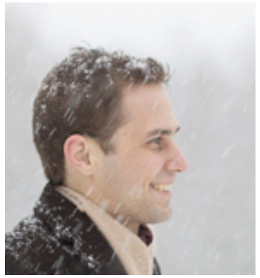
Learning the Preferences of Ignorant, Inconsistent Agents

Owain Evans (Oxford), Andreas Stuhlmueller (Stanford),
Noah Goodman (Stanford)

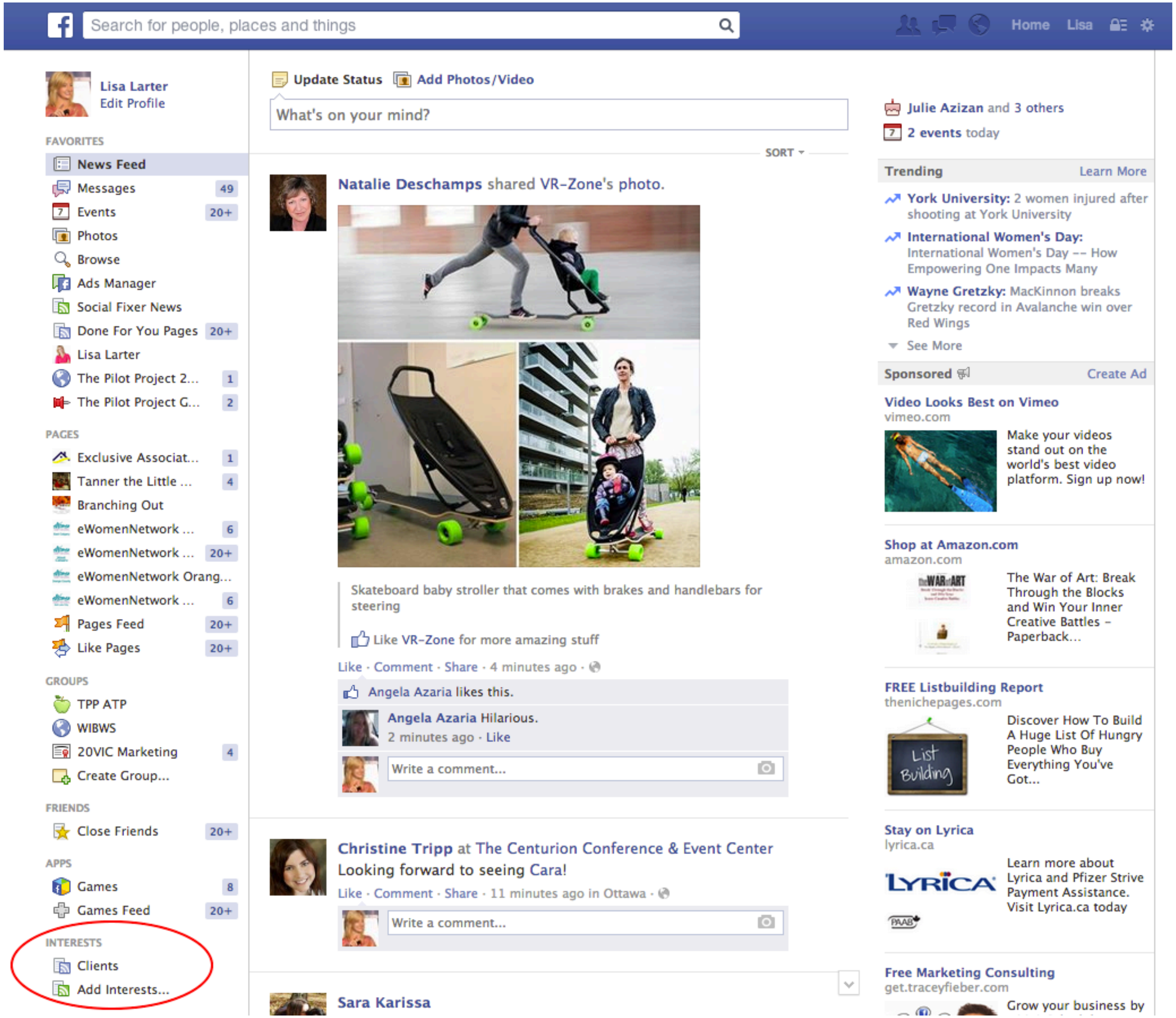


1. Motivation for learning human preferences

- Scientific (economics, psychology): how do people value work vs. leisure, short-term vs. long-term, country vs. friends & family?

- Machine learning (applications): recommendation (movie, job, dating), create tailored content.

- Machine learning (long-term goal): the more systems **understand** our preferences, the more they can help us make **high stakes** decisions in **novel** circumstances.



Search for people, places and things
Home
Lisa
Lisa Larter
Edit Profile
FAVORITES
News Feed
Messages 49
Events 20+
Photos
Browse
Ads Manager
Social Fixer News
Done For You Pages 20+
Lisa Larter
The Pilot Project 2... 1
The Pilot Project G... 2
PAGES
Exclusive Associat... 1
Tanner the Little ... 4
Branching Out
eWomenNetwork ... 6
eWomenNetwork ... 20+
eWomenNetwork Orang...
eWomenNetwork ... 6
Pages Feed 20+
Like Pages 20+
GROUPS
TPP ATP
WIBWS
20VIC Marketing 4
Create Group...
FRIENDS
Close Friends 20+
APPS
Games 8
Games Feed 20+
INTERESTS
Clients
Add Interests...
Update Status
Add Photos/Video
What's on your mind?
SORT
Natalie Deschamps shared VR-Zone's photo.
Skateboard baby stroller that comes with brakes and handlebars for steering
Like VR-Zone for more amazing stuff
Like · Comment · Share · 4 minutes ago ·
Angela Azaria likes this.
Angela Azaria Hilarious.
2 minutes ago · Like
Write a comment...
Christine Tripp at The Centurion Conference & Event Center
Looking forward to seeing Cara!
Like · Comment · Share · 11 minutes ago in Ottawa ·
Write a comment...
Sara Karissa
Julie Azizan and 3 others
2 events today
Trending
Learn More
York University: 2 women injured after shooting at York University
International Women's Day: International Women's Day -- How Empowering One Impacts Many
Wayne Gretzky: MacKinnon breaks Gretzky record in Avalanche win over Red Wings
See More
Sponsored
Create Ad
Video Looks Best on Vimeo
vimeo.com
Make your videos stand out on the world's best video platform. Sign up now!
Shop at Amazon.com
amazon.com
The War of Art: Break Through the Blocks and Win Your Inner Creative Battles - Paperback...
FREE Listbuilding Report
thenichepages.com
List Building
Discover How To Build A Huge List Of Hungry People Who Buy Everything You've Got...
Stay on Lyrica
lyrica.ca
LYRICA
Learn more about Lyrica and Pfizer Strive Payment Assistance. Visit Lyrica.ca today
Free Marketing Consulting
get.traceyfieber.com
Grow your business by

2. Learning preferences with IRL

Inverse Reinforcement Learning (AI) / Structural Estimation (Econ):

- Unsupervised learning, assumed model is MDP, POMDP, RL.
- Learn from sequences of choices in complex environments (cf. Netflix)
- Learn utility/reward function not policy: enduring cause not contingent effects.
- People act on their preferences without ability to report them quantitatively (driving skill, detailed vacation plan)

3. The problem of systematic error

- IRL: infer preferences from observed actions … assuming human fits (MDP/POMDP) model up to random (softmax) errors.

- But human make **systematic** errors! Person smokes every day but regrets it.

- Behavioral economics (hyperbolic discounting, Prospect Theory)

- Bounded cognition (forgetting, limited computational ability, etc.)

4. Learning from ignorant, inconsistent agents

Our approach:

1. build flexible generative models to capture a range of biases and cognitive bounds (while maintaining tractability)

2. jointly infer **biases** (or lack thereof) and **preferences** from behavior

3. if successful, can help humans overcome biases

5. Human bias: Time inconsistency

- Intuition: tonight you want to rise early but tomorrow you want to sleep in.
- Most prominent bias: addiction, procrastination, impulsiveness, will-power / pre-commitment.
- Formally, any non-exponential discounting implies time-inconsistency.

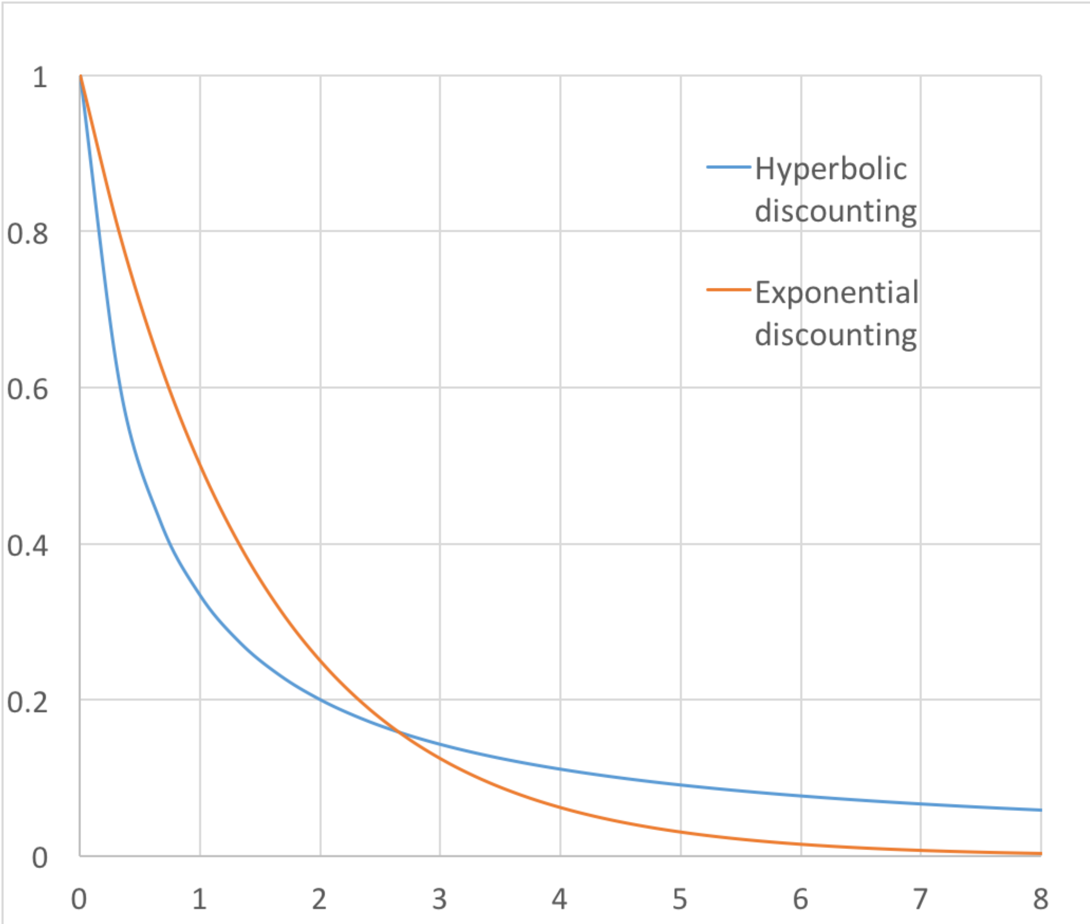
5. Human bias: Time inconsistency

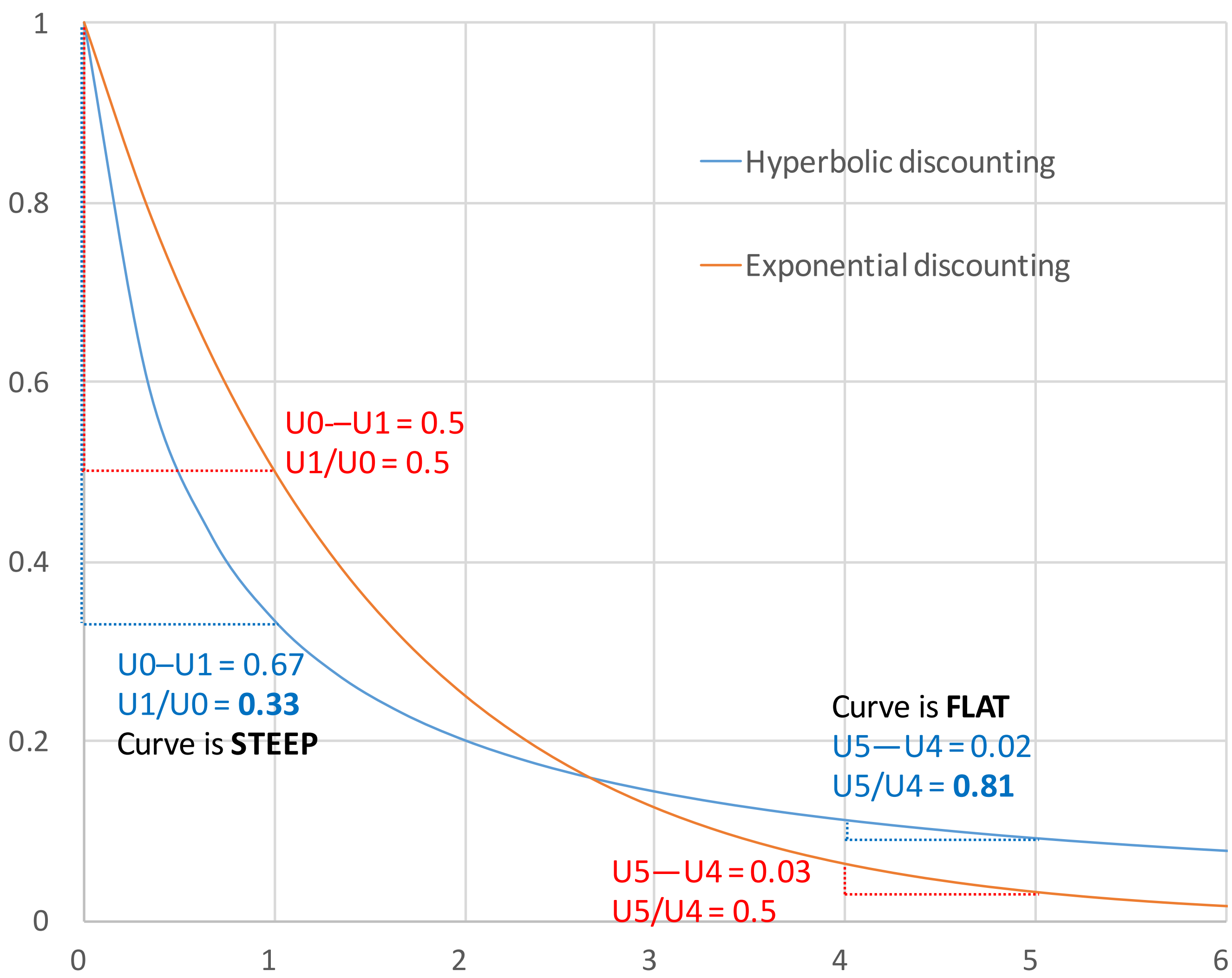
Hyperbolic discounting

Discount factor = 1/(1+kt)

At t=0, you prefer $80 at t=8 to $70 at t=7 (curve **shallow**)

At t=7, you re-evaluate and prefer $70 now to $80 tomorrow (curve **steep)**.




Hyperbolic discounting
Exponential discounting
U0-–U1 = 0.5
U1/U0 = 0.5
U0–U1 = 0.67
U1/U0 = **0.33**
Curve is **STEEP**
Curve is **FLAT**
U5—U4 = 0.02
U5/U4 = **0.81**
U5—U4 = 0.03
U5/U4 = 0.5
1
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
6

5. Model for biased agent

MDP model:

$$\mathrm{EU}_s\left[a\right] = U(s,a) + \mathop{\mathbb{E}}_{s',a'}\left[\mathrm{EU}_{s'}\left[a'\right]\right]$$

with $s' \sim T(s,a)$ and $a' \sim C(s')$

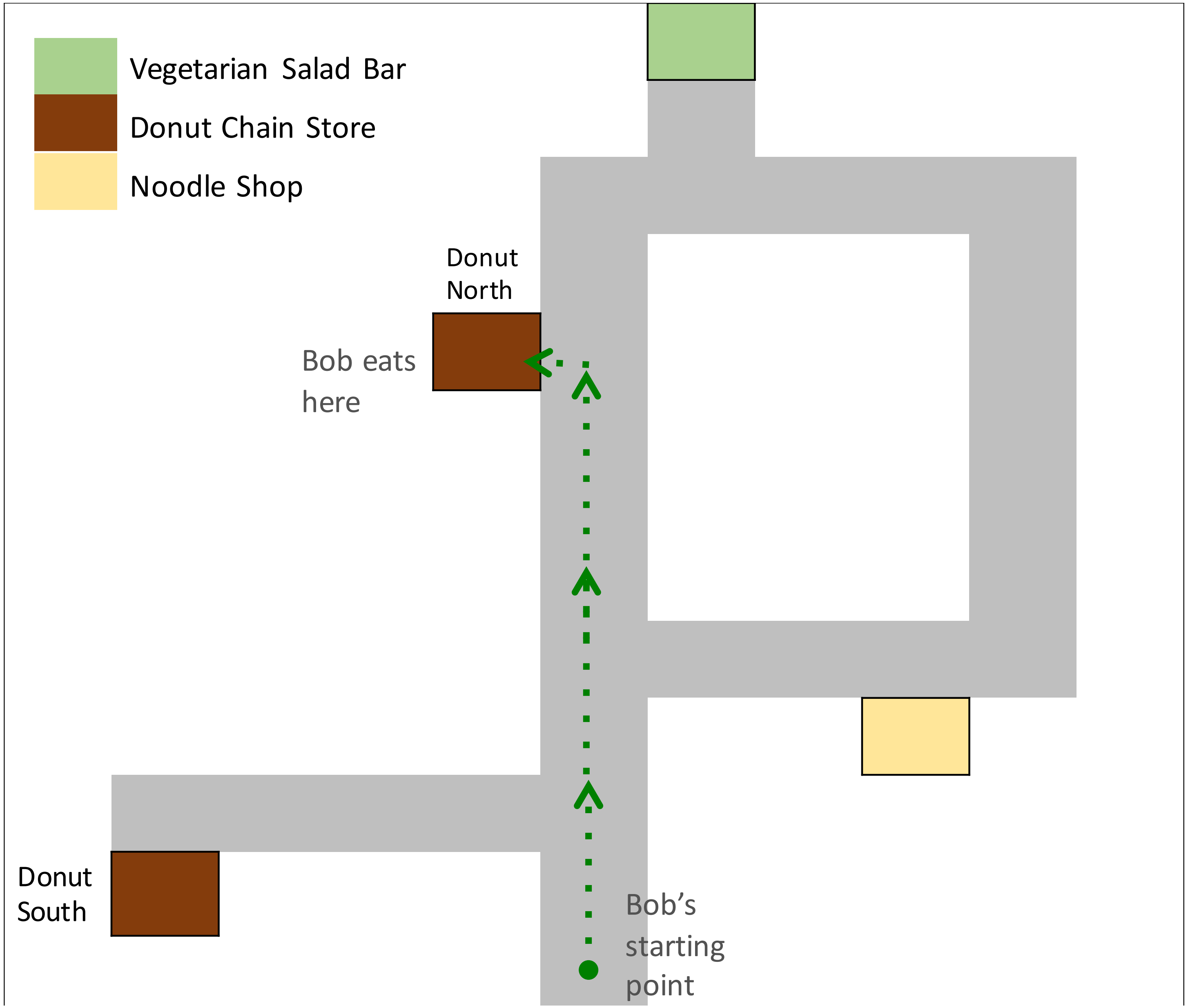
MDP + Hyberbolic discounting (variable *d* for "delay" measures how far in the future the action *a* would take place):

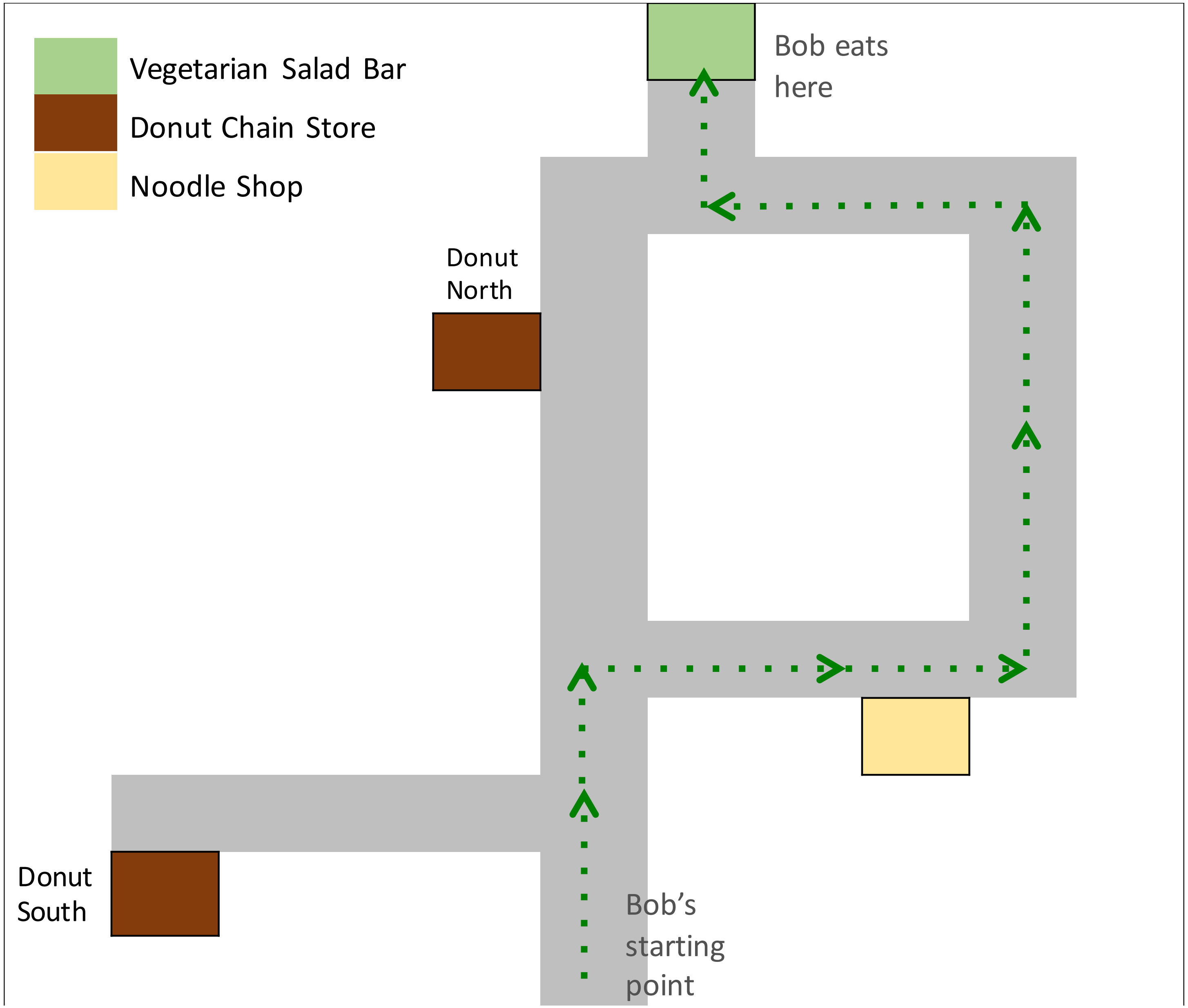
$$\mathrm{EU}_{s,d}\left[a\right] = \frac{1}{1+kd}U(s,a) + \mathop{\mathbb{E}}_{s',a'}\left[\mathrm{EU}_{s',d+1}\left[a'\right]\right]$$

6. Goal for examples and experiments

- Show that ignoring biases (assuming optimality) leads to mistakes in learning preferences

- Mistakes occur in simple, uncontrived, everyday scenarios.


Vegetarian Salad Bar
Donut Chain Store
Noodle Shop
Donut
North
Bob eats
here
Donut
South
Bob's
starting
point


Vegetarian Salad Bar
Donut Chain Store
Noodle Shop
Bob eats here
Donut North
Donut South
Bob's starting point

5. Model for biased agent - NAIVE

MDP model:

$$\mathrm{EU}_s[a] = U(s,a) + \mathop{\mathbb{E}}_{s',a'}[\mathrm{EU}_{s'}[a']]$$

with $s' \sim T(s,a)$ and $a' \sim C(s')$

MDP + Hyberbolic discounting (variable *d* for "delay" measures how far in the future the action *a* would take place):

$$\mathrm{EU}_{s,d}[a] = \frac{1}{1+kd}U(s,a) + \mathop{\mathbb{E}}_{s',a'}[\mathrm{EU}_{s',d+1}[a']]$$

$$a' \sim C(s', d+1)$$

5. Model for biased agent - SOPHISTICATED

MDP model:

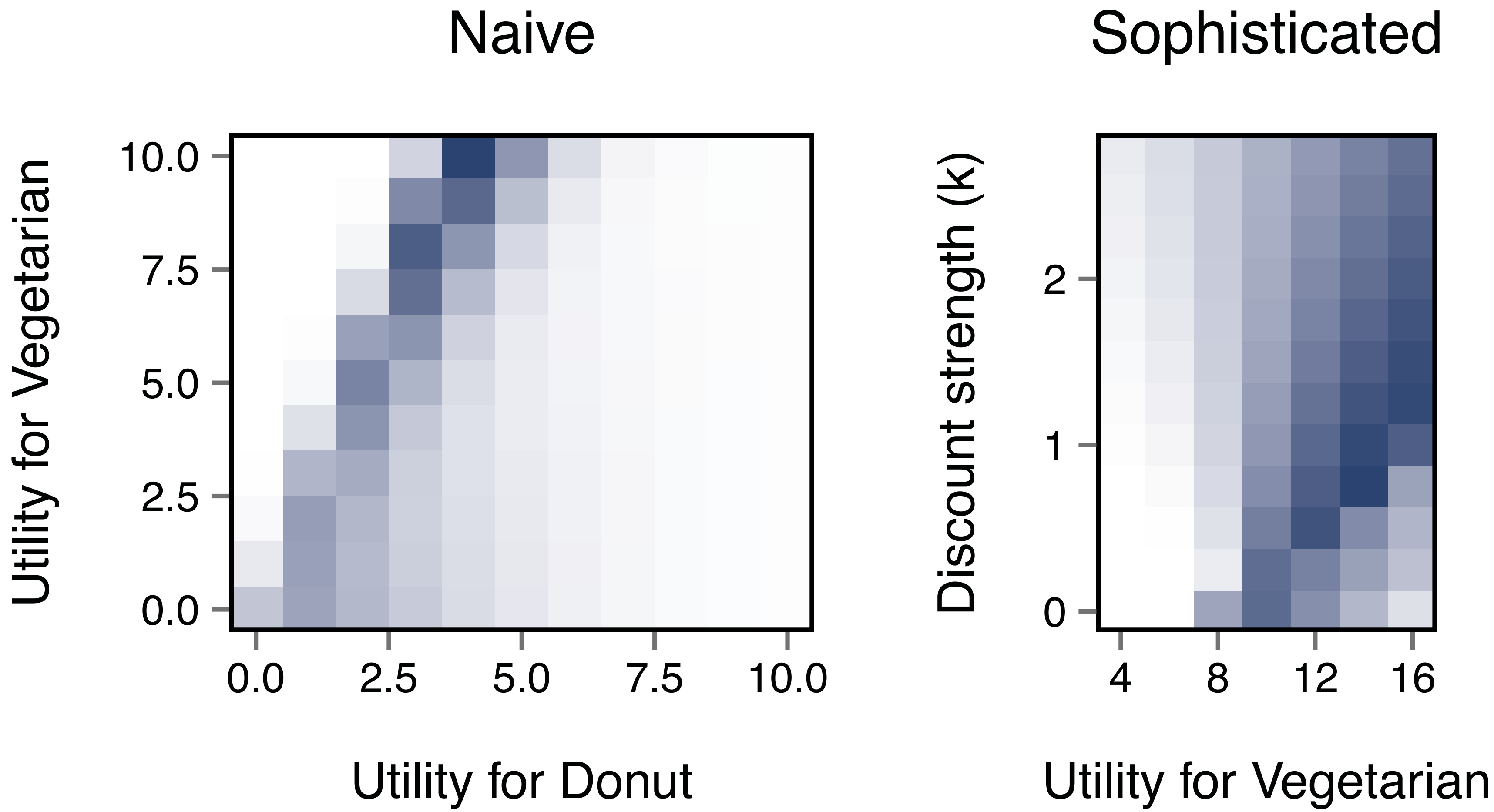
$$\mathrm{EU}_s[a] = U(s,a) + \mathop{\mathbb{E}}_{s',a'}[\mathrm{EU}_{s'}[a']]$$

with $s' \sim T(s,a)$ and $a' \sim C(s')$

MDP + Hyberbolic discounting (variable *d* for "delay" measures how far in the future the action *a* would take place):

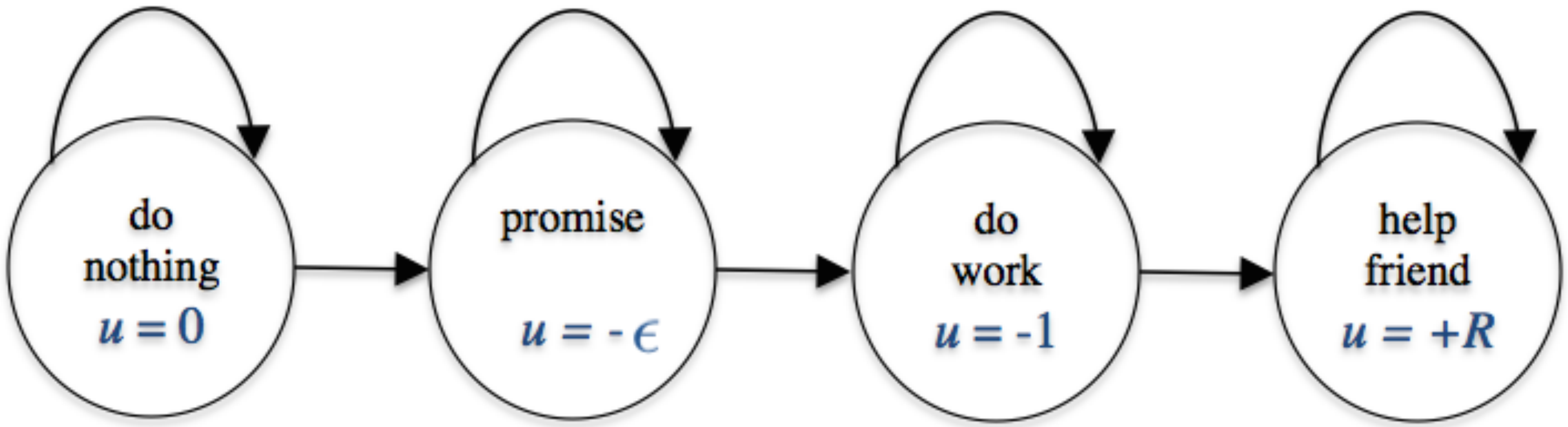
$$\mathrm{EU}_{s,d}[a] = \frac{1}{1+kd}U(s,a) + \mathop{\mathbb{E}}_{s',a'}[\mathrm{EU}_{s',d+1}[a']]$$

$$a' \sim C(s', 0)$$

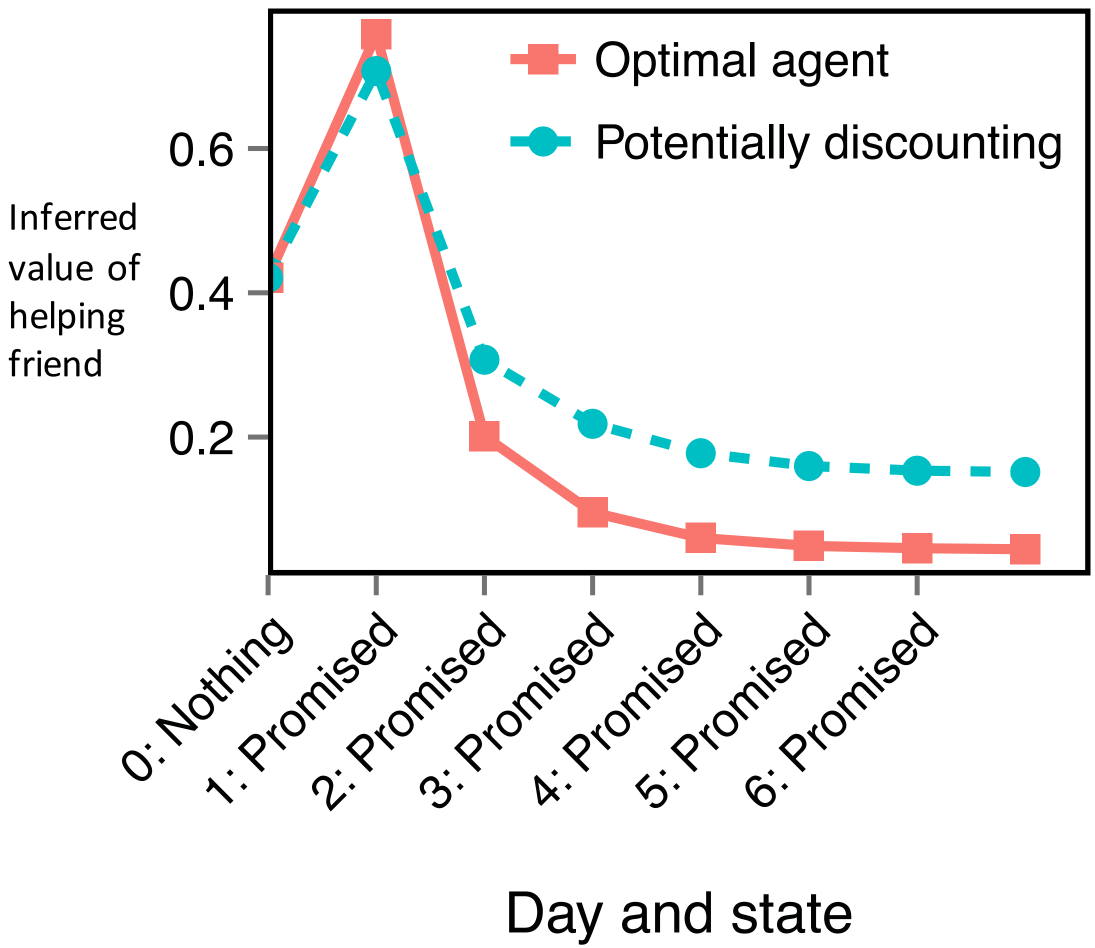


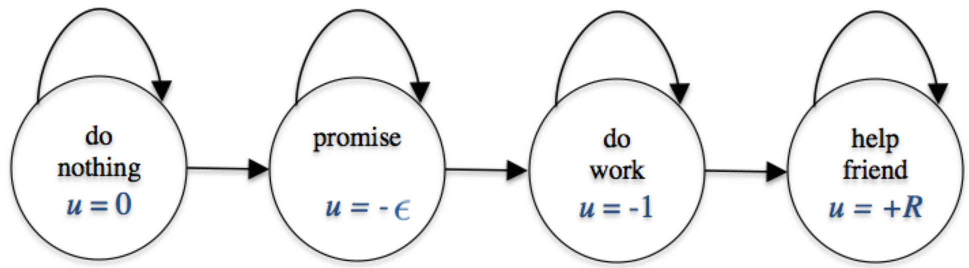
Naive
Sophisticated
Utility for Vegetarian
10.0
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
10.0
Utility for Donut
Discount strength (k)
2
1
0
4
8
12
16
Utility for Vegetarian

6. Model for biases agent: Procrastination



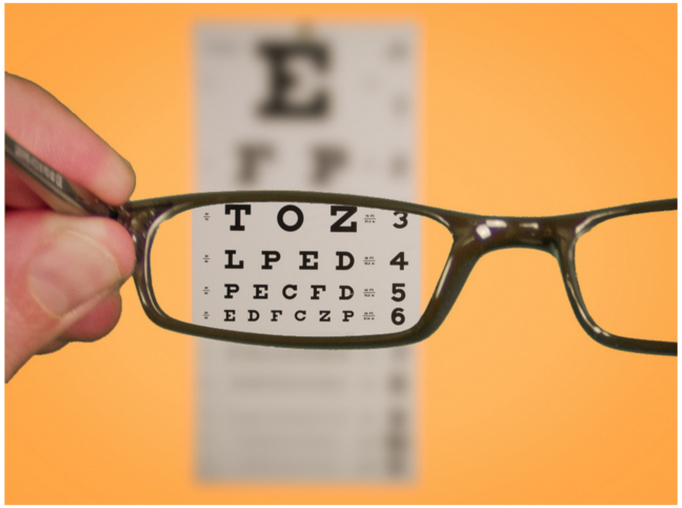
6. Model for biased agent: Procrastination





7. Model for biased agent: Myopia

- **Simple myopia (near sighted)**: ignore any rewards or costs after time $k1 > 0$ (even though you'll still be alive).
- **Bounded Value-of-Information:** ignore the value of information gained after time $k2 > 0$ (even though you will still get benefits from information).



7. Model for biased agent: Myopia

agentmodels.org

Interactive, online tutorial and open-source library for constructing this kind of model (Work in progress).

Main sections:

- Agent models for one-player sequential problems (MDPs, POMDPs, RL), where agent can be biased
- Inference (IRL) for a large space of possible agents
- Multi-agent interactions: coordination, group preferences.

Tom's *decision rule* is to take action a that maximizes utility, i.e., the action

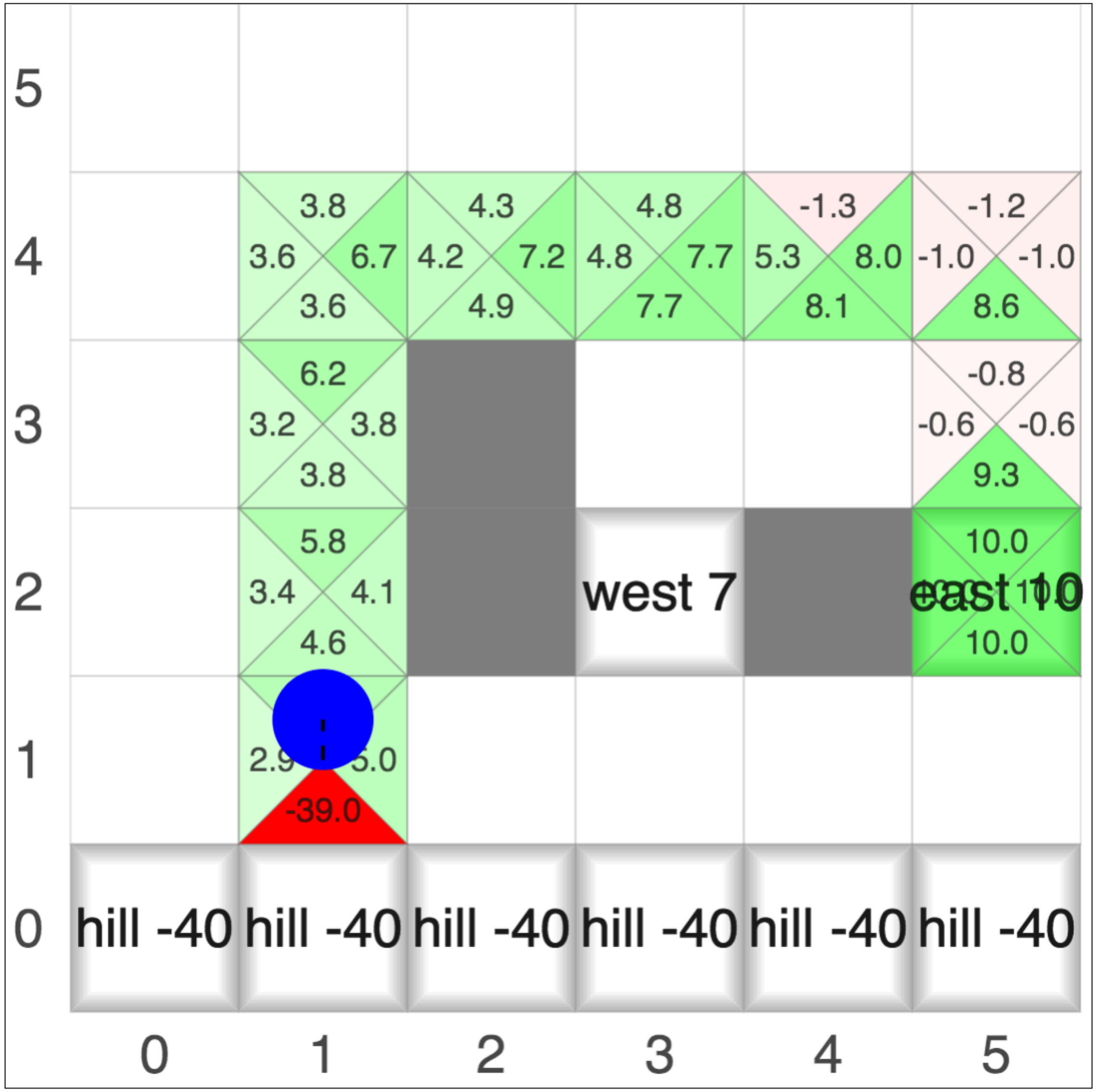
$$\arg\max_{a \in A} U(T(s, a))$$

In WebPPL, we can implement this utility-maximizing agent as a function `maxAgent` that takes a state $s \in S$ as input and returns an action. For Tom's choice between restaurants, we assume that the agent starts off in a state `"default"`, denoting whatever Tom does before going off to eat. The program directly translates the decision rule above using the higher-order function `argMax`.

```
// Choose to eat at the Italian or French restaurants
var actions = ['italian', 'french'];

var transition = function(state, action){
  return (action === 'italian') ? 'pizza' : 'steak frites';
};

var utility = function(state){
  return (state === 'pizza') ? 1 : 0;
};

var maxAgent = function(state){
  return argMax(
    function(action){
      return utility(transition(state, action));
    },
    actions);
};

print("Agent chooses: " + maxAgent("default"));
```

run

```
Agent chooses: french
```



5
4
3
2
1
0
3.8
3.6
6.7
3.6
4.3
4.2
7.2
4.9
4.8
4.8
7.7
7.7
-1.3
5.3
8.0
8.1
-1.2
-1.0
-1.0
8.6
6.2
3.2
3.8
3.8
-0.8
-0.6
-0.6
9.3
5.8
3.4
4.1
4.6
west 7
10.0
east 10
10.0
2.9
5.0
-39.0
hill -40
hill -40
hill -40
hill -40
hill -40
hill -40
0
1
2
3
4
5

Acknowledgments

- Future of Life Institute
- ONR
- DARPA
- Future of Humanity Institute, University of Oxford
- Department of Psychology, Stanford University