Algorithms for Understanding People: Computational Theory of Mind with POMDPs

Felix Sun, Julian Jara-Ettinger, Joshua Tenenbaum, Leslie Kaelbling
MIT EECS - eBay Inc. Undergraduate Research and Innovation Scholar

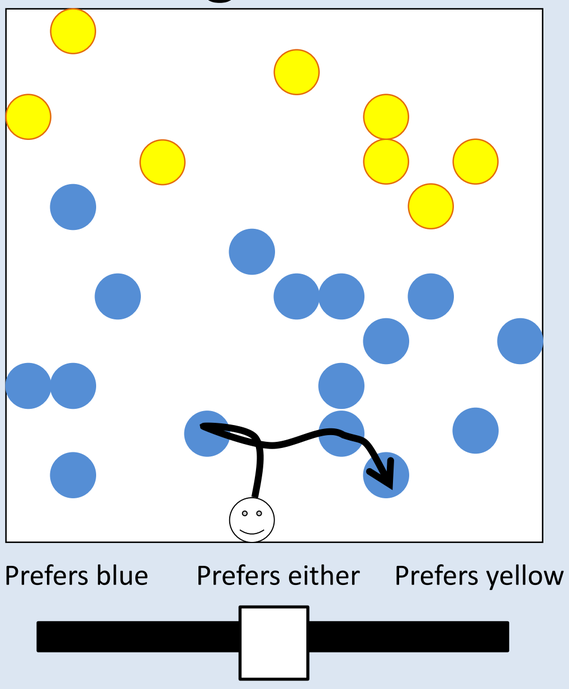
MIT EECS ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

SuperUROP

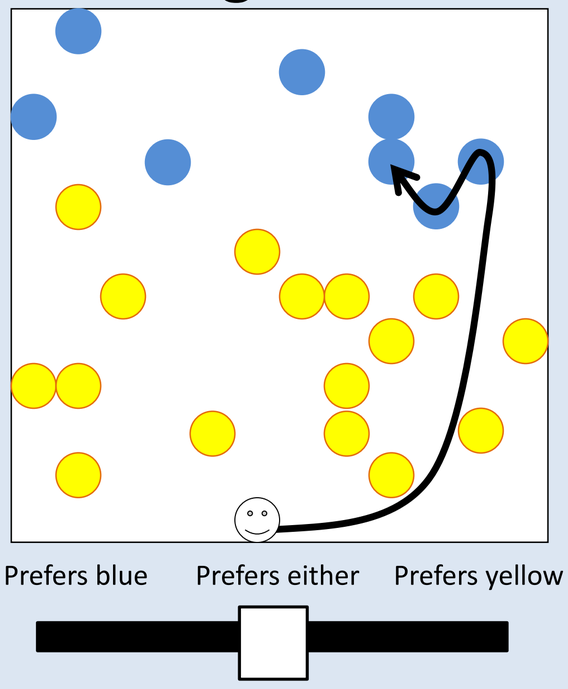
The problem: We observe an agent doing things. What were his goals?

In this game, the agent is allowed to pick up three balls. Your job is to figure out what color he prefers.

Agent 1

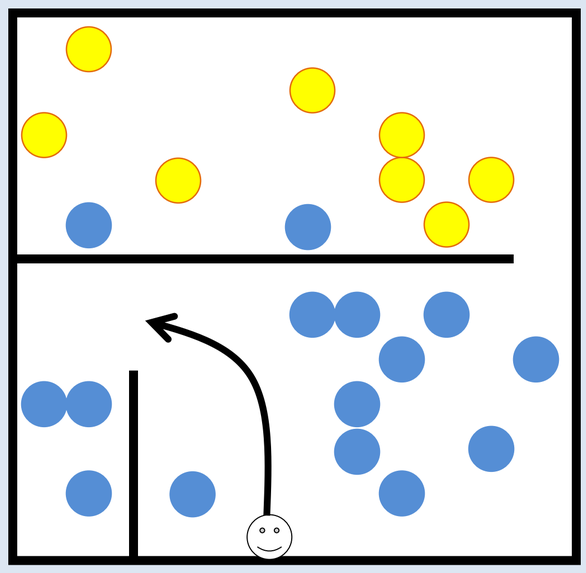


Agent 2



Most people think that agent 2 likes blue balls more than agent 1 does. (Why?) How do we make an AI program that reaches the same conclusions?

Humans can also infer the agent's beliefs:



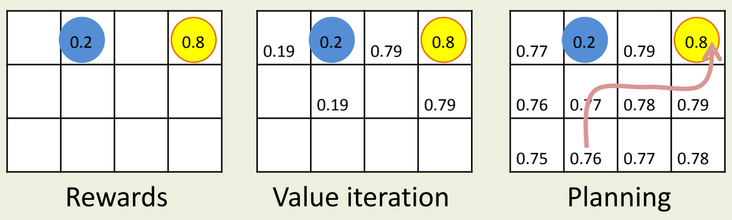
Background

Theory of Mind (ToM) psychology

- People can reason about others' goals and false beliefs by the age of 4. [Carey 2011]
- Recent Bayesian explanations of ToM: the observer calculates (via intuition) the probability that the agent has each belief. [Baker, CogSci 2011]
- POMDPs used in Baker, other papers.

(PO-)MDPs

- A model of planning actions, popular in robotics.
- State + action -> new state.
- Some state-action combinations have rewards.

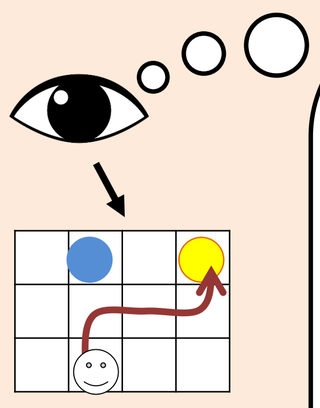


Rewards | Value iteration | Planning

- In a POMDP, the agent can have partial or false information (i.e. walls that block vision). The planning algorithm becomes more complicated – we use 3rd party packages to solve these problems

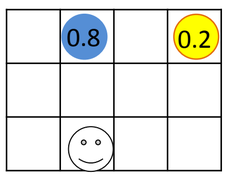
ToM algorithm

The observer "simulates" the agent in its own mind.



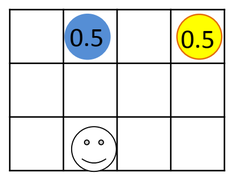
It considers many hypotheses of what the agent might prefer.

Hypothesis 1



P(observed path) = 0.001

Hyp. 2



P(observed path) = 0.05

Hyp. 3

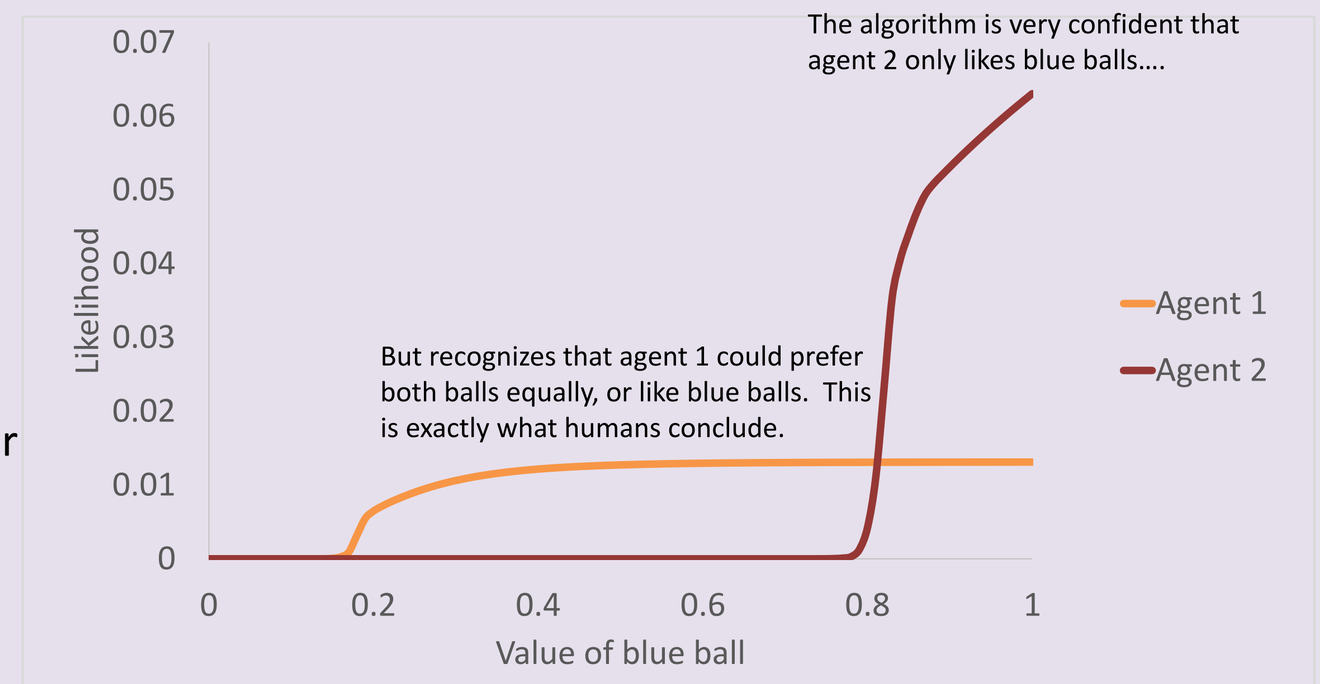
...

And computes the probability that the agent takes the observed actions, given each hypothesis.

Finally, the observer uses Bayes' rule to get the likelihood of each hypothesis, and finds the most plausible one.

Computational results

Our algorithm can infer the preferences of agents 1 and 2 (from above). Shown below is a probability distribution over the agents' preferences.



Next steps

- Human experiments on Amazon Mechanical Turk, to verify the judgment of our model.
- Publication of results at CogSci 2015.
- Apply model to more complex situations, involving inference of beliefs as well as goals, etc.
- Recursive theory of mind: "I know that he doesn't know that I know his secret".