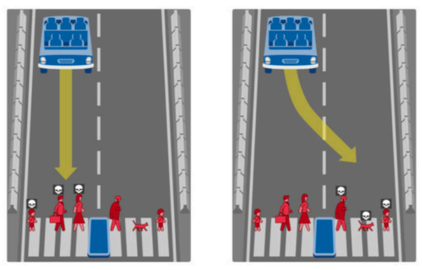
Towards Ethical Moral Machines

Torty Sivill

Supervisor: Carl Henrik EK

Motivation

- Can computational models tell us more about human moral theory?
- Do we need intelligent systems to make ethical decisions? Do we **want** them to?

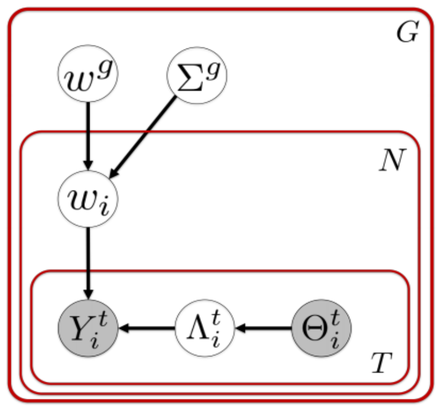


Research Aims

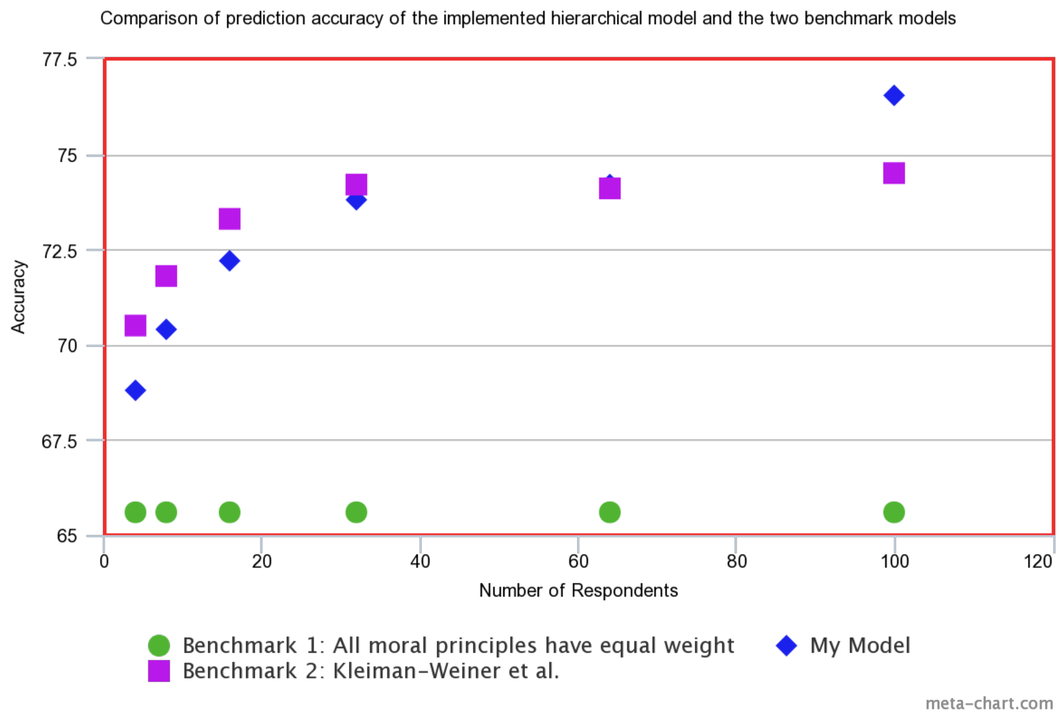
- **Evaluate** current models of common sense moral theory
- **Implement** a computational model of commonsense moral decision making to be tested on extended trolley problem dataset
- **Extend** the model to address ethical challenges

A Computational Model for Learning Moral Theory

- Model describes moral dilemmas as utility functions and social structures as a hierarchical Bayesian model
- Model uses autonomous vehicle domain to learn societal preferences over abstract features
- Bayesian inference implemented in PyMC3 using Markov chain Monte Carlo methods



Intermediate Results



How Can We Learn Moral Theories Without a Ground Truth?

- Use computational social choice to aggregate a society's preferences over ethical data
- Use K Mixture Models to aggregate individual moral principles ensuring global utility maximization

Does Response Time Reflect Moral Confidence?

- Incorporate a drift diffusion model to expand implementation to describe confidence and error in moral decision making
- Use the "two systems of moral judgment" model to reason about the use of response time to weight decision making

Reasons as Defaults

- Discuss the Generalism vs. Particularism argument and use default logic to argue that the two are not mutually exclusive.
- How could default theory be applied to create a system capable of reasoning about decisions?

Beyond the Trolley Problem

- Use Non-Parametric Bayesian processes to extend the mechanism that maps observable data onto the abstract feature space