Research Proposal: Cooperation among Self Interested Agents

Reshef Meir

1 Introduction

In the well known Prisoner's Dilemma, two people that are following the only *rational* behavior end up in the worst possible outcome. Unfortunately, this example is a useful analogy for many situations in real life, where (individually) rational behavior leads to a disaster for the society.

With the rapid delegation of decision making to automated agents, the role of game theory within artificial intelligence is becoming increasingly important. In particular, game-theoretical principles must be taken into account in the design of systems and environments in which agents operate (human and automated alike).

My research focuses on *mechanism design* (see [Nisan and Ronen, 2001] for background). More specifically, on ways to *incentivize self-interested agents* to cooperate in a way that will benefit the entire society. This cooperation arises not by forcing them or by relying on their good intentions, but by changing the "rules of the game" so that *the best individual decision* would be to cooperate. The research is multi-disciplinary in nature, involving tools and ideas from economics, computer science, mathematics, artificial intelligence, and cognitive science.

This proposal briefly describes my recent work on prompting cooperation in two related domains, and outlines some future directions. I will conclude with some remarks on the strong assumption of rationality that underlies standard game-theoretic analysis and how it can be relaxed in the quest for cooperation.

2 Strategyproof Classification

An essential part of the theory of machine learning deals with the *classification problem*: a setting where a decision maker must classify a set of input points with binary labels, while minimizing the expected error. In contrast with the standard assumption in machine learning, we handle situations in which the labels of the input points are reported by self-interested agents, rather than by a credible expert. Agents might lie in order to obtain a classifier that more closely matches their own opinion, thereby creating a bias in the data; this motivates the design of *truthful* mechanisms that discourage false reports. Such mechanisms are called *strategyproof* (SP).

We designed various mechanisms for this purpose, and studied their limitations (e.g., [Meir *et al.*, 2009]); we keep improving results in the field with the introduction of better algorithms [Meir *et al.*, 2011a].

We recently observed that certain problems in SP classification can be treated within a unified framework along with seemingly unrelated problems in the fields of judgment aggregation on binary domains (see [Dokow and Holzman, 2010]) and facility location [Alon *et al.*, 2010]. Our initial results indicate that techniques from SP classification are useful in these other domains as well, and will help us to better understand the underlying connections between them.

3 The Cost of Stability

Cooperative games are a rapidly developing branch of game theory, which aims to describe and predict the coalitions that are most likely to arise in certain interactions, and how their members distribute the gains from cooperation (see [Peleg and Sudhölter, 2003] for an overview). When the agents are self-interested, the latter question is obviously of great importance. Indeed, the *total* utility generated by the coalition is of little interest to individual agents; rather, each agent aims to maximize her own utility. Thus, a *stable* coalition can be formed only if the gains from cooperation can be distributed in a way that satisfies all agents.

The model of cooperative games attracted much attention in AI research due to the increasing ubiquity of automated agents, and the complex computations required to answer some natural questions in the model.

The most prominent solution concept that aims to formalize the idea of stability in cooperative games is the *core*. Informally, this is an allocation of the total profits such that every coalition is allocated at least what it can gain by itself (and thus has an incentive to participate). However, this concept has an important drawback: the core of a game may be empty. In games with empty cores, any outcome is unstable, and therefore there is always a group of agents that is tempted to abandon the existing plan. This observation has triggered the development of alternative solution concepts in several directions. These include relaxations of the core such as the least core and cores in social contexts; and different notions of stability, such as the Nucleolus and the Bargaining Set [Peleg and Sudhölter, 2003].

In a series of recent papers we approach this issue from a different perspective (see [Meir *et al.*, 2010; 2011b], and references therein). Specifically, we examine the possibility of stabilizing the outcome of a game using an external subsidy. Under this model, an external party, which can be seen as a central authority interested in a stable outcome of the system, is willing to provide a supplemental payment if *all* agents cooperate. The minimal subsidy that can stabilize a game is known as its *Cost of Stability*. Previous work in economics focused on other aspects of subsidies in coalitional games [Jain and Vazirani, 2001; Bejan and Gòmez, 2009].

In our papers, we study bounds on the Cost of Stability in various games and suggest algorithms to compute it efficiently, when possible. We are also interested in the relation of the Cost of Stability with other solution concepts such as those mentioned earlier.

4 Beyond rational agents

Standard game theory (including my own work thus far) typically makes the assumption that behavior of agents is *rational* in the sense that agent are not only self-interested, but also *maximizing their utility*, where this "utility" follows well defined mathematical principles [von Neumann and Morgenstern, 1944]. In particular, agents are assumed to be risk-neutral and games are invariant under certain simple changes.

Evidence from psychological studies in the last four decades suggests that human decision makers are subject to consistent biases that can be measured and predicted. Such biases have been thoroughly investigated in the context of a single decision maker (e.g., prospect theory by Kahneman and Tversky [1979]). Following experiments in decision making, many empirical findings in games played by human players have been collected by Camerer [2003], and show similar biases.

While early observations date back to the 19th century, Camerer and others have also made efforts to treat cognitive and behavioral findings within the formal framework of game theory, in what has been termed *behavioral game theory*. Within AI, similar ideas have been advocated under the title of *bounded rationality*, termed by Herbert Simon [1957]. Nevertheless, mainstream work within game theory has remained largely unaffected by this progress. Observed biases are usually ignored, especially when one treats game theory as a branch of mathematics rather than a social science.

I believe that while game theory can be studied purely from a mathematical perspective, much of its appeal is derived by the perception that it does help us to understand and predict human behavior in situations of conflict, and to design appropriate mechanisms. In future studies, I intend to better understand how actual players behave and cooperate in various interactions ("games"), in light of the abundant theoretical and experimental findings. These behaviors should be either explained by classical solution concepts (Nash equilibrium, the Core, the Minmax value, etc.), or induce the development of new ones.

In particular, new solution concepts will shed a new light on the design of mechanisms that will increase cooperation between actual people in real-world situations.

References

[Alon *et al.*, 2010] Noga Alon, Michal Feldman, Ariel D. Procaccia, and Moshe Tennenholtz. Strategyproof approximation of the minimax on networks. *Mathematics of Operations Research*, 35(3):513–526, 2010.

[Bejan and Gòmez, 2009] Camelia Bejan and Juan C. Gòmez. Core extensions for non-balanced TU-games. *Int. journal of game theory*, 38:3–16, 2009.

[Camerer, 2003] Colin F. Camerer. *Behavioral game theory: experiments in strategic interaction*. Princeton uni. press, 2003.

[Dokow and Holzman, 2010] Elad Dokow and Ron Holzman. Aggregation of binary evaluations. *Journal of Economic Theory*, 145:495–511, 2010.

[Jain and Vazirani, 2001] Kamal Jain and Vijay V. Vazirani. Applications of approximation algorithms to cooperative games. In *Proc. of 43rd STOC*, pages 364–372, 2001.

[Kahneman and Tversky, 1979] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, XLVII:263–291, 1979.

[Meir *et al.*, 2009] R. Meir, A. D. Procaccia, and J. S. Rosenschein. Strategyproof classification with shared inputs. In *Proc. of 21st IJCAI*, pages 220–225, 2009.

[Meir *et al.*, 2010] Reshef Meir, Yoram Bachrach, and Jeffrey S. Rosenschein. Minimal subsidies in expense sharing games. In *Proc. of 3rd SAGT*, pages 347–358, 2010.

[Meir *et al.*, 2011a] Reshef Meir, Shaul Almagor, Assaf Michaely, and Jeffrey S. Rosenschein. Tight bounds for strategyproof classification. In *Proc. of 10th AAMAS*, 2011. To appear.

[Meir *et al.*, 2011b] Reshef Meir, Enrico Malizia, and Jeffrey S. Rosenschein. Subsidies, stability, and restricted cooperation in coalitional games. In *Proc. of 22nd IJCAI*, 2011. to appear.

[Nisan and Ronen, 2001] Noam Nisan and Amir Ronen. Algorithmic mechanism design. *Games and Economic Behavior*, 35(1–2):166–196, 2001.

[Peleg and Sudhölter, 2003] Bezalel Peleg and Peter Sudhölter. *Introduction to the Theory of Cooperative Games*. Kluwer Publishers, 2003.

[Simon, 1957] Herbert Simon. A behavioral model of rational choice. New York: Wiley, 1957.

[von Neumann and Morgenstern, 1944] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton Univ. Press, 1944.