Project: Enabling Passenger Comfort by Increasing Behavior Transparency

Pls: Anca Dragan, Pieter Abbeel
Student: Sandy Huang

Abstract
The overarching goal of this work is to efficiently enable end-users to correctly anticipate a robot’s behavior in novel situations. Since a robot’s behavior is often a direct result of its underlying objective function, our insight is that end-users need to have an accurate mental model of this objective function in order to understand and predict what the robot will do. While people naturally develop such a mental model over time through observing the robot act, this familiarization process may be lengthy. We strive to reduce this time by having the robot model how people infer objectives from observed behavior, in order to then select those behaviors that are maximally informative.

Background
Imagine riding in a self-driving car that needs to quickly change lanes to make a right turn. The car suddenly brakes in order to merge safely behind another car, because it deems it unsafe to speed up and merge in front of that car. A passenger who knows this self-driving car is defensive and that it values safety much more than efficiency would be able to anticipate this behavior. But passengers less familiar with the car would not anticipate this sudden braking, so they may be surprised and possibly frightened.

There are many reasons why it is beneficial for humans to be able to anticipate a robot’s movements, from subjective comfort to ease of coordination when working with and around the robot. Our goal is to enable end-users to accurately anticipate how a robot will act, even in novel situations that they have not seen the robot act in before---like a new traffic scenario, or a new placement of objects on a table that the robot needs to clear.

In many cases, a robot’s behavior is a direct consequence of the objective (or reward) function the robot is optimizing. Whether this objective function is hard-coded or learned, it captures the trade-offs the robot makes between features relevant to the task. For instance, a car might trade off between features related to collision avoidance and efficiency, with more “aggressive” cars prioritizing efficiency at the detriment of, say, distance to obstacles.

Our insight is that the key to end-users being able to anticipate what a robot will do in novel situations is having a good understanding of the robot’s objective function.

Users will naturally improve their mental model of how a robot acts, given examples of the robot behaving optimally. However, not all examples are equally informative. For example, an autonomous car driving down a highway with no cars nearby will drive at the speed limit and stay in its lane, regardless of its trade-off between efficiency versus staying far away from other
cars. Another example is when an autonomous car changes lanes without interacting with any other cars (figure, right). An end-user mainly exposed to these types of behavior will have difficulty forming an accurate mental model of the robot’s objective function and anticipating how the robot will behave in more complex scenarios. On the other hand, suppose an autonomous car speeds up to merge in front of another car, cutting it off (figure, left). This scenario more clearly illustrates the trade-offs this car makes regarding safety versus efficiency.

We focus on enabling robots to purposefully choose such informative behaviors that actively communicate the robot’s objective function. We envision a training phase for interaction, where the robot showcases informative behavior to quickly teach the end-user what it is optimizing for.

**Progress**
In order to choose the most informative example behaviors for communicating a robot’s objective function to humans, we take an algorithmic teaching approach: we model how humans make inferences about the robot’s objective function from examples of its optimal behavior, and use this model to generate examples that increase the probability of humans inferring the correct objective function.

We introduce two factors to define candidate models for human learning in this setting, and analyze them in a user study in the autonomous driving domain. We show that certain models lead to the robot generating example behaviors that better enable users to anticipate what it will do in novel situations.

Our results also suggest, however, that additional research is needed in modeling how humans extrapolate from examples of robot behavior. In particular, we found that teaching performance correlates with covering the full space of strategies that the robot is capable of adopting. For instance, the teaching algorithm cannot just show the car cutting others off; it should also show an example where it is optimal to brake and merge behind. We show the best results are obtained by a coverage-augmented algorithm that leverages an approximate-inference user model while encouraging full coverage of all possible driving strategies.