Automated Agents for Human Persuasion

Amos Azaria
Department of Computer Science,
Bar-Ilan University, Ramat Gan, Israel
azariaa1@cs.biu.ac.il
http://azariaa.com

1. INTRODUCTION

Advanced technology allows computer systems to take an increasingly active role in people's decision-making tasks, whether as proxies for individuals or organizations (e.g., automated bidder agents in e-commerce [10]), or autonomous agents that work alongside people (e.g, training systems for diplomatic negotiation [6]). The participants of these heterogeneous human-computer applications share common goals, but each of the participants also has its own incentives.

Often an automated agent faces a human who is required to make decisions, and to choose an action to take among several actions. Usually, some actions are preferable for the agent than others. One of the most challenging problems for an agent facing a human is to persuade him to choose an action which is better for the agent. The agent may either face a single human or multiple humans and have either a single interaction or repeated interactions.

For example, investor person may consult with his investment consultant regarding the best way to invest his capital. The investment consultant has many options from which to choose, but besides helping the investor, the investment consultant may have some preferences. Although he may not outright lie regarding any investment opportunities, he may not want to give the investor the full information regarding each opportunity. He may also advise the person to invest in an opportunity which may or may not be best for him.

Another example may be a young person receiving career advice. The counselor may have been advised of government preferences which jobs are more needed for the society.

It is well known that people often follow suboptimal decision strategies due to irrationalities attributed to: lack of knowledge of own preferences, the effects of the task complexity, framing effects, the interplay between emotion and cognition, the problem of self control, the value of anticipation, future discounting, anchoring, risk aversion and many more effects ([14, 1, 3, 9]).

Efficient interaction with humans requires understanding and modeling of their behavior. For example, while equilibrium strategy is theoretically considered the most rational one, agents using such strategies often perform poorly in practice [8, 5, 2]. Since humans commonly do not use equilibrium strategy themselves, replying with such a strategy can be suboptimal. Thus, using machine learning techniques and based on psychological factors and human decision-making theory, one should develop a good model of the true human behavior in order to optimize the performance of agents interacting with these humans [4, 7, 12, 8]. The learned model should be generalized to new environments as well as different people. Once a model is created, one can use search methods or decision-making modules in order to conclude which action will be best for the agent.

I am currently dealing with the following three different cases:

- Information disclosure: The agent has information unknown to the human, and can reveal full or partial information to encourage the human to take a certain action.
- Advice provision: The agent may advise the human to take a certain action. In this case the system may either be exposed to more information than the human, or uses its computational advantage. Otherwise, the human will have no incentive to follow the agent's advice.
- Reward giving: The agent may suggest a reward to the human if he takes the preferred action. In this case the agent will need to minimize the expected sum of rewards given.

Research in multi-agent systems primarily encompasses systems composed of automated agents with three different cooperation levels. Cooperative systems are usually described by a single utility function which all agents attempt to maximize. Competitive systems, on the other hand, may be designed and analyzed, for example, as zero sum games where the gain of one agent is the loss of another. I focus on systems composed of both automated agents and human users. Although in general these interactive systems are cooperative, users and machines may have different interests. Each party may want to optimize different parameters, not necessarily at the expense of the other. In particular, I study automated agents interested in persuading their users to perform actions that increase the agent's utility. In my study I intend to focus on these cases (i.e. cases where the agent and the human utility functions differ but do not contrast). That is because in the case where the agent and the human utility functions are identical, the agent will reveal all information (as long as sending information is free), or advise the human to take the best for both sides and the human is most likely to accept. In a case where the agent and the human utility

Appears in: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), June, 4–8, 2012, Valencia, Spain.

functions are in contrast, the agent is most likely to reveal no information, and the human is most likely to ignore any advice given to him by the agent. Consider for example a centralized traffic control system that provides congestion information to commuters. The system and drivers both share the goal of getting commuters to their destination as quickly as possible. However, the system may also wish to increase the amount of tolls collected from drivers, while drivers may wish to minimize this amount.

As of January 30th 2012 my publications are:

- Azaria A, Aumann Y, Kraus S
  Automated Strategies for Determining Rewards for Human Work, Submitted to AAAI 2012
- Azaria A, Rabinovich Z, Kraus S, Goldman CV, Gal Y
  Strategic Advice Provision in Repeated Human-Agent Interactions, Submitted to AAAI 2012
- Rosenfend A, Zukerman I, Azaria A, Kraus S
  Combining Psychological Models with Machine Learning to Better Predict People's Decisions, submitted to Synthese (SYNT)
- Azaria A, Rabinovich Z, Kraus S, Goldman CV, Tsimhoni O,
  Giving Advice to People in Path Selection Problems, to be published in AAMAS, 2012 (also presented in AAAI 2011 WS-11-13)
- Azaria A, Rabinovich Z, Kraus S, Goldman CV,
  Strategic Information Disclosure to People with Multiple Alternatives,AAAI, 2011, 51-58 (also presented in BISFAI 2011)

All publications can be downloaded from my website at: http://azariaa.com/Home/Publications.

2. FUTURE WORK

I intend on focusing on the following two subjects:

- Self-interested recommender systems: Models for predicting users' ratings have been proposed that are used by recommender systems to advise their users (See Ricci et al. [11] for a recent review). Most works in this realm have only considered the utility of the system and have not modeled the user's reactions to its actions over time. An exception is the work by Shani et al. [13], which uses a discrete-state MDP model to maximize the system utility function taking into account the future interactions with their users. I intend to extend this work by considering the possible effects of the recommendations on users' future behavior.
- Automated agents for helping people in decision making processes: As mentioned above, people often follow suboptimal decision strategies. Some of the effects causing people to follow those suboptimal strategies are caused by true subjective utility functions, where the people themselves do not consider their action as being irrational, even after receiving proper explanation (such as the risk aversion effect). However, many people would like to eliminate other effects disturbing their decision making process. I intend to create an agent that will help people in decision making and will be measured by peoples' satisfaction. I will begin my research using the domain of gambling where a human faces an unknown sequence of bets and must make sequential decisions on each bet.

3. REFERENCES

[1] D. Ariely, G. Loewenstein, and D. Prelec. Coherent arbitrariness: Stable demand curves without stable preferences. *Journal of Business Research*, 59(10-11):1053–1062, 2006.

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

[3] C. F. Camerer. *Behavioral Game Theory. Experiments in Strategic Interaction*, chapter 2. Princeton University Press, 2003.

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

[5] P. Hoz-Weiss, S. Kraus, J. Wilkenfeld, D. R. Andersend, and A. Pate. Resolving crises through automated bilateral negotiations. *Artificial Intelligence journal*, 172(1):1–18, 2008.

[6] S. Kraus, P. Hoz-Weiss, J. Wilkenfeld, D. R. Andersen, and A. Pate. Resolving crises through automated bilateral negotiations. *Artificial Intelligence*, 172(1):1–18, 2008.

[7] Y. Oshrat, R. Lin, and S. Kraus. Facing the challenge of human-agent negotiations via effective general opponent modeling. In *Proceedings of AAMAS 2009*, 2009.

[8] N. Peled, Y. Gal, and S. Kraus. A study of computational and human strategies in revelation games. In *Ninth International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, Taipei, 2011.

[9] J. W. Pratt. Risk aversion in the small and in the large. *Econometrica*, 32(1/2):122–136, 1964.

[10] D. Rajarshi, J. E. Hanson, J. O. Kephart, and G. Tesauro. Agent-human interactions in the continuous double auction. In *Proc. of IJCAI-01*, 2001.

[11] F. Ricci, L. Rokach, B. Shapira, and P.B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.

[12] A. Rosenfeld and S. Kraus. Using aspiration adaptation theory to improve learning. In *Ninth International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, Taipei, 2011.

[13] Guy Shani, David Heckerman, and Ronen I. Brafman. An mdp-based recommender system. *J. Mach. Learn. Res.*, 6:1265–1295, 2005.

[14] A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.