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
- Unrealistic option of opting out
- Respect of autonomy: consent is necessary but not enough (analogy: consenting yet unhealthy/abusive relationship between humans)

Acquiring Info about Actions

2 Epistemic Strategies for Learning about Two Kinds of Actions (Moran 2001)

[**Non-Autonomous Actions**] Suppose you want to know what you would do when you are drunk. Action patterns ➜ Prediction

[**Autonomous Actions**] Suppose you want to know whether you will get milk tonight. Pros vs. Cons for getting milk ➜ Decision

Predictive Strategy: for non-autonomous actions

Rational Engagement Strategy: for autonomous actions

Explanation: Value vs. Desire

[Willing Addict] Teddy is addicted to drugs. So he wants to do drugs. Furthermore, he is happy with the fact that he wants drugs. He thinks that those who judges drug users are snobs.

[Unwilling Addict] Jim is addicted to drugs. So he wants to do drugs. But he is not happy with the fact that he craves drugs. He wants to be the kind of person who doesn't want to do drugs.

Teddy is **autonomous** in the sense that we hold him morally accountable for taking drugs.

Jim is **non-autonomous** in the sense that we do not hold him morally accountable for taking drugs.

Explanation: Value vs. Desire

Autonomy as Value Driven Actions

Autonomy =/= being able to do otherwise

Autonomy =/= doing what one desires

Autonomy = acting in a way that is driven by one's values

What is it to value something? (Bratman)

One values X

= One desires X ***and*** considers that desire for X to be **a good reason** to obtain X.

Ethical Demands…

Respect Autonomy requires…

1. Use the Rational Engagement Strategy to learn about users' actions
2. Users' values/reasons must be part of the driving force behind the choices [we won't get to this]

Tracking reason **at least** consists of: (a) sensitivity to **order** of choices; (b) sensitive to what the user considers **valuable**.

Semi-resolved Issues…

1. First Person vs. Third Person
2. How to get a user's value to be part of the driving force?
3. Should **weakness of the will** be respected? Raz argued no, but it can be rather controversial.

   Consider this: It seems wrong to give the unwilling addict more drugs. And the explanation we offered is that it violates his autonomy. **BUT** we don't seem to feel the same way for some other cases, e.g., medical decisions. *Movie selection seems to belong to the former category.*

Gaps Within Machine Learning

- Questions about high level ethics
- Questions about what can be done
- Questions about accuracy of results
- Questions about what we want to do

- Fewer questions about how to do those things

Introduction to Inverse Reinforcement Learning

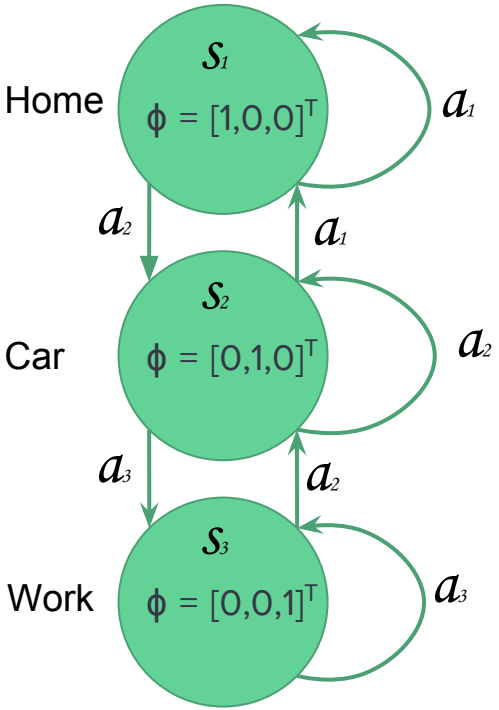
An algorithmic approach to determine environmental rewards for behavior

- Input
  - Markov decision process (MDP/R)
    - $\mathcal{S}$ -- States
    - $\mathcal{A}$ -- Actions
    - $\mathcal{T}$-- Transition Probabilities,
    - γ -- Discount Factor
  - Behavior traces
    - $\mathcal{D}$ -- Trajectories
  - Reward features (similar to state)
    - ϕ-- Feature Vector

- Outputs
  - Reward function, $\mathcal{R}$: $\mathcal{S}$ x $\mathcal{A} \mapsto \mathbf{R}$
  - Behavioral rules, π

The behavior rule outputs (π) are given to agents in the newly constructed MDP models

Quick Example

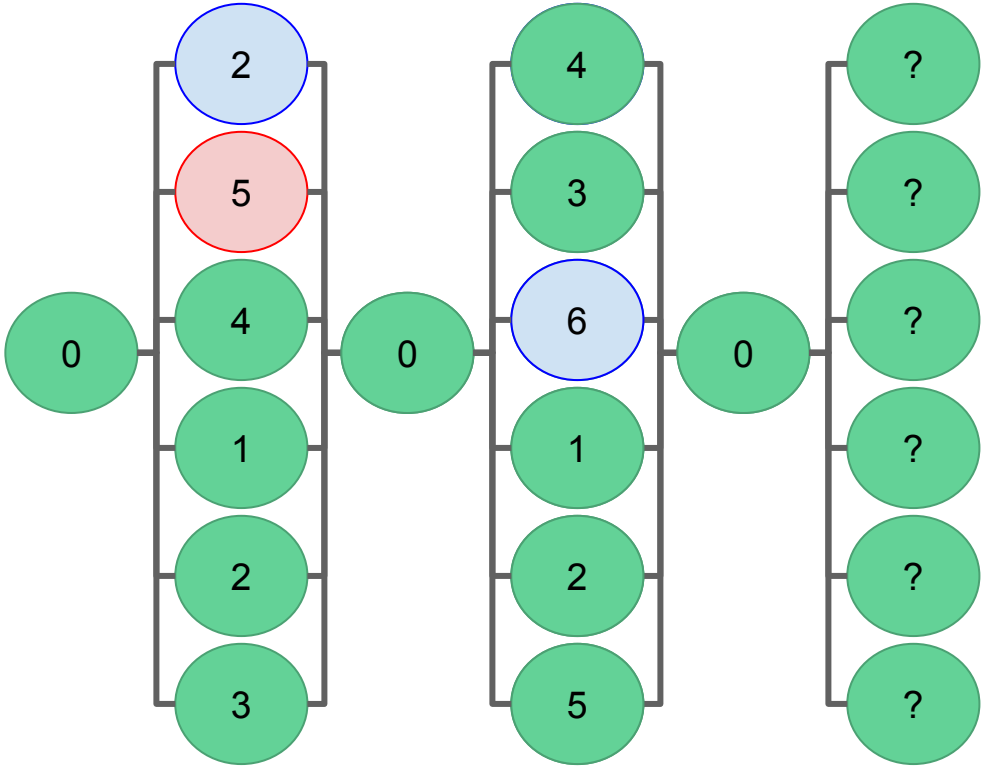


$$\mathcal{S} = \{s_1 = (\text{home}),\ s_2 = (\text{car}),\ s_3 = (\text{work})\}$$
$$\mathcal{A} = \{a_1 = (\text{go to home}),\ a_2 = (\text{go to car}),\ a_3 = (\text{go to work})\}$$
$$\phi(s) = [\phi_1 = (\text{is at home}),\ \phi_1 = (\text{is in car}),\ \phi_1 = (\text{is at work})]^T$$

$$\mathcal{R}(s) = w^T \phi(s)$$
$$\mu(s,\pi) = \sum_{a \in \mathcal{A}} \phi(s) \cdot \pi(s,a) \sum_{s' \in \mathcal{S}} \mathcal{T}(s,a,s') \cdot \gamma \mu(s',\pi)$$
$$V^\pi(s) = \sum_{a \in \mathcal{A}} \mathcal{R}(s) \cdot \pi(s,a) \sum_{s' \in \mathcal{S}} \mathcal{T}(s,a,s') \cdot \gamma V^\pi(s')$$

$$f(s,\pi) = \mathbb{E}[(\%\text{ time in } s_1), (\%\text{ time in } s_2), (\%\text{ time in } s_3) | s_0 = s, \pi]$$
$$\mu(s,\pi,\mathrm{T}) = \phi \cdot f(s,\pi) \cdot \mathrm{T}$$
$$V^\pi(s,\mathrm{T}) = w^T \mu(s,\pi,\mathrm{T})$$

Our MDP Structure and IRL



State Features

- Theater (3)
- Time of day (3)
- Day of week (7)
- Movie seen before (2)

|States| = 3x3x7x2 = 126

Natural Affinities To Human Thought

- Recent choices have more bearing on recommendations than previous
- Non-competing states don’t affect the rating of one another
- Learned model has a natural interpretation with human thought
- The interaction of features is important to the reward of a state

Next Steps

- Incorporate question: “Are you glad you watched that movie?”
- Provide feedback on how rewarding certain features are to a user
- Design and run a usability experiment comparing to other algorithms

Thank You!

2018 Kernel Projection ▼ Feedback

ons

Time
ouse

A Wrinkle in Time
Alamo Drafthouse
@20:30

W

Night
fthouse
05

ma

21:55

23:25

21:20

20:30 22:30

Death Wish

12:15 14:45 17:15 19:45 22:15

Regal Stonefield

Black Panther

13:00 14:40

21:20 22:30

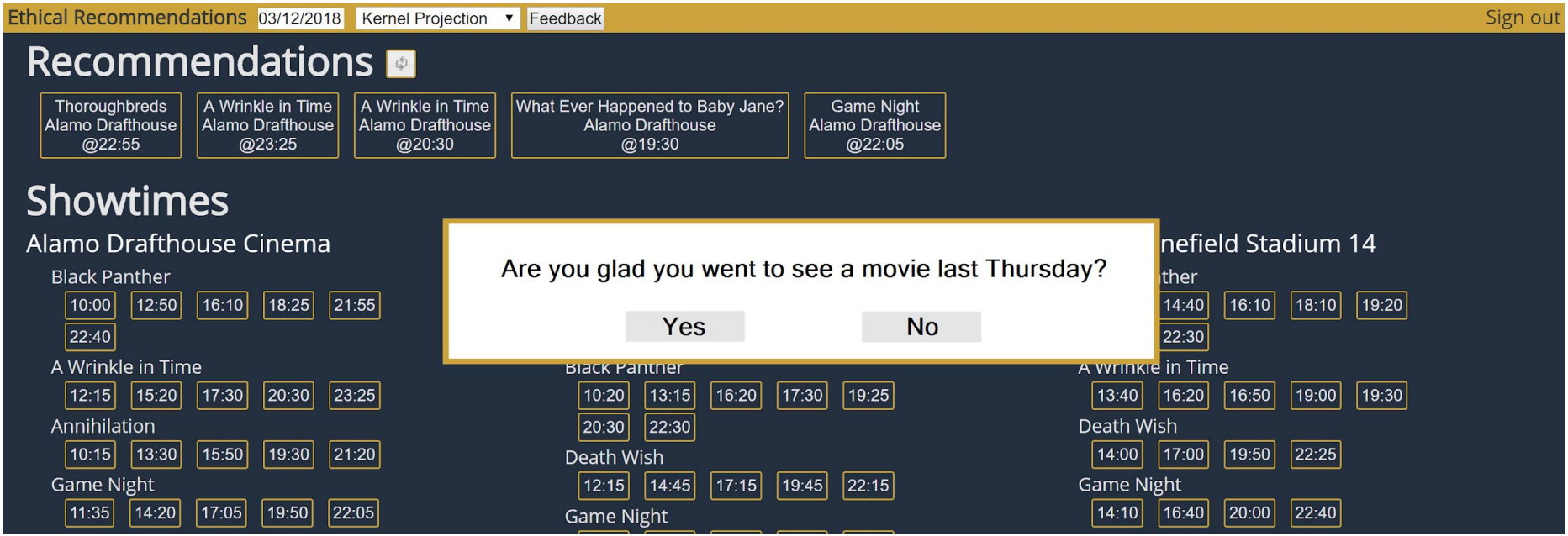
A Wrinkle in Time

13:40 16:20

Death Wish

14:00 17:00

Game Night


Ethical Recommendations
03/12/2018
Kernel Projection
Feedback
Sign out
Recommendations
Thoroughbreds
Alamo Drafthouse
@22:55
A Wrinkle in Time
Alamo Drafthouse
@23:25
A Wrinkle in Time
Alamo Drafthouse
@20:30
What Ever Happened to Baby Jane?
Alamo Drafthouse
@19:30
Game Night
Alamo Drafthouse
@22:05
Showtimes
Alamo Drafthouse Cinema
Black Panther
10:00 12:50 16:10 18:25 21:55
22:40
A Wrinkle in Time
12:15 15:20 17:30 20:30 23:25
Annihilation
10:15 13:30 15:50 19:30 21:20
Game Night
11:35 14:20 17:05 19:50 22:05
Are you glad you went to see a movie last Thursday?
Yes
No
Black Panther
10:20 13:15 16:20 17:30 19:25
20:30 22:30
Death Wish
12:15 14:45 17:15 19:45 22:15
Game Night
nefield Stadium 14
ther
14:40 16:10 18:10 19:20
22:30
A Wrinkle in Time
13:40 16:20 16:50 19:00 19:30
Death Wish
14:00 17:00 19:50 22:25
Game Night
14:10 16:40 20:00 22:40