

16 במאי 2011

רשף מאיר ת.ז. Reshef Meir 040097503

בית הספר להנדסה ומדעי המחשב, האוניברסיטה העברית

נושא העבודה:

## Better for Everyone: Cooperation among Self-Interested Agents

---

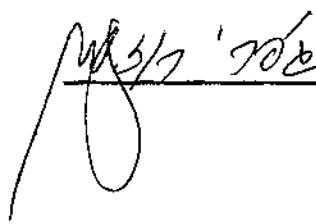
יותר טוב לכולם: שיתוף פעולה בין סוכנים  
אנוכיים

---

מנחה: ג'פרי רוזנשיין

הריני מאשר את הנושא ואת התוכנית, ומסכים להדריך את  
המועמד בביצוע עבודה זו.

חתימת המנחה: ג'פרי רוזנשיין



# Research Proposal

Reshef Meir

May 2, 2011

## 1 Preface

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 is multi-disciplinary in nature, involving tools and ideas from Economics, Computer science, Mathematics, Artificial Intelligence, and Cognitive science. My work so far consists of several projects in various domains. While all project ultimately aim for a better understanding of cooperation in games, they study different problems and use different theoretical tools. These span from Voting Theory to Cooperative Games to Machine Learning and the suggestion of new equilibrium concepts. Indeed, I believe that there is no one way to model cooperation, as this is an abstract concept whose realization strongly depends on the domain and the underlying assumptions.

The proposal is organized as follows. In Section 2 is focused on one project that is described in detail. This project was selected since it has both substantial results that had already been published, and some very promising directions that I currently work on. In the Section 3, other projects will be briefly described. In the closing section, I will discuss the strong assumption of rationality that underlies standard game-theoretic analysis and how it can be relaxed in the quest for cooperation.

It is important to note that the work described here has been accomplished by joint work with many other researchers. Their names are explicitly mentioned when discussing unpublished work (coauthors of published work are visible in the list of publications).

A short version of this proposal was published in the proceedings of IJCAI-2011 [25].

## 2 Strategyproof Classification

### 2.1 Background

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 mechanisms that discourage false reports.

To describe the problem in a more precise way, we are interested in the design and analysis of classification mechanisms that are *strategyproof* (SP). A mechanism is SP if none of the involved agents (the experts in our case) can benefit by lying. It turns out that there is an inherent tradeoff between our requirements, as strategyproofness can be achieved by ignoring some of the input, but this comes at the expense of suboptimal classification results. We therefore seek for mechanisms that are both SP, and guarantee good approximation results, compared to the optimal classifier.

**Motivating example** Consider a large organization that is trying to fight the congestion in an internal email system by designing a smart spam filter. Due to reasons of organizational uniformity, the filter cannot be personalized. That is, all employees will be affected by the same filter. In order to train the system, managers are asked to review the last 1000 emails sent to the “all employees” mailing list and classify them as either “work-related” (positive label) or “spam” (negative label). Whereas the managers will likely agree on the classification of some of the messages (e.g., “Buy Viagra now!!!” or “Christmas Bonus for all employees”), it is likely that others (e.g., “Joe from the Sales department goes on a lunch break”) would not be unanimously classified. Moreover, as each manager is interested in filtering most of what he sees as spam, a manager might try to compensate for the “mistakes” of his colleagues by misreporting his real opinion with respect to some cases. For example, the manager of the R&D department, believing that about 90% of the Sales messages are utterly unimportant, might classify *all* of them as spam in order to reduce the congestion. The manager of Sales, suspecting the general opinion on her department, might do the exact opposite to prevent her e-mails from being filtered.

Similar examples for biases are common in aggregating data from Internet polls or sales data from local retailers (see cited papers for details). In the remaining of this section, I will survey background literature (including results from my M.Sc. Thesis), present the formal framework and some of the prominent results.

## 2.2 Related and previous work

Research on Strategyproof Learning typically followed two primary directions. The first is the design and analysis of specific strategyproof mechanisms that guarantee optimal or near-optimal results for certain learning problems. Such mechanisms have been proposed for example in the supervised regression domain [37, 12].

A second, complementary, direction explores the limits of such mechanisms, by showing the best possible approximation ratio that SP mechanisms can guarantee on certain problems. The study of classification mechanisms was initiated in my Masters thesis (see [24, 29, 30, 31]), and there have also been results of similar flavor for unsupervised learning models [38]. Together, these two directions gradually compose a full picture of the restrictions and assumptions that allow for effective SP learning.

While our results in SP classification have been motivated by problems in the area of machine learning, SP classification is just part of the rapidly developing subfield of mechanism design called *approximate mechanism design without money* (AMDw/oM). Such mechanisms are being studied in multiple domains, including voting, resource allocation, matching problems, and even money-free auctions [2, 14, 17, 18, 19]. These domains are often interconnected, as results and techniques from one domain can be applied in others; proofs in different combinatorial settings, such as voting, matching, and labeling, often tackle similar issues like continuity and private information.

We recently observed that certain problems in SP classification, Judgment aggregation on binary domains [22, 13] and in facility location [1, 23, 39] can be treated within a unified framework. Our initial results indicate that techniques from SP classification are useful in these other domains as well. See more on this application in Section 2.6.

### 2.3 Model

Let  $\mathcal{X}$  be an input space, which we assume to be either a finite set or some subset of  $\mathbb{R}^d$ . A *classifier* or *concept*  $c$  is a function  $c : \mathcal{X} \rightarrow \{+, -\}$  from the input space to the *labels*  $\{+, -\}$ . A *concept class*  $\mathcal{C}$  is a set of such concepts. For example, the class of linear separators over  $\mathbb{R}^d$  is the set of concepts that are defined by the parameters  $\mathbf{a} \in \mathbb{R}^d$  and  $b \in \mathbb{R}$ , and map a point  $\mathbf{x} \in \mathbb{R}^d$  to  $+$  if and only if  $\mathbf{a} \cdot \mathbf{x} + b \geq 0$ .

Denote the set of *agents* by  $I = \{1, \dots, n\}$ ,  $n \geq 2$ . The agents are interested in a (finite) set of  $k$  data points  $X \in \mathcal{X}^k$ . We typically assume that  $X$  is *shared* among the agents, that is, all the agents are equally interested in each data point in  $X$ .<sup>1</sup> This plausible assumption, as we shall see, allows us to obtain surprisingly strong results. Naturally, the points in  $X$  are common knowledge.

Each agent has a private *type*: its labels for the points in  $X$ . Specifically, agent  $i \in I$  holds a function  $Y_i : X \rightarrow \{+, -\}$ , which maps every point  $x \in X$  to the label  $Y_i(x)$  that  $i$  attributes to  $x$ . Each agent  $i \in I$  is also assigned a *weight*  $w_i$ , which reflects its relative importance; by normalizing the weights we can assume that  $\sum_{i \in I} w_i = 1$ . Let  $S_i = \{(x, Y_i(x)) : x \in X\}$  be the partial *dataset* of agent  $i$ , and let  $S = \langle S_1, \dots, S_n \rangle$  denote the complete *dataset*.  $S_i$  is said to be *realizable* w.r.t. a concept class  $\mathcal{C}$  if there is  $c \in \mathcal{C}$  which perfectly separates the positive samples from the negative ones. If  $S_i$  is realizable for all  $i \in I$ , then  $S$  is said to be *individually realizable*. Figure 1 shows an example of a dataset with a shared set of points  $X$ .

We use the common 0-1 loss function to measure the error. The *risk*, or negative utility, of agent  $i \in I$  with respect to a concept  $c$  is simply the relative number of errors that  $c$  makes on its dataset. Formally,

$$R_i(c, S) = \frac{1}{k} \sum_{(x,y) \in S_i} [c(x) \neq y] = \frac{1}{k} \sum_{x \in X} [c(x) \neq Y_i(x)], \quad (1)$$

where  $[A]$  denotes the indicator function of the boolean expression  $A$ . Note that  $S_i$  is realizable if and only if  $\min_{c \in \mathcal{C}} R_i(c, S) = 0$ . The *global risk* is defined as

$$R_I(c, S) = \sum_{i \in I} w_i \cdot R_i(c, S) = \frac{1}{k} \sum_{i \in I} \sum_{x \in X} w_i \cdot [c(x) \neq Y_i(x)]. \quad (2)$$

<sup>1</sup>Some of our previous work relaxes this assumption [31].

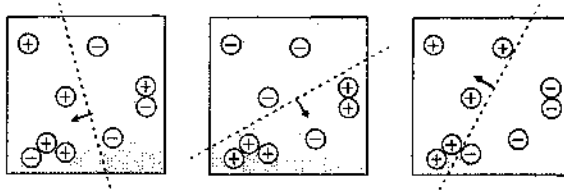


Figure 1: An instance with shared inputs. Here,  $\mathcal{X} = \mathbb{R}^2$ ,  $\mathcal{C}$  is the class of linear separators over  $\mathbb{R}^2$ , and  $n = 3$ . The data points  $X$  of all three agents are identical, but the labels, i.e., their types, are different. The best classifier from  $\mathcal{C}$  with respect to each  $S_i$  is also shown (the arrow marks the positive halfspace of the separator). Only the rightmost dataset is realizable.

## 2.4 Mechanism Properties

A *deterministic mechanism*  $\mathbf{M}$  receives as input a dataset  $S$ , and outputs a classifier  $c \in \mathcal{C}$ . We implicitly assume that the information on the weights of the agents is contained in the dataset.

A *randomized mechanism* is identified with a probability distribution  $p_{\mathbf{M}}$  over  $\mathcal{S} \times \mathcal{C}$ . We restrict our attention to probabilities with a finite support. That is, for every dataset  $S$ , the mechanism  $\mathbf{M}$  returns  $c \in \mathcal{C}$ , with a probability of  $p_{\mathbf{M}}(c|S)$ .

When measuring the risk, we are interested in the *expected* number of errors that the mechanism makes on the given dataset. Formally,  $R_i(\mathbf{M}(S), S) = \mathbb{E}_{p_{\mathbf{M}}} [R_i(c, S) | S]$ , and the global risk is defined analogously.

For any (complete or partial) dataset  $S' \subseteq S$ , the best available classifier with respect to the dataset  $S'$  is referred to as the *empirical risk minimizer* – a common term in machine learning literature. Formally,  $\mathbf{erm}(S') = \operatorname{argmin}_{c \in \mathcal{C}} \sum_{(x,y) \in S'} \mathbb{1}[c(x) \neq y]$ .

For the complete dataset, we denote the best classifier by  $c^*(S)$ , and its risk by  $r^*(S)$  (or simply  $c^*, r^*$  if  $S$  is clear from the context). That is,  $c^*(S) = \mathbf{erm}(S)$  and  $r^*(S) = R_I(c^*(S), S)$ .

The simple mechanism that always computes and returns  $c^*(S)$  is referred to as the **ERM** mechanism. Similarly, a mechanism which returns the best classifier with respect to a partial dataset of a specific agent (e.g.  $\mathbf{erm}(S_1)$ ) is called a *dictator mechanism*.

We measure the quality of the outcome of a mechanism using the standard notion of *multiplicative approximation*.

**Definition 2.1.** A mechanism  $\mathbf{M}$  is an  $\alpha$ -approximation mechanism if for any dataset  $S$  it holds that  $R_I(\mathbf{M}(S), S) \leq \alpha \cdot r^*(S)$ .

We emphasize that an agent may report different labels than the ones indicated by  $Y_i$  (i.e. lie). We denote by  $\bar{Y}_i : X \rightarrow \{+, -\}$  the reported labels of agent  $i$ . We also denote by  $\bar{S}_i = \{(x, \bar{Y}_i(x)) : x \in X\}$  the reported partial dataset of agent  $i$ , and by  $\bar{S} = \langle \bar{S}_1, \dots, \bar{S}_n \rangle$  the reported dataset.

*Strategyproofness* implies that reporting the truthful types is a dominant strategy for all agents. For a dataset  $S$  and  $i \in I$ , let  $S_{-i}$  be the complete dataset without the partial dataset of agent  $i$ .

	All Classes (shared inputs)		Binary decision
	general datasets	realizable datasets	( $ C  = 2$ )
<b>WRD</b>	3	2	3
<b>CRD</b>	$3 - \frac{2}{n}$	$2 - \frac{1}{n}$	2
<b>RRD</b>	$\geq 3$	$2 - \frac{2}{n}$	3
best upper bound	$3 - \frac{2}{n}$ ( <b>CRD</b> )	$2 - \frac{2}{n}$ ( <b>RRD</b> )	2 ( <b>CRD</b> )
lower bound	$3 - \frac{2}{n}$	1	2

Table 1: Summary of results.

**Definition 2.2.** A (deterministic or randomized) mechanism  $M$  is strategyproof (SP) if for every dataset  $S$ , for every  $i \in I$ , and  $\bar{S}_i$ ,  $R_i(M(S), S) \leq R_i(M(\bar{S}_i, S_{-i}), S)$ .

Our goal is to design mechanisms that are both SP and guarantee a low worst-case approximation ratio (in expectation).

There is an inherent tradeoff between strategyproofness and good approximation. The **ERM** mechanism (which always returns  $\text{erm}(S)$ ), for example, is a 1-approximation mechanism, but is not SP (as we show in the next section). On the other hand, a mechanism that selects agent 1 as a *dictator*, and returns  $\text{erm}(S_1)$  is clearly SP but in general may give a very bad approximation (e.g. if all other agents disagree with agent 1).

## 2.5 Results

Regarding deterministic mechanisms, it can be shown that no SP mechanism can guarantee a finite approximation ratio without using randomization [30, 31]. We therefore only present our results regarding randomized mechanisms. For reasons of brevity, all results are summarized in a single table, rather than presented as theorems. Proofs are available on the cited papers.

**The mechanisms** The most simple mechanism is *weighted random dictator* (**WRD**) mechanism, which works as follows: W.p. of  $w_i$ , the mechanism select agent  $i$  as a dictator, and returns  $\text{erm}(S_i)$ . It is easy to see that this mechanism, as well as any other randomization of dictators, is SP. This is since no agent has an incentive to lie, whether he is selected or not.

An improvement of the **WRD** is the *convex-weight random dictator* (**CRD**), which selects agent  $i$  w.p. proportional to  $\frac{w_i}{1-w_i}$ . The *realizable-weight random dictator* (**RRD**) is a variation of it, which was designed to handle realizable datasets (its full description is omitted). Clearly both mechanisms are SP. Their approximation ratios are available in Table 1.

**Lower bounds** In the bottom row of the table, it can be seen that in the general (non-realizable) case, the **CRD** mechanism is the best possible SP mechanism. The proof involves several interesting techniques, and relies on impossibility results from social choice theory. We conjecture that the upper bound for realizable datasets is tight, but this remains as a question for future research.

**Generalization from samples** In the standard supervised learning model, the learning algorithm is used on data sampled from some distribution, in order to be applied on the same distribution in the future. We would like to know that our mechanism still guarantee good approximation w.r.t. to the distribution and not just w.r.t. the training set. We provide several PAC-style bounds, which show that the expected error of our mechanisms can be arbitrarily close to the approximation bounds in Table 1. We discuss the exact game-theoretic assumptions required for the mechanisms to work, and prove that a polynomial number of samples is sufficient for the bounds to hold.

## 2.6 Discussion and current research

Part of the importance of our results lies in their implications on the related problems mentioned in Section 2.2. As one example, the described mechanisms can be generalized to random-selection mechanisms in arbitrary metric spaces, thereby solving facility location problems [1]. Also, our impossibility proof tackles general issues, such as continuity and private information. This will be relevant to the study of lower bounds in other domains.

The application of our tools and results to these domains also contributes to the expansion of the AMDw/oM approach. Within judgment aggregation, there is beginning to be a characterization of strategyproof rules, analogous to the trend that existed in the domain of facility location before the explicit introduction of AMDw/oM. Concepts that have played an important role in our strategyproof classification work, such as the emphasis on social welfare and approximation, seem to be pertinent in judgment aggregation as well.

Interestingly, the techniques we used thus far seem insufficient for proving lower bounds (both deterministic and randomized) in realizable scenarios. Since realizability is common and sometimes necessary (e.g. in judgment aggregation), it is important to close these gaps, and we are currently studying this problem.

**Future directions** New models of strategic learning that better encapsulate the practical challenges of learning theory and game theory should be developed. One example to a variation of the current model is a more realistic description of public and private information, as well as the information that agents are assumed to have on the preferences of other agents. Other variations may include the use of various loss functions, that are commonly used in the machine learning literature, as well as in actual off-the-shelf learning algorithms (for example SVM).

Another important line of study is an empirical validation of the actual bounds of SP mechanisms (which might prove better than the theoretical bounds). Such experiments should also explicitly compare SP mechanisms to naïve algorithms, either on synthetic data assuming specific well-defined types of strategic behavior (in the spirit of [37]); or with actual human experts acting strategically.

## 3 Other Projects

### 3.1 Cooperative games and the Cost of Stability

In many settings, such as online auctions and other types of markets, agents act individually. In this case, the standard notions of noncooperative game theory, such as *Nash equilibrium* or *dominant-strategy equilibrium*, provide a credible prediction of the outcome of the interaction. However, another frequently occurring type of scenario is that agents need to form teams to achieve their individual goals. In such domains, the focus turns from the interaction between single agents to the capabilities of subsets, or *coalitions*, of the agents.

*Cooperative games* (a.k.a. coalitional 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 e.g. [36] for an overview). When the agents are selfish, 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 most prominent solution concept that aims to formalize the idea of stability in coalitional 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 [8, 42, 3].

In a line of recent papers we approach this issue from a mechanism design perspective (see [6, 5, 40, 26, 34], and a follow up on our work by Aziz et al. [4]). 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 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 [20, 7].

In our papers, we study bounds on the Cost of Stability in various games and suggest algorithms to compute it efficiently, when possible.

**Current and future work** Our current line of study focuses on gaining a better understanding of the Cost of Stability, and in particular its relations with other solution concepts such as those mentioned earlier.

### 3.2 Social Choice

Social choice theory is perhaps the oldest field in game theory, with concrete roots back in the 18th century [10]. Its applications are not restricted to political elections, as collective decision making in the modern world occurs everywhere and involves committees, firms, interest groups and even computerized agents. Much of the theory is dealing with the issue of *manipulation*,



when voters report false preferences in an attempt to bias the elections. Such strategic behavior is typically considered hazardous, as it means that the outcome of the elections does no longer reflect the true preferences of the society (similarly to the problem that we challenged in the strategyproof classification project). The severity of this problem was accentuated in the 70's, when it was proved that in every reasonable voting system some voters may be motivated to lie [16, 41].

Researchers in economics and political sciences have been suggesting various solutions to manipulations, where in the last two decades an unexpected assistance arrived from the field of *artificial intelligence* (see an overview in [15]), an effort that I also contributed to [28, 32].

However, even if individual preferences are known and voters are truthful, the fact that there are many different voting systems suggests that the "preferences of the society" can be interpreted in multiple ways.<sup>2</sup> Our initial results indicate situations in which strategic behavior in the common Plurality voting system in fact promotes candidates that are more acceptable (according to other systems). Other studies support similar conclusions in various contexts [35, 44].

In our study [27], we considered a voting scenario where voters can change their vote according to the current score of each candidate (which is a public knowledge). The winner is determined when no voter is willing to change her vote. We assumed that voters are unaware of the hidden incentives of other voters, and that they act myopically, i.e. thinking only one step ahead. Such behavior can be justified in this context, as the voter always has the option to change the vote again later. We proved conditions under which convergence is guaranteed, and analyzed the rate of convergence.

**Current and future directions** Further study of the principles that guide the decision making of (human) voters, without necessarily tagging it as a "negative" behavior, would help us to better understand the nature of collective decision making and voting processes, and possibly to improve them. More specifically, we are interested in comparing our (simple) behavioral assumptions with actual human behavior (see also the closing section), and enhance our formal model accordingly.

### 3.3 Better for Everyone: Improving leasing agreements

While most of my work deals with abstract models of interaction, this approach can most certainly be applied to improve our everyday life. One problem that can be tackled using better mechanisms, is the problem of excessive use in public resources. In a recent paper [33], we analyzed the standard car leasing agreements where the fuel of an employee is included in the deal, and is paid for by the hiring company. It is known from previous studies that such arrangements significantly increase the total mileage, which bares grave effects on road congestion, air pollution, and prompts other environmental and social hazards [9]. This is a clear case where a rational behavior on the part of the individual (using free fuel) is destructive for the society.

Using a game-theoretic analysis, we showed that an alternative leasing model, in which the employee pays for her own fuel, induces a new equilibrium between the company and its

---

<sup>2</sup>This is in contrast to the problem of strategyproof classification, where there is a more explicit standard for system performance.

employees. In the new outcome everybody gains (compared to the current equilibrium), as both the company and the employee save money and enjoy lighter congestion. We further discussed possible explanations for the unpleasant fact that paid-for fuel is still prevalent, and how alternative agreements can be promoted. In this work we do not make any assumptions on the actual price of fuel, work, etc., but only assume that there is *decreasing marginal utility* from driving more, which is a reasonable assumption.

The described situation is not unique to the leasing world, and is in fact common whenever there is a valuable resource given free of charge. This is since in many cases the resource (fuel in our case) is not really free. There are explicit costs (production and transfer) and implicit costs (environmental effects), that are being externalized on the the other employees and on the public. The use of mechanisms, such as the alternative leasing model described here, guarantee that each user bears the costs of his own used resource, and will therefore refrain from excessive use.

**Current and future directions** In addition to natural future directions, such as extending our model to consider taxation effects, we are collaborating with “Transportation Today and Tomorrow”, an organization that promotes sustainable transportation solutions. Hopefully, theoretic results such as ours can be used to accentuate the conclusions of field research regarding paid-for fuel. A joint effort may be able to persuade decision makers that replacing the standard leasing model will be better for everyone.

### 3.4 New solution concepts

Solution concepts in game theory are intended to predict the expected behavior of agents in various interactions, under certain assumptions on their capabilities. The best known example is *Nash equilibrium*, which assumes that every single agent will follow his best possible strategy, but agents cannot (or are not allowed to) make joint deviations from the equilibrium profile. Many other solution concepts for normal form games have been suggested, for example to deal with such joint deviations (strong equilibrium) or with multiple equilibria (Pareto efficiency). More equilibrium concepts have been formed for extensive form games (subgame perfect), coalitional games (see Section 3.1), and other types of games.

Choosing which concept to apply depends ultimately on the underlying assumptions one makes on the game and the participating agents. This is particularly important in mechanism design, where the designer strives to implement a certain outcome or a desired property (e.g. maximizing revenue or social welfare). It is easier, for example, to guarantee that a some action will be a Nash equilibrium, than making it a strong equilibrium. However, if players in the considered game have no way to communicate then joint deviations are unlikely, and thus strong equilibria might be too strong a requirement. A rich repertoire of solution concepts is therefore a necessary tool in mechanism design.

In a recent paper with Feldman and Tennenholtz (in submission), we introduced *stability scores* – a quantitative measure to the stability of an outcome based on the counting the coalitions that might deviate. This measure allows us to compare multiple Nash equilibria for example, and identify the one more stable against coalitional deviations. As an example application, we analyzed the stability of two common ad auction mechanism, showing that under certain assumptions on bidders’ valuations one of the mechanisms is far more stable.

**Current and future directions** Other than applying the stability score measure to various domains, we are also working on certain refinements of other equilibrium concepts in normal form and extensive form games.

## 4 Beyond rational agents

Standard game theory 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.<sup>3</sup> In particular agents are assumed to be risk-neutral and games are invariant under certain simple changes. Moreover, agents are assumed to have perfect knowledge of their environment, and to make the best rational decisions based on this knowledge. Most of my own work thus far is using the same standard assumptions. The iterated voting project is the only partial exception, where we tried to make minimal assumptions on the rationality and knowledge available to the voters.

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 [21]). Following experiments in decision making, many empirical findings in games played by human players have been collected by Camerer [11], 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 [43]. 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 order to have a real scientific value, a theory cannot ignore findings in the world that consistently contradict its predictions. In fact, correctly formalizing cognitive biases and treat them within the theory poses a great challenge that may also lead to substantial theoretical breakthroughs. 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 on single decision makers. 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.

That said, behavioral game theory is new to me, and I am still in the reading phase. Therefore it is hard for me to predict at this stage to what extent such ideas will shape my PhD thesis.

---

<sup>3</sup>These have been explicitly stated by von-Neumann and Morgenstern [45].

## References

- [1] N. Alon, M. Feldman, A. D. Procaccia, and M. Tennenholtz. Strategyproof approximation of the minimax on networks. *Mathematics of Operations Research*, 35(3):513–526, 2010.
- [2] I. Ashlagi, F. Fischer, I. Kash, and A. D. Procaccia. Mix and match. In *Proceedings of the 11th ACM Conference on Electronic Commerce (ACM-EC)*, pages 305–314, 2010.
- [3] R. Aumann and M. Maschler. The bargaining set for cooperative games. In M. Dresher, L. S. Shapley, and A. Tucker, editors, *Advances in game theory, Annals of mathematical study*, volume 52, pages 443–476. Princeton University press, 1964.
- [4] H. Aziz, F. Brandt, and P. Harrenstein. Monotone cooperative games and their threshold versions. In *Proceedings of the 10th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1017–1024, May 2010.
- [5] Y. Bachrach, E. Elkind, R. Meir, D. Pasechnik, M. Zuckerman, J. Rothe, and J. S. Rosenschein. The cost of stability in coalitional games. In *Proceedings of the 2nd International Symposium on Algorithmic Game Theory (SAGT)*, pages 122–134, October 2009.
- [6] Y. Bachrach, R. Meir, M. Zuckerman, J. Rothe, and J. S. Rosenschein. The cost of stability in weighted voting games (extended abstract). In *Proceedings of the 8th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1289–1290, Budapest, Hungary, May 2009.
- [7] C. Bejan and J. C. Gómez. Core extensions for non-balanced TU-games. *Int. journal of game theory*, 38:3–16, 2009.
- [8] J. M. Bilbao. *Cooperative Games on Combinatorial Structures*. Kluwer Publishers, 2000.
- [9] W. R. Black. *Sustainable Transportation: Problems and Solutions*. The Guilford Press, 2010.
- [10] J.-C. d. Borda. Memoires sur les elections au scrutin, 1781. *Paris: Histoire de l’Academie Royale des Sciences. Translation in Alfred de Grazia, 1953, “Mathematical Derivation of an Election System”*. *Isis* 44:42-51.
- [11] C. F. Camerer. *Behavioral game theory: experiments in strategic interaction*. Princeton uni. press, 2003.
- [12] O. Dekel, F. Fischer, and A. D. Procaccia. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76:759–777, 2010.
- [13] E. Dokow and R. Holzman. Aggregation of binary evaluations. *Journal of Economic Theory*, 145:495–511, 2010.
- [14] S. Dughmi and A. Ghosh. Truthful assignment without money. In *Proceedings of the 11th ACM Conference on Electronic Commerce (ACM-EC)*, pages 325–334, 2010.

- [15] P. Faliszewski and A. D. Procaccia. AI's war on manipulation: Are we winning? *AI Magazine*, 31:53–64, 2010.
- [16] A. Gibbard. Manipulation of voting schemes. *Econometrica*, 41:587–602, 1973.
- [17] M. Guo and V. Conitzer. Strategy-proof allocation of multiple items between two agents without payments or priors. In *Proceedings of the 9th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 881–888, 2010.
- [18] M. Guo, V. Conitzer, and D. Reeves. Competitive repeated allocation without payments. In *Proceedings of the 5th International Workshop on Internet and Network Economics (WINE)*, pages 244–255, 2009.
- [19] B. P. Harrenstein, M. M. de Weerd, and V. Conitzer. A qualitative vickrey auction. In *Proceedings of the 10th ACM Conference on Electronic Commerce (ACM-EC)*, pages 197–206, New York, NY, USA, 2009. ACM.
- [20] K. Jain and V. V. Vazirani. Applications of approximation algorithms to cooperative games. In *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing (STOC)*, pages 364–372, 2001.
- [21] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, XLVII:263–291, 1979.
- [22] C. List and C. Puppe. Judgement aggregation: A survey. In P. Pattanaik, P. Anand, and C. Puppe, editors, *The Handbook of Rational and Social Choice*. Oxford University Press, 2009.
- [23] P. Lu, X. Sun, Y. Wang, and Z. A. Zhu. Asymptotically optimal strategy-proof mechanisms for two-facility games. In *Proceedings of the 11th ACM Conference on Electronic Commerce (ACM-EC)*, pages 315–324, 2010.
- [24] R. Meir. Strategy proof classification. Master's thesis, The Hebrew University of Jerusalem, 2008. Available from: <http://www.cs.huji.ac.il/~reshef24/spc.thesis.pdf>.
- [25] R. Meir. Research proposal: Cooperation among self interested agents. In *Proceedings of the 22th International Joint Conference on Artificial Intelligence (IJCAI)*, 2011. to appear.
- [26] R. Meir, Y. Bachrach, and J. S. Rosenschein. Minimal subsidies in expense sharing games. In *Proceedings of the 3rd International Symposium on Algorithmic Game Theory (SAGT)*, pages 347–358, 2010.
- [27] R. Meir, M. Pofukarov, J. S. Rosenschein, and N. Jennings. Convergence to equilibria of plurality voting. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI)*, pages 823–828, Atlanta, Georgia, July 2010.

- [28] R. Meir, A. D. Procaccia, and J. S. Rosenschein. A broader picture of the complexity of strategic behavior in multi-winner elections. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 991–998, 2008.
- [29] R. Meir, A. D. Procaccia, and J. S. Rosenschein. Strategyproof classification under constant hypotheses: A tale of two functions. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*, pages 126–131, 2008.
- [30] R. Meir, A. D. Procaccia, and J. S. Rosenschein. Strategyproof classification with shared inputs. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 220–225, 2009.
- [31] R. Meir, A. D. Procaccia, and J. S. Rosenschein. On the limits of dictatorial classification. In *Proceedings of the 9th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 609–616, 2010.
- [32] R. Meir, A. D. Procaccia, J. S. Rosenschein, and A. Zohar. Complexity of strategic behavior in multi-winner elections. *Journal of Artificial Intelligence Research*, 33:149–178, 2008.
- [33] R. Meir and J. S. Rosenschein. A game-theoretic approach to leasing agreements can reduce congestion. In *The Sixth Workshop on Agents in Traffic and Transportation (ATT'10), at The Ninth International Joint Conference on Autonomous Agents and Multiagent Systems*, Toronto, 2010.
- [34] R. Meir, J. S. Rosenschein, and E. Malizia. Subsidies, stability, and restricted cooperation in coalitional games. In *Proceedings of the 22th International Joint Conference on Artificial Intelligence (IJCAI)*, 2011. to appear.
- [35] R. B. Myerson and R. J. Weber. A theory of voting equilibria. *The American Political Science Review*, 87(1):102–114, 1993.
- [36] B. Peleg and P. Sudhölter. *Introduction to the Theory of Cooperative Games*. Kluwer Publishers, 2003.
- [37] J. Perote and J. Perote-Peña. Strategy-proof estimators for simple regression. *Mathematical Social Sciences*, 47:153–176, 2004.
- [38] J. Perote-Peña and J. Perote. The impossibility of strategy-proof clustering. *Economics Bulletin*, 4(23):1–9, 2003.
- [39] A. D. Procaccia and M. Tennenholtz. Approximate mechanism design without money. In *Proceedings of the 10th ACM Conference on Electronic Commerce (ACM-EC)*, pages 177–186, 2009.
- [40] E. Resnick, Y. Bachrach, R. Meir, and J. S. Rosenschein. The cost of stability in network flow games. In *Proceedings of the 34th International Symposium on Mathematical Foundations of Computer Science (MFCS)*, pages 636–650. Springer, August 2009.

- [41] M. Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10:187–217, 1975.
- [42] D. Schmeidler. The nucleolus of a characteristic function game. *SIAM journal on applied mathematics*, 17:1163–1170, 1969.
- [43] H. Simon. A behavioral model of rational choice. New York: Wiley, 1957.
- [44] F. D. Sinopoli. Sophisticated voting and equilibrium refinements under plurality rule. *Social Choice and Welfare*, 17:655–672, 2000.
- [45] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton Univ. Press, 1944.