Self-Driving Cars

Module 6 of a course on *Ethical Issues in AI*

Prepared by

John Hooker

Emeritus Professor, Carnegie Mellon University

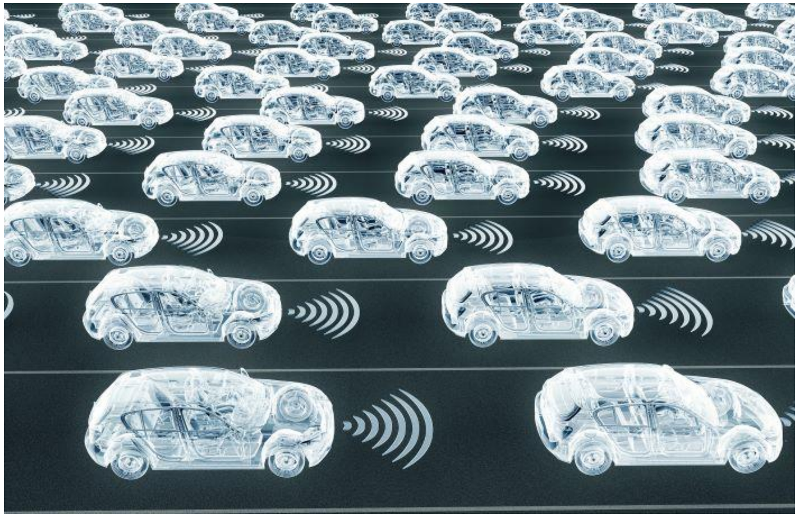
Chautauqua, June 2024

The Future of Cars

- Cars have **taken over** the world.
  - *Chronic* ***congestion*** *everywhere.*
  - *Shockingly* ***unsafe****.*
    - 2022: in U.S. alone, **42,514 deaths** in motor vehicle crashes, including **7522 pedestrians**.
    - **2.38 million injured.**

The Future of Cars

- A **self-driving** fleet offers enticing solutions
    - *Travel without stop lights or traffic jams.*
        - Due to sophisticated scheduling and coordination.
    - *High degree of safety.*
        - Comparable to airline safety?
        - By removing human irresponsibility and misjudgments.



The Future of Cars

- A **difficult challenge** for AI
  - *Progress has **stalled**.*
    - Projections overoptimistic, as in much of AI history.
    - Current projection: fully AVs by **2035**.
  - *To make progress, at some point we must put AVs **on the road**…*

Two Issues

- Should self-driving cars be **on the road**?
  - *If so, under what conditions?*
- How can we **teach ethics** to self-driving cars?
  - *Using “value alignment”?*

AVs on the Road

- **Utilitarian** principle
  - *If there is any reasonable possibility of improving traffic safety…*
    - There is a **strong utilitarian imperative** to develop the technology.

AVs on the Road

- **Utilitarian** principle
  - *If there is any reasonable possibility of improving traffic safety…*
    - There is a **strong utilitarian imperative** to develop the technology.
  - *How about mishaps along the way?*
    - AV development still passes utilitarian test if it maximizes **discounted long-term benefit**.
    - The utilitarian principle takes account of future generations, after suitable discounting to reflect uncertainty, etc.

AVs on the Road

- **Utilitarian** principle
  - *If there is any reasonable possibility of improving traffic safety…*
    - There is a **strong utilitarian imperative** to develop the technology.
  - *How about mishaps along the way?*
    - AV development still passes utilitarian test if it maximizes **discounted long-term benefit**.
    - The utilitarian principle takes account of future generations, after suitable discounting to reflect uncertainty, etc.
  - But we must satisfy **other principles** in the meantime.

AVs on the Road

- **Autonomy** principle
  - *This is the **big one**.*
  - *We are rationally constrained to believe that experimental AVs will cause **injury and death**.*
    - They **already do**.
  - *So, AVs on the road **violate** the autonomy principle…*
    - Unless we can show that victims **give informed consent to the risk**.

AVs on the Road

- **Autonomy** principle
  - *We **already consent** to risk posed by traffic accidents*
    - …whenever we get into a car on walk on a street
    - …assuming that drivers exercise a normal degree of caution.
    - We **know** that cars are dangerous.

AVs on the Road

- **Autonomy** principle
  - *We **already consent** to risk posed by traffic accidents*
    - …whenever we get into a car on walk on a street.
    - …assuming that drivers exercise a normal degree of caution.
    - We **know** that cars are dangerous.
  - *But do we consent to risk posed by AVs?*
    - We don’t necessarily know there are AVs on the road.
    - So, maybe we don’t consent to the risk they pose.

AVs on the Road

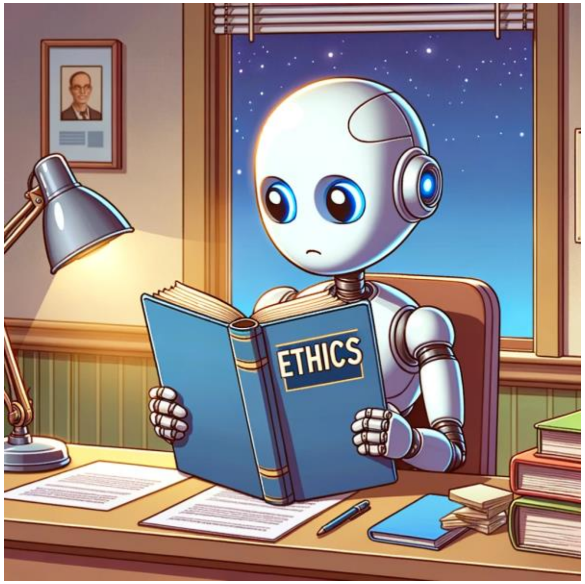
- **Autonomy** principle
  - *We **already consent** to risk posed by traffic accidents*
    - …whenever we get into a car on walk on a street.
    - …assuming that drivers exercise a normal degree of caution.
    - We **know** that cars are dangerous.
  - *But do we consent to risk posed by AVs?*
    - We don't necessarily know there are AVs on the road.
    - So, maybe we don't consent to the risk they pose.
  - *However, if AVs pose **no greater risk** than other cars…*
    - …then we consent to the **level of risk** posed by their presence.
    - This enough to **pass the autonomy test**.

AVs on the Road

- **Conclusions**
  - ***Utilitarian*** *principle – There is a strong imperative to* ***develop AVs***
    - …and test them **on the road** when necessary
    - …if there is a **reasonable chance** of future success.
    - But we must satisfy **other principles** in the meantime.
  - ***Autonomy*** *principle – Experimental AVs must be* ***no more dangerous*** *than other traffic.*
    - More precisely, we must not be rationally constrained to believe otherwise.
    - This guideline that can apply to **technology development in general**.

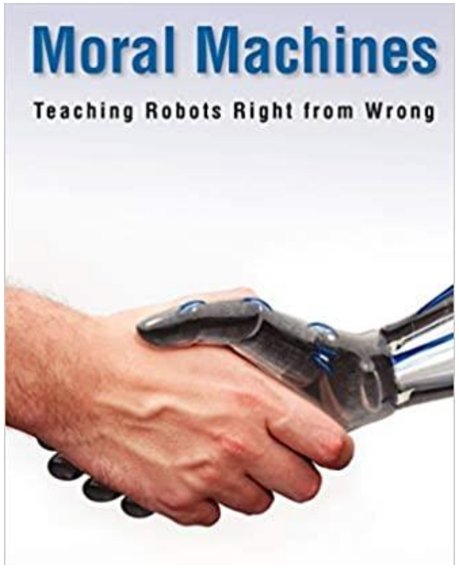
Teaching ethics to machines

- How do we teach AVs to drive ethically?
- AI community immediately saw it as a problem of **value alignment**.



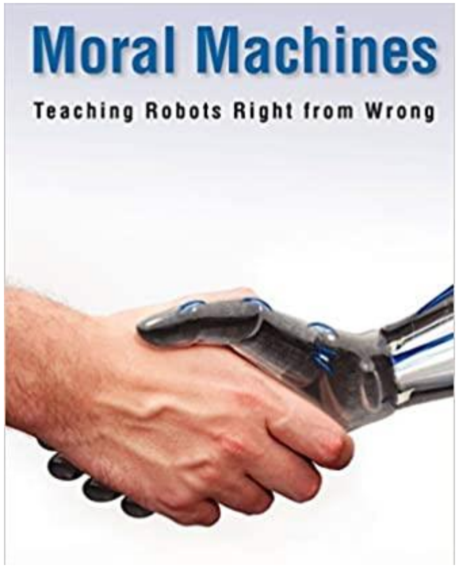
Value alignment

- Value alignment tries to teach **ethics** to **machines**.
  - *“**Align**” machine values with human values.*
  - *Based on **crowd sourcing**.*



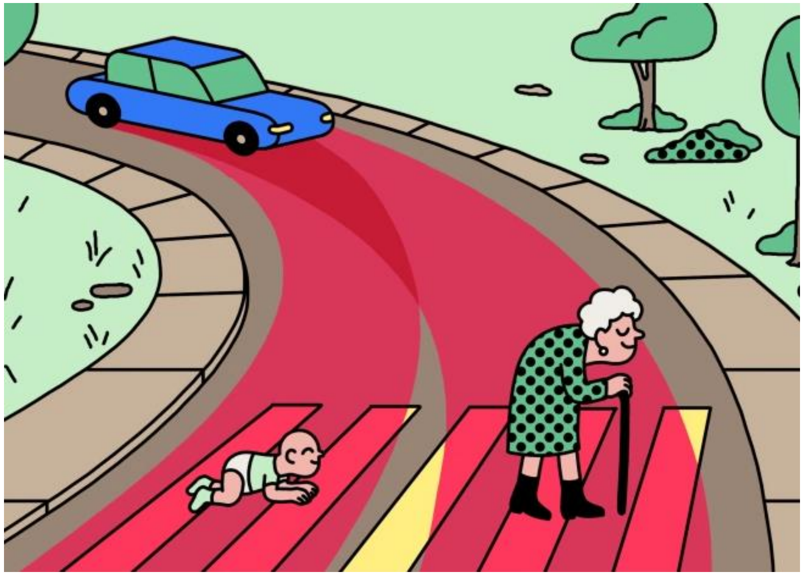
Value alignment

- Value alignment tries to teach **ethics** to **machines**.
  - *“**Align**” machine values with human values.*
  - *Based on **crowd sourcing**.*
- Problem:
  - *“Values” is **ambiguous**.*
    - What humans value (fact)
    - What is valuable (ethics)
  - *Value alignment trades on this ambiguity*.



The Moral Machine

- Developed by MIT’s Media Lab
  - *Crowd-source 1000s of responses to trolley-car type driving dilemmas*
  - *Derive **ethical rules for self-driving car.***

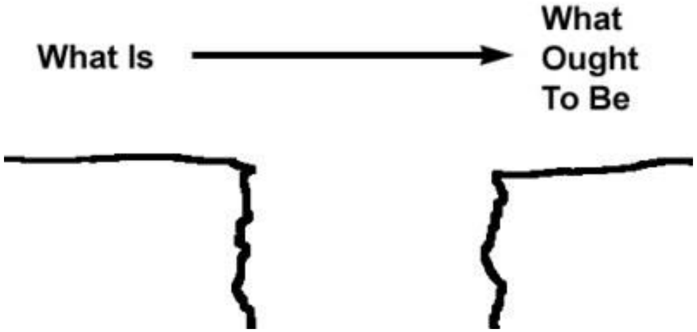


The Moral Machine

- Two problems:
    - *This type of dilemma rarely if ever occurs in practice.*
        - People don’t have meaningful “values” for such cases.

The Moral Machine

- Two problems:
  - *This type of dilemma rarely if ever occurs in practice.*
    - People don’t have meaningful “values” for such cases.
  - This commits **naturalistic fallacy**.
    - *We can’t infer “values” from “values.”*
    - *We can’t infer* ***ethical driving rules*** *from driving* ***opinions*** *and* ***behavior****.*



What Is

What Ought To Be

Value alignment

- To avoid naturalistic fallacy:
  - *We need an **ethical premise**.*
  - *Such as, "We should drive the way most people think we should drive."*
- No such premise seems reasonable.
  - *Designers of Moral Machine had 2nd thoughts.*

> "A word of warning: the preferences we found are not meant to instruct car programmers as to how they *should* regulate AVs.... The public can be ill-informed and biased, and some of the preferences we report are troubling."
>
> Edmond Awad, "Your (future) car's moral compass," *Behavioral Scientist,* Feb 11, 2019.

Value alignment

- There is no substitute for ethical principles.
  - *Driving practices and norms are **relevant**, of course.*
  - *But they alone don’t determine what is ethical.*

Value alignment

- Ethical principles can be **incorporated** into AI technology.
  - *For example, by using **rule-based** AI – already a trend.*
  - *We know how to build huge, complicated rule bases.*
  - *Non-self-driving cars are already regulated by >100,000 lines of code.*

©2001 HowStuffWorks