Postdoc proposal (1 pager)

Game-Theoretic Analysis and Development for Hybrid Intelligence

Basic facts

- Erman Acar research question, research activity, capability + level (all from proposal text)
- Postdoc position

What will you do?

We will develop a game-theoretic framework which can capture the cooperation scenarios in heterogeneous groups (a mix of human and artificial agents). Such a study will require the development of necessary game-theoretic concepts; amongst them, refinement of existing equilibrium concepts, bounded rationality models and provably good guarantees.

Why is it important (for HI)?

As mentioned in HI proposal, **collaborative** aspect (C of CARE) stand highly important for the project. Moreover, collaborative aspect is tightly connected twitho **responsibility** aspect (R of CARE), since one agent's gain in one outcome can easily cause another agent's loss, or even risk group efficiency. Therefore, the study of collaboration requires a high-level perspective; a game-theoretic lens, in order to understand and steer human-machine interactions as a system. In particular, a game-theoretic approach will serve as a compass in designing systems robust against unexpected scenarios with provably good guarantees.

How will you do it?

We will employ (cooperative/non-cooperative) game theory and further develop existing solution concepts, to be able to capture various behaviours of such interactions observed in lab experiments. One important aspect is the use (or development) of a general-enough notion of bounded rationality, since different agents can often misinterpret or miscalculate signals or other agents' intentions, differently, due to environment and the physical nature of interaction. A good starting point would be the solution concepts with "mistakes and imprecision" introduced in our recent work [*Distance-based Equilibrium in Normal Form Games, Acar and Meir,* AAAI 2020]. Although, general enough for a good starting point, the theory demands further developments: It needs to address the turn-based nature of interactions i.e., "Extended-Form Games". Moreover, in its current form the theory lacks the epistemic dimension; we ideally want a theory of games that can easily be integrated with the so-called "theory of mind" (at least up to "level 3").

How will you know you are done?

We will create a lot of test settings, both in virtual and physical environments with various incremental scenarios and with different instantiations in heterogeneous (human + artificial) agents set-ups. This will serve as a standard. The theory developed has to predict the expected average of the outcome with satisfying accuracy. Moreover, using the theory we need to be able to predict the convergence rate, and a bound estimate of

achieving a given task e.g., w.r.t various quality parameters such as time and efficiency cost, price of anarchy, etc.

What's your goal at the end of year 1?

We will have the initial theoretical framework with possible extensions and simple empirical tests. We aim to have it published on a top quality conference (e.g., AAMAS, IJCAI, AAAI, EC) or ready for a submission to a good journal.