A Culture Sensitive Computer-Agent in a Non-binding Multi-round Bilateral Negotiation

Galit Haim[1]

[1]Computer Science Department, Bar Ilan University

ABSTRACT

The ability to negotiate successfully is critical in many social interactions. The dissemination of applications such as the Internet across geographical and ethnic borders is opening up opportunities for computer agents to negotiate with people of diverse cultures. These automated negotiators should be able to proficiently interact with their human partners. This paper presents an agent design for negotiating with people in settings where participants repeatedly need to agree on the exchange of scarce resources, but agreements are not binding. Such settings characterize many interactions in the real world. A major challenge of modeling people's behavior in such settings is that people retaliate or reward each other's actions over time, despite the absence of direct material benefit. In addition, people's behavior is highly dependent on social and cultural factors. To meet this challenge, our approach combines machine learning techniques of people's negotiation behavior in a game with decision theory. Our Personality Adaptive Learning (PAL) agent explicitly reasons about the tradeoff between being reliable and generous towards people and the ramifications of its actions on its future success, given its model of how people retaliate and reward its actions. In our empirical investigation we performed extensive human subject trials which were used to train models of people's behavior. We compared the performance of PAL with new people. Our preliminary results show that when learning models from subjects in Israel, PAL was able to outperform new people in Israel in all dependency relationships. Our hypothesis is that these results will also generalize to models learned from data collected in the U.S. and Lebanon.

1. INTRODUCTION

Negotiation is a tool widely used by humans to resolve disputes in settings as diverse as business transactions, diplomacy and personal relationships. Computer agents that negotiate successfully with people have profound implications: They can negotiate on behalf of individual people or organizations (e.g., bidders in on-line auctions (Kamar et al., 2008; Rajarshi et al., 2001)); they can act as training tools for people to practice and evaluate different negotiation strategies in a lab setting prior to embarking on negotiation in the real world (e.g., agents for negotiating a simulated diplomatic crisis (Lin et al., 2009)); or work autonomously to reach agreements for which they are responsible (e.g., computer games, systems for natural disaster relief (Schurr et al., 2006; Murphy, 2004)). (Haim et al., 2010) shows that people's cultural diversity affects the accuracy of prediction models on human negotiation behavior.

The purpose of this paper was to investigate the role of culture in learning and adapting people's negotiation behavior with computer agents. Our goal was to be able to build a computer agent that can learn and adapt in order to negotiate proficiently with people across different cultures. Culture is a key determinant of the way people interact and reach agreements in different social settings. Advancing technology such as the Internet requires that computer systems negotiate proficiently with people across geographical and ethnic boundaries. It is thus important to understand the decision-making strategies that people of different cultures deploy when computer systems are among the members of the groups in which they work, and to determine their responses to the different kinds of decision-making behavior employed by others.

This paper investigates the hypothesis that explicitly representing behavioral traits that vary across cultures will improve the ability of computer agents to learn and adapt human negotiation behavior. We expect that in turn this will improve the performance of computer agents when negotiating with people. To evaluate this hypothesis, we developed the Personality Adaptive Learning (PAL) agent that combines machine learning of people's negotiation behavior in the game with decision theory. The PAL agent explicitly reasons about the tradeoff between being reliable and generous towards people and the ramifications of its actions on its future success, given its model of how people retaliate and reward its actions. It learned separate models from human negotiation behavior data. This data was collected in laboratory conditions from three different countries: Israel, Lebanon and the U.S. We used an identical negotiation scenario in each country which required people to complete a task by engaging in bilateral negotiation rounds with non-binding agreements. The negotiation protocol included alternating take-it-or-leave-it offers for the exchange of resources. Agreements were not binding, and participants were free to choose the extent to which they fulfilled their commitments. The main contribution of this work is that it suggests a new paradigm of automatic agents which can negotiate with people across cultures by combining the individual learner models into an agent's decision-making model, and by adapting the learning to a specific culture.

2. RELATED WORK

There is a body of work in the psychological and social sciences that investigates cross-cultural behavior among human negotiators (De Dreu and Van Lange, 1995; Gelfand and Dyer, 2001; Gelfand et al., 2002). However, there are scant computational models of human negotiation behavior that reason about cultural differences. In artificial intelligence, past works have used heuristics, equilibrium strategies and opponent modeling approaches toward building computer agents that negotiate with people. For a recent comprehensive review, see Lin and Kraus (2010). Within repeated negotiation scenarios, Kraus et al. (2008) modeled human bilateral negotiations in a simulated diplomatic crisis characterized by time constraints and deadlines in settings of complete information. They adapted equilibrium strategies to people's behavior using simple heuristics, such as considering certain non-optimal actions. Jonker et al. (2007) designed computer strategies that involve the use of concession strategies to avoid impasses in the negotiation. Byde et al. (2003) constructed agents that bargain with people in a market setting by modeling the likelihood of acceptance of a deal and allowing agents to renege on their offers. Kenny et al. (2007) constructed agents for the training of individuals to develop leadership qualities and interviewing capabilities.

Recent approaches have used learning techniques to model the extent to which people exhibit different social preferences when they accept offers in one-shot and multiple interaction scenarios (Gal et al., 2009; Oshrat et al., 2009; Lin et al., 2008). Learning techniques have also been applied to model gender differences (Katz and Kraus, 2006) and the belief hierarchies that people use when they make decisions in one-shot interaction scenarios (Gal and Pfeffer, 2007; Ficici and Pfeffer, 2008).

To date, all work on human-computer negotiation assumes that agreements are binding and have relied on prior data of people's negotiation behavior. A notable exception is the work by Kraus and Lehmann (1995) where they proposed an agent for negotiating with multiple participants which may renege on agreements. This work, however, was restricted to the specific domain of the game of diplomacy.

This research extends the human-computer negotiation in its focus on situations where agreements are not binding.

3. IMPLEMENTATION USING THE COLORED TRAILS TEST-BED

Our study was based on the Colored Trails (CT) game (Grosz et al. (2004)), a test-bed for investigating decision-making in groups comprising people and computer agents. Colored Trails is a free software and is available for download at `http://www.eecs.harvard.edu/ai/ct`. The CT configuration we used consisted of a game played on a 7x5 board of colored squares with a set of chips. One square on the board was designated as the goal square. Each player's icon was initially located in one of the non-goal positions, eight steps away from the goal square. To move to an adjacent square a player needed to surrender a chip in the color of that square. Players were issued 24 colored chips at the onset of the game.

Figure 1 shows the CT board game, in which there are two players, "me" and "O". The board game is shown from the point of view of the "me" player. The relevant path from the point of view of the "me" player is outlined. Figure 2 shows the chips that both players possess at the onset of the game. Both the "me" and "O" players are missing three chips to get to the goal. The "me" player is lacking three yellow chips, while the "O" player is lacking three grey chips. In addition, each player has the chips that the other player needs in order to get to the goal. For example, the "me" player has ten grey chips. Figure 3 shows an example of a proposal made by the "me" player to give two grey chips to the "O" player in return for two of its yellow chips.

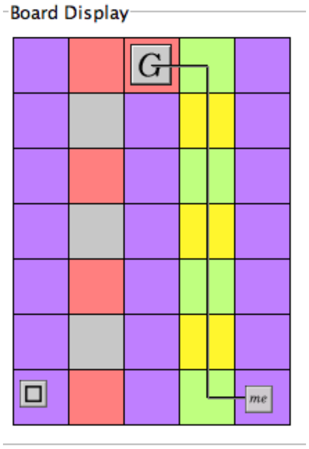
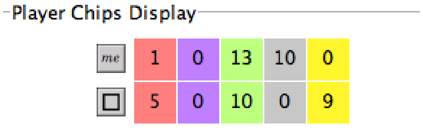


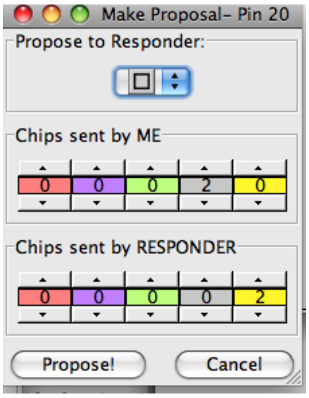
Figure 1: An example of a CT Board



me	1	0	13	10	0
□	5	0	10	0	9

Figure 2: Chip Display Panel (showing the chips possessesd by both participants)

At the onset of the game, one of the players was given the role of proposer, while the other was given the role of responder. The interaction proceeded in a recurring sequence of phases. In the *communication* phase, the player designated as the proposer could make an offer to the other player, who was designated the responder. In turn, the responder could accept or reject the offer. If the offer was rejected, then players switched roles: the responder became the proposer and the proposer became the responder. This sequence of alternating offers continued until an offer was accepted, or the time limit for the communication phase was up. In the *transfer phase*, both players could choose chips to transfer to each other. The transfer action was done simultaneously, such that neither player could see what the other player transferred until the end of the phase. In particular, players were not required to fulfill their commitments to an agreement reached in the communication phase. A player could



Make Proposal- Pin 20

Propose to Responder:

Chips sent by ME: 0 0 0 2 0

Chips sent by RESPONDER: 0 0 0 0 2

Figure 3: Communication Panel (used by participants to make offers)

choose to transfer more chips than it agreed to, or any subset of the chips it agreed to, including transferring no chips at all. In the *movement phase*, players could manually move their icons across one square on the board by surrendering a chip in the color of that square. At the end of the movement phase, a new communication phase began. Players alternated their roles, such that the first proposer in the previous communication phase was designated as a responder in the next communication phase, and vice versa. These phases repeated until the game ended, which occurred when one of the following conditions held: (1) at least one of the participants reached the goal square; or (2) at least one of the participants remained dormant and did not move for three movement phases. When the game ended, both participants were automatically moved as close as possible to the goal square, and their score was computed as follows: 100 bonus points for getting to the goal square, 5 bonus points for any chip left in a player's possession; a 10 point penalty for each square left in the path from a player's final possession to the goal square.

These parameters were chosen so that getting to the goal was by far the most important component, but if a player could not get to the goal it was preferable to get as close to the goal as possible. Note that players had full view of the board and each others' chips, and thus they had complete knowledge of the game situation at all times during the negotiation process.

One of the advantages of using CT for cross-cultural studies is that it provides a realistic analog to task settings, highlighting the interaction among goals, tasks required to achieve these goals and resources needed for completing tasks. In CT, chips correspond to agent capabilities and skills required to fulfill tasks. Different squares on the board represent different types of tasks. A player's possession of a chip of a certain color corresponds to having the skill available for use at that time. Not all players possess chips in all colors, much as different agents vary in their capabilities. Traversing a path through the board corresponds to performing a complex task whose constituents are the individual tasks represented by the colors of each square.

CT is thus particularly suitable for modeling negotiation that occurs between people of different cultures, in which negotiation processes are conducted within task contexts and involve the exchange of resources (for example, within diplomatic negotiations for trade agreements or peace treaties). In addition, it has been shown that people who use CT generally display more cooperative behavior than identical decision-making scenarios that involve more abstract representations, such as payoff matrices or decision-trees (Gal et al., 2007). This incentive for cooperation may allow both parties in negotiation to reach agreements more quickly. Both of these are important qualities to multi-cultural disputes that are often volatile.

4. THE PAL AGENT

As already discussed, in order to enable computer-agents to negotiate proficiently with people across geographical and ethnic boundaries, it is important to understand the decision-making strategies that people of different cultures employ when they negotiate with computer systems. Therefore, we developed a new agent named PAL (Personality Adaptive Learning). The PAL agent is based on:

1. Learning: Data collection was used to build predictive models of human negotiation behavior:

    - Reliability Model - the extent to which a person was reliable in the negotiation.
    - Acceptance Model - the likelihood of accepting a given proposal.
    - Agent Reached Goal Model - the likelihood of the agent getting to the goal.

    Data was collected from an identical negotiation scenario of human negotiation behavior under laboratory conditions in different countries.

2. A utility function that used these models. The PAL aims to maximize its expected utility.

For a detailed explanation of the PAL learning models and utility functions, see appendix A.

4.1 Potential Features

For the various learning models we used the the following set of potential. Each feature is described from the point of view of a general player in the game. We also considered the symmetrical feature from the point of view of the opponent player.

- The current round in the game.
- The current score of a player in a specific round measures the score in the game given its current set of chips.
- The resulting score of a player in specific round measures the score that the player would receive in the case that both players sent all of the promised chips according to the proposal.
- The score-base-reliability of a player in the game in a specific round is the extent to which the player fulfilled an agreement. It computes reliability measures solely for negotiation rounds in which agreements were reached.
- The weighted score-base-reliability is a weighted average of the score-base-reliability.

- The generosity of a player in a specific round. This feature clustered the chip sets offered by both players in a specific round into three classes measuring the difference in the number of chips proposed.

- The dependency role of a player in a specific round can be one of two classes: task independent, task dependent.

- Missing chips: this feature includes the total number of chips that a player needed to get to its goal given its position on the board in a specific round.

- The dormant round of a player in a specific round is the player's current number of no consecutive movement. When the player moves, the number is re-initialized to zero. If this number exceeds 3, the game is over. Note that according to the data we collected for building the agent reached-goal model, people in Lebanon played the game but did not move. In this case, this feature was very important.

Table 1 (left side) lists the set of optimal features chosen to be the potential features of the learning models. The selection process for the features was carried out by hand on a held-out test-set that was not used to evaluate the learning modules. This means that we evaluated the learning models from the 90% of the collected data we have, and we tested these models on the remaining 10% of the data.

4.2 Decision-making

This section provides a high-level description of how PAL makes decisions in the game. PAL combines backward induction with predictive models of human behavior. It uses an Expectimax tree with two types of nodes. Decision nodes in the tree correspond to decisions by PAL (what offers to make, how reliable to be, whether to accept an offer). Edges emanating from decision nodes are labeled with possible actions for PAL's decisions. Chance nodes in the tree correspond to decisions made by people. Edges emanating from chance nodes are labeled with probabilities that predict people's behavior. An example of a decision tree for making offers in the game is shown in Figure 4.

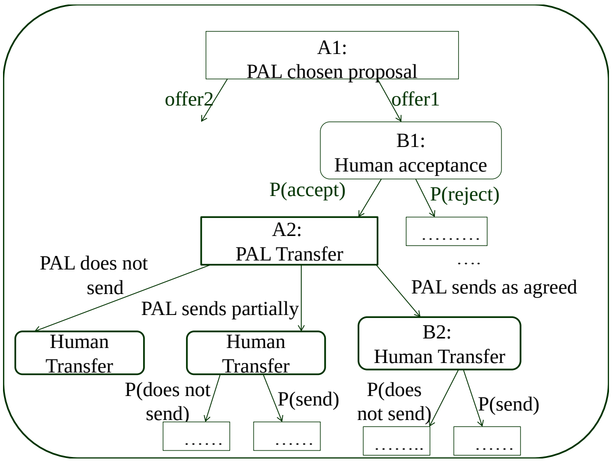


Figure 4: One round of a decision tree for PAL making offers.

Let s be the current state in the game that encompasses players' positions on the board, their chips, and the history of past offers. Leaves in the tree are labeled with PAL's expected utility given its model of reaching the goal at s. Appendix A.1.1 describes how the agent reaches the goal model. This represents PAL's valuation function.

We define a round as a sequence including an offer made by a proposer player at s, a response made by a responder player and transfer decisions for both players. (Note that chips can be transferred regardless of whether an offer is accepted. In case of rejection, this is a way of being generous, for instance). Suppose that PAL is a proposer at the beginning of a round. Then for each possible offer of PAL in state s (e.g., node A1), it considers the probability that the person accepts the offer at s (e.g., node B1). PAL then decides whether to send its promised chips at s (e.g, node A2) given its model of people's reliability (node B2). At the end of a round, PAL transitions to state s' which extends the history to include the results of the round. At this point PAL can consider decisions for the next round or it can generate a leaf as described above. PAL chooses the offer that maximizes its expected utility given the decision tree.

It is easy to see that the size of the tree is exponential to the number of rounds. For tractability, PAL considers only two rounds. In addition, it only considers all-or-nothing transfers when reasoning about how many chips to transfer. The approach for choosing whether or not to accept offers and whether or not to transfer chips is similar and is omitted for brevity.

4.3 Learning Algorithm

Our study was based on the Weka framework (`http://www.cs.waikato.ac.nz/ml/weka/`), a repository of machine learning algorithms that is freely available on the web. We used the Multilayer-perceptron (MLP) learner to model all prediction tasks. With its remarkable ability to derive meaning from complicated or imprecise data, MLP can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computational techniques. In addition, MLP has the ability to learn how to do tasks based on the data given for training or initial experience. It does not make any assumption regarding the underlying probability density functions or other probabilistic information about the pattern classes under consideration in comparison to other probability based models. In turn, it yields the required decision function directly via training.

5. EMPIRICAL METHODOLOGY

5.1 Data Collection

In order to build culture-based models for the PAL agent, we collected data from three countries: Lebanon, the U.S. and Israel. In each country a human played against different computer agents and also played human versus human. (Haim et al., 2010) describes in detail the collected data where a human played against the specific agent named (Personality, Utility Rule Based) PURB agent.

These models were used to play against PAL. 63 subjects from the Information System Department of Ben Gurion University in Israel played the CT game against the PAL agent. Each participant was given an identical 30 minute tutorial on CT. This tutorial consisted of a written description of the CT game, as well as a short movie that explained

Table 1: Features used for Learning Modules

Model	Feature set
Reliability	CS, RS, WPR, NC and OC
Acceptance	CS, RS, OG, NC and OC
Agent reached Goal	CS, PR, R and DR(only in Lebanon)

Feature Key	Description (FBP=For Both Players)
CS	Current Score(FBP)
RS	Resulting Score(FBP)
PR	Previous Reliability(FBP)
NC	Needed chips to reach the goal (FBP)
OC	Other chips (FBP)
WPR	Weighted Previous Reliability(FBP)
OG	Offer Generosity
R	player's Role(FBP)
DR	Dormant Round Number(FBP)

the rules of the game using a different board than those used in the study. Participants were seated in front of terminals for the duration of the study and could not speak to each other or see the other terminals. All participants played one or two games with the PAL agent, but were told they would be playing with different people. Authorization for this slight deception was granted by the ethics review board of the institutions that participated in the study. Subjects were given an extensive debriefing at the end of the study which revealed this fact and explained the study.

The study included the CT game. We used three different types of boards. In all of these boards, there was a single distinct path from each participant's initial location to its goal square. One of the board types exhibited a symmetric dependency relationship between players: neither player could reach the goal given its initial chip allocation, and there existed at least one exchange such that both players could reach the goal. We referred to players in this game as task co-dependent. The other board types exhibited an asymmetric task dependency relationship between players: one of the players, referred to as *task independent*, possessed the chips it needed to reach the goal, while the other player, referred to as *task dependent*, required chips from the task-independent player to get to the goal and vice versa. An example of the co-dependent board is shown in Figure 1. In this game both "me" and "O" players were missing three chips to get to the goal. The relevant path from the point of view of the "me" player is outlined.

To standardize our experiments, people were designated as first proposers and the PAL agent was designated as the first responder in all of the CT games we ran. Each subject was randomly assigned one of the following dependency roles: a task co-dependent participant that was paired with a task co-dependent PAL agent; a task independent participant that was paired with a task dependent PAL agent; or a task dependent participant that was paired with another task independent PAL agent.

5.2 Evaluation Criteria

The criteria we used to evaluate the PAL agent is the benefit criteria. The benefit criteria is the final score minus the initial score, which is the same for each board. We compared the PAL agent's benefits versus the opponent's benefits.

6. RESULTS AND DISCUSSION

Our preliminary results show that when learning models from subjects in Israel, PAL was able to outperform new people in Israel in all dependency relationships. Table 2 lists, for each game condition, the number of games in this condition and the benefits of PAL versus opponent's benefits. In case of Agent-TI, where PAL was task independent and the opponent was dependent, PAL's benefit was 39.45 points more than the opponent. In case of Human-TI, where PAL was task dependent and the opponent was independent, PAL's benefit was 80.27 points more than the opponent. In the last game condition, BOTH-DD, where both PAL and the opponent were task dependent, PAL's benefit was 44.82 points more than the opponent.

Table 2: PAL vs Israeli human benefit results

	Number of games	PAL benefit	Opponent benefit
Agent-TI	18	27.5	-11.95
Human-TI	18	82.77	2.5
Both-DD	27	78.7	33.88

We hypothesize that these results will also generalize to models learned from data collected in the U.S. and Lebanon.

7. CONCLUSIONS

This paper presents our approach of combining machine learning of people's negotiation behavior in the game with decision theory. Our PAL agent explicitly reasons about the tradeoff between being reliable and generous towards people and the ramifications of its actions on its future success, given its model of how people retaliate and reward its actions. The agent learned separate models for whether or not people accept offers, the extent to which they commit to agreements, and the effect of its own reliability on the agent's future success.

It focused on a repeated negotiation setting in which participants need to accrue and exchange resources in order to complete their individual goals, and agreements were not binding. This setting was implemented using the Colored Trails game that consists of a computer board game which provided a task analogous to the types of interactions that occur in the real world.

Our results showed preliminary results for a cohesive agent such as PAL, that has a good chance of negotiating successfully with people from disparate cultures.

References

A. Byde, M. Yearworth, K.Y. Chen, C. Bartolini, and N. Vulkan. Autona: A system for automated multiple

1-1 negotiation. In *Proceedings of the 4th ACM conference on Electronic commerce*, pages 198–199. ACM New York, NY, USA, 2003.

C.K.W. De Dreu and P.A.M. Van Lange. The impact of social value orientations on negotiator cognition and behavior. *Personality and Social Psychology Bulletin*, 21: 1178–1188, 1995.

S. G. Ficici and A. Pfeffer. Simultaneously modeling humans' preferences and their beliefs about others' preferences. In *Proc. 7th International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, 2008.

Y. Gal and A. Pfeffer. Modeling reciprocity in human bilateral negotiation. In *Proc. 22nd National Conference on Artificial Intelligence (AAAI)*, 2007.

Y. Gal, B. Grosz, S. Shieber, A. Pfeffer, and A. Allain. The effects of task contexts on the decision-making of people and computers. In *Proce. of the Sixth International Interdisciplinary Conference on Modeling and Using Context, Roskilde University, Denmark*, 2007.

Y. Gal, S. DâĂŹsouza, P. Pasquier, I. Rahwan, and S. Abdallah. The effects of goal revelation on computer-mediated negotiation. In *Proceedings of the Annual meeting of the Cognitive Science Society (CogSci), Amsterdam, The Netherlands*, 2009.

M. Gelfand and N. Dyer. A cultural perspective on negotiation: Progress, pitfalls, and prospects. *Applied Psychology*, 49(1):62–99, 2001.

M.J. Gelfand, M. Higgins, L.H. Nishii, J.L. Raver, A. Dominguez, F. Murakami, S. Yamaguchi, and M. Toyama. Culture and egocentric perceptions of fairness in conflict and negotiation. *Journal of Applied Psychology*, 87(5):833–845, 2002.

B. Grosz, S. Kraus, S. Talman, and B. Stossel. The influence of social dependencies on decision-making. Initial investigations with a new game. In *Proc. 3rd International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, 2004.

G. Haim, Y. Gal, and S. Kraus. Learning Human Negotiation Behavior Across Cultures. Group Decision and Negotiation, 2010.

C.M. Jonker, V. Robu, and J. Treur. An agent architecture for multi-attribute negotiation using incomplete preference information. *Autonomous Agents and Multi-Agent Systems*, 15(2):221–252, 2007.

E. Kamar, E. Horvitz, and C. Meek. Mobile Opportunistic Commerce: Mechanisms, Architecture, and Application. *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, page 1087, 2008.

R. Katz and S. Kraus. Efficient agents for cliff edge environments with a large set of decision options. In *Proc. 5th International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, 2006.

P. Kenny, A. Hartholt, J. Gratch, W. Swartout, D. Traum, S. Marsella, D. Piepol, and CA Marina Del Rey. Building interactive virtual humans for training environments. In *Proceedings of I/ITSEC*. NTSA, 2007.

S. Kraus and D. J. Lehmann. Designing and building a negotiating automated agent. *Computational Intelligence*, 11:132–171, 1995.

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.

R. Lin and S. Kraus. Can automated agents proficiently negotiate with humans? *Commun. CACM*, 53(1):78–88, 2010.

R. Lin, S. Kraus, J. Wilkenfeld, and J. Barry. Negotiating with bounded rational agents in environments with incomplete information using an automated agent. *AIJ*, 172(6-7):823–851, 2008.

R. Lin, Y. Oshrat, and S. Kraus. Investigating the benefits of automated negotiations in enhancing people's negotiation skills. In *Proc of AAMAS-09*, pages 345–352, 2009.

R. Murphy. HumanâĂŞrobot interaction in rescue robotics. *IEEE Transactions On Systems, Man, and CyberneticsâĂŤPart C: Applications and Reviews*, 34(2), May 2004.

Y. Oshrat, R. Lin, and S. Kraus. Facing the challenge of human-agent negotiations via effective general opponent modeling. In *AAMAS*, pages 377–384, 2009.

D. Rajarshi, J. E. Hanson, J. O. Kephart, and G. Tesauro. Agent-human interactions in the continuous double auction. In *Proc. 17th International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

N. Schurr, P. Patil, F. Pighin, and M. Tambe. Using multiagent teams to improve the training of incident commanders. In *Proc. 5th International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, 2006.

APPENDIX

A. PAL UTILITY FUNCTIONS

A.1 General features

In this section we define a general set of features for the PAL agent to be used by the various potential utility function model formulas. Below we present a description of the layout of the features' notations.

Without loss of generality, for any two participants i and j, where $i, j \in \{a, h\}$, a refers to the agent, h refers to the human, within a specific game board type brd, let pos_i denote the position of player i on the board, ${pos_i}^*$ denotes a new position (after a player moves), and $\bar{d} = (d_i, d_j)$ denotes the dormant rounds of the players in the game.

Let c_i denote the set of chips that i possesses at phase n in the game. An action that is performed in the game by any player is noted as ac. ac can be a single action or a sequence of actions. Let ${mo^k}_i$ denote i making an offer, where k is the round type, $k = 1, 2$, where 1 relates to making an offer and 2 relates to making a counter offer.

An offer proposed by the proposer i is defined as $\bar{o}_i = (o_i, o_j)$, where $o_i \subseteq c_i$ is the set of chips that i proposes to send to j, and let $o_i^+ \subseteq c_i$ be the set of chips actually sent by i following an agreement $\bar{o}_i$.

The set of offers that i can propose is defined as $\widehat{o}_i$. This phase is called the communication phase. In this phase, the player playing the role of the "Proposer" is entitled to make an offer to the opponent. A player can offer to send and receive as many chips as he wants as long as each player has at least that amount of chips. The player playing the role of the "Responder" can accept or reject the proposal, but in this round cannot make an offer of its own. After an offer is proposed, let ${ro^k}_j$ denote that j responded to an offer. The response can be $accept_j$ when j accepts i's proposal, or $reject_j$ when j rejects i's proposal.

Then, the next phase is the exchange phase defined by ex, and after exchanging chips, the movement phase defined by mv begins. Let ${c_i}^*$ denote the chip i uses to move. i's move to another position given the chip needed for this move is defined by $move_i({pos_i}^*, {c_i}^*)$. In general, $\alpha \in \{{mo^k}_i, {ro^k}_i, mv, ex\}$ is defined as the phase in the game. Let $s_i(brd, c_i, pos_i, \alpha, d_i, r_i)$ denote the state of i given the tuple (board, chips, position, phase, dormant, round), and let $\bar{s} = \{s_i, s_j\}$ be the state of both players. When a specific parameter within the state s_i shall not be changed by an action, it will be noted as $NULL$.

An additional required parameter for the utility function is defined as $\bar{v}_i$, which relates to the reliability parameters within the models that need to be recalculated when executing the expected score formula. These reliability parameters are the score-based-reliability and the weighted score-based-reliability. The score-based-reliability of a player at specific round is the extent to which a player fulfilled an agreement at this round. It computes reliability measures solely for negotiation rounds in which agreements were reached. This is defined as the ratio between the score to j given the chips that i actually transferred, and the score that j would receive if i fulfilled its agreement. We also consider the symmetrical notations from j's perspective.

Before describing the utility functions, we present some general notations that are used within these formulas:

- "$\|$" refers to events that occur in parallel.
- ";" refers to actions that occur sequentially.
- "\`" refers to a parameter whose value will be changed.
- "$\rightarrow$" refers to a change that is done on the left side of the $\rightarrow$ and results on the right side of the $\rightarrow$.

We first present a detailed list of how states are changed based on the actions taken by the players.

- $\bar{s}[ac_1; ac_2] = (\bar{s}[ac_1])[ac_2]$
  The state resulting from the performance of two sequential actions is equivalent to executing the first action on the state and then the second one. For example, $\bar{s}[\bar{o}_i; accept_j]$ denotes the state after a player i proposes an offer, and then player j accepts it.
- $\bar{s}[\bar{o}_i] \rightarrow (\grave{\alpha} = {ro^k}_j \text{ when } \alpha = {mo^k}_i)$
  Proposing an offer $\bar{o}_i$ in a given state s causes the phase to be changed to a phase of response to an offer.
- $\bar{s}[accept_i] \rightarrow (\grave{\alpha} = ex)$
  Accepting an offer causes the phase of the state to be changed to chip exchanges.
- $\bar{s}[o_j^+ \| o_i^+] \rightarrow (\grave{c}_i = c_i \cup {o_j}^+ \setminus {o_i}^+, \grave{c}_j = c_i \cup {o_i}^+ \setminus {o_j}^+; \grave{\alpha} = mv, i \neq j; i, j \in \{a, h\}$.
  Chip exchange in a given state $\bar{s}$ causes both players' chips to be updated according to the ones sent and received. For example, assume a has 24 chips, 5 are grey and none are yellow, and assume a sent 2 grey chips and received 1 yellow chip. This will cause a's set of chips to be updated to a total amount of 23 chips, 3 grey and 1 yellow. The same update is done to the opponent's chip set.
- $\bar{s}[\bar{o}_i; accept; o_j^+ \| o_i^+] \equiv (\bar{s}[o_j^+ \| o_i^+])$.
  Making an offer, accepting it and exchanging chips is equivalent to the state resulting from chip exchanges.
- $$\bar{s}[reject_i] \rightarrow \begin{cases} \grave{\alpha} = mo_j^2 & \alpha = {ro_i}^1, i \neq j; i, j \in \{a, h\} \\ \grave{\alpha} = ex & \alpha = {ro_i}^2, i \neq j; i, j \in \{a, h\} \end{cases}$$

  Rejecting a proposal causes the phase of the state to be changed to placing a counter offer if it was the 1st response to an offer, otherwise the phase is changed to chip exchange.
- $\bar{s}[NULL] = (\bar{s})$
  No action does not change the state.
- $\bar{s}[move_i(pos_i^*, c_i^*), \neg(move_j)] \rightarrow ((\grave{c}_i = c_i \setminus {c_i}^*), \grave{d}_i = 0, \grave{d}_j = d_j + 1, i \neq j; i, j \in \{a, h\})$
  A move by player i and NOT by player j causes the set of chips of player i to be updated and its dormant type to become zero. The dormant type of player j is increased by 1 since he has not moved.
- $\bar{s}[move_i, move_j] \rightarrow ((\grave{c}_i = c_i \setminus {c_i}^*), \grave{d}_i = 0, (\grave{c}_j = c_j \setminus {c_j}^*), \grave{d}_j = 0, i \neq j; i, j \in \{a, h\})$
  A move by both players causes their sets of chips to be updated and their dormant type to become zero.

We now provide a detailed description of changing $\bar{v}$, which causes the weighted previous reliability and previous reliability parameters to be re-calculated for both players.

- $\bar{v}[ac_1; ac_2] = (\bar{v}[ac_1])[ac_2]$
  The change in $\bar{v}$ after two sequential actions is the same as the change in $\bar{v}$ after the first action is performed, and then, re-calculate according to the second action.

- $$\bar{v}[\bar{o_i}; accept; o_j^+ || o_i^+] \rightarrow \begin{cases} w\grave{p}r_i = & \bar{o_i} = lao, lao \neq -1 \\ (0.7 \times pr\grave{_i}\, o_i) + & \\ (0.3 \times wpr_{i_{\bar{o_i}}}) & \\ w\grave{p}r_i = 1 & lao = -1 \end{cases}$$

  Proposing an offer, accepting it and exchanging chips causes parameter wpr_i (Weighted Previous Reliability) to be updated. If it is the 1st accepted offer, this parameter will be a constant, otherwise, it will be calculated according to the specified formula.

A.1.1 Agent reach the goal model

Based on the definitions above, let $ES_i(\bar{s}, ac, \bar{v}, \bar{d})$ be the expected score of i given the tuple (state, action, reliability traits, dormant). This is the expected score that will be used by the utility function. In addition, let $P(G_i|\bar{s}, \bar{v})$ denote the probability that i will reach the goal given a state $\bar{s}$ and $\bar{v}$, and let $P(\neg G_i|\bar{s}, \bar{v})$ denote the probability that i will not reach the goal. The expected score of reaching the goal or not reaching the goal is calculated by a heuristic function h_i.

A.2 Utility of chip transfer

The transfer model is used to predict the opponent's transfer behavior, and thereby determine the PAL transfer strategy. We assume that the possible values that will be considered for o_j^+ (the opponent's actually transferred chips) will be $\{o_j\}$ or $\{\emptyset\}\}$. This means that the opponent will send all the chips as agreed, or he will send nothing from the agreed proposal. Note that all the probabilities within these utility functions are calculated using the learned models.

1. $ES_a(\bar{s}, NULL, \bar{v}) = P(G_a|\bar{s}, \bar{v}) \times h_a(\bar{s}, G_a, \bar{v}) +$
   $(1 - P(G_a|\bar{s}, \bar{v})) \times h_a(\bar{s}, \neg G_a, \bar{v})$
   This is the basis of the recursion function. It denotes the expected score when the state and the phase do not change.
   This is measured as the probability of reaching the goal when the agent is in the given state multiplied by a heuristic function to reach the goal plus the probability of not reaching the goal when the agent is in the given state multiplied by a heuristic function not to reach the goal.

2. $ES_a(\bar{s}, \bar{o}_i; accept_j; o_a^+, \bar{v}) =$
   $\sum_{o_h^+ \subseteq c_h} P(o_h^+|s[\bar{o}_i; accept_j], \bar{o}_i, \bar{v}) \times$
   $ES_a(\bar{s}[\bar{o}_i; accept_j; o_h^+||o_a^+], NULL, \bar{v}[o_i; accept_j; o_h^+||o_a^+])$
   Denotes the expected score of a proposal i offered and j accepted, and $o_j^+; o_i^+$ were sent.
   This is measured as the sum of the probabilities of the opponent to send o_j^+ multiplied by the
   expected score of the state after $o_h^+; o_a^+$ were sent.
   Note: this formula is a general one. Actually, in PAL, the only possible values that will be considered for o_j^+ will be o_j^+ or $\{\emptyset\}\}$.

3. Let o_a^* be $argmax_{o_a^+ \subseteq o_a}(ES_a(\bar{s}, \bar{o}_i; accept_j; o_a^+, \bar{v}))$
   o_a^* will be the actual chips that will be transferred, and is calculated by performing the expected score of the agent from each subset of the agent's chips of the offer. The o_a^+ that yields the highest expected score value is the o_a^*.

A.3 Utility of accepting an offer

This model is used by PAL to decide whether to accept or reject a specific offer that was proposed by the opponent.

1. $ES_a(\bar{s}, \bar{o}_h; accept_a; \bar{v}) = ES_a(\bar{s}, \bar{o}_h; accept_a; o_a^*; \bar{v})$
   Denotes the expected score for accepting the proposed offer. This is calculated by using the expected utility of transferring chips given the following parameters: the current state within the game, the opponent's proposed offer, PAL accepting it, transferring o_a^* and the previous reliability parameters that need to be recalculated to predict this expected score.

2. $ES_a(\bar{s}, \bar{o}_h; reject_a; \bar{v}) = ES_a(s[\bar{o}_h; reject_a], NULL, \bar{v})$
   Denotes the expected score for rejecting the proposed offer. This is calculated by using the utility formulas of the transfer model given the following parameters: the current state of the game, the opponent's proposed offer, PAL rejecting it (and therefore no chips are sent which is referred to as the NULL parameter), and the previous reliability parameters that need to be re-calculated to predict this expected score.

3. Let ac^* be $argmax_{ac_a \subseteq \{accept_a, reject_a\}}(ES_a(\bar{s}, \bar{o}_h; ac_a; \bar{v}))$
   Deciding on whether to accept or reject an offer by PAL is done by taking the action that yields the highest expected score.

A.4 Utility of an offer

The purpose of the following utility function is to decide which offer PAL will propose the opponent. This is done by estimating the expected score of each possible offer.

1. $ES_a(\bar{s}, \bar{o}_a, \bar{v}) = P(accept_h|\bar{s}[\bar{o}_a], \bar{o}_a, \bar{v}) \times$
   $ES_a(\bar{s}, \bar{o}_a; accept_h; o_a^*; \bar{v}) + (1 - P(accept_h|\bar{s}[\bar{o}_a], \bar{o}_a, \bar{v})) \times$
   $ES_a(\bar{s}[\bar{o}_a, reject_h], NULL, \bar{v})$
   The expected score of a specific optional offer is calculated by using the probability that the opponent will accept this offer multiplied by the expected score of PAL given the state, the proposed offer, accepting the offer by the opponent, the actual chips PAL will transfer if this is the chosen offer, and the reliability parameters that might be recalculated. In addition, the probability that the opponent will reject this offer and the expected score in such a case need to be calculated.

2. Let $\bar{o}_a^*$ be $argmax_{\bar{o}_a \subseteq \hat{o}_i}(ES_a(\bar{s}, \bar{o}_a, \bar{v})))$
   The actual offer that will be proposed will be the one that gives the highest expected score in comparison to the other optional offers.