Welcome!

Moral Decision-making in Robotics

Rohan Chaudhari
IR Seminar
16-12-2019

https://www.facebook.com/photo.php?fbid=2547509228674030&set=gm.1450648995101979&type=3&theater

Outline

- What is “moral” decision making?
- Why is it important?
- What’s my goal here?

5 mins

Kinds of machine morality:

Ethical Law

3 mins

Kinds of machine morality:

Machine Learning

3 mins

Research:
A Computational Model of Commonsense Moral Decision-making

10 mins

Future work and Closing Thoughts

4 mins

https://media.giphy.com/media/6901DbEbbm4o0/giphy.gif

What is “moral” decision making?

- Multiple courses of action to choose from
- Decision is based on qualitative judgements

Why do we care?

http://www.thecomicstrips.com/subject/The-Ethical-Comic-Strips-by-Speed+Bump.php

- Clear ethical goals give direction
- Can we? ≠ Should we?
- Safeguards are good, but can we be proactive?

What's my goal here?

I will not:

- Delve into AI and existential risk...but come find me later!
- Argue for/against any decision-making strategy

I will (try to):

- Show how nuanced this topic is
- Explain how current decision-making strategies work
- Show why these strategies fall short
- Present avenues for further work

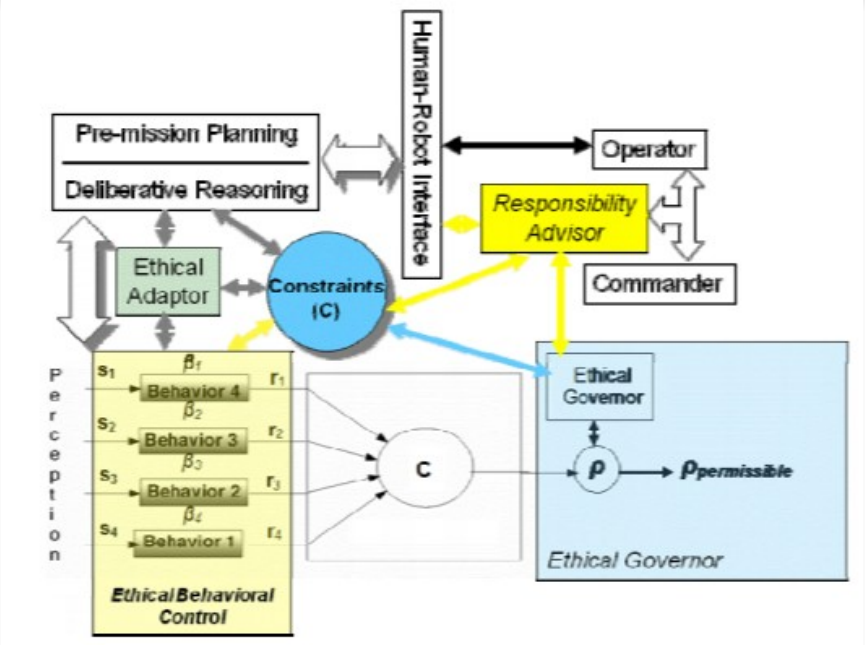
Kinds of Machine Morality

- **Operational** —> Preprogrammed responses for specific scenarios **(not "intelligent")**
- **Functional** —> Perform reasoning based on set of laws/rules
- **Full** —> Learn from prior actions and develop a moral compass

https://robotise.eu/wp-content/uploads/2018/02/robot-ethics-3.jpg

Kinds of Machine Morality: Ethical Law

- Give the robot guidelines for what it can/cannot do
- Top-down approach
- Early intelligent systems used this approach
  - “Ethical Governor” by Arkin et al. [1]

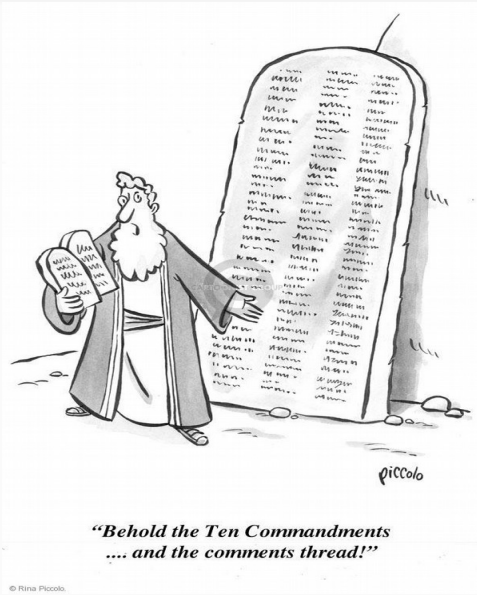


[1]

Kinds of Machine Morality: Ethical Law

Problems with this strategy:

- Raises more social and philosophical issues than it solves
- Makes dilemmas black and white
- Which ethical law do you follow?
  - There is no “universal” value system
    —> Moral imperialism

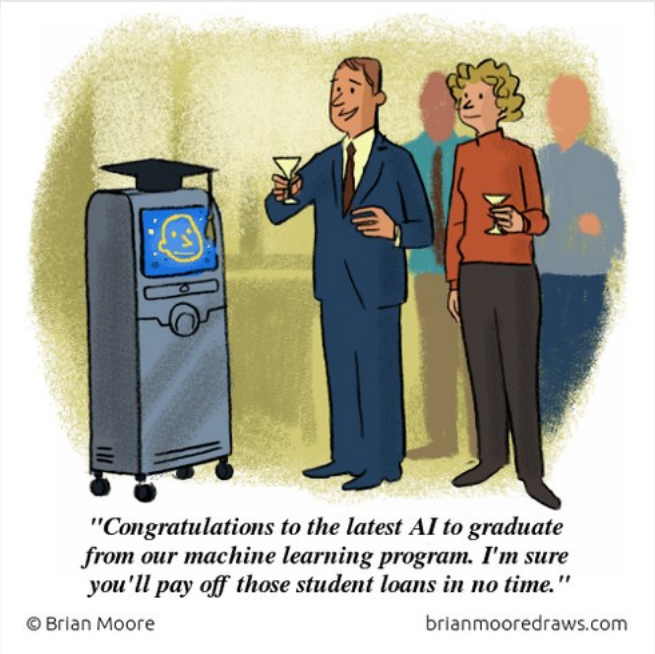


http://www.cartoonistgroup.com/properties/piccolo/art_images/cg52484c367907a.jpg

Kinds of Machine Morality: Ethical Law

...and perhaps the biggest problem of them all:

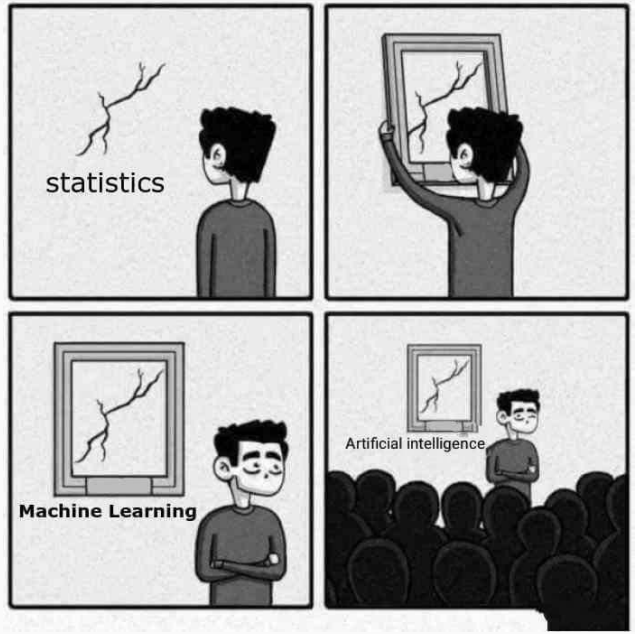
- Makes robots decide like humans
  - but we do not expect them to, as Malle et al. [2] point out
  - we want robots to do things and get the answers that we cannot; applying our normative views on robots only hinders this endeavor



https://img.deusm.com/informationweek/2016/03/1324681/ubm0313machineloan_final.png

Kinds of Machine Morality: Machine Learning

- This is the frontier in decision-making today
- Bottom-up approach
- Make decisions using inductive logic
  - The goal is not to find a right decision, but to eliminate the wrong ones



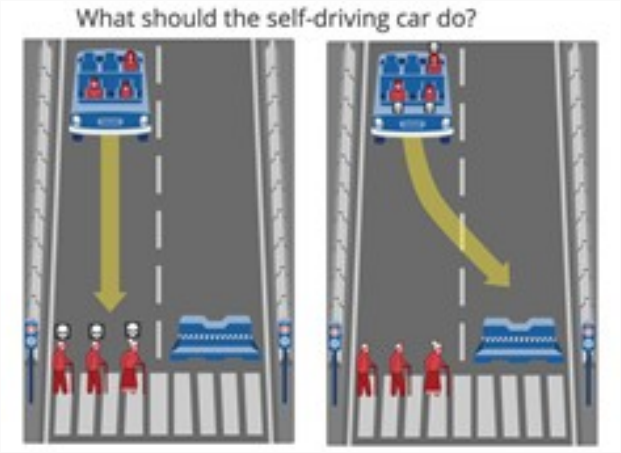
https://miro.medium.com/max/700/1*x7P7gqjo8k2_bj2rTQWAfg.jpeg

Research: A Computational Model of Commonsense Moral Decision-making [CMCMD] by Kim et al. (MIT 12/01/2018) [3]

- Key idea: incorporate people's moral preferences into informative distributions that encapsulate scenarios where decisions need to be made
  - Heavily context dependent
- Goal is to develop a "moral backbone"
  - The means, and not just the end, is of value
  - Instead of a greedy algorithm, relies on Bayesian dynamic statistical analysis

Research: CMCMD
The Data

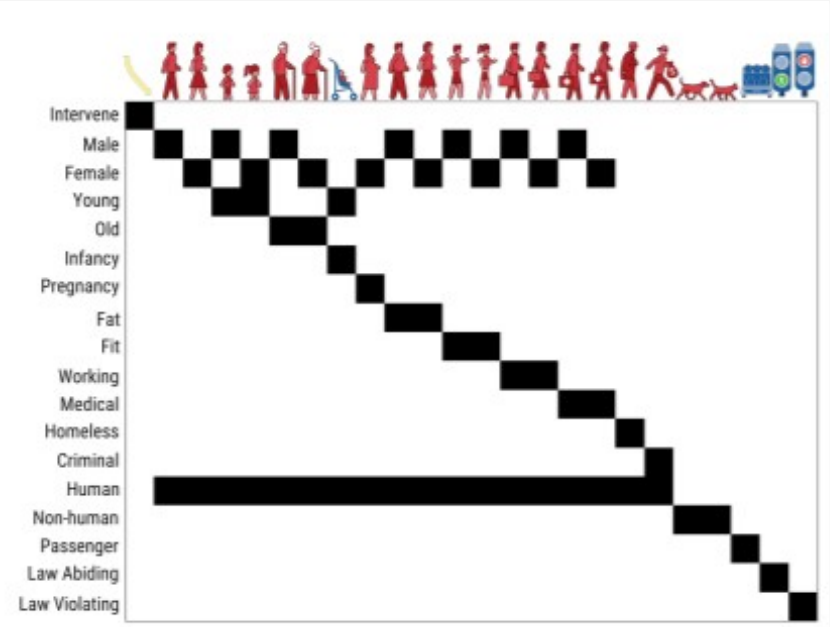
- Uses MIT's Moral Machine Dataset
  - **30 million gamified responses for various "trolley problem" binary scenarios**
  - characters have abstract features stored in a binary matrix
  - responses are not lab-controlled
  - responses themselves are unanalyzed/unqualified



Moral Machine interface. An example of a moral dilemma that features an AV with sudden brake failure, facing a choice between either not changing course, resulting in the death of three elderly pedestrians crossing on a "do not cross" signal, or deliberately swerving, resulting in the death of three passengers; a child and two adults. [3]

Research: CMCMD
The Data

- Uses MIT's Moral Machine Dataset
  - 30 million gamified responses for various "trolley problem" binary scenarios
  - **characters have abstract features stored in a binary matrix**
  - responses are not lab-controlled
  - responses themselves are unanalyzed/unqualified



[3]

Research: CMCMD
The Data

- Uses MIT's Moral Machine Dataset
  - 30 million gamified responses for various "trolley problem" binary scenarios
  - characters have abstract features stored in a binary matrix
  - **responses are not lab-controlled**
  - **responses themselves are unanalyzed/unqualified**

Research: CMCMD
2. Learning Strategy

- Goal is not to develop a “wire-heading” algorithm that maximizes utility
- Goal is a “virtuous” machine
  - Bayesian model that constantly updates decision function with new information
  - The utility value of a state: $u(\Theta_i) = w^\top F(\Theta_i)$
  - The better choice in the scenario is based on sigmoid function of net utility:

$$P(Y = 1|\Theta) = \frac{1}{1 + e^{-U(\Theta)}}$$

where

$$U(\Theta) = u(\Theta_1) - u(\Theta_0).$$

Research: CMCMD
3. Making Predictions

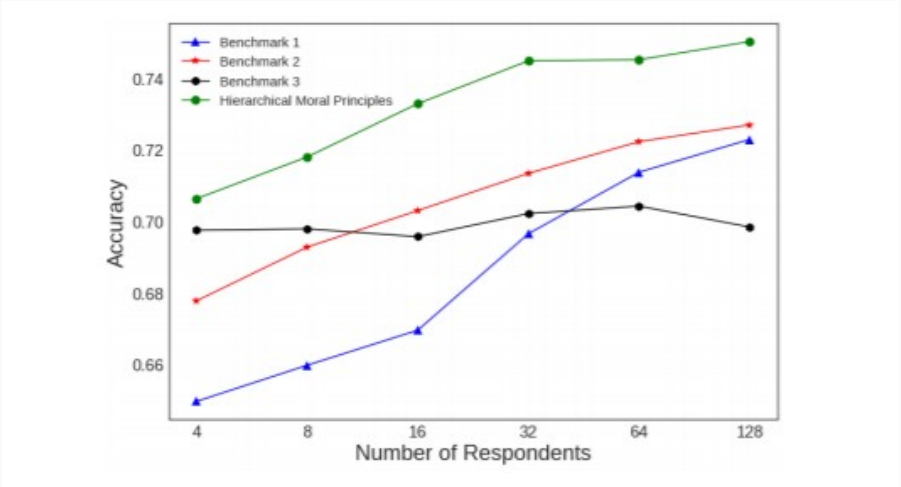
- Let Σ represent the covariance matrix that represents differences in responses over abstract principles
- Let w be the set of abstract principles learned from N responses
- Let Y be the decision made by the respondent
- Let Θ represent the state from T scenarios

Given this, the posterior distribution: $P(\boldsymbol{w}, w^g, \Sigma^g | \boldsymbol{\Theta}, \boldsymbol{Y}) \propto P(\boldsymbol{\Theta}, \boldsymbol{Y} | \boldsymbol{w}) P(\boldsymbol{w} | w^g, \Sigma^g) P(w^g) P(\Sigma^g)$

And the likelihood of decisions: $P(\boldsymbol{\Theta}, \boldsymbol{Y}) | \boldsymbol{w}) = \prod_{i=1}^{N} \prod_{t=1}^{T} p_{ti}^{y_i^t} (1 - p_{ti})^{(1-y_i^t)}$

Research: CMCMD
4. Getting Results

- Trained algorithm over 5000 samples, of which 1000 were tuning samples
- Compared results against
  - Benchmark 1 —> Pre-defined moral principle $u(\Theta) = w^{c\top}\Theta$
  - Benchmark 2 —> Multiple equally weighted abstract principles
  - Benchmark 3 —> Greedy algorithm where the values of one agent give no insight into the values of another



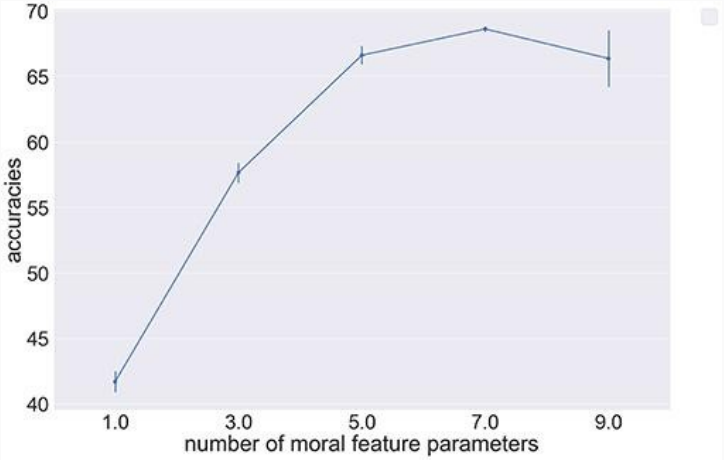
[3]

Research: CMCMD Discussion

- Issues with Dataset
  - Sivill [4] posits using Autonomous Vehicle Study Dataset (much smaller) which has lab-controlled data collection for more reliability
- Issues with the decision strategy
  - Abstract features are equally weighted —> is this how it should be?
  - Is learning the decisions people make in a scenario enough to understand how people make decisions?
- Issues with run-time

Research: Ethical and Statistical Considerations in Models of Moral Judgements by Sivill (University of Bristol 16/08/2019) [4]

- Recreates Kim's experiment with the Autonomous Vehicle Study Dataset
  - much smaller (216 responses)
  - lab-controlled survey
- Tries to apply Kim's model to new domains
  - main challenge is revamping the character vectors
  - found that the accuracy starts falling as the number of indefinite parameters increases past 7



[4]

General Discussion: Machine Learning

- Inductive logic is a process of elimination that gives us a “likely” choice
  - not necessarily the “right” choice
- Context specific
- Big-Data will always have shortcomings
- Real decision-making is not linear
  - Need more advanced strategies to emulate cognitive deliberation

So where does this leave us?

- We are far, far, far, far away from implementing full moral agency
  - Many scientists and philosophers believe General AI is unattainable

- Machine Morality today tries to model specific, isolated scenarios to make individual judgements
  - But even this is extremely challenging

Possible Avenues for Future Work

- Accurate, scenario-encompassing data-collection
  - Using real-world sources like traffic cameras —> ...more ethical concerns?
- When should the robot act and when should it be a bystander?
- How does a robot adapt to a fluid moral landscape?
- Hybrid approach that combines top-down and bottom-up strategies
- Combining intelligent decision-making with quantum-computing

Summary

- Why ethics and moral decision-making matter
- The ways in which robots can make decisions
- Ethical law and how it falls short
- Research that shows how ML is the more promising option
- Discussed the shortcomings of ML and some avenues for future work

References

1. Arkin, Ronald C., Patrick Ulam, and Brittany Duncan. “An Ethical Governor for Constraining Lethal Action in an Autonomous System:” Fort Belvoir, VA: Defense Technical Information Center, January 1, 2009. https://doi.org/10.21236/ADA493563.
2. Malle, Bertram F., Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. “Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents.” In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI ’15, 117–24. Portland, Oregon, USA: ACM Press, 2015. https://doi.org/10.1145/2696454.2696458.
3. Kim, Richard, Max Kleiman-Weiner, Andres Abeliuk, Edmond Awad, Sohan Dsouza, Josh Tenenbaum, and Iyad Rahwan. “A Computational Model of Commonsense Moral Decision Making.” ArXiv:1801.04346 [Cs], January 12, 2018. http://arxiv.org/abs/1801.04346.
4. Sivill, Torty. “Ethical and Statistical Considerations in Models of Moral Judgments.” Frontiers in Robotics and AI 6 (August 16, 2019): 39. https://doi.org/10.3389/frobt.2019.00039.

Thank You!

I’m no expert, but if this topic fascinates you, check out:

- Martin Heidegger- The Question Concerning Technology
- Isaac Asimov- Foundation
- Nick Bostrom- Superintelligence
- John Leslie Mackie- Inventing Right and Wrong
- Hubert Dreyfus- Thinking in Action: On the Internet
- David Kaplan- Readings in the Philosophy of Technology