Helper Artificial Intelligent agent

1 Introduction

The project will focus on mechanism design for human and artificially and intelligent agents. The aim is to develop mechanisms that enables their co-operation. The project will take place within the context of a research program in cooperation with Harvard university, exploring the role of mechanism design, multi-agent dynamical models, and privacy preserving algorithms, in promoting the emergence of beneficial AI, for example, social-welfare maximizing AI, in multi-agent systems, and especially in multi-agent systems in which the AIs are built through reinforcement learning.

Overall, we study two specific multi-agent learning or planning problems, both situated within the formalism of Markov decision processes. The first is *experiment design*, typically formalized as a *multi-armed bandit process* [Chernoff, 1959], which we intend to study in a multi-agent, privacy-preserving setting. The second is the more general problem of learning to act in Markovian *dynamical systems*. The third, is the more ambitious problem of aligning the incentives of multiple agents in either framework.

The work done so far in this project has focused on the problem faces by an AI that needs to collaborate with a single human. As the AI and human's views on reality disagree, the AI must take into account the human's beliefs Dimitrakakis et al. [2016].

2 Project 1. Experiment design.

One classic problem area of interest is experiment design. As an example, drug companies want to design a drug with certain properties, and each company has an AI to plan experiments into the efficacy of drugs. There is a plethora of compounds that may be useful, and not all can be tested. However, there exist large scale databases of drug toxicity and activity. Each AI is able to use data from previous clients to do better planning at a lower cost. The AIs can post drug descriptions, *in vitro* results, simulations for *in silico* experiments, and the results of clinical trials.

An important question in this context is how to align incentives so that the joint plan is beneficial to society (in terms of access to useful therapies), while simultaneously balancing the computational and human cost associated with designing, performing and analyzing experiments. An additional challenge relates to privacy– not only in regard to individual's concern about their own data, but in regard to AIs, for example looking to minimize information revealed in order to avoid a "ratchet effect" where other AIs can take advantage of this in the future, in the context of this competitive, market-based mechanism.

2.1 Project 2. Dynamical systems.

For the more general dynamic setting, an illustrative example is the smart grid, where each AI acts on behalf of a household, and decides when to consumer power, how to allocate power to different devices, and how to control the set-points of devices. The AIs interact with people through preference elicitation (e.g., temperature of house, importance of charging an electric vehicle, importance of dry clothes) and with other agents because of competition for a scare resource with a variable price.

2.2 Project 3. Agent collectives.

Human-agent collectives hold great promise, but many challenges remain. For example, different human and AI agents make decisions based on a specific belief system and their beliefs will be unlikely to be perfectly aligned. This is firstly because they do not have the same information. Secondly, and perhaps more importantly, it is because they do not have a common prior belief or belief system, perhaps due to differing computational capabilities. How should information about beliefs be priced and communicated? Is it possible to design a general mechanism that near-optimally allocates effort among the different agents, that takes privacy into account?

The main focus of this thesis will be privacy-preserving, incentive-compatible mechanism design for multi-agent systems. The choice of topic will depend highly on the candidate's skills. Any candidate to have a good grasp of at least one of the following topics:

- Markov decision processes.
- Bayesian statistics.

- Game theory and mechanism design.
- Differential privacy.

Good programming skills are a bonus, but not essential.

References

Herman Chernoff. Sequential design of experiments. *Annals of Mathematical Statistics*, 30(3):755–770, 1959.

Christos Dimitrakakis, Firas Jarboui, David Parkes, and Lior Seeman. Multi-view sequential games: The helper-agent problem. Technical Report hal-01408294, 2016. URL `https://hal.archives-ouvertes.fr/hal-01408294/`.