Socially Rational Artificial Agents
A Key to Human-Machine Collaboration

Overview *Under the assumption that humans are, perhaps boundedly but nonetheless ideally, socially rational creatures, we propose to design and build socially rational artificial agents that learn via repeated interactions, with the aim being for such agents to collaborate effectively with humans.*

Keywords: reinforcement learning, stochastic games, behavioral experiments, social preferences.

Intellectual Merit Much of human life occurs in contexts where people must coordinate their actions with those of others. From party planning to space exploration to grant-proposal evaluation, groups of people have accomplished great things by reasoning as a team and engaging in jointly intentional behavior. Indeed, some have argued that, because other animals lack the capacity to work adaptively as a cohesive unit across many domains, team reasoning may be the hallmark of human sociality.

We call optimal decision making, when agents can hold social preferences and employ team reasoning as appropriate, **socially rational behavior**. Socially rational agents attempt to optimize a social utility function (i.e., a representation of social preferences), which is sufficiently rich to incorporate perceived societal benefits. We assume that social utilities can be broken down into two components—an objective component, which is usually a direct function of the rules of interaction, and a subjective component, which captures notions of distributivity and reciprocity.

Given these assumptions, we propose an iterative computational model in which socially rational artificial agents construct social preferences from observed histories of repeated interactions with other, potentially human, agents, and then decide how to optimize them. We contend that socially rational behavior, in which agents can jointly optimize a learned social utility function that reflects constructed social preferences, is a promising avenue for orchestrating collaborations between humans and machines.

Central to our computational model is the notion of inverse reinforcement-learning (IRL), whereby an agent is shown demonstrations of behavior and based on which it infers utilities that motivate that behavior. Nearly all existing IRL algorithms to date assume the demonstrations are generated by an expert. In our setting, in contrast, the demonstrations will be past interactions among agents who are not necessarily skilled at the task at hand, but, rather, are learning. Consequently, we are proposing to develop new IRL technology that learns social utility functions from both bad and good examples of behavior. This technology will enable humans to give agents both positive and negative feedback while learning to collaborate.

In reality, when an agent arrives on the playing field, it does not know whether the other agents it faces err on the side of being social (team reasoners) or selfish (best-repliers). While evaluating other agents' behavior, it behooves a socially rational agent to decide upon a strategy for itself—social or selfish? We will also develop algorithms that classify others as social or selfish, and adopt the corresponding stance.

Broader Impact This project is part of Brown University's Humanity Centered Robotics Initiative (HCRI) and its ongoing efforts to design robotic systems that interact with people and support independent living tasks (e.g., gerontechnological support for aging in place). For an elderly person to trust and collaborate on tasks with a machine effectively, the machine must act in a manner that the elderly person expects. Our proposed project is foundational for these important applications.

To engage a wider group in these efforts, we will create an undergraduate course called "Social autonomous driving" to be offered as part of Brown's new robotics course sequence. Students will develop robot cars that drive around a test environment, making sure they interact smoothly with other robotic and remote-controlled cars. We also plan to integrate our work on this project into Artemis, a free summer program that introduces rising 9th grade girls to computational thinking by having the Artemis girls teach a robot to collaborate with them directly on routine tasks.

Our deliverables include an open-source publicly accessible toolkit for implementing human-machine collaborative-learning tasks via reinforcement learning. Further, we will maintain a database of machine–machine, human–machine, and human–human interactions, which can serve as a benchmark for future researchers who also seek to build artificial agents that increasingly achieve human-like behavior. Finally, we expect to publish the results of the proposed research in top-tier archival conference proceedings and journals with high impact factors, and to present our work at innovative, non-archival workshops (e.g., the AAAI symposia).