A Cultural Decision-Making Model for Negotiation based on Inverse Reinforcement Learning

Elnaz Nouri (nouri@ict.usc.edu)

Kallirroi Georgila (kgeorgila@ict.usc.edu)

David Traum (traum@ict.usc.edu)

Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Playa Vista, CA 90094, USA

Abstract

We learn culture-specific weights for a multi-attribute model of decision-making in negotiation, using Inverse Reinforcement Learning (IRL). The model takes into account multiple individual and social factors for evaluating the available choices in a decision set, and attempts to account for observed behavior differences across cultures by the different weights that members of those cultures place on each factor. We apply this model to the Ultimatum Game and show that weights learned from IRL surpass both a simple baseline with random weights, and a high baseline considering only one factor of maximizing gain in own wealth in accounting for the behavior of human players from four different cultures. We also show that the weights learned with our model for one culture outperform weights learned for other cultures when playing against opponents of the first culture. We conclude that decision-making in negotiation is a complex, culture-specific process that cannot be explained just by the notion of maximizing one's own utility, but which can be learned using IRL techniques.

Keywords: cultural decision-making; negotiation; ultimatum game; inverse reinforcement learning.

Introduction

Social scientists have often observed that people from different cultures behave differently in interactive situations (Camerer, 2003; Roth, Prasnikar, Okuno-Fujiwara, & Zamir, 1991). There are several different possible explanations for this, including

1. one culture is better than another at optimizing outcomes;
2. there is some kind of convention (Lewis, 1969) or equilibrium at work, such that people behave differently because the context is different, particularly their expectations about how others will behave. For example people in Japan or England drive on the left while people from America and Europe drive on the right, because that is the safest, most efficient way given how other drivers will behave, even though the goals of safety and efficiency are the same, and neither is innately better at achieving these goals;
3. the cultures have different goals, which lead to their optimizing different functions.

Most classical economic game-theory accounts of decision-making, e.g. (Neumann & Morgenstern, 1944), look at a monolithic notion of utility and maximizing expected utility as the key to rationality. This, in effect, denies the third explanation above. For very simple games, where it is relatively easy to calculate the payoffs, the first possibility seems hard to believe, thus we are left with the hypothesis that differences in behavior are based on applying common utility principles to different problems. Others, e.g. (Gal, Pfeffer, Marzo, & Grosz, 2004), have claimed that there are many factors that contribute to the behavior of humans in social situations. This makes the third explanation plausible, if people from different cultures have different relative weights for the different factors. But this leads to a further question of how to determine those different weights. In (Nouri & Traum, 2011) we presented one such model of decision-making that culture-specific virtual agents were able to use to play the Ultimatum Game (see the following section) with each other or with people. The model used Hofstede's multi-dimensional model of culture (Hofstede, 2001) to determine the relative weights of different factors. However, in that work the weights were set manually using our intuitions about how to apply the literature, which involved a number of relatively arbitrary decisions.

In this paper we attempt to learn the weights using Inverse Reinforcement Learning (IRL) (Abbeel & Ng, 2004). To our knowledge no one has used IRL before in the Ultimatum Game or generally to learn patterns of behavior in negotiation. We also perform two experiments to try to get at the question above of what is the best explanation for the observed behavioral differences across cultures. On one account, it is the different goals that lead to different behavior. In this case we would predict that we learn different goals for different cultural patterns and that these goals would be better at generating observed behavior than other possible goals. On another account, we would expect the same set of goals to be satisfactory for any population, and differences in behavior to result from the different environments that are encountered. Our results show that the learned weights are better able to match observed distributions of culture-specific behavior than either arbitrary weights, a simple model based on economic gain, or in most cases the weights learned for other cultures. This suggests that cultures vary in goals, not just conventional circumstances but also that we can successfully use IRL techniques to learn population-specific goals for this type of game.

The structure of the paper is as follows. First we briefly present the Ultimatum Game and studies that show different behaviors for different culture groups. Then we describe our decision-making model and we present an overview of Reinforcement Learning (RL) and IRL. After that we talk about our experimental setup and present our results. Then we discuss our results and propose ideas for future work, and finally we conclude.

Culture and the Ultimatum Game

We use the Ultimatum Game as a testbed for our model. The Ultimatum Game involves two players bargaining over a certain amount of money (in our experiments, $100). One player, the proposer, proposes a division, and the second player, the responder, accepts or rejects it. If the responder accepts, each player earns the amount specified in the proposal, and if the responder rejects, each player earns zero. At perfect equilibrium, according to economic game theory, the proposer receives all or almost all of the money and the responder accepts all offers made to them. This classic experimental economics game has received a great deal of attention since the initial experiment by (Güth, Schmittberger, & Schwarze, 1982). Results from these studies often deviate from the predictions of game theory (Henrich, 2000; Camerer, 2003). In fact there is considerable variation of offers and rejection rates across studies (Henrich, 2000; Buchan, Croson, & Johnson, 1999), and it has been reported that people from different cultures behave differently in this game. For example (Roth et al., 1991) studied the Ultimatum Game in four countries (US, Japan, Israel, and former Yugoslavia). They found that the offers in US and Yugoslavia were higher than the offers in Japan which were higher than the offers in Israel. (Henrich, 2000) compared the behavior of 18-30 year old Machiguenga men of the Peruvian Amazon with UCLA students and found significant differences, i.e. the offers of the latter were higher than the offers of the former. (Buchan et al., 1999) studied the differences in comparable student populations in Pennsylvania and Tokyo and observed that the offers of the former were lower than the offers of the latter.

All the above studies clearly show that culture can play an important role in negotiation and in particular in the Ultimatum Game. The question however is what role: different goals, or different conventions, and whether we can learn to emulate culture-specific behavior.

Our Decision-Making Model

Our decision-making model presented in (Nouri & Traum, 2011) considers a number of different metrics for evaluating a given situation, even for something as simple as division of money in an economic game such as the prisoner's dilemma (Camerer, 2003) or the Ultimatum Game (Güth et al., 1982). Each of the metrics can be calculated from a basic payoff matrix. The metrics we considered for the Ultimatum Game include: *Self* (the agent's own gain); *Other* (the gain of another); *Self/Other* (the relative gain of the negotiators); *Minimum* (lower bound of any participant - the aim of Rawls' theory of justice (Rawls, 1971)). Each of these metrics can be given one or more valuations, choosing an optimum point and scale. The agent has a vector of weights, one per valuation, indicating the relative importance of that valuation. The total value for each choice is the sum of the product of values and weights for each valuation as shown in equation (1):

$$Value(Choice_i) = \sum_{j=1}^{n} (W_j * V_j(Choice_i)) \quad (1)$$

An advantage of this multi-valuation approach is that it can model an agent who cares (possibly to different extents) about different aspects of the situation, such as self-interest, collective interest, and fairness. In (Nouri & Traum, 2011) we also adapted this model to take into account Hofstede's dimensions (Hofstede, 2001), i.e. *Individuality* (IDV), *Power Distance* (PDI), *Long Term Orientation* (LTO), *Masculinity* (MAS), *Uncertainty Avoidance* (UAI). Thus our generalized model shown in (2) breaks down the elements of the weight vector into one component per dimension, and thus an overall matrix of n valuations and m (=5) dimensions.

$$Value(Choice_i) = \sum_{j=1}^{n} ((\prod_{d=\mathrm{IDV}}^{\mathrm{UAI}} W_{j,d}) * V_j(Choice_i)) \quad (2)$$

In this paper our focus is to learn the weights of (1) but our ultimate goal is also the learning of the weights of equation (2) that take into account Hofstede's dimensions (see the discussion section).

Reinforcement Learning and Inverse Reinforcement Learning

An agent's policy is a function from contexts to (possibly probabilistic) decisions that the agent will make in those contexts. Reinforcement Learning (RL) is a machine learning technique used to learn the policy of an agent (Sutton & Barto, 1998). For an RL-based agent the objective is to maximize the reward it gets during an interaction. Because it is very difficult for the agent, at any point in the interaction, to know what will happen in the rest of the interaction, the agent must select an action based on the average reward it has previously observed after having performed that action in similar contexts. This average reward is called *expected future reward*. RL is used in the framework of Markov Decision Processes (MDPs). An MDP is defined as a tuple (S, A, P, R, γ) where S is the set of states (representing different contexts) which the agent may be in, A is the set of actions of the agent, $P : S \times A \rightarrow P(S, A)$ is the set of transition probabilities between states after taking an action, $R : S \times A \rightarrow \Re$ is the reward function, and γ a discount factor weighting long-term rewards. At any given time step i the agent is in a state $s_i \in S$. When the agent performs an action $\alpha_i \in A$ following a policy $\pi : S \rightarrow A$, it receives a reward $r_i(s_i, \alpha_i) \in \Re$ and transitions to state s_{i+1} according to $P(s_{i+1}|s_i, \alpha_i) \in P$. The quality of the policy π followed by the agent is measured by the expected future reward also called Q-function, $Q^{\pi} : S \times A$

$\rightarrow \Re$. Details are given in (Sutton & Barto, 1998). There are several algorithms for estimating the Q-function and we use Q-learning (Sutton & Barto, 1998). However, Q-learning requires thousands of interactions between the agent and the environment in order to learn the optimal policy. In the case of a multi-party interaction, such as dialogue or the Ultimatum Game, the environment also needs to represent the decisions and actions of another participant. For this reason we need to build another agent, called a simulated user (SU) (Georgila, Henderson, & Lemon, 2006), that will behave as part of the environment and will interact with the policy for thousands of iterations to generate data in order to explore the search space and thus facilitate learning. Note that the SU generates a variety of actions for each state based on a probability distribution but does not learn from the interaction.

With RL, the reward function should be defined. Designing a good reward function is not trivial and not always possible. There are tasks where it is not clear what constitutes a good reward function. Inverse Reinforcement Learning (IRL) (Abbeel & Ng, 2004) aims to learn a reward function (not necessarily the true reward function) from a set of data recording interactions between the agent and the environment. This data is called *expert data*. The reward function R can be expressed as follows:

$$R_w(s,\alpha) = w^T \phi(s,\alpha) = \sum_{i=1}^{k} w_i \phi_i(s,\alpha) \quad (3)$$

where s is the state that the agent is in and α the action that it performs in this state, and w^T is a vector of weights w_i for the feature functions $\phi_i(s, \alpha)$. Note that these feature functions are specified manually and the weights w_i are estimated by IRL.

In particular we use the imitation learning algorithm (Abbeel & Ng, 2004). The imitation learning algorithm is an iterative process. Initially we have a random policy π_i that by interacting with the SU generates data. Then this data is compared with the expert data and the weights w_i are calculated. Based on these weights a reward function is estimated and RL is performed to learn a new policy π_{i+1} which generates a new set of data by interacting with the SU. Then this new data is compared with the expert data and new weights are calculated, a new reward function is computed and so forth. The iteration stops when the distance between the data generated from the interaction of the latest policy with the SU and the expert data is lower than an empirically set threshold.

Experimental Setup

We use data of the distribution of offers and acceptances or rejections for four different cultures (US, Japan, Israel, and former Yugoslavia) reported in (Roth et al., 1991). For each culture, we generate SU-proposers and SU-responders by using probability functions that match the reported data. (Roth et al., 1991) provide this data for the first and last round of the game. In our setup the game lasts 5 rounds. For the rounds in between we interpolate the first and last round values using weights that vary depending on the round. For example, for round 4 we give a higher weight to the last round values and for round 2 a higher weight to the first round values. For each culture we generate "expert" data by having the SU-proposer interact with the SU-responder for that culture. We then apply IRL to learn weights of different motivational factors for each of these cultures and roles (proposer and responder), by iteratively playing against the appropriate SU. We then use the weights as a reward function, using RL, to learn policies for a proposer and responder for each culture. We evaluate success of the learned policies by how closely they match the expert data. We compare our learned policies with two baselines: RL models trained with either a random reward function or a reward function based on maximizing the wealth of the agent. We also compare the policies learned for a particular culture with the policies learned for the other cultures and the human expert data of the other cultures.

Our state definition includes information about the accumulated wealth gain of the agent (AccSelf), i.e. the wealth gain that the agent has gathered starting from the first round of the game, the accumulated wealth gain of the SU (AccOther), the wealth gain of the agent in the current round (Self), the wealth gain of the SU in the current round (Other), and also different representations of their relative gain (Self/Other) and the minimum gain (Min). We also take into account the round of the game. There are 11 actions that the proposer can perform (offer=0, offer=10, ..., offer=100). The initial context can be different for each round depending on the accumulated wealth of the agents, and the resulting reward is uncertain, depending on the action of the responder. For the responder, there are only two actions (accept, reject), but again there are many possible different start states to consider depending on the accumulated wealth of the agents (the reward is deterministic based on the state and action chosen).

The feature functions that we use are binary, i.e. the value of the feature function $\phi_i(s, \alpha)$ is 1 when ϕ_i is true for state s and action α. So to form the feature functions $\phi_i(s, \alpha)$ each feature is paired with all the available actions. Table 1 lists the features that we use to represent the type of context that we consider in each state. Thus for the proposer the feature function Self$\geq$10-offer=10 is 1 when the self gain of the proposer is $\geq$10 and the proposer has made an offer of 10, which means that this feature function is going to be 0 at the time of the offer (because at that point Self is always 0), 1 after this offer has been accepted, and 0 after this offer has been rejected. We also use additional features related to the accumulated wealth that are not depicted in Table 1 due to space constraints. In fact every possible value of AccSelf or AccOther can form a feature, e.g. AccSelf=150, AccOther=200, etc. Thus for the proposer the feature function AccSelf=150-offer=20 is goint to be 1 when the accumulated wealth of the proposer is 150 and the proposer has made an offer of 20.

As we can see from the previous discussion our model is considerably different from the original SU model (human data) that just uses a probability distribution per round. First,

Table 1: Features used for IRL.

Self≥0	Other≥0	Self/Other>2
Self≥10	Other≥10	Self/Other>1
Self≥20	Other≥20	Self/Other=1
Self≥30	Other≥30	Self/Other<1
Self≥40	Other≥40	Self/Other<1/2
Self≥50	Other≥50	Min(Self,Other)=0
Self≥60	Other≥60	Min(Self,Other)=10
Self≥70	Other≥70	Min(Self,Other)=20
Self≥80	Other≥80	Min(Self,Other)=30
Self≥90	Other≥90	Min(Self,Other)=40
Self=100	Other=100	Min(Self,Other)=50

our model is deterministic for each state but keeps track of additional state information, such as accumulated gains for each side. Thus we can still get a range of different offers and responses from our agents, depending on the learned policy for each state (including the accumulated gain) and the probability of those states. Second, our model for a specific culture includes a reward function, which is specific to that culture distribution. Third, the reward function could potentially be applied to other problems (see the discussion section), whereas we would have to collect human data to create a SU for a new problem.

We perform two experiments. The goal of the first experiment is to show that the reasoning behind the actions of the proposer is better modelled as a complex tradeoff of multiple goals, and cannot be explained merely by learning the behavior patterns of the partners. Thus for the proposer and the responder and the 4 cultures we learn 3 policies using RL; one based on a random reward function that assigns arbitrary weights (weak baseline), one where the reward function is based only on wealth (strong baseline), and one based on IRL. If only the data patterns mattered and not the reward function, we should see comparable performance between policies trained using the weak baseline reward functions and policies using the learned ones. Surpassing this weak baseline would be evidence that reward functions matter. The strong baseline follows classical economic game theory predictions. If everyone really does have this as a reward function and differences in behavior are due to learned differences in convention rather than goals, we should see this reward function able to match the observed behavior of different populations. On the other hand, if the IRL reward functions lead to better models than the strong baseline, that is evidence that multiple factors are taken into consideration.

The purpose of the second experiment is to show that the weights learned with IRL really are culture-dependent, i.e. that they work better for the culture that the weights were learned from than models learned for other cultures. To show that we use IRL to learn the reward function for the 4 cultures and then we use these reward functions to learn policies for each culture (for example for the US culture we have policy-rewardUS-trainUS, policy-rewardJapan-trainUS, policy-rewardIsrael-trainUS, policy-rewardYugoslavia-trainUS). Then we test the 4 policies against SUs from the same culture that they were trained on (in this case, US). If the goals for different cultures really are different, then one would expect that policy-rewardUS-trainUS would better match the expert US data than policies learned using weights from other cultures.

To measure how closely the distributions generated with the 3 models match the human expert data we use Kullback-Leibler divergence.[1] The Kullback-Leibler (KL) divergence between two probability distributions P and Q is defined as follows:

$$D_{KL}(P||Q) = \sum_{i=1}^{n} P(i) log_2 \frac{P(i)}{Q(i)} \quad (4)$$

where n is the number of points in the distribution that we consider. Because KL divergence is asymmetric we calculate $D_{KL}(P||Q)$ and $D_{KL}(Q||P)$ and then we take the average. The lower the KL divergence the closer the distributions.

Results

In Table 2 we can see the KL divergences that we get when we compare our model and the two baselines with the human expert data for the proposer and responder policies of the 4 cultures. To avoid local optima or just being lucky with the random rewards, we ran both our model and the weak baseline (based on a random reward) multiple times and for each run we calculated the KL divergence. In Table 2 we report the median value of all computed KL divergences. In Figures 1 and 2 we can also see a graphical representation of our comparisons for the Japan proposer policy and the US responder policy. As we can see in all cases our IRL-based model outperforms both the weak and strong baselines. This verifies our hypothesis that decision-making is a complex process that cannot be attributed just to reacting to data or the sole factor of self-gain. It also shows the power of IRL for accurately modelling negotiation.

Table 2: KL divergences for IRL and the two baselines for all cultures and roles.

	Proposer			Responder		
	random	wealth	IRL	random	wealth	IRL
US	3.95	19.82	**2.84**	0.61	0.37	**0.10**
JP	4.01	4.86	**0.74**	0.64	0.25	**0.16**
IS	3.68	16.11	**1.29**	0.58	0.27	**0.13**
YU	9.28	3.49	**1.73**	0.57	0.26	**0.11**

The next question is whether our models are really capturing performance of people from the cultures that they were

[1]We also looked at Cartesian distance, but in all cases the best matching policy for the expert data was the same, so we report only KL-divergence, due to space restrictions.

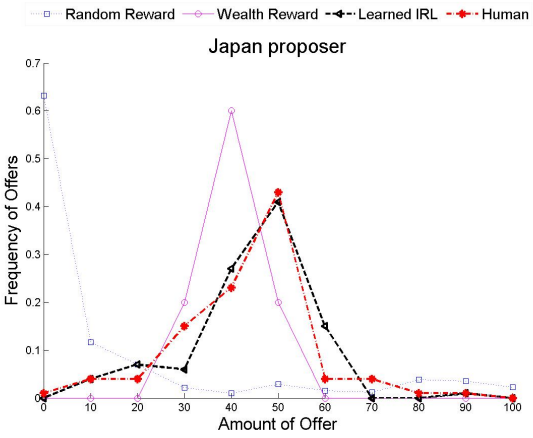

Figure 1: Comparison of random reward, wealth reward, IRL-based reward and human data for the Japan proposer policies tested with Japan SU-responders.

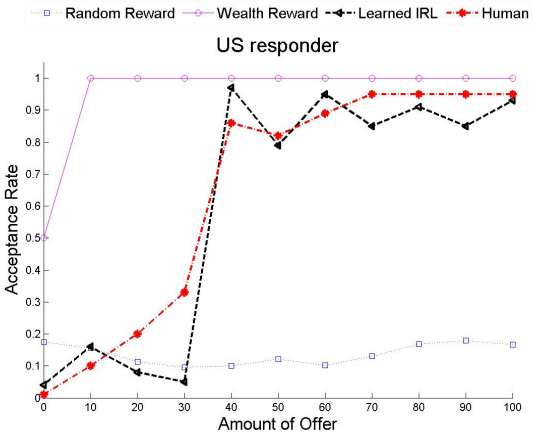

Figure 2: Comparison of random reward, wealth reward, IRL-based reward and human data for the US responder policies tested with US SU-proposers.

trained for. We examine this question in two ways. First we look at the KL divergences between all learned models and all original data sets. This is shown in Table 3. We can see that most of the time the model for each culture matches the data set from that culture better than other data sets. On the other hand there are several exceptions, for example, US proposers do better on Yugoslavia data than US data, and US responders perform well on all human data. We can also look at this table from a different perspective, as a way to compare various models (learned from data of different cultures) with the same human data. Here we can see that in most cases the data set is best modelled by the culture trained on it. However, there are a few exceptions, for example, Israel proposers are a better model of the Israel data than US and Yugoslav proposers, but a worse model of the Israel data than Japanese proposers. US proposers are not a very good model of the US data. US responders are the best model for US data, Japanese responders are a good model of the Japan data (equally good to US and Israel responders), Israel and Yugoslav responders are a good model of the Israel and Yugoslavia data respectively, but not as good as US responders. These results are encouraging and show that our models do not just beat the weaker baselines of wealth and random rewards, but also in most cases learn to model a culture better than models learned with different cultures. As we saw there are a few cases in which the results were not optimal. We believe that there could possibly be some convergence issues, even though our IRL algorithms ran for over 1000 iterations and our KL divergences are based on many runs, or perhaps, we need a larger set of features and constraints between features. Given that we take into account in our state accumulated wealth as well as rounds, our state space is fairly large. These are issues for further investigation.

Table 3: Cross-culture results, comparison with human data from different cultures (KL divergences). Best values are in bold (horizontally) and italics (vertically).

	Proposer				Responder			
	Human Data				Human Data			
	US	JP	IS	YU	US	JP	IS	YU
US	2.84	3.11	4.61	**2.71**	*0.10*	*0.13*	*0.08*	***0.06***
JP	*1.05*	***0.74***	*1.06*	1.96	0.27	**0.16**	0.18	0.24
IS	1.82	2.04	**1.29**	4.27	0.25	0.14	**0.13**	0.20
YU	2.21	2.83	5.76	***1.73***	0.15	0.27	0.20	**0.11**

In Table 4 we can see the results of experiment 2, where we use weights learned with one culture to learn policies by training on other cultures.[2] The results generally verify our hypothesis that the learned weights are culture-specific: with only two exceptions, the policy based on the reward function learned for that culture outperforms policies based on reward functions for all the other cultures. In the case of the US and Japanese responders, it appears that the policy trained with the Israel reward performs just as well as the policy using the learned reward function for US and Japanese responders, respectively. However the converse does not hold: the Japan and US reward functions do not work well for the Israel policies. These issues need to be investigated further.

Discussion

Our results show clearly that there are various factors that may affect one's decision and these factors may vary significantly depending on the culture of the decider. They also show the power of IRL for uncovering the decision-making mechanism of negotiators. (Turan, Dudik, Gordon, & Weingart, 2011) argue about the potential advantages of using IRL for learning the goals and motives of negotiation participants.

[2]We used only some of the many reward functions for each culture to learn policies for other culture data. In this table we show results for the reward functions that are closest to the median values reported in Table 2, but in some cases they are not identical.

Table 4: Cross-culture results, learning policies using rewards calculated from different cultures (KL divergences).

Policy/Role	Reward functions			
	US	JP	IS	YU
US Proposer	**2.84**	7.32	14.13	11.78
US Responder	**0.08**	6.77	**0.08**	19.10
JP Proposer	7.71	**1.58**	4.89	14.25
JP Responder	14.94	**0.11**	**0.11**	15.25
IS Proposer	3.92	5.93	**1.27**	19.52
IS Responder	7.62	6.95	**0.10**	13.08
YU Proposer	3.32	7.90	19.84	**1.73**
YU Responder	7.51	8.16	0.12	**0.06**

They use scenarios from group negotiation research and discuss how IRL could hypothetically be applied to such scenarios, but they have not actually used IRL for negotiation.

With IRL we calculated the weights for a number of features (see Table 1). These weights can be used in equation (1). However, in order to use equation (2) we need to find some kind of mapping between these weights and Hofstede's dimensions. Our ultimate future goal is once we have Hofstede's dimensions for a culture to be able to calculate these weights automatically. That would be very useful in cases where we do not have data to calculate the weights directly from but it is indeed a very ambitious goal. Other future work involves examining whether the reward function learned for one game or role can transfer to another. We also aim to experiment with larger numbers of RL and IRL iterations and runs, different exploration parameters, different state representations, and different features.

Conclusion

We used IRL to learn a model for cultural decision-making in negotiation. This model takes into account multiple individual and social factors for evaluating the available choices in a decision set. Our model assigns different weights to these factors based on the modelled culture. We applied this model to the Ultimatum Game and we showed that weights learned from IRL surpass both a weak baseline with random weights, and a strong baseline that only seeks to maximize the agent's own gain. Our model outperformed both baselines by generating behavior that was closer to the behavior of human players of the game in 4 different cultures. We also showed that the weights learned with our model for one culture outperform weights learned for other cultures when playing against opponents of the first culture.

Our results verify our hypothesis that decision-making in negotiation is a complex, culture-specific process that cannot be explained just by the notion of maximizing one's own utility. We showed that cultures vary in goals, not just in conventional circumstances but also that we can successfully use IRL techniques to learn culture-specific goals.

Acknowledgments

This work was funded by the NSF grant IIS-1117313 and a MURI award through ARO grant number W911NF-08-1-0301.

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning (ICML).*

Buchan, N. R., Croson, R. T. A., & Johnson, E. J. (1999). Understanding what's fair: Contrasting perceptions of fairness in ultimatum bargaining in Japan and the United States. In *Discussion paper, University of Wisconsin.*

Camerer, C. F. (2003). *Behavioral game theory - Experiments in strategic interaction*. Princeton University Press.

Gal, Y., Pfeffer, A., Marzo, F., & Grosz, B. J. (2004). Learning social preferences in games. In *Proceedings of the 19th National Conference on Artificial Intelligence* (p. 226-231).

Georgila, K., Henderson, J., & Lemon, O. (2006). User simulation for spoken dialogue systems: Learning and evaluation. In *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH-ICSLP).*

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367-388.

Henrich, J. (2000). Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *American Economic Review*, *90*, 973–979.

Hofstede, G. H. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. Thousand Oaks, CA: SAGE.

Lewis, D. K. (1969). *Convention: A philosophical study*. Harvard University Press.

Neumann, J. V., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

Nouri, E., & Traum, D. (2011). A cultural decision-making model for virtual agents playing negotiation games. In *Proceedings of the International Workshop on Culturally Motivated Virtual Characters.*

Rawls, J. (1971). *A theory of justice*. The Belknap Press of Harvard University Press.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *American Economic Review*, *81*(5), 1068-95.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Turan, N., Dudik, M., Gordon, G., & Weingart, L. R. (2011). Modeling group negotiation: Three computational approaches that can inform behavioral sciences. In *E. A. Mannix, M. A. Neale, and J. R. Overbeck, eds., Negotiation and Groups (Research on Managing Groups and Teams)* (Vol. 14, p. 189-205).