Automated Agents' Behavior in the Trust-Revenge Game in Comparison to Other Cultures

(Extended Abstract)

Amos Azaria
Dept. of Computer Science
Bar Ilan University, Israel

Ariella Richardson
Dept. of Industrial Engineering
Jerusalem College of Technology, Israel

Avshalom Elmalech
Dept. of Computer Science
Bar Ilan University, Israel

Avi Rosenfeld
Dept. of Industrial Engineering
Jerusalem College of Technology, Israel

ABSTRACT

Agents that interact with humans are known to benefit from modeling them. Therefore, when designing agents intended for interaction with automated agents, it is crucial to model the other agents. However, little is known about how to model automated agents and in particular non-expert agents. Are automated agents to be modeled the same way that an agent models humans? Or does a separate model for interacting with automated agents need to be developed? We evaluate automated agent behavior (for non-expert agents) 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 sub game-perfect equilibrium, however, very rarely do people follow it. We compared the behavior of automated agents to that of human actions in several demographic groups (including a group which is similar but not identical to the designers of the automated agents). We show that differences between automated agents' and humans' behavior are similar to differences between different human cultures.

Categories and Subject Descriptors

I.2.m [**Computing Methodologies**]: ARTIFICIAL INTELLIGENCE—*Miscellaneous*

Keywords

Automated Agents, Behavior Modeling, Trust Game

1. INTRODUCTION

Automated agents are integrated into countless environments, such as electronic commerce, web crawlers, military agents, space exploration probes and automated drivers. Scientifically designed automated agents or automated agents designed by experts often implement a fully rational strategy. However, the rise in computer science education along with the increase in software development tools available to the public have caused a significant rise in software and thus automated agents developed by non-experts. Nowadays, many pieces of software and thus automated agents are written by programmers with no more than a Bachelors' degree in computer science.

How should designers plan their agents when opponent modeling is unavailable? Can any general assumptions be made about automated agents and used for agent design? Research into peoples' behavior has found that people often do not make strictly rational decisions but instead behave differently. Many studies have shown that psychological factors and human decision-making theory are needed to develop a good model of true human behavior, which in turn is required for optimizing the performance of agents which interact with humans [8, 12, 3, 1, 4, 13, 2, 11]. Research attempting to compare strategies used by people and agents designed by non-experts often point to some differences in the strategies used by the two [6, 5].

A good agent must be embedded with the capability to interact well with different cultures [7, 10], but would it be required to be equipped with a special ability to model and interact with other amateur automated agents? Can automated agents be treated as another culture, or do differences between automated agents' and humans' behavior span much more than the differences between different human cultures?

In this paper we analyze the behavior of automated agents (developed by non-experts) in the Trust-Revenge game [9]. we compare it to the behavior of a group of humans which come from the same background and demographic group as the automated agents' programmers as well as to groups of humans from different cultures. We determine whether automated agents can be described as a separate human culture or whether their behavior is too different from human behavior to model them the same way humans are.

2. TRUST-REVENGE GAME

In this work, we studied a two-player game 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 - by which the

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

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 - 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 known to both players at the beginning of the game. In this game there is a clear, unique sub game-perfect equilibrium (SPE) strategy. In the revenge stage, there is no rational reason for Player A to 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 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 do not revenge, do not reciprocate and do not trust. We examine the agents' behavior and to what extent is this typical human behavior embedded in the strategy of the agents. We also examine to what extent they differ from the behavior of human cultures and in particular the culture of their designers.

3. EXPERIMENTAL EVALUATION

Throughout the experiments we used 5 different settings for the Trust-Revenge Game: *Investment*, *Dictator*, *TR 1*, *TR 2* and *TR 3*.

A set of 36 undergraduate computer science students from Israel composed automated agents for the Trust-Revenge Game (the *Agents* group). Another group of 35 undergraduate computer science students from the same culture and demographic group as those who composed the agents, played the Trust-Revenge game with each other. 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 and players played 10 consecutive games.

Figure 1 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 blend in nicely with all the other groups.

4. CONCLUSION

Since automated agents are known to benefit from modeling their opponents we investigated how similar non-expert agents are to humans. We evaluated automated 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. The agents' behavior did not deviate from the standard behavior of the different cultures. We deduce that when playing against non-expert agents it is reasonable to assume that agents can be construed as an additional human culture.

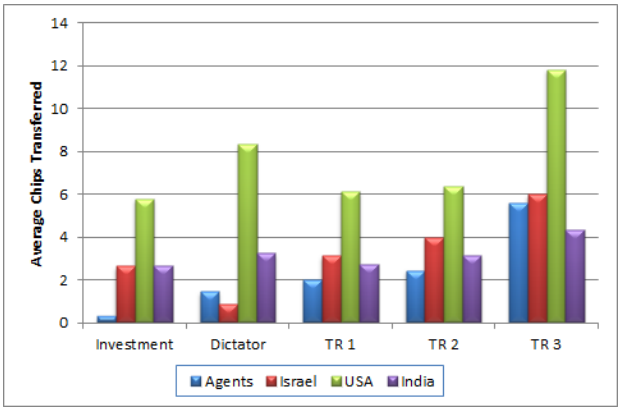

Figure 1: Average Action in Trust Stage (in chips)

5. REFERENCES

[1] A. Azaria, Y. Aumann, and S. Kraus. Automated strategies for determining rewards for humanwork. In *AAAI*, 2012.

[2] A. Azaria, Z. Rabinovich, S. Kraus, and C. V. Goldman. Strategic information disclosure to people with multiple alternatives. In *AAAI*, 2011.

[3] A. Azaria, Z. Rabinovich, S. Kraus, C. V. Goldman, and Y. Gal. Strategic advice provision in repeated human-agent interactions. In *AAAI*, 2012.

[4] A. Azaria, Z. Rabinovich, S. Kraus, C. V. Goldman, and O. Tsimhoni. Giving advice to people in path selection problems. In *AAMAS*, 2012.

[5] Michal Chalamish, David Sarne, and Raz Lin. The effectiveness of peer-designed agents in agent-based simulations. *Multiagent and Grid Systems*, 8(4):349–372, 2012.

[6] Avshalom Elmalech and David Sarne. Evaluating the applicability of peer-designed agents in mechanisms evaluation. In *Proc. of Intelligent Agent Technology 2012*, pages 374–381, 2012.

[7] Y. Gal, S. Kraus, M. Gelfand, H. Khashan, and E. Salmon. An adaptive agent for negotiating with people in different cultures. *ACM Transactions on Intelligent Systems and Technology*, 3(1):8, 2011.

[8] Y. Gal and A. Pfeffer. Modeling reciprocity in human bilateral negotiation. In *AAAI*, 2007.

[9] A. Gneezy and D. Ariely. Don't get mad get even: On consumers' revenge. *manuscript*, 2010.

[10] Galit Haim, Ya'akov Kobi Gal, Michele Gelfand, and Sarit Kraus. A cultural sensitive agent for human-computer negotiation. In *Proc. of AAMAS'12*, pages 451–458, 2012.

[11] T. Nguyen, R. Yang, A. Azaria, S. Kraus, and M. Tambe. Analyzing the effectiveness of adversary modeling in security games. In *Conf. on Artificial Intelligence (AAAI)*, 2013.

[12] A. Rosenfeld and S. Kraus. Using aspiration adaptation theory to improve learning. In *AAMAS*, Taipei, 2011.

[13] Avi Rosenfeld, Inon Zuckerman, Amos Azaria, and Sarit Kraus. Combining psychological models with machine learning to better predict peopleŠs decisions. *Synthese*, 189(1):81–93, 2012.