Collaborative AI

Spring 2023
Class location: TBD
Lecture Meeting time: Tuesday 09:30-11:30
Tutorial: Tuesday 11:30-12:30

**Teaching Staff:**

Instructor: Sarah Keren sarahk@cs.technion.ac.il
Office Hours: By appointment

TA: TBD
Office Hours: By appointment

**Prerequisites:** 236501 (Introduction to Artificial Intelligence) or 236609 (AI and Robotics) with a final grade of at least 85.
**Co-requisites:** NA
**Courses Without Credit:** NA

**Credits:** 3

**Study hours per week:** Lecture-2 Tutorial-1 (Lectures and tutorials will be in English)

**Course Goals and Description**

Historically, most AI research has dealt with the canonical setting consisting of a single agent confronting a possibly non-social environment. Among the frameworks that do account for multiple agents, most efforts have been devoted to adversarial settings and zero-sum games, in which achievements can be made only at the expense of others. While such settings are relatively easy to model and benchmark, they are rare in the real world. Instead, recent global events and technological advancements have given rise to the need for AI agents that can interact and collaborate with each other and with humans in dynamic and uncertain environments. In accordance with this agenda, the course will be dedicated to understanding the science of collaborative and cooperative AI. The focus will be on promoting cooperative behaviors between AI agents, even when their incentives are not fully aligned.

In order to promote effective collaborations between self-interested and possibly partially informed agents, there is a need to understand their decision-making procedures. The course will therefore start by covering the main common AI approaches to sequential decision making under uncertainty for a single agent. This includes both planning, for settings in which the agent has a complete model of the underlying environment, and reinforcement learning (RL), for setting in which the model of the environment is only partially known. After gaining a deep understanding of the different elements that influence the decision making of an agent that is assumed to be alone in the environment, we will explore the different complexities that arise in the presence of other agents. Focusing on non-adversarial settings and sequential social

dilemmas, we will explore different methods that have been suggested in the literature for computing joint policies in settings with limited resources, limited and noisy communication, and individual reward functions.

The course will include learning the theoretical aspects of various single-agent and multi-agent AI frameworks, as well as practical work with different systems, using Python based implementations. Students will be required to run various single-agent and multi-agent planning, RL and deep-RL algorithms on different domains and analyze the performance of the different approaches. In addition, students will be required to suggest a novel approach for cooperative AI. All students will work together on two standard RL domains and an additional domain of their choice.

**Learning Outcomes**

- Knowledge of various AI frameworks for modeling single-agent and multi-agent settings.
- Understanding the theoretical guarantees and limitations of different AI algorithms for single and multi-agent planning and RL.
- Acquiring practical experience using AI tools and implementing them in various AI domains.
- Experience in analyzing the performance of different AI approaches.
- Offering new algorithms for collaborative AI.

**Course Content/Topics**

The tools and models we will explore are based in a variety of AI fields including automated planning, sequential decision making under uncertainty, model-based reasoning, game theory, multi-agent systems, reinforcement learning and more.

The course agenda is as follows.

- Single-agent planning:
    - Classical planning.
    - Planning in fully observable stochastic environments.
    - Planning in partially observable stochastic environments.
- Single-agent reinforcement learning (RL)
    - Tabular methods vs. approximate methods.
    - Policy gradient vs. value-based methods.
    - Model-based vs. model free RL.
- Multi-agent planning:
    - Planning in adversarial, cooperative, and collaborative multi-agent settings.
    - Communication and resource allocation in multi-agent systems.
- Multi-agent RL:
    - Learning in the presence of others.
    - Emergent coordination and cooperation.

- Efficient communication in noisy partially known environments.
- Automated design of AI environments to promote collaboration.

**Assignments and Grading Procedures**

The course will require the submission of small-scale weekly assignments, writing a tutorial to the class, and a final project.

Each week, 1-2 students will prepare a tutorial about one or more existing algorithms for single-agent and multi-agent planning and RL which we will discuss in class. This will include the preparation of a python notebook with code that will run on several benchmarks which will be used throughout the course.

The final project will involve one or more of the following (the project structure will be formulated together with the course staff):

- suggesting a novel multi-agent benchmark (or a novel adaptation of an existing benchmark)
- offering a novel algorithm (or a novel adaptation of an existing approach) for a cooperative setting
- analyzing the formal properties of the suggested approach
- evaluating the suggested approach on the set of domains, including a novel domain which was not used in class.

Grading:
85% - assignments
15% - participation.

80% physical attendance in class is mandatory.

**Text book(s) and/or other materials**

- Reinforcement Learning: An Introduction (second edition). By Richard S. Sutton and Andrew G. Barto. 2015. MIT Press.
- Rules of Encounter: Designing Conventions for Automated Negotiation. By Jeffrey S. Rosenschein and Gilad Zlotkin. 2014. MIT Press
- Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations by Yoav Shoham and Kevin Leyton-Brown. Cambridge University Press. 2008
- An Introduction to MultiAgent Systems by Michael Wooldgridge John. Wiley & Sons. 2009.
- A concise introduction to models and methods for automated planning by Hector Geffner and Blai Bonet. 2013 *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan Claypool Publishers.