Better Environments for Better AI

Sarah Keren

Technion - Israel Institute of Technology, Taub Faculty of Computer Science

Most past research aimed at increasing the capabilities of AI methods has focused exclusively on the AI agent itself, i.e., given some input, what are the improvements to the agent's reasoning that will yield the best possible output. In my research, I take a novel approach to increasing the capabilities of AI agents via the design of the environments in which they are intended to act. My methods for automated design identify the inherent capabilities and limitations of AI agents with respect to their environment and find the best way to modify the environment to account for those limitations and maximize the agents' performance.

From a theoretical perspective, the kind of design problems that I study take as input an initial environment, a description of agents (human users or artificial autonomous agents) that are active within it, a set of applicable design interventions, and a design objective. The interventions vary between domains, and may include physical changes to the environment, changes to the information that is provided to the agents, or changes to the ways agents can interact. The challenge is to efficiently search through the space of design options for a set of interventions that maximizes the targeted design objective. Finding an optimal solution is difficult because the design space is typically very large, and because it can be costly to estimate the effect of even a single change.

My research addresses these challenges by formulating the design problem as a state space search and specifying theoretical conditions under which heuristic methods remain admissible. My work also shows how to exploit *model-based reasoning and decision making under uncertainty* to increase efficiency as well as help interpret the results. I am also increasingly making use of methods from *game theory*, *multi-agent systems*, and *reinforcement learning*.

The research projects that are conducted at the Collaborative AI and Robotics (CLAIR) lab I lead at the Technion vary in their design objective, in the AI methodologies that are applied for finding optimal designs, and in the real-world applications to which the settings correspond. In spite of this diversity, all approaches use automated design to support effective collaborations in multi-agent AI systems.

One kind of application we are working on involves a group of robots that need to collaborate and coordinate their behavior to achieve their objective in an unfamiliar environment. The robots have limited sensing capabilities and a limited communication channel. The challenge is to find a sensor distribution within the environment that is sufficient for task completion and to formulate an information sharing protocol that allows the robots to collaborate effectively. Another project focuses on smart-grids modeled as multi-agent reinforcement learning settings in which multiple self-interested agents can produce and consume electricity while sharing the electrical grid and a set of limited resources. We apply automated design to find the best distribution of grid resources to maximize individual revenue while maintaining electricity quality.

The future will bring an ever increasing set of interactions between people and automated agents, whether at home, at the workplace, on the road, or across many other everyday settings. Autonomous vehicles, robotic tools, medical devices, and smart homes, all allow ample opportunity for human-robot and multi-agent interactions. In these settings, recognizing what agents are trying to achieve, providing relevant assistance, and supporting an effective collaboration are essential tasks, and tasks that can all be enhanced via careful environment design. However, the increasing complexity of the systems we use and the environments in which we operate makes devising good design solutions extremely challenging. This stresses the importance of developing automated design tools to help determine the most effective ways to apply change and enable robust AI systems.

My long-term intention is to use automated design to promote effective multi-agent collaboration and to enhance the way robots and machines interact with humans. This agenda is driven by various real-world multi-agent scenarios for which current AI methods are insufficient. In my current work, I have made progress by adopting a separate account of different design objectives, such that, for example, the objective was either to enhance goal recognition or maximize agent performance. Looking forward, my intention is to create a general and adaptable framework that supports and combines multiple design objectives. This will present new theoretical challenges and stress the need to produce solutions that come not only with efficiency guarantees but also with an explanation that helps humans understand their nature and necessity.