Ethical Paternalism

Derek Lam, Mark Rucker

Philosophy Meets Mathematical Models

Cyber Human Systems

- Fitting numbers to human environments
- Digital machines have more power to influence the world
- People experience much of their world through digital systems

Ethics of Action and Autonomy

- Ethics about search engines has been focusing almost exclusively on biases
- Forgotten problem of paternalism: algorithms making choices for our best interest (restricting our option is a choice made)
- Respect of autonomy: consent is necessary but not enough (analogy: consenting yet unhealthy/abusive relationship between humans)

Gaps Within Machine Learning

- Questions about high level ethics
- Questions about what can be done
- Questions about how accurate results are
- Questions about what we want to do

- Fewer questions about how to do those things

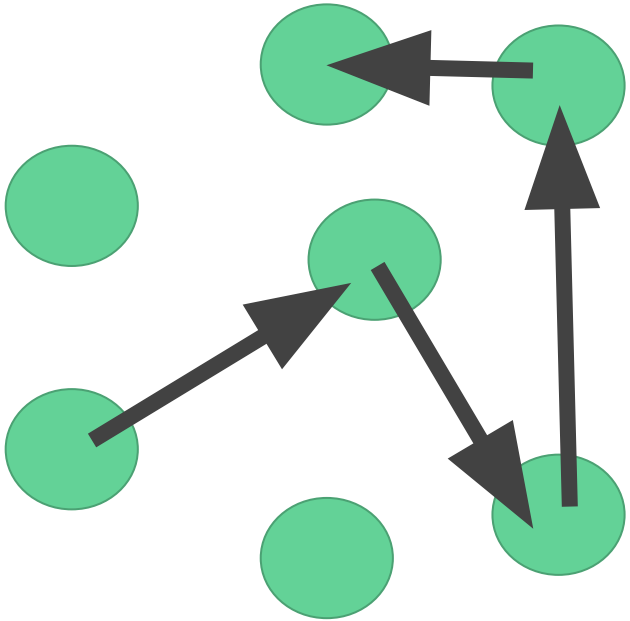
State Space

Primary Question

Where is the user most likely to go next?

- Content Filtering
  - External Data Feeds
- Collaborative Filtering
  - Latent Factor Models
  - Neighborhood Formations

State Space



Primary Question

Why has the user gone to places?

- Rational reward maximizer
  - Rational Choice Theory
  - Theory of Planned Behavior
- Implicit motivations
  - Implicit Association Tests
- Bounded Memory Models

What is Autonomy?

- Limiting Options and Respecting Autonomy
- Inadequacy of Simple Desire-Based Theory
- Mesh-Theories & Desire-Value Distinction (unwilling drug addict example)
- What is value?
  - Mental State
  - A kind of desire that the subject treats as justification for the desired action
  - Not all desires are values, e.g., desires for one's guilty pleasure satisfaction)
- Mark of value?
  - Willing to share publicly (Publicity of Justification)
  - Some stability over time (Private Language Argument)
  - Self-control

The Transparency Thesis

- Do *you* believe that 23 times 65 equals 1485? Do you believe that there is a goat outside?
- To answer these questions, do you “look inward”?
- According to many philosophers: No. [FYI: Brie Gertler (UVa) thinks yes]

- **Transparency Thesis**: (for the kind mental states that we can decide to have) We typically know about their existence by attending to the reasons/justifications for having them instead of “looking for them”

Transparency & Recommender System

- Knowing about mind: (1) predictive approach vs. (2) deliberative approach
- Richard Moran: “[O]nly if I can see my own belief as somehow ‘up to me’ will it make sense for me to answer a question as to what I believe about something by reflecting exclusively on that very thing.” (2001: 66)

- **Stronger Claim**: For a mental state **m**, if I consider whether or not I have **m** is “up to me”, my knowledge about whether or not I actually have **m** must be *largely* based on a *deliberative* approach.

Transparency & Recommender System

Applying to 3rd Person

- A recommender system that solely aims to ***predict*** a person's values based on past pattern alone is not engaging with the person's values as something that are "up to that person".
- 3rd Personal *Deliberative* Approach
  - Learning whether a person **S** has the mental state **m**: Going through the *reasons for and against* having **m** together with **S**.
  - Analogy: Dialogue between a healthy couple (e.g. South Park season 21 "Put It Down")
- Using personalized reason-oriented and deliberative questions to reach recommendations **with** the user

What We've Done

- Literature review
- Syncing ideas, expectations, vocabulary and vision
- Traveling to conferences in both disciplines to learn more
- Beginning to connect ideas to mathematical structures
    - Transition probabilities
    - Deep learning vs feature selection
    - State space representation
    - Reward learning rather than supervised learning algorithms