Noname manuscript No.
(will be inserted by the editor)

Autonomous Agents and Human Cultures in the Trust-Revenge Game

Amos Azaria · Ariella Richardson · Avi Rosenfeld

the date of receipt and acceptance should be inserted later

Abstract Autonomous agents developed by experts are embedded with the capability to interact well with people from different cultures. When designing expert agents intended to interact with autonomous agents developed by Non Game Theory Agents (NGTE), it is beneficial to obtain insights on the behavior of these NGTE agents. Is the behavior of these NGTE agents similar to human behavior from different cultures? This is an important question as such a quality would allow an expert agent interacting with NGTE agents to model them using the same methods that are used to model humans from different cultures.

To study this point, we evaluated NGTE agents behavior using a game called the Trust-Revenge game, which is known in social science for capturing different human tendencies. The Trust-Revenge game has a unique subgame-perfect equilibrium strategy profile, however, very rarely do people follow it. We compared the behavior of autonomous agents to the actions of several human demographic groups—one of which is similar to the designers of the autonomous agents. We claim that autonomous agents are similar to human players from various cultures. This enables the use of approaches, developed for handling cultural diversity among humans, to be applied for interaction with NGTE agents. This paper also analyzes additional aspects of autonomous agents behavior and whether composing autonomous agents affects human behavior.

1 Introduction

Autonomous agents are integrated into numerous environments, such as electronic commerce, web crawlers, military agents, space exploration probes, autonomous drivers, disabled assistance and social interaction [1–5]. Autonomous agents designed by game theory

Amos Azaria
Department of Machine Learning, Carnegie Mellon University, Pittsburgh PA
E-mail: azariaa@cs.cmu.edu

Ariella Richardson
Department of Industrial Engineering, Lev Academic Center, Israel
E-mail: ariellarich@gmail.com

Avi Rosenfeld
Department of Industrial Engineering, Lev Academic Center, Israel
E-mail: rosenfa@gmail.com

experts often implement a fully rational strategy. However, the rise in computer science education along with the availability of software development tools has increased the number of autonomous agents developed by non game theory experts (NGTE) and amateurs. As a result, when designing autonomous agents one must consider that they will also interact with the NGTE agents.

Modeling agents is beneficial for agent-agent interaction [6]. However, building an exact model is often infeasible. There are many studies on agents that interact with fully rational agents [7] or humans [8–18]. However one cannot assume that the NGTE autonomous agents are rational [19], or that they exhibit the same behavior as the population that designed them [20,21]. This raises a challenge for an agent interacting with NGTE agents.

Recent autonomous agents developed by experts are embedded with the capability to interact well with different cultures [22,23]. These agents were built to explicit reason about and react to the specific cultures they considered in several ways. Our work particularly follows Haim et al.'s negotiation agent that considers different cultures. Their PAL agent contains a meta-framework that uses a combination of Machine Learning and Decision Theory to tune parameters that fit each culture [23]. Specifically, they posit that general social factors regarding generosity and reliability exist across all cultures, but to varying degrees. The goal of the agent is then to learn the exact levels of these social factors in a given culture so that a more effective negotiation agent can be built.

The main focus of this study is to determine whether the behavior of NGTE agents is similar to human behavior from different cultures. We posit that although NGTE agents use different strategies than people these differences can be effectively modeled as differences between cultures. Once one realizes that NGTE agents can be modeled as a distinct culture, then methods developed for interacting with distinct cultures (such as [22,23]) can be applied to better interact with NGTE agents. This is a non-trivial result, since if NGTE agents were to behave very differently of humans (e.g. if they were nearly fully rational), then using methods developed for interaction with multiple cultures would likely fail, since such methods usually rely on human properties known from social science [24]. We use the term culture loosely, to refer to demographic groups, as is common in the literature.

In order to experimentally confirm our hypothesis that the behavior of autonomous agents falls within cultural diversities we use the Trust-Revenge game. This game is an extension of the investment game introduced by Berg [25]. The investment game was developed in order to study investment and reciprocity in an economic setting, and has generated much interest since its inception [26]. This game was later enhanced with an extra stage where the player can choose to take revenge [27].

In this paper we analyze the behavior of NGTE autonomous agents and compare it to the behavior of groups of humans from three different cultures. We introduce a quantitative measure, and use it to show that in the trust-revenge game the autonomous agents average action is similar to the human cultures. We therefore determine that NGTE autonomous agents can be described as a separate human culture and conclude that modeling NGTE autonomous agents the same way human cultural variations are modeled, is likely to be a generally useful approach. The findings of this research have practical implications for designing agents that interact with autonomous agents designed by NGTE. An additional contribution of this study is, that we determine whether the programming of autonomous agents has an impact on the composers in our game.

2 Related Work

This paper's significant contribution lies in its empirical finding that NGTE agents can be modeled as a distinct culture within a general trust-revenge game involving an opponent. Opponent modeling is of great importance when designing autonomous agents which interact with other agents [6]. Carmel and Markovitch [28] showed that if an agent can build a model of its opposing agent, this model could be used to improve its own performance. McCracken and Bowling [29] proposed a method for modeling an opponent in the Rock-Paper-Scissors domain. Lazaric et al. [30] proposed a method for opponent modeling in Kuhn Poker (a degenerated version of poker). However, in more sophisticated games these methods become intractable.

Another field which raised significant interest is whether people develop agents similar to their own behavior [31–35]. While we will survey some of these works below, the overall trend within these works is that while there are some similarities between the agents developed and their developers' behavior, they are not identical [20,21,35]. In this paper, we study whether these differences are within the differences observed among different cultures. Grosz et al. [21] introduced the Colored Trails game specifically in order to investigate decision-making strategies in multi-agent situations. They showed that human players and agents do not play in the same way. Their results indicate that people design autonomous agents that don't play as well as human players, possibly because they cooperate less than people. It is interesting to note that the agents also do not adhere to the equilibrium strategy. Unlike Grosz et al. [21] who studied a mixed human-agent environment, since this paper studies understanding agent behavior, we required that agents play with agents and humans play with humans. Manisterski et al. [36] explored the evolution of autonomous agents in a negotiation environment. Designers were given a chance to improve the agents based on previous performance. Their findings showed that autonomous agents seem to perform better than equilibrium agents but not as well as Pareto-optimal agents. However, they did not compare the behavior of agents to that of humans. Rosenfeld and Kraus [19] demonstrated that autonomous agents for search domains use key elements of bounded rationality known from behavioral economics (such as the Aspiration Adaptation Theory). This work may be an important indication to our initial questioning. However, unfortunately, they did not compare the behavior of agents to that of humans either. Chalamish et al. [37] tested the similarity between human strategies and those of autonomous agents. They concluded that human behavior is often similar to that of their own autonomous agent. Unfortunately, in their study, the human player subjects were also the composers of the autonomous agents and were requested to play many games before they composed their agents. Therefore, when composing the agents, the subjects may have been heavily influenced by how they played the game earlier. Work by Lin et al. [35] evaluated if NGTE agents—which they term Peer Designed Agents (PDA) could be used instead of people to evaluate expert negotiation agents. They found that while significant differences did exist between PDAs and people's behaviors, PDAs did provide strong indications about how the people would behave. Nonetheless, their work did not attempt to quantify when these similarities and differences exist as our work does. Note that PDA's are usually used to simulate the behavior of humans. In our paper we use the term NGTE agents, rather than PDA's, to emphasize that we study interaction with autonomous agents developed by NGTE in general (as is becoming widespread), not specifically PDA's which are used for simulations.

Many previous studies have examined how culture affects players' behavior. Examples include studies designed to compare cultures [38,39] and meta-studies [40,26,41]. Hofstede [42] studied the causes of differences between cultures, while we are interested in observing

these differences and not in explaining them. Willinger et al. [38] compared French and German players of the investment game, and showed that German players invested more than French players, while reciprocity does not differ between the two groups. Johnson and Mislin [26] performed a meta-analysis of the investment game by collecting studies conducted in various settings with different cultures and comparing the results. They found that demographic groups do indeed play differently, and less was spent in investment games played in Africa than in North America. As these studies show, there are differences between different cultures or demographic groups. This motivates us to study whether autonomous agents display behavior that falls within the diversity of human cultures.

Studies on expert agents that interact with different human cultures provide motivation to our study. Haim et al. [23] rely on measures adapted from psychology [43] to build their PAL agent using machine learning. Gal et al. [22] also build an expert agent that is designed to interact with different cultures. This agent although being adaptive is Rule Based. Our results imply, that these agents should be able to interact well with NGTE agents as well.

Interaction with different cultures has raised interest in the domain of virtual agents as well [44,45]. Mascarenhas et al. [44] propose a model that allows agents to adapt their relational behavior to different cultures. They propose a model where agents can be given different social interaction dynamics, that represent different cultures. These agents can be used to train humans on interaction with different cultures. Kistler et al. [45] explain that humans from different cultures practice varying non-verbal behavior. They introduce a virtual agent that can display culture-specific behaviors, enabling a comfortable interaction between the agent and people. However, our results may have milder impact on agents using virtual human characters since NGTE agents are less likely than humans to be influenced by the behavior of a virtual human character.

In this paper we also test whether the simple act of composing agents has any impact on the behavior of the composers. This exact question was raised by Elmalech et al. [46]. They studied the 'door game', which is a single player game developed by Shin and Ariely [47]. In this game, a player is faced with three doors, each associated with a different distribution of payoffs. The distribution of each door is a priori unknown to the player. The player first chooses with which door to begin, and from that point on, any additional click on that door will yield a reward drawn from that distribution. At any time, the player may pay some cost and switch to any other door by clicking on it. In their work they showed that agent composition alters the behavior of the composers, i.e. that subjects which composed agents behaved differently than subjects which did not compose any agents. Our findings are described in Section 5.

3 Trust-Revenge Game

The Trust-Revenge Game, which will be described shortly, is a two player game designed to arouse three different behaviors. We term these behaviors - trust, reciprocation (fear) and revenge. Although trust and fear are not behaviors, we use these terms to describe behaviors (and actions) that might be performed as a result of trust or fear.

Research with human subjects on trust, reciprocation and revenge (or punishment) has been conducted in the past. The *investment game* was first introduced by Berg [25]. In the *investment game* there are two types of players. Each player is given 10 chips at the beginning of the game. Players of type A are told that they can give some or all of their chips to a player of type B (this is the trust stage). The number of chips that Player A decides to give is multiplied by 3. Then Player B can give back some or all of what he was given

(the reciprocation stage). The subgame-perfect equilibrium for this game is for both types of players to send nothing. Berg conducted the experiment with students (human subjects). As expected, the human subjects did not act according to the subgame-perfect equilibrium, and chips were transferred by both types of players.

Gneezy and Ariely [27] used a variant of the investment game which included an additional revenge phase. In their experiment, each of the two players started off with $\$10$. The first player had to decide whether he wanted to end the game or to pass the full amount to the second player. If he decided to pass his money, the second player received an additional $\$40$ for a total of $\$50$. Then the second player had to decide whether to keep all of the money or to give half of the money back to the first player. If the second player decided to keep all of the money for herself, the first player could decide to take revenge on her and pay any amount from his own private money (up to $\$25$); this amount was multiplied by 2 and subtracted from the second player's revenue. The results of Gneezy and Ariely's experiment show that the first player often takes revenge on the second player (when the second player keeps all of the money).

We use a variant of the game used by Gneezy and Ariely: the Trust-Revenge Game. This game is composed of three stages: *Trust*, *Reciprocate* and *Revenge*. This game is a "one-shot" game, i.e. after the three stages are completed, the game terminates (there are no repeated interactions). There are two types of players (A and B) in the game. At the beginning of the game Players A and B are both given a certain number of chips. The first stage is the *Trust* stage, where Player A is able to give any portion of his chips to Player B. There is a factor - the Trust Rate (**tr**) - by which the number of chips is multiplied when they are passed from Player A to Player B. The second stage is *Reciprocate*: after the chips have been transferred to Player B, Player B can decide how many chips to transfer back to Player A. Player B can transfer any number of chips (which she acquires) to player A. The third and final stage is *Revenge*: Player A plays another round in which he may pay any number of chips he has to the operator. Note that the chips are not transferred to anyone, merely subtracted from Player A's stack. However, in this round, Player B must pay a factor - Revenge Rate (**rr**) - on the number of chips Player A chose for revenge. Again, the chips are not transferred to anyone but merely subtracted from Player B's stack. Both the Trust Rate and the Revenge Rate are common knowledge and are revealed to both players at the beginning of the game.

Consider the following example: Assume that the Trust Rate, **tr**, equals 4 and that the Revenge Rate, **rr**, equals 6. Assume that both players started with 10 chips. Suppose that Player A gives 5 chips to Player B in the Trust stage. After applying the Trust Rate, Player B will receive 20 chips ($5 \cdot$ **tr**). At the end of this stage Player A has 5 chips and Player B has 30. Suppose Player B decides to give 7 chips to Player A in the Reciprocate stage. At the end of this stage Player A has 12 chips ($5 + 7$) and Player B has 23 chips ($30 - 7$). Suppose Player A revenges 3 chips at the Revenge stage. At the end of this stage (which ends the game) Player A has 9 chips ($12 - 3$), and Player B has 5 chips ($23 - 3 \cdot$ **rr**). Since this is a "one-shot" game, after the game ends, Player A and Player B have no further interactions.

In this game there is a clear, unique subgame-perfect equilibrium (SPE) strategy. In the revenge stage, there is no rational reason for Player A to take revenge, therefore in the SPE there is no revenge. In the reciprocation stage there is no reason for Player B to reciprocate since she assumes that Player A is rational and that he will not take revenge, therefore in the SPE there is no reciprocation. As a result, in the trust stage there is no rational reason for Player A to trust Player B since he knows that she will not reciprocate. Consequently the SPE is not to take revenge, not to reciprocate and not to trust.

Settings	Player A Initial	Player B Initial	Trust Rate	Revenge Rate
Investment	10	10	3	0
Dictator	20	0	1	0
TR 1	10	10	3	3
TR 2	10	10	6	6
TR 3	20	0	6	6

Table 1 Settings Used in the Trust-Revenge Game

We did not expect human subjects to follow the SPE and we presumed they transfer chips in all three stages (in accordance with the literature [25–27]). We also believed we would find differences in behavior between the different cultures. We examine the NGTE agents' behavior to see if it is similar enough to humans from various cultures to be considered within cultural diversity.

4 Experimental Setup

4.1 Game Settings

Throughout the experiments we used 5 different settings of the Trust-Revenge Game, shown in Table 1. These settings were chosen meticulously to capture the following behaviors: The *Investment* setting, is analogous to a simple classical trust (investment) game, with no revenge. In such a game, Player A cannot enforce or threaten Player B in any way, and any amount returned by Player B may be associated with reciprocation rather than fear of revenge. The *Dictator* setting is analogous to a dictator game, where one player starts with all the money and may contribute any amount to the second player. In this setting, Player A cannot attempt to gain anything from transferring chips over to Player B in the trust stage, since player B may only return in the reciprocate stage some (or all) the money received from Player B, but not more. Thus, any amount transferred in this setting may be attributed to generosity. The *TR 1* setting is a classical Trust-Revenge setting, in which both the trust and the revenge rates are set to 3. The *TR 2* setting doubles both the trust and the revenge rates; we expected this setting to increase average trust and revenge. The *TR 3* setting gives Player A double chips and Player B no chips; we expected this setting to increase average trust even more.

4.2 Subjects

A set of 36 undergraduate computer science students from Israel composed NGTE agents for the Trust-Revenge Game (the *Agents* group). 6 agents were removed since they either did not compile or raised exceptions in run time, this resulted in a total of 30 agents. These developers were asked to provide detailed documentation explaining their design of the agent.

This group of developers also played all the five settings of the Trust-Revenge game themselves (as humans) after they composed the agents. However, since developing an agent may have had impact on the way that these developers played the game themselves, we do not consider these games in the results section until we explicitly question whether developing an agent has impact on the developers behavior at the end of Section 5.

Group name	Role	Country	Type	Motivation	Num. of subjects	Avg. age	Stdev age	Female percent	Total number of games
Agents	Agent design	Israel	Students	Grade	36(30)	27.7	6.8	19.4%	4350
Israel	Human player	Israel	Students	Grade	35	27.4	5.5	5.7%	175
USA	Human player	USA	AMT	Monetary	50	29.3	7.6	40%	250
India	Human player	India	AMT	Monetary	46	30.3	6.5	35.4%	230

Table 2 Summary of the four different groups of subjects, including their demographic details

A different group of 35 undergraduate computer science students from Israel played the Trust-Revenge game with each other (these subjects did not compose agents). Two additional sets of players, one from the USA (50 subjects) and the other from India (46 subjects), played the game with each other (USA players played with other USA players and players from India played with other players from India). These players were recruited using Amazon's Mechanical Turk (AMT) [48] (these subjects did not compose agents either). Table 2 summarizes the different groups of subjects and provides some demographic details on each of the groups.

4.3 Number of Games and Motivation

Each autonomous agent (in the *Agent* group) played twice against all of the other agents, once as Player A and once as Player B in each of the 5 different settings. Since there were 30 agents and each agent played against all other 29 agents twice in each of the 5 settings, in total, each agent played $29 \cdot 5 \cdot 2 = 290$ summing up in a total of $290 \cdot 30 \cdot \frac{1}{2} = 4350$ games for all agents (we divide by 2 since every game requires two agents). Since the game is designed as a one-shot game, the agents were initialized every game and thus were unable to save information from game to game.

All human subjects (including the Israeli students who did not develop agents and the USA and India subjects) played 10 consecutive games[1]. To preserve the characteristic of a one-shot game, we ensured that the players were paired with different partners every game. The subjects were fully aware of this fact. All human subjects were required to pass a short quiz to ensure that they fully understood the game.

The two groups of students (those who composed agents and those who played as humans) were motivated by their grades to compose good agents and play well. In the group that composed agents, the students' grades depended on their agents' performance and its documentation. In the Israeli students who did not compose agents were motivated by their grades as well, since they were offered a bonus proportionate to their performance which was added to the grade of one of their assignments. The subjects in the USA and India groups which were recruited via AMT were motivated by a monetary payment which was proportionate to the final stake in each of the games.

The performance of all subjects and agents (used for granting the promised motivation) was measured according to its final result (relative to the number of chips) and did not depend on the opponent's performance or on the average performance. All subjects and agent designers were explicitly informed of this policy. The designers of the autonomous agents knew that their agents would play with other autonomous agents designed by their

[1] All human subjects played all five settings as Player A exactly once in random order, and played five more games as player B. Unfortunately, due to matching difficulties, we could not ensure that all human subjects played each of the five settings as player B.

Fig. 1 Screen-shot of trust-revenge game. In this game, the player is playing the role of player A and is in the final revenge stage. Note that the revenge rate is 0.

classmates. The students playing on the web knew that they were playing against a different student from their class, however they did not know against whom they were playing and we assured them that this information would remain confidential. The AMT subjects were also informed that they were playing with humans from their own culture. The human interface was fully web-based. Figures 1 and 2 present screen-shots of the trust-revenge interface which the subjects played with in different settings and stages of the game.

5 Results

Recall that the main question of this paper is to determine whether the behavior of NGTE agents falls within cultural diversities. That is, whether the differences between NGTE agents and humans are quantitatively similar to the differences typically found between humans from different cultures. In order to answer this question we define the following quantitative measure: A group A, is considered to fall among the diversity of other groups $\mathcal{B} = \{B_1, B_2, B_3, ...\}$ if its average is between the average of the other groups minus their standard deviation, and the average of the other groups plus their standard deviation, i.e. $avg(A) \in avg(\cup\mathcal{B}) \pm stdev(\{avg(B_1), avg(B_2), avg(B_3), ...\})$. Note that the standard deviation is calculated on the groups $(B_1, B_2, B_3, ...)$ and not on the individuals inside these

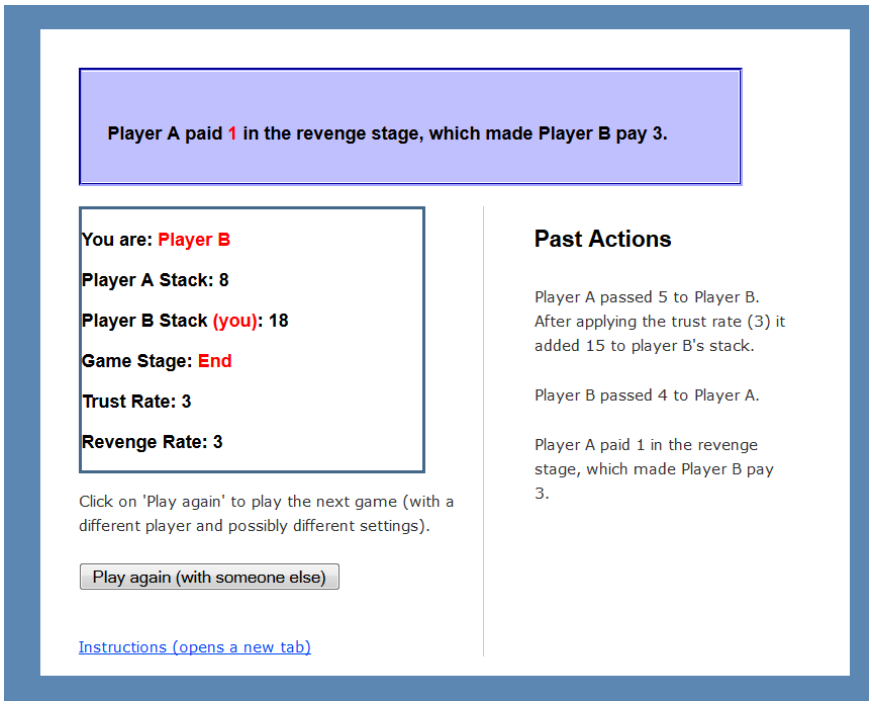

Fig. 2 Screen-shot of trust-revenge game after the game ended. The player will now click the "play again (with someone else)" button.

groups[2]. We will show that in the trust revenge game, the group of autonomous agents falls within the average action of the three human cultures plus or minus the standard deviation among these three human cultures. Note, that assuming a normal distribution, only 68% of the population is expected to fall within 1 standard deviation of the mean. Clearly, if a group is not part of the population, the profitability of it falling within this diversity would be low. Nevertheless, we show that the autonomous agents, fall within this range in all three game actions.

We begin by presenting the average results for the strictly Trust-Revenge settings (*TR 1*, *TR 2* and *TR 3*). Figure 3 presents the number of chips transferred (or paid) on average at each stage of the game. The black bars show the confidence range ($\alpha = 0.05$). (Note that the confidence range is much smaller for the agents since they played many more games.) We conducted a Multivariant test with the trust, reciprocate and revenge amounts in each of the games played as the dependent variables and the culture and the settings as fixed variables. As expected, both culture and settings had significant impact on the behavior ($p < 0.0001$), with $F = 57.964$ and $F = 27.542$ respectively. Also the two-way interaction between culture and settings (culture $\times$ settings) is statistically significant ($p < 0.0001$)

[2] We also considered a different criterion which tests whether the difference between the group A and $\cup\mathcal{B}$ are statistically significant. However, such a criterion would be sensible to the size of the data since, theoretically, any two groups which have different averages would differ statistically significantly if enough data is gathered. Futhermore, we do not claim that group A's distribution or average is similar to that of $\cup\mathcal{B}$, but simply that it falls among the diversity of the groups in $\mathcal{B}$. We therefore did not test this method on our data.

Group	Agents	Israel	USA	India
Agents	-	**0.035**	**0.000**	1.00
Israel	**0.035**	-	**0.000**	0.189
USA	**0.000**	**0.000**	-	**0.000**
India	1.00	0.189	**0.000**	-

Table 3 Scheffe post-hoc p-value results on Trust amounts.

Group	Agents	Israel	USA	India
Agents	-	**0.048**	**0.000**	0.989
Israel	**0.048**	-	**0.000**	0.302
USA	**0.000**	**0.000**	-	**0.000**
India	0.989	0.302	**0.000**	-

Table 4 Scheffe post-hoc p-value results on Reciprocate amounts.

Group	Agents	Israel	USA	India
Agents	-	0.309	0.969	**0.000**
Israel	0.309	-	0.354	0.365
USA	0.969	0.354	-	**0.002**
India	**0.000**	0.365	**0.002**	-

Table 5 Scheffe post-hoc p-value results on Revenge amounts.

Stage	Agents	Israel	USA	India	$mean$	$stdev$	$mean - stdev$	$mean + stdev$
Trust	**3.34**	4.36	8.07	3.38	5.27	2.48	**2.8**	**7.75**
Reciprocate	**4.09**	6.49	19.4	4.36	10.08	8.14	**1.94**	**18.22**
Revenge	**1.26**	1.69	1.16	2.23	1.69	0.53	**1.16**	**2.23**

Table 6 Average action at each stage (in chips). As can be seen in the table, the agents fall within one standard deviation of the three human cultures.

with $F = 4.861$. Tables 3,4 and 5 show the Scheffe post-hoc test p-value results for each of the stages of the game. Results with p-value smaller than 0.05 (which are considered statistically significant) are presented in bold font. Note that since the agents group included more games (and thus more data), differences reach statistical significance in many cases.

As we hypothesized, the activity of the agents falls within one standard deviation of the average of the three human cultures (see Table 6 for the exact details). This indicates that autonomous agents built by NGTE can indeed be treated within cultural diversities.

The percentage of players that gave away chips at each stage is shown in Figure 4. As depicted in Figures 3 and 4, subjects from all cultures along with the autonomous agents transferred (or paid) chips at all stages of the game. Although this does not comply with the subgame-perfect equilibrium, this result was expected in human behavior. Another interesting result is that the autonomous agents' activity is close to that of the activity shown by their own culture, but to a slightly lesser extent. Namely, the autonomous agents' behavior is slightly closer to that of the subgame-perfect equilibrium.

We found that using the SPE when playing with humans or autonomous agents does not yield the highest outcome. As illustrated in Figure 3, on average Player B reciprocated more than Player A trusted, with both humans and autonomous agents. No Player B ever reciprocated if Player A did not trust; therefore on average Player A gained from trusting (unlike the SPE which requires that there be no trust at all). It is fair to assume that the agent designers thought it might be beneficial to display trust as their agent might be rewarded

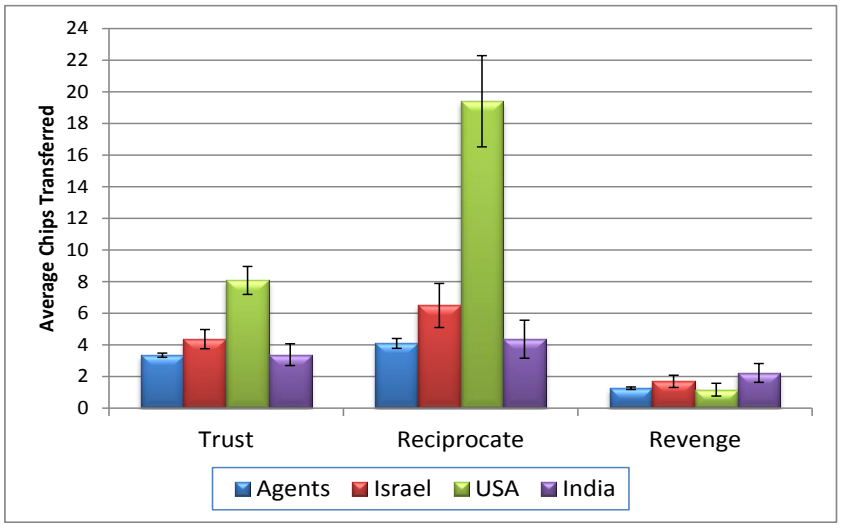

Fig. 3 Average action at each stage (in chips). Averaging on all subjects in each group, in the three Trust-Revenge games (TR 1, TR 2 and TR 3).

by other agents during the reciprocation stage. Revenging, on the other hand, is clearly not an optimal behavior and thus an agent designed to take revenge may be attributed to its developer seeking for justice or fairness (in [27] it was shown that revenge is linked with enjoyment). The documentation collected from the students who designed the agents indicates that they understood that revenge would lower their final profit and that it was not beneficial in this single-shot game (74% of the students explicitly stated this in their documentation). One of the players who sometimes took revenge, explicitly stated that: "...people must know that if they don't play fair they will get punished.", but also said that "...after all player B can't turn the clock back.". It was interesting to see that among students who chose not to take revenge some felt the need to justify this choice of not punishing their opponent, as one of them stated: "Take it easy. What was was.".

From the documentation we conclude that 38% of the agents were explicitly designed to never take revenge, while 44% of the agents decided whether to take revenge or not, based on the other player's stake (24%) or reciprocation amount (20%), and the remaining 18% based their decision of whether to take revenge and its extent, solely on other variables, such as the game settings, their own stake, and random variables. We found that a common strategy with respect to the revenge stage, which was played by 17% of the agents, was to take revenge if the other player had more chips than the agent, and if so to try and equalize the two players' stakes.

Although the fact that autonomous agents were designed to take revenge may be surprising to game theorists, it clearly shows that the actions of autonomous agents are similar to those of humans and can therefore be modeled using the same methods.

We proceeded by analyzing all the game settings (*investment*, *dictator*, *TR 1*, *TR 2* and *TR 3*). Figure 5 presents the average chip transfer in the trust stage for each of the groups in each of the settings. As can be seen in the figure, in all but the *investment* setting, the agents blended in nicely with all the other groups. However, in the *investment* setting the agents displayed very little trust. Although this behavior is different than that demonstrated by the

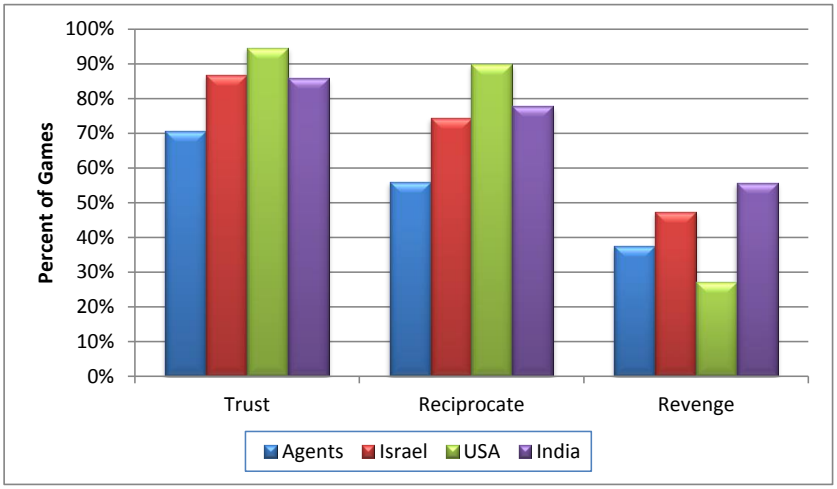

Fig. 4 Percent of games for each of the groups in which the subjects performed actions greater than zero in each of the stages (trust, reciprocate and revenge).

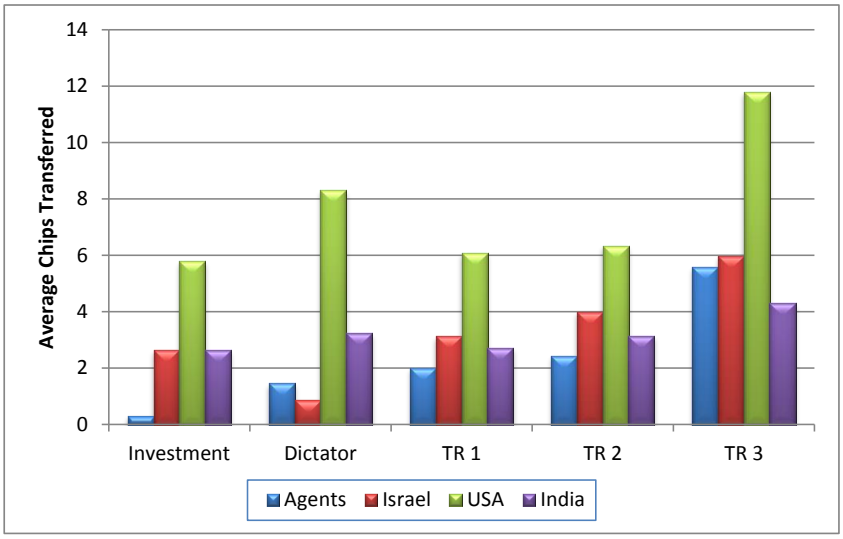

Fig. 5 Average action in trust stage (in chips) for each of the groups and for all settings separately.

other groups, it turned out to be a better approach performance-wise, since on average the subjects in the same culture, *Israel*, lost on average 0.99 chips, or 63% of their trust amount (the agents in the *Agent* group who did trust, received on average 27% less than their trust amount). Interestingly, performance was not the major concern of the agents in the *dictator* settings, as they transferred more chips than their equivalent culture, when clearly little or no return could be expected. Figure 6 shows the percent of games in which the trust transfer was greater than 0.

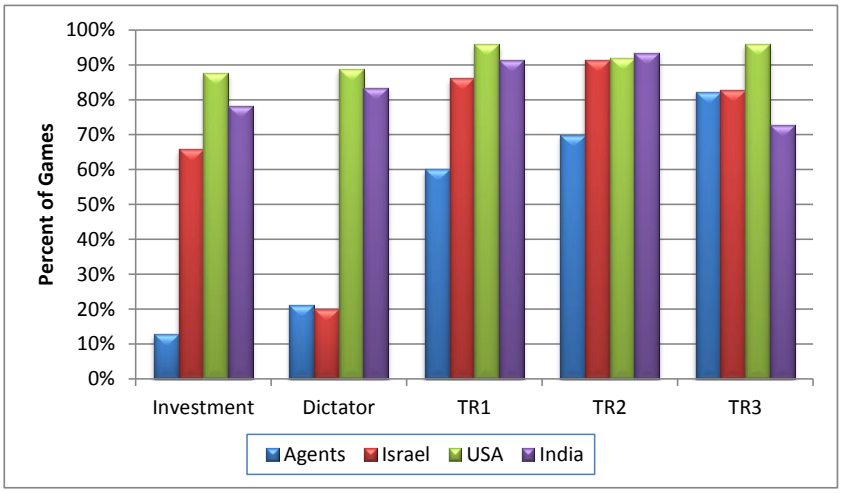

Fig. 6 Percent of games in which the trust transfer was greater than zero.

Figure 7 presents the average chip transfer among all groups in each of the settings for each of the stages of the game. As illustrated by the figure, in all games with a positive revenge rate (*TR 1*, *TR 2* and *TR 3*), the players transferred (or paid) higher amounts in all three stages. It is also apparent that the players seemed to treat the *dictator* setting, similar to the dictator game, as the only major action was trust. Similarly, in the *investment* settings, only a negligible sum was actually revenged, showing that the players did treat this setting similarly to the investment game. In the *investment* or *dictator* settings, there was a revenge rate of 0. This means that if Player A decided to take revenge, chips were subtracted from Player A's stack. However since the revenge rate was 0, player B paid nothing, and thus only Player A was affected by revenge in these games and lost chips. Therefore, we refer to revenge in these two settings (which have a revenge rate of 0) as human error. The *Agents* took revenge in these two settings only 0.088 on average, which is only 22.6% of the average revenge of the *Israeli* group in these settings (these differences are statistically significant with $p < 0.05$) and only 10.7% of the total human revenge in these settings ($p < 0.001$). This indicates that the agents tended to make less errors.

Although we weren't generally interested in the performance of the autonomous agents or the humans, we compared the average final stacks obtained by the players at the end of the game (averaging throughout all five settings). In Figure 8, the SPE column shows theoretical game results between subgame perfect equilibrium agents. Since all subjects played both as Player A and as Player B the combined stack best represents the average success. The subgame perfect equilibrium (SPE) clearly does not provide the highest profit. When considering the final performance, it seems that the autonomous agents achieved a very similar score to that of the group of subjects from their own culture, both as Player A and Player B. However, performance is by no means a scale for measuring behavior.

A composer of a culture dependent expert agent for the Trust-Revenge variants (for Player A), who's only interest is performance, should follow these guidelines:

1. Obviously, if the only interest is performance - one should never take revenge;
2. Similarly, in the *dictator* setting - one should not transfer any chips;

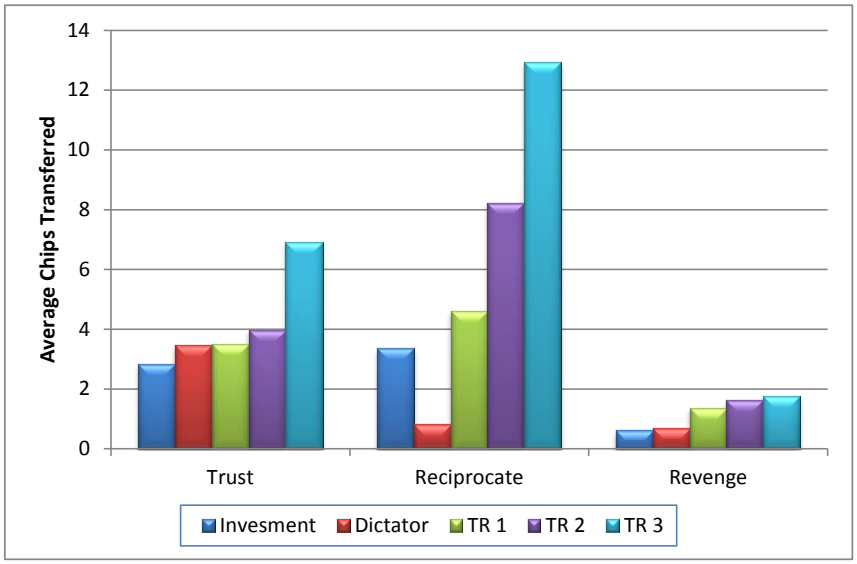

Fig. 7 Average action in each stage (in chips), comparing all five different settings (averaging on all groups).

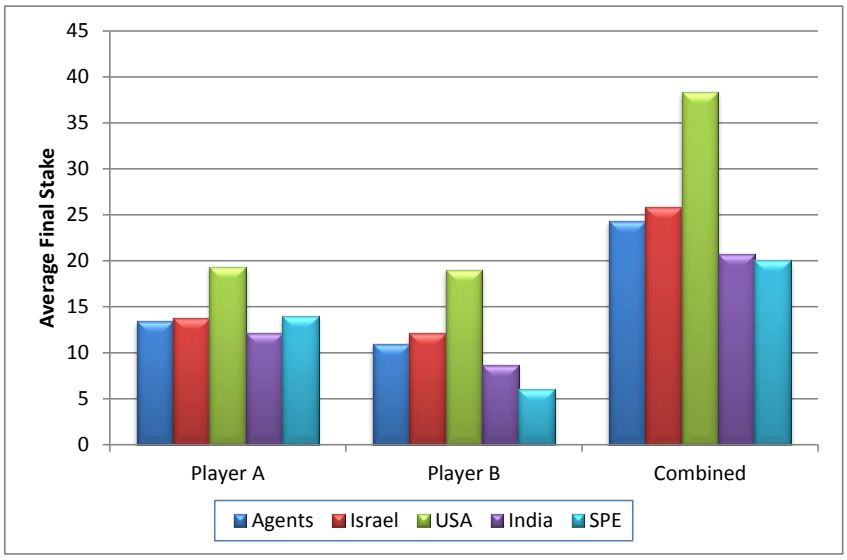

Fig. 8 Average Final Stacks (in chips) for each of the four groups and the theoretical subgame-perfect equilibrium agent result.

3. In the *investment* setting one should not trust, unless Player B is from the USA (which yields an average gain of 60%);
4. In the *TR 1* setting, one should trust only human players from the USA and India - this yield an average gain of 75% for players from the USA and 10% for players from India;
5. In settings *TR 2* and *TR 3*, one should trust all 4 groups, producing a 94% average gain;

We conclude this section by answering another interesting question, of whether (and to what extent) the actual activity of building autonomous agents influences human behavior. After the students in the *Agents* group composed their autonomous agents, they were asked to play the game themselves (the students were not aware that they would need to play the game, when they composed the agents, and we did not publish the performance of their agents until the students played the game themselves). In this experiment, the motivation was also based on bonus points and was identical to the motivation of the subjects in the *Israeli* group. We call this group of players the *Composers*.

Figure 9 compares the average number of chips transferred (or paid) for settings (*TR 1*, *TR 2* and *TR 3*) by players in the *Israeli* group to that of the *Composers*. We also added the *Agents* group for reference. The black bars show the confidence range ($\alpha = 0.05$). As depicted in the figure, there is no statistical significance in the behavior of the players in the two human groups. Furthermore, the behavior of the subjects among the *Composers* does not seem by any means to be closer to that of their own agents than the subjects in the *Israeli* group. However, when examining the revenge in the *investment* and *dictator* settings (which is attributed to human error for these settings), the *Composers* did indeed take revenge on average 62% less than the *Israeli* group but still 70% more than their own agents. However, since the human error is small in the first place, these results do not reach statistical significance, though they are very close ($p = 0.08$). Note that a similar achievement of reducing human error could possibly be obtained also by different means (such as training). Therefore, it seems that the only impact that building agents had on the *Composers* was the reduction of human error.

This result seems to challenge the work by Elmalech et al. [46] which shows that composing autonomous agents does have impact on the composers' behavior. However, unlike our work, Elmalech et al. examined a single player game, which was not designed to capture any emotion. Any deviation from the optimal strategy may be assigned to human error. Therefore, Elmalech et al.'s results, that show how composing autonomous agents has impact on composers' behavior, may be limited to the impact on human error, which aligns with our own findings.

6 Discussion

This paper exhibits four findings using the trust-revenge game, the first, which is the main focus of this paper is that expert agents that interact with NGTE agents can use the same models developed for modeling cultural diversities within humans, for modeling the NGTE agents; second, the NGTE agents' behavior is closer to that of the subgame-perfect equilibrium; third NGTE agents are less prone to error; and finally, composing agents has no impact on human behavior aside of possibly reducing error rate.

The trust-revenge game encompasses many of human prototype behavior, such as reciprocation and revenge. Emotions such as trust, fear, kindness, anger, envy and pity, also take a role in this game. Being a two player game (as opposed to a single player game) the trust-revenge game is a good platform to test human and NGTE behavior. To establish our work, different types of games should be studied, possibly including games which capture

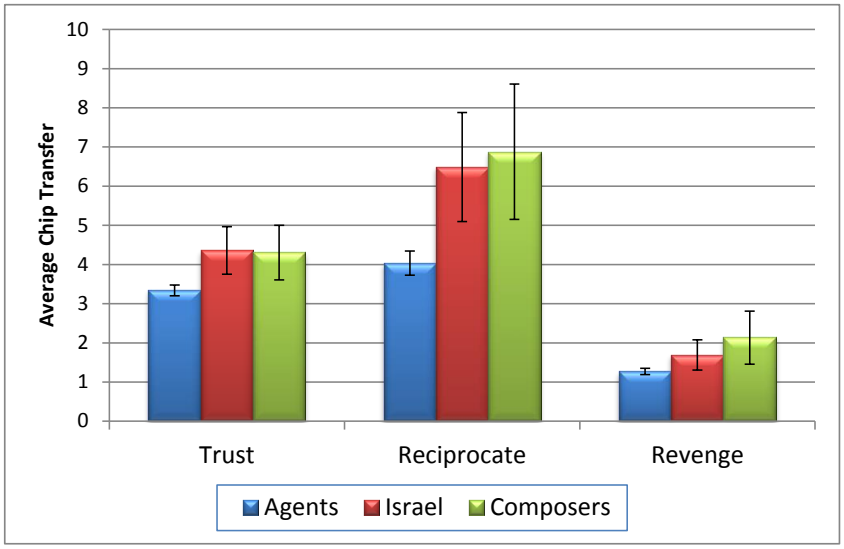

Fig. 9 Average action in each stage (in chips), comparing the agents (Agents), the Israeli subjects who played after composing the agents (Composers) and those who did not compose agents at all (Israel).

behavior such as anticipation, surprise and dishonesty. In our work we ensured that every group of human culture would only play with humans from its own culture, and that the group of agents would only play with agents. This was done, to enable all player to reason about the behavior of the second player and not require the players to be aware of possible culture diversity. More research is required in order to determine human and agent behavior in inter-cultural interaction along with interactions which include more than two players (in which some may be agents while others may be human possibly from different cultures). We did not explore whether there are hidden motivations for the subjects behaviors. Such behaviors might include the students being careful not to harm one another. This type of behavior could be evident even though the subject identity was hidden from the players, as they knew they were playing against other students, and might feel compelled to "be nice". On the other hand they may have been strongly competitive because of this. The subjects recruited through Mechanical Turk may have had similar compulsions, since they were explicitly told that they will be playing with subjects from their own culture. Further study of these issues remains for future work.

In this work, all autonomous agents were developed by participants having the same culture or cultural background (Israeli students). Our results post evidence for our general claim that NGTE agents behavior falls within the diversity of different human cultures, allowing expert agent developers to use methods currently used to model human cultures also with NGTE agents. However, to strengthen our claim, it would be interesting if our experiment could be repeated using a different agent developer culture. Our results could also benefit from being replicated with different human cultures and using different games.

Although having milder impact on expert agent designers, an interesting topic for future work would be to analyze how different might the behavior of NGTE agents developed by humans from a different culture, such as USA or India, be from the behavior presented by the Israeli students' agents. Having both USA and Indian students develop agents for the

trust revenge game would allow us to determine the similarity among the three different groups of agents. Performing between subject analysis with both agents and humans from different cultures, will allow us to determine whether the behavior of NGTE agents from different cultures is more similar to one another or to the culture of the developer which composed them.

7 Conclusion and Future Work

Since autonomous agents are known to benefit from modeling their opponents we investigated how similar agents developed by NGTE are to humans. We evaluated autonomous agent behavior in the Trust-Revenge game, using 5 different variants of the game, and compared it to the behavior of people from three different cultures. We found that, like humans, the NGTE agents do not follow the subgame-perfect equilibrium when playing the game. The agents' behavior did not deviate from the standard behavior of the different cultures, and, in each of the stages of the game, the average action performed by the NGTE agents was within one standard deviation of the average action of the three human cultures. We deduce that when playing against agents developed by NGTE it is reasonable to assume that agents can be considered as an additional human culture and thus, expert agents that interact with humans from different cultures may use the same methods to interact with NGTE agents.

The revenge stage of the game is particularly interesting, as there is clearly no motivation to take revenge. Taking revenge seems to be attributed to emotional human behavior or the search for justice, therefore, finding it embedded in the agents may seem unexpected. We found that even in the revenge stage the agents behave in a similar fashion to the humans, and differences in the revenge extent between the agents and the humans are not larger than the differences found between various cultures. The only exception we found to this rule was that the agents are less prone to error, and thus scarcely took revenge when the revenge rate was 0.

When building an expert agent intended for interaction with autonomous agents composed by NGTE, it is important to know whether we need to develop new models to describe the NGTE. Our findings indicate that these autonomous agents should use the same models developed for interaction with humans from various cultures. For example if one wants to design an expert agent that is intended to play Poker on the Internet, it is fair to assume that this agent will play against autonomous agents designed by NGTE as well as human players from different cultures. It is very likely that the agents will exhibit attributes that humans use when playing, such as over-bluffing. Our study implies that these agents can be treated in the same manner as the various human cultures are treated. The same holds for electronic commerce, where autonomous agents are likely to be influenced by anchoring or the "sunk cost" effect. An expert agent which interacts with such NGTE agents, may exploit such behavior, just as an expert agent exploits such behavior in humans. For example, Hajaj et al. [49] present a comparison shopping agent in the domain of electronic commerce which interacts with humans and exploit human behavior such as the anchoring effect in order to encourage the humans not to query a different comparison shopping agent. However, whether using the exact same method used for a certain population (or culture) will be beneficial for an agent interacting with NGTE agents may depend on how well this agent may be extended to interacting with other cultures. That is, according to our results, if the agent can be extended to efficiently interact with other human cultures, it is very likely to be able to interact with NGTE agents using this same extension. If the agent has a model which can

be trained on new data, such an extension may be achieved simply by retraining the same model used by the agent but on NGTE agents (rather than humans). Note, that this is not trivial since the human model usually relies on human behavior (known from psychology and social science) [24] and simply retraining the model is likely to fail if used, for instance, when interacting with a perfectly rational agent. Clearly, an agent which is already embedded with a capability of interacting with humans from different cultures, would be able to efficiently interact with NGTE agents and would require no additional changes.

We focused on agents developed by NGTE. However, it is known that agents designed by experts behave more rationally than agents developed by NGTE, and that humans who acquire vast experience in a game become more rational players [50,51]. This encourages further study to examine whether treating NGTE agents as a culture may be extended to expert agents as well.

8 Acknowledgments

We thank Shira Abuhatzera for her help. A very preliminary version of this paper appears in Azaria et al. [52].

References

1. R. H. Guttman, A. G. Moukas, P. Maes, Agent-mediated electronic commerce: a survey, The Knowledge Engineering Review 13 (02) (1998) 147–159.
2. A. Heydon, M. Najork, Mercator: A scalable, extensible web crawler, World Wide Web 2 (4) (1999) 219–229.
3. J. Markoff, Google cars drive themselves, in traffic, The New York Times 10 (2010) A1.
4. R. Bemelmans, G. J. Gelderblom, P. Jonker, L. De Witte, Socially assistive robots in elderly care: A systematic review into effects and effectiveness, Journal of the American Medical Directors Association 13 (2) (2012) 114–120.
5. B. Robins, P. Dickerson, P. Stribling, K. Dautenhahn, Robot-mediated joint attention in children with autism: A case study in robot-human interaction, Interaction studies 5 (2) (2004) 161–198.
6. P. Riley, M. Veloso, Planning for distributed execution through use of probabilistic opponent models., in: AIPS, 2002, pp. 72–82.
7. M. Wooldridge, An introduction to multiagent systems, John Wiley & Sons, 2009.
8. Y. Gal, A. Pfeffer, Modeling reciprocal behavior in human bilateral negotiation, in: Proceedings of the National Conference on Artificial Intelligence, Vol. 22, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007, p. 815.
9. K. Hindriks, D. Tykhonov, Opponent modelling in automated multi-issue negotiation using bayesian learning, in: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1, International Foundation for Autonomous Agents and Multiagent Systems, 2008, pp. 331–338.
10. Y. Oshrat, R. Lin, S. Kraus, Facing the challenge of human-agent negotiations via effective general opponent modeling, in: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1, International Foundation for Autonomous Agents and Multiagent Systems, 2009, pp. 377–384.
11. A. Rosenfeld, S. Kraus, Using aspiration adaptation theory to improve learning, in: The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 1, International Foundation for Autonomous Agents and Multiagent Systems, 2011, pp. 423–430.
12. N. Peled, Y. Gal, S. Kraus, A study of computational and human strategies in revelation games, Autonomous Agents and Multi-Agent Systems 29 (1) (2015) 73–97.
13. A. Azaria, Y. Gal, S. Kraus, C. Goldman, Strategic advice provision in repeated human-agent interactions, Autonomous Agents and Multi-Agent Systems (2015) 1–26.
14. A. Azaria, Y. Aumann, S. Kraus, Automated strategies for determining rewards for humanwork, in: AAAI, 2012.

15. A. Azaria, Z. Rabinovich, S. Kraus, C. V. Goldman, O. Tsimhoni, Giving advice to people in path selection problems, in: AAMAS, 2012.
16. A. Azaria, S. Kraus, C. Goldman, O. Tsimhoni, Advice provision for energy saving in automobile climate control systems, in: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1391–1392.
17. A. Azaria, A. Hassidim, S. Kraus, A. Eshkol, O. Weintraub, I. Netanely, Movie recommender system for profit maximization, in: RecSys, ACM, 2013, pp. 121–128.
18. A. Azaria, Y. Aumann, S. Kraus, Automated agents for reward determination for human work in crowdsourcing applications, Autonomous Agents and Multi-Agent Systems 28 (6) (2014) 934–955.
19. A. Rosenfeld, S. Kraus, Modeling agents through bounded rationality theories., in: IJCAI, Vol. 9, 2009, pp. 264–271.
20. A. Elmalech, D. Sarne, Evaluating the applicability of peer-designed agents in mechanisms evaluation, in: Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 02, IEEE Computer Society, 2012, pp. 374–381.
21. B. J. Grosz, S. Kraus, S. Talman, B. Stossel, M. Havlin, The influence of social dependencies on decision-making: Initial investigations with a new game, in: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 2, IEEE Computer Society, 2004, pp. 782–789.
22. Y. Gal, S. Kraus, M. Gelfand, H. Khashan, E. Salmon, An adaptive agent for negotiating with people in different cultures, ACM Transactions on Intelligent Systems and Technology 3 (1) (2011) 8.
23. G. Haim, Y. K. Gal, M. Gelfand, S. Kraus, A cultural sensitive agent for human-computer negotiation, in: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1, International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 451–458.
24. A. Rosenfeld, I. Zuckerman, A. Azaria, S. Kraus, Combining psychological models with machine learning to better predict peoples decisions, Synthese 189 (1) (2012) 81–93.
25. J. Berg, J. Dickhaut, K. McCabe, Trust, reciprocity, and social history, Games and economic behavior 10 (1) (1995) 122–142.
26. N. D. Johnson, A. A. Mislin, Trust games: A meta-analysis, Journal of Economic Psychology 32 (5) (2011) 865–889.
27. A. Gneezy, D. Ariely, Don't get mad get even: On consumers' revenge, manuscript,Duke University (2010).
28. D. Carmel, S. Markovitch, Opponent modeling in multi-agent systems, in: Adaption and Learning in Multi-Agent Systems, Springer, 1996, pp. 40–52.
29. P. McCracken, M. Bowling, Safe strategies for agent modelling in games, in: AAAI Fall Symposium on Artificial Multi-agent Learning, 2004.
30. A. Lazaric, M. Quaresimale, M. Restelli, On the usefulness of opponent modeling: the kuhn poker case study, in: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 3, International Foundation for Autonomous Agents and Multiagent Systems, 2008, pp. 1345–1348.
31. M. Chalamish, D. Sarne, S. Kraus, Programming agents as a means of capturing self-strategy., in: AAMAS, 2008, pp. 1161–1168.
32. A. Elmalech, D. Sarne, Evaluating the applicability of peer-designed agents in mechanisms evaluation, in: Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 02, WI-IAT '12, 2012, pp. 374–381.
33. M. Chalamish, D. Sarne, R. Lin, Enhancing parking simulations using peer-designed agents., IEEE Transactions on Intelligent Transportation Systems (1) 492–498.
34. M. Mash, R. Lin, D. Sarne, Peer-design agents for reliably evaluating distribution of outcomes in environments involving people, in: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems, AAMAS '14, 2014, pp. 949–956.
35. R. Lin, S. Kraus, Y. Oshrat, Y. K. Gal, Facilitating the evaluation of automated negotiators using peer designed agents (2010).
36. E. Manistersky, R. Lin, S. Kraus, The development of the strategic behavior of peer designed agents, Lecture Notes in Computer Science 8001 (I) (2013) 180–196.
37. M. Chalamish, D. Sarne, R. Lin, The effectiveness of peer-designed agents in agent-based simulations, Multiagent and Grid Systems 8 (4) (2012) 349–372.
38. M. Willinger, C. Keser, C. Lohmann, J.-C. Usunier, A comparison of trust and reciprocity between france and germany: experimental investigation based on the investment game, Journal of Economic Psychology 24 (4) (2003) 447–466.

39. S. Gächter, B. Herrmann, Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment, Philosophical Transactions of the Royal Society B: Biological Sciences 364 (1518) (2009) 791–806.
40. C. Engel, Dictator games: a meta study, Experimental Economics 14 (4) (2011) 583–610.
41. H. Oosterbeek, R. Sloof, G. Van De Kuilen, Cultural differences in ultimatum game experiments: Evidence from a meta-analysis, Experimental Economics 7 (2) (2004) 171–188.
42. G. Hofstede, G. J. Hofstede, M. Minkov, Cultures and organisations-software of the mind: intercultural cooperation and its importance for survival, McGraw-Hill New York, NY, 1991.
43. C. K. De Dreu, P. A. Van Lange, The impact of social value orientations on negotiator cognition and behavior, Personality and Social Psychology Bulletin 21 (11) (1995) 1178–1188.
44. S. Mascarenhas, R. Prada, A. Paiva, G. J. Hofstede, Social importance dynamics: A model for culturally-adaptive agents, in: Intelligent Virtual Agents, Springer, 2013, pp. 325–338.
45. F. Kistler, B. Endrass, I. Damian, C. T. Dang, E. André, Natural interaction with culturally adaptive virtual characters, Journal on Multimodal User Interfaces 6 (1-2) (2012) 39–47.
46. A. Elmalech, D. Sarne, N. Agmon, Can agent development affect developer's strategy?, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2014, pp. 923–929.
47. J. Shin, D. Ariely, Keeping doors open: The effect of unavailability on incentives to keep options viable, Management Science (2004) 575–586.
48. Amazon, Mechanical Turk services, http://www.mturk.com/ (2013).
49. C. Hajaj, N. Hazon, D. Sarne, Ordering effects and belief adjustment in the use of comparison shopping agents, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2014.
50. R. Selten, R. Stoecker, End behavior in sequences of finite prisoner's dilemma supergames a learning theory approach, Journal of Economic Behavior & Organization 7 (1) (1986) 47–70.
51. C. Camerer, K. Weigelt, Experimental tests of a sequential equilibrium reputation model, Econometrica 56 (1) (1988) 1–36.
52. A. Azaria, A. Richardson, A. Elmalech, A. Rosenfeld, Automated agents' behavior in the trust-revenge game incomparison to other cultures, in: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1389–1390.