1 Introduction

As a human, making sense of other people’s behavior involves reasoning about their unobservable mental states such as beliefs and desires. Because people cannot see other people’s minds, they must infer them.

In recent years, machine learning has made remarkable progress in recognition, categorization, and planning. Yet, algorithms that can effectively predict other people’s actions, even in the simplest scenarios, have lagged behind. In this area, human-inspired machine learning, machines that learn and think like people do, may be the answer (Lake, Ullman, Tenenbaum, & Gershman, 2016). By studying how people infer other people’s mental states, and using these findings to decide how to interact with them, it is possible to engineer systems that can better interpret people’s behavior.
2 Background

In recent years, researchers have developed human inspired artificial intelligence systems which are capable of inferring underlying intentions and desires based on observable actions. These systems rely on frameworks developed for robotics: Markov Decision Processes. In the study of robotics, researchers know what they want robots to accomplish, and their task is finding a set of actions to consistently reach that goal. To do so, researchers formalize the robot’s task using a reward function, which, combined with the actions that the robot can take and a model of the world (represented as a state transition matrix), produces the optimal policy – a function that maps each state to the action the robot ought to take.

Computational models of mental-state inference reverse this process, observing a set of actions and attempting to find the underlying goal. This is done by applying Bayesian inference to invert a Markov Decision Process model. That is, the task of inferring other people’s goals can be formalized as the problem of seeing an agent’s policy (the actions that they take), and finding a latent reward function that generates the observed policy. The models are then run on situations, and their outputs compared to human behavior in the same situation. In this way, it is possible to test cognitive theories, and better understand how humans make sense of the social and physical world.

Computational models using these ideas have been successfully applied to reach human-level performance in relatively simple tasks that involve inferring other people’s beliefs (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017), and their desires (Velez-Ginorio et al., 2017; using a probabilistic context free grammar to express more complex reward functions with temporal and logical structure).
3 Project

I will be working with Professor Jara-Ettinger to extend these mental-state inference models so they can infer other people’s relative knowledge or ignorance in both spatial context (such as inferring that an agent does not know where X is when the agent walks in the wrong direction) and non-spatial contexts (inferring that an agent does not know how to use an object X when they use it in an inefficient way). I will start by implementing a Markov model for a simple solver that bets money so that it has the best chance of winning, to better understand the structure and code so I am able to build more complex models. Then, I will build a more complex model for one of the issues the lab is currently studying. In particular, I am interested in modeling what information a student knows and doesn’t know in an educational setting, to best choose what information to teach them. I will compare the models to human behavior in simple behavioral tasks.

I will consider how the findings may apply to education technology. This proposal may change as my own understanding develops.

The deliverables of this project will include all of the code that I have written, and a report detailing my contributions.

References
