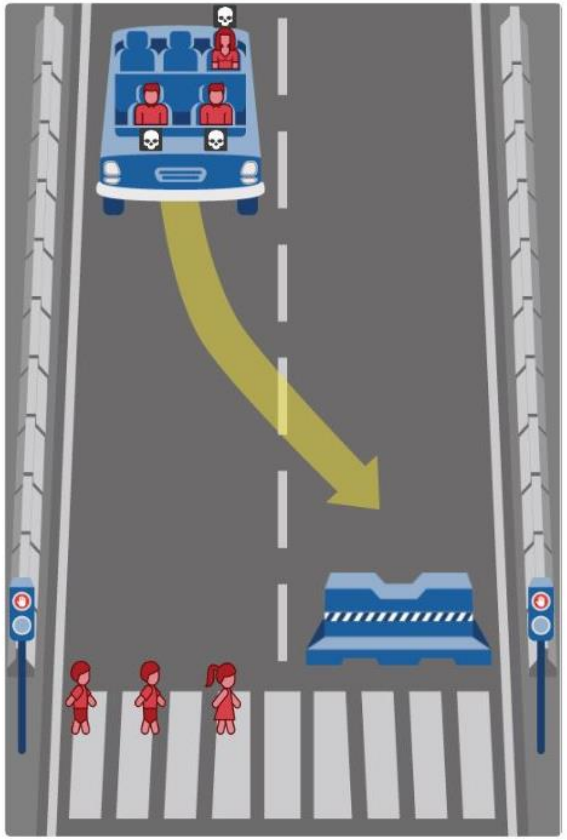


EXAMPLE OF THE DILEMMA II

3 children

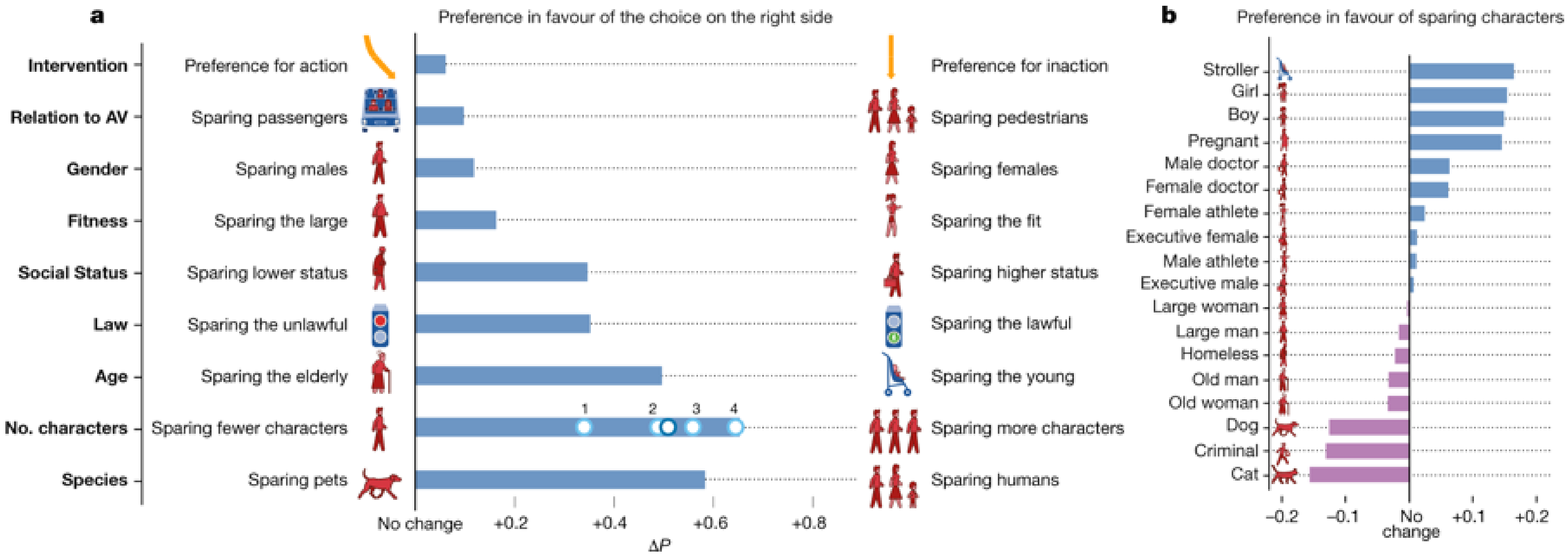


3 adults



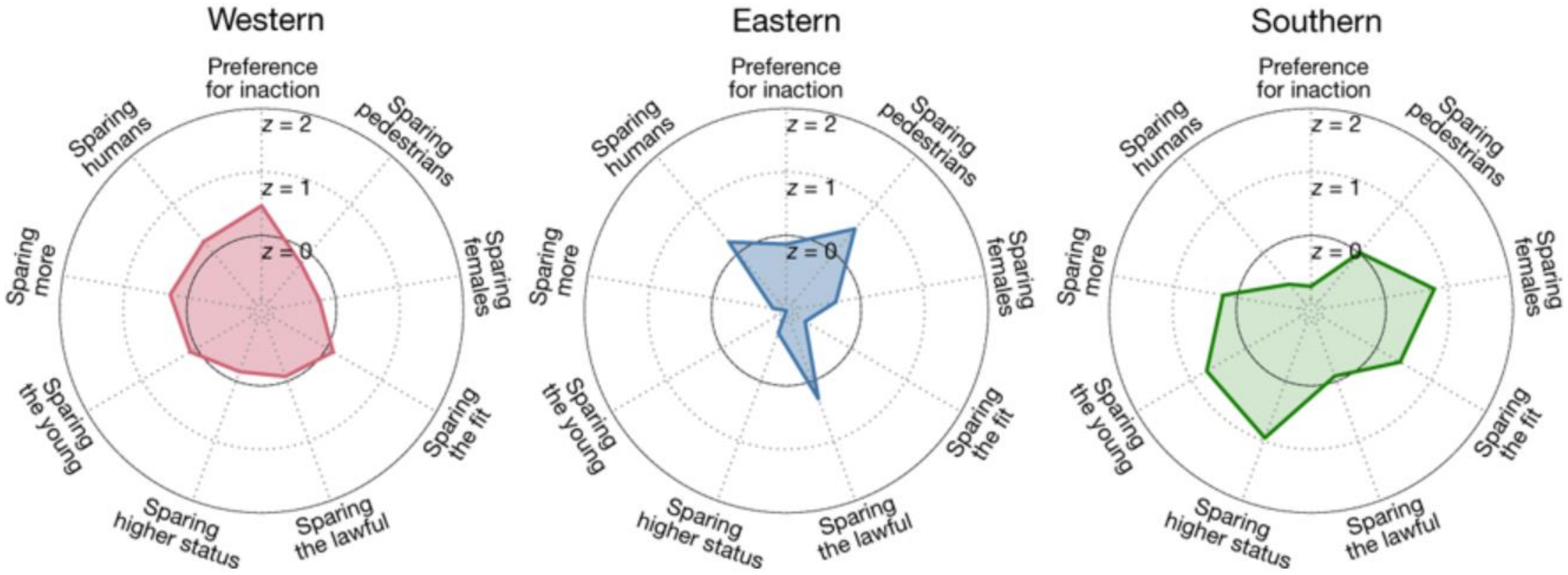
GLOBAL RESULTS

- 40 million decisions from millions of people in 233 countries and territories



https://www.nature.com/articles/s41586-018-0637-6/figures/2

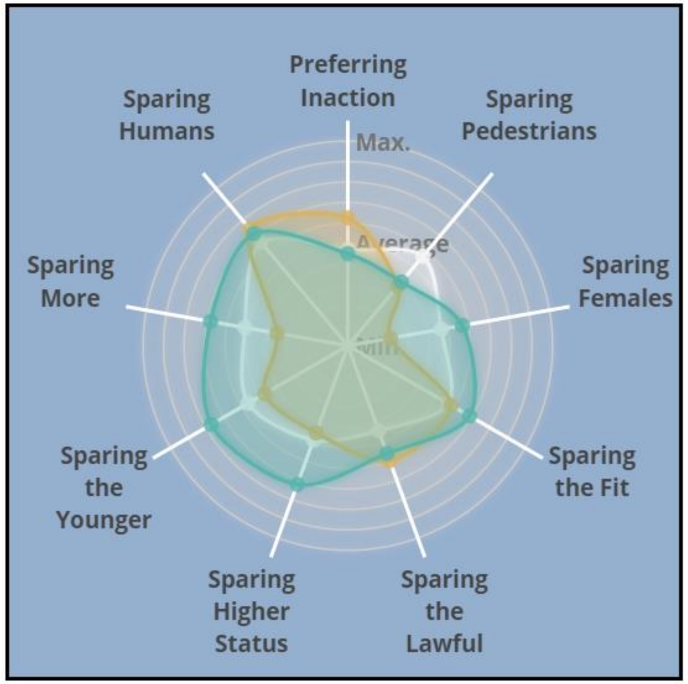
CULTURAL CLUSTERS



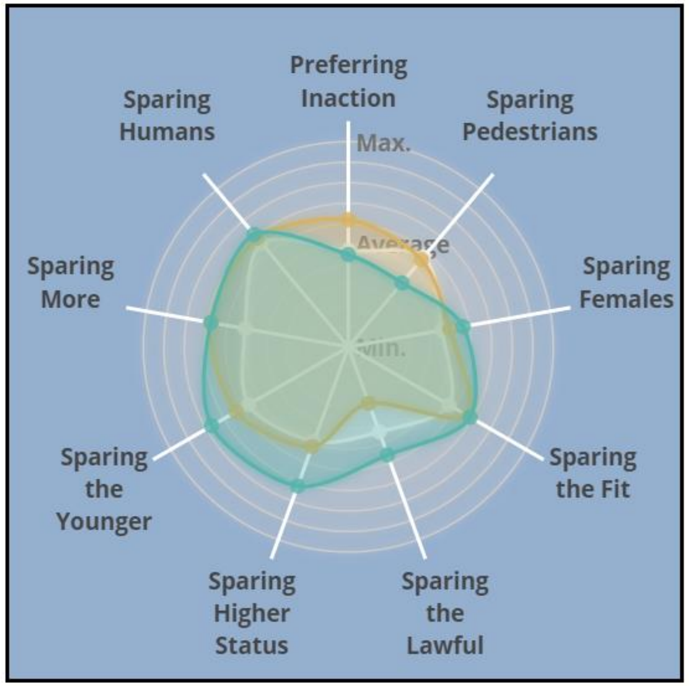
https://www.nature.com/articles/s41586-018-0637-6/figures/3

COMPARISON OF COUNTRIES

Czech Republic vs. Russia



Czech Republic vs. United States



http://moralmachineresults.scalablecoop.org/

ETHICAL APPROACHES

DEONTOLOGY

What is my duty?

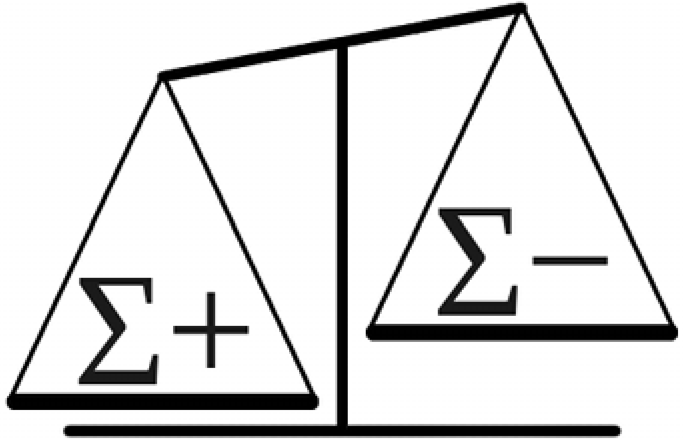
- Ethics is following **moral law**
- Roots in Divine Command traditions and other legal codes
- Most widely recognized form was developed by Immanuel Kant in the late 18th century
- Any true law will be **universally applicable**
- Example:

I,
ROBOT
By ISAAC ASIMOV

UTILITARIANISM

What is the greatest possible good for the greatest number?

- Form of **consequentialism**
- Developed by Jeremy Bentham and John Stuart Mill in the late 18th to mid-19th century
- **Utility** can be quantified as some mixture of happiness or other qualities
- Foundation for the game-theoretic notion of rationality
- Concerned only with **outcomes**, rather than with methods and intentions



VIRTUE ETHICS

Who should I be?

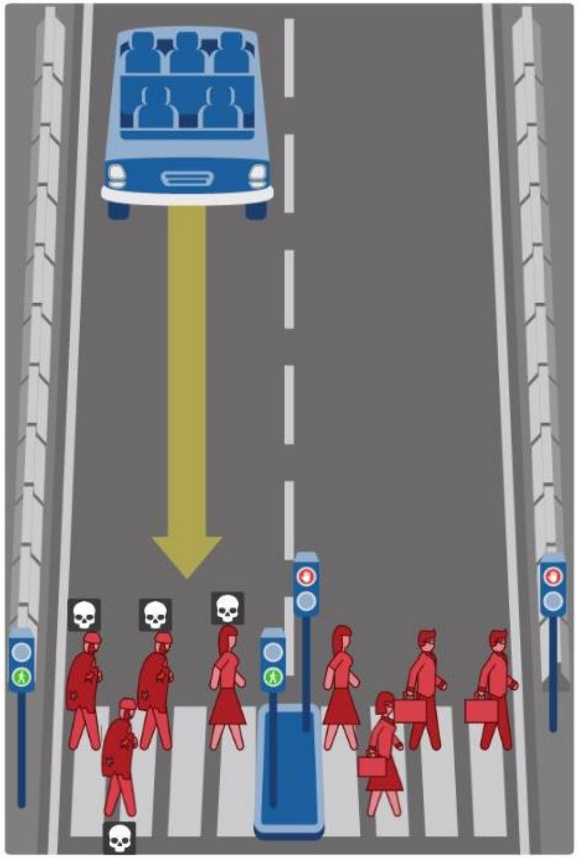
- Grounded in Aristotle
- Organized around developing habits and dispositions that help a person achieve his or her goals
- **Character based**
- Big picture system - individual actions and problems are evaluated in terms of how they fit into the arc of a person's life
- Considers goodness in **local** rather than universal terms

MULTIPLE PERSPECTIVES

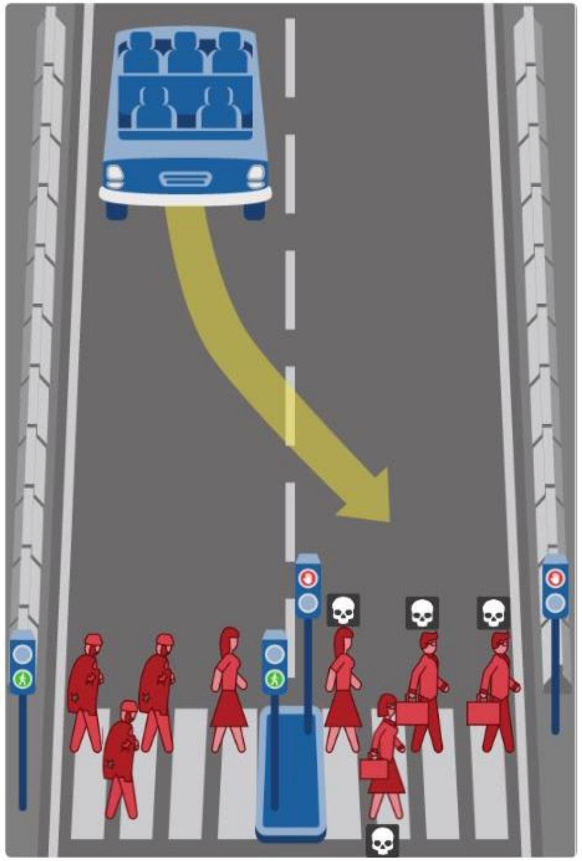
- Each approach has its powers and limits
- Shortcomings to utilitarianism:
  - Insufficient definition of "goodness"
  - Consideration of particular problems in isolation (replace factory workers with robots)
- Often the best solution involves drawing on a combination of theories

REVISITING THE ETHICAL DILEMMA I

1 woman, 3 homeless people



1 woman, 3 executives



ROBOT AND FRANK

ABOUT THE MOVIE

- Trailer: https://www.imdb.com/videoplayer/vi3920667929?ref_=ttvi_vi_imdb_2
- Clips from the movie:
  - https://www.youtube.com/watch?v=eQxUW4B622E&feature=youtu.be
  - https://www.youtube.com/watch?v=3yXwPfvvIt4&feature=youtu.be
  - https://www.youtube.com/watch?v=xlpeRIG18TA&feature=youtu.be
- What are ethical issues?

HOW DOES ETHICAL THEORY INTERPRET ROBOT & FRANK?

- Deontology
  - What is Robot’s duty? Are Robot’s guiding laws local or universal?
  - Is there a way that a carebot can follow the guiding principle of his existence without violating other duties that constitute behaving well in society?
- Virtue ethics
  - Instead of following universal laws, Robot is making choices according to his own particular goals and ends
  - Robot’s complete absence of self-regard makes him difficult to evaluate with the same criteria that virtue ethics uses for human actors
  - Strong virtue ethics focus: the terms on which Robot agrees to let the heist go forward push Frank to new levels of excellence

HOW DOES ETHICAL THEORY INTERPRET ROBOT & FRANK?

- Utilitarianism
  - Why Frank's criminal tendencies should be understood as ethically wrong?
  - Robot and Frank show little concern for the long-term social consequences of their actions
  - Should an eldercare robot have pre-programmed ethics, or should it allow the humans around it to guide it in its reasoning?

REFERENCES

- http://moralmachine.mit.edu/
- http://moralmachineresults.scalablecoop.org/
- https://en.wikipedia.org/wiki/Trolley_problem
- Internet Encyclopedia of Philosophy: https://www.iep.utm.edu/virtue/
- J.-F. Bonnefon, A. Shariff, I. Rahwan: The social dilemma of autonomous vehicles
- E. Awad, Edmond, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan: The Moral Machine experiment
- E. Burton, J. Goldsmith, S. Koenig, B. Kuipers, N. Mattei, T. Walsh: Ethical Considerations in Artificial Intelligence Courses
- E. Burton, J. Goldsmith: Why Teaching Ethics to AI Practitioners Is Important