Moral Machines

" . . . self-driving cars . . . The technology is essentially here. We have machines that can make a bunch of quick decisions that could drastically reduce traffic fatalities, drastically improve the efficiency of our transportation grid, and help solve things like carbon emissions that are causing the warming of the planet. But . . . what are the values that we're going to embed in the cars? There are gonna be a bunch of choices that you have to make, the classic problem being: If the car is driving, you can swerve to avoid hitting a pedestrian, but then you might hit a wall and kill yourself. It's a moral decision, and who's setting up those rules?"

– Barack Obama, 2016

Moral Machines

- Trolley problems are classical experiments in moral choice. People have to decide on which hypothetical scenario for a runaway trolley (streetcar / tram) is preferred.
- In a modern variant, the moral machines experiment asked millions of people from 233 countries about what autonomous vehicles (self-driving cars) should do in various circumstances.
  `https://www.moralmachine.net`
  `https://dx.doi.org/10.1038/s41586-018-0637-6`
- Suppose there is a self-driving car with sudden brake failure, and it has to choose:
  - It can go straight ahead which will result in the death of a man and a baby who are flouting the law by crossing on a red signal.
  - It can swerve which will result in the death of a pregnant woman who was abiding by the law.

  What should it do?

Moral Machines

Alternative scenarios included

- number of deaths
- people versus animals,
- men versus women,
- young versus old
- lawful versus unlawful,
- fit versus unfit.
- social status (eg, doctor versus homeless drug addict)

Some preferences seemed universal (across cultures):

- preference for humans over animals
- fewer lives over more lives
- young over old

Some were culturally specific.

Moral Machines

> *We can embrace the challenges of machine ethics as a unique opportunity to decide, as a community, what we believe to be right or wrong; and to make sure that machines, unlike humans, unerringly follow these moral preferences.*
>
> *– Awad at al., [2018]*

Issues:

- Some principles are universal, but some are culturally specific.
- The outcomes were all well defined; there was no uncertainty.
- Some outcomes will tend to be controversial to implement; E.g., people thought it was more important to save pedestrians than to save people in the vehicle; so drivers should not be the ones giving the preferences for their car!