AI & Soc (2017) 32:51–63
DOI 10.1007/s00146-014-0570-7

ORIGINAL ARTICLE

Culture-specific models of negotiation for virtual characters: multi-attribute decision-making based on culture-specific values

Elnaz Nouri · Kallirroi Georgila · David Traum

Received: 16 January 2014 / Accepted: 6 October 2014 / Published online: 29 October 2014

Abstract We posit that observed differences in negotiation performance across cultures can be explained by participants trying to optimize across multiple values, where the relative importance of values differs across cultures. We look at two ways for specifying weights on values for different cultures: one in which the weights of the model are hand-crafted, based on intuition interpreting Hofstede dimensions for the cultures, and one in which the weights of the model are learned from data using inverse reinforcement learning (IRL). We apply this model to the Ultimatum Game and integrate it into a virtual human dialog system. We show that weights learned from IRL surpass both a weak baseline with random weights and a strong baseline considering only one factor for 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.

Keywords Cultural decision-making · Negotiation · Virtual agents · Ultimatum Game · Inverse reinforcement learning

Abbreviations

MARV	Multi-attribute relational value
RL	Reinforcement learning
IRL	Inverse reinforcement learning
MDP	Markov decision process
SU	Simulated user
KL divergence	Kullback–Leibler divergence

E. Nouri (✉) · K. Georgila · D. Traum
Institute for Creative Technologies, University of Southern California, 12015 Waterfront Drive, Playa Vista, CA 90094, USA
e-mail: nouri@ict.usc.edu
URL: http://nld.ict.usc.edu/group/

1 Introduction

Decision-making is an important part of social interaction. In the most general case, a decider must consider not only the impact on his own utility, but also the impact on others, including individuals, groups, and society as a whole. There are also differences in how individuals value the options in a decision space as well as broad similarities in the values of individuals from the same cultural group. Social scientists have often observed that people from different cultures behave differently in interactive situations (Camerer 2003; Roth et al. 1991). There are several 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 the UK 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. For example, typically, Western individuals are individualistic, whereas Eastern individuals are collectivistic and more willing to

sacrifice their short-term gain, while negotiating, in order to establish better relationships in the long-term (Brett and Gelfand 2006).

Most classical economic game theory accounts of decision-making, e.g., Neumann and 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 et al. (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 this paper, we present a multi-attribute model of decision-making that can be used by virtual agents to make a wide variety of decisions in a social setting, such as game play or negotiation. This 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 (see Sect. 3) and perform two studies, using different methods for setting the weights on different values.

In our first study, we use Hofstede's multi-dimensional model of culture (Hofstede 2001) in order to determine the relative weights of different factors. For each culture that we are interested in, we set the weights of the model manually based on Hofstede dimensions for this culture and our intuitions about how to relate the score in a dimension to the relative importance of each of the factors. This approach makes sense when there is no data available from which we could potentially learn the weights automatically. Our results show that setting the weights of the model manually does not always lead to realistic behaviors, which is not surprising given that hand-crafting these weights may involve a number of relatively arbitrary decisions.

To overcome the limitations of manually setting our model's weights, in our second study, we attempt to learn the weights of our model from data using inverse reinforcement learning (IRL) (Abbeel and Ng 2004). To our knowledge, this is the first use of IRL in the Ultimatum Game or more generally to learn patterns of behavior in negotiation. We perform two experiments to try to get at the above question 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: In Sect. 2, we describe our multi-attribute relational value (MARV) model for decision-making. In Sect. 3, we present a concrete example, the Ultimatum Game, in which a fair amount of data has been gathered on decision-making, and to which we adapt our model. In Sect. 4, we present our first study, in particular, we describe how our model can be adapted to cover cultural differences using Hofstede's multi-dimensional model of culture (Hofstede 2001) and how it has been integrated into a virtual human dialog system, allowing the virtual agents to play the Ultimatum Game with each other and with humans, using natural language dialog interaction. We also evaluate this model by comparing behavior generated by our model to observations of humans from different cultures. Section 5 presents our approach to automatically learning the weights of our model from data using IRL. In this second study, we perform two experiments. The goal of the first experiment is to show that the reasoning behind the actions of an agent is better modeled as a complex trade-off of multiple goals, and cannot be explained merely by learning the behavior patterns of the negotiation partners. 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 from other cultures. Section 6 concludes with a discussion of our findings and ideas for future work.

2 The multi-attribute relational value (MARV) model

The MARV model takes into account 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). It is thus a multi-attribute model. Most of the metrics are relational, in that they are a function of multiple, more primitive metrics. The metrics we have considered so far include:

1. Self Utility (the agent's own utility);
2. Other Utility (the utility of another);
3. Total Utility (sum of individual utilities of all participants);
4. Average Utility (may not be derivable from Total Utility when the number of participants is variable);
5. Relative Utilities (viewed in several ways, such as self/total, self-other, self/other, self/average);
6. Minimum Utility (lower bound of any participant—maximizing this is the aim of Rawls' theory of justice (Rawls 1971));
7. Uncertainty (variation among possible outcomes);
8. For each of the above, minimum outcome versions, denoting the worst case rather than expected utility (in cases where payoffs are probabilistic rather than guaranteed for a particular choice).

Each of these metrics can be given one or more valuations, choosing an optimum point and scale. For example, the optimum for self/average might be infinity, when considering relative self-interest, 1 when considering fairness, or below 1 if trying to satisfy a vow of poverty. Each individual agent has a vector of weights, one per valuation, indicating the relative importance of that valuation. If the weight is zero, then the valuation is not considered in the overall utility computation for that individual. The total value for each choice is the sum of the product of values and weights for each valuation as shown in Eq. (1). For every decision, our agent calculates the utility of all its possible choices and selects the one that has the highest overall valuation (according to the agent's knowledge and ability to calculate or estimate these values).

$$\text{Value}\,(\text{Choice}_i) = \sum_{j=1}^{n} \left(W_j \times V_j(\text{Choice}_i)\right) \quad (1)$$

An advantage of MARV 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. Our belief is that the MARV model can improve on unitary models of utility, by more accurately simulating the kinds of decisions that people make in different situations. MARV considers multiple factors, which can also account for systematic differences in decisions made by different individuals (who have a different weight vector), as well as subtle changes in decision-making, based on changes in the context of the decision, where changes affect only some dimensions but not others.

3 Culture and the Ultimatum Game

We use the Ultimatum Game as an initial test bed 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. According to economic game theory, a rational responder interested only in maximizing own gain will accept any non-zero offer, and thus, at perfect equilibrium, the proposer makes very low offers and keeps almost all of the money. This classic experimental economics game has received a great deal of attention since the initial experiment by Güth et al. (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 et al. 1999). While some have seen no differences in the amounts of offers or rejection rates between subjects, e.g., Okada and Riedl (1999) and Oosterbeek et al. (2004), others have reported that people from different cultures behave differently in this game. For example, Roth et al. (1991) studied the Ultimatum Game in four countries: USA, Japan, Israel, and former Yugoslavia. They found that the offers in USA and Yugoslavia were higher than the offers in Japan which were higher than the offers in Israel. Buchan et al. (1999) studied the differences in perception of what is a fair offer between USA and Japan among comparable student populations in Pennsylvania and Tokyo and demonstrated that national culture interacts with power to yield differing beliefs about what is fair among buyers and sellers in Japan versus those in the USA. In the USA, participants believe that it is fair that the party with greater power takes a larger share of the surplus. In Japan, participants believe that it is fair that the party with greater power earns a smaller portion of the surplus, sharing more of it with the weaker partner. Henrich (2000) compared the behavior of 18- to 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.

The findings presented above clearly show that culture can play an important role in negotiation and in particular, in the Ultimatum Game. The question, however, is what role it can play: different goals or different conventions? Another question is whether we can use the MARV model to emulate culture-specific behavior. In the next two sections, we report on two studies, each of which includes two experiments that use the MARV model to represent values and decision procedures for different cultures that lead to behaviors for these cultures, which approximate the observed data.

4 First study: cultural value norms based on Hofstede's multi-dimensional model of culture

The MARV model in Eq. (1) can capture cultural differences among populations (e.g., the relative value of utility

given to self, a group, or the whole population) by setting the weights for different valuations to be congruent with the norms of specific cultures for those valuations. We use Hofstede's model of culture (Hofstede 2001) as our basis for cultural modeling of decision-making because it has the following advantages:

- Explicit dimensions of cultural norms that can be tied to valuation;
- Multiple ways in which cultures can be similar or differ;
- Data on dimension values for a large range of (national) cultures.

Hofstede's model has five dimensions: individuality (IDV), power distance index (PDI), long-term orientation (LTO), masculinity (MAS), uncertainty avoidance (UAI). In theory, each of these dimensions could contribute to the relative weight of any of the valuations. Thus, our generalized MARV model, shown in Eq. (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.

$$\text{Value}(\text{Choice}_i) = \sum_{j=1}^{n} \left(\left(\prod_{d=\text{IDV}}^{\text{UAI}} W_{j,d} \right) \times V_j(\text{Choice}_i) \right) \tag{2}$$

In practice, however, not every dimension value will impact every valuation, so a number of these weights will be 1, and not modify the resulting valuation. The issue then becomes how to assign the weights for individuals from particular cultures, in specific circumstances. Individuals could vary quite a bit from the cultural norms; however, we would expect a weighted average of individuals from the culture to roughly match the norms for that culture. If we have no other information about the preferences of an individual agent, we can use the cultural norms of the society that the agent belongs to as an approximation of his cultural profile. If we have information about the personality, power position, and the status of the membership of the agent to different groups, we can also take into account the variations it imposes on the weights of the valuation functions. The variation among agents also introduces uncertainty in prediction of how another agent may choose, even when following the above deterministic decision process.

In our first study, the weights of Eq. (2) are chosen by trying to match intuitions about the meanings of Hofstede dimensions to the values that high and low points in the dimension would place on each of the metrics, under different circumstances. For the Ultimatum Game, given the simplicity of the task and symmetries coming from splitting a fixed amount with one other player, we use only the following four values: Self Utility, Other Utility, Self/Other Utility, and Minimum Utility. Decisions are made by selecting the choice that will maximize the value for Eq. (2).

This approach of using Hofstede dimensions for virtual human decision-making is inspired by the work of Mascarenhas et al. (2009). However, we find that the model of Mascarenhas et al. (2009) is limited in several aspects, notably that it only covers two dimensions, it can be applied to only very specialized types of decisions, and it seems hard to generalize the utility equations. Moreover, it has not been implemented within virtual humans that can interact with people (only with other virtual humans). Unlike the model of Mascarenhas et al. (2009), the MARV model can be used to make a variety of decisions in different contexts. As described in Sect. 4.2, it has been integrated into a virtual human dialog system that supports both interactions between virtual agents, and interactions between virtual agents and humans using natural language.

4.1 Hofstede dimensions and assumed values

We briefly discuss each of Hofstede dimensions and how we assign weights for different valuations, based on these scores. We give initial weights for all possible cases of interaction between agent A from one country and agent B from another country (or the same country). For each country, its corresponding score in each dimension is tagged as either "high" or "low", e.g., high IDV, low PDI, etc. We also take into account whether the agents belong to the same group (in-group) or different groups (out-group), and their difference in power position. The appropriate preset weights are used for the calculation of the utility. For a given dimension d, weights are shown as a vector for that dimension: ($W_{\text{self},d}$, $W_{\text{other},d}$, $W_{\text{self/other},d}$, $W_{\text{lowerbound},d}$). These weights are set according to our interpretation of the findings in the literature regarding that dimension. The weight values are selected from the set of {1, 2, 4}, and the magnitude of the weight shows the degree of importance of the metric (e.g., self utility) for that dimension (e.g., high IDV). Details are provided in the following sections.

4.1.1 The individualism–collectivism dimension (IDV)

The assumption is that in collectivistic societies where people have low individualism scores, people tend to care more about other individuals and give higher weight to group rather than self utility. The distinction between in-group and out-group is essential in collectivistic cultures (Hofstede 2001), with value being placed on utility of others only within the group. In our MARV model, we define group IDs for individuals, and each individual considers his/her in-group or out-group relationships when

Table 1 Weights for individualism and collectivism

Low IDV in-group	Low IDV out-group	High IDV
(2, 2, 1, 4)	(1, 1, 1, 1)	(2, 1, 1, 1)

Table 2 Weights for power distance

PDI score	Low PDI	High PDI
High to low power position	(2, 1, 2, 2)	(2, 1, 2, 1)
Low to high power position	(1, 2, 1, 4)	(2, 4, 1, 1)

he/she wants to calculate the utility function of the decision. High individualism cultures put high weight on self utility, while low individualism puts less weight on self utility. Low individualism cultures put high weight on other utility for group members and high weight on total utility, when everyone considered is in the group. Table 1 below shows our initial weights for high individualism and for collectivism, considering both in-group and out-group partners.

4.1.2 Power distance index (PDI)

Power distance is the tendency to accept that more powerful individuals should have more resources. In a low power distance culture, more weight is put on fairness (minimum utility and average utility), regardless of the status of the participants. For high power distance cultures, the value of self or other is proportional to the power or status of the individual, and the ideal value for relative utility is in accordance with the relative power.

In the context of the Ultimatum Game, the assumption is that in societies with low PDI, both parties would expect a more even split, while in high power distance societies, it would be natural to allocate more to the more powerful party. Our initial weights for this dimension are shown in Table 2, considering the power relationship as well as the cultural dimension value.

4.1.3 Masculinity dimension (MAS)

The MAS index in general refers to differentiation among gender roles. However, high MAS is also correlated with higher assertiveness and competitiveness, especially for male members of the culture. For low MAS cultures, the values are more similar across gender roles and tend more toward caring and general well-being. In terms of our valuations, high MAS cultures have higher weights for relative utility (self/average) and self utility, while low MAS cultures have higher weights for other utility, average utility, and minimum utility. Our model of the MAS dimension contrasts fairness and cares for everyone versus competitive self-interest. It is summarized in Table 3.

Table 3 Weights for masculinity

Low MAS	High MAS
(2, 2, 1, 4)	(4, 1, 4, 1)

4.1.4 Uncertainty avoidance index (UAI)

The uncertainty avoidance indices in societies show how much people do not want to tolerate uncertainty and ambiguity. This dimension of the culture brings up another example of violation of classic game theory measurement of utility. Allais (1953) showed that many people would prefer a risk-free decision over a decision with higher average payoff but uncertainty involved. We capture this aspect by looking not at expected utility of each valuation, but at a more qualitative representation of the decision space, looking at the range of different possible outcomes that are not under the control of the player. In the simplest case, this degenerates to looking at the worst-case options.

The main source of uncertainty in the Ultimatum Game is only for the proposer, considering how likely the responder is to accept. Proposers with a high uncertainty avoidance score try to come up with decisions that would minimize the chance of rejection. The probability of acceptance is calculated by looking at the likelihood of a random responder accepting, considering the decision that each cultural model would make and the frequency of each type of culture assumed to be in the population of players. For low uncertainty avoidance, expected utility is used. For high uncertainty avoidance, we assume a deterministic decision based on how the majority would decide.

4.1.5 Long-term orientation (LTO)

Long-term orientation can be viewed as different ways of assigning utility, depending on the timescale of the utility and discount rate for future payoff. It can also be viewed as the tendency to view a set of decisions together as a policy toward an ultimate goal, rather than looking at each decision in its own right. Since the Ultimatum Game used for our experiments is a short-term game (not more than 10 rounds), there is no effect of modeling LTO, and we omit this from our models, for simplicity.

4.2 Integration with virtual humans

We use the MARV model developed above to control virtual human decision-making in playing the Ultimatum

Fig. 1 Virtual human Ultimatum Game players: Utah (*left*) and Harmony (*right*)

Game. We integrated an implementation of the decision-making protocol described above within the tactical questioning architecture (Gandhe et al. 2008), which facilitates rapid development of virtual human dialog systems. The authoring environment of this architecture was used to construct domain knowledge and textual realizations for natural language understanding and generation of a range of speech acts.

Two agents were also developed, each of which can play the roles of the proposer and the responder in the context of the Ultimatum Game. For convenience, we reused two available bodies for these agents from the Gunslinger environment (Hartholt et al. 2009) including a nineteenth-century Western US saloon, the background art, and the bartender (Utah) and bargirl (Harmony) (see Fig. 1). The dialog model was extended so that the dialog manager would consult the decision-making module before deciding on which offer to give (for proposer) or whether or not to accept an offer (for responder). Other dialog moves (e.g., greetings, closings, explanations) were handled by the existing dialog networks in (Gandhe et al. 2008). For each character, we can use any of the possible cultural models and instantiate relationship variables, such as whether the players are from the same group, the relative power status, and the population set considered for calculating probability of refusals.

4.3 Evaluation

In this study, we ran two experiments to test our virtual human culture models. In the first experiment, we evaluated the culture models by testing virtual humans playing against humans, to test whether we can embed different culture models within virtual human dialog and use MARV to play the Ultimatum Game. In the second experiment, we played virtual humans against each other to evaluate the culture models and compare the behavior of agents following those models with observed human player data.

4.3.1 Experiment 1: virtual humans versus humans

We first evaluated the virtual human dialog integration by having the virtual human play against human players. An example is shown in Fig. 2.

Our virtual humans were tested against humans playing both proposer and responder roles. The goal was to demonstrate that the system can play the game successfully, including recognizing a range of human utterances as needed to participate. An example dialog is shown in Table 4. Three different users acted as both proposer and responder against a range of cultural models. The average dialog was composed of three iterations of the game. On average, three offers were made by the user proposer before they were accepted by the virtual human responder, and vice versa. The maximum and minimum offers made to the virtual human were $100 and $0, respectively, while the virtual human minimum offer was $10. Our initial results are successful in that humans can play the game by engaging in unrestricted (text) natural language for the domain. In future work, we plan a more controlled study to quantify performance levels.

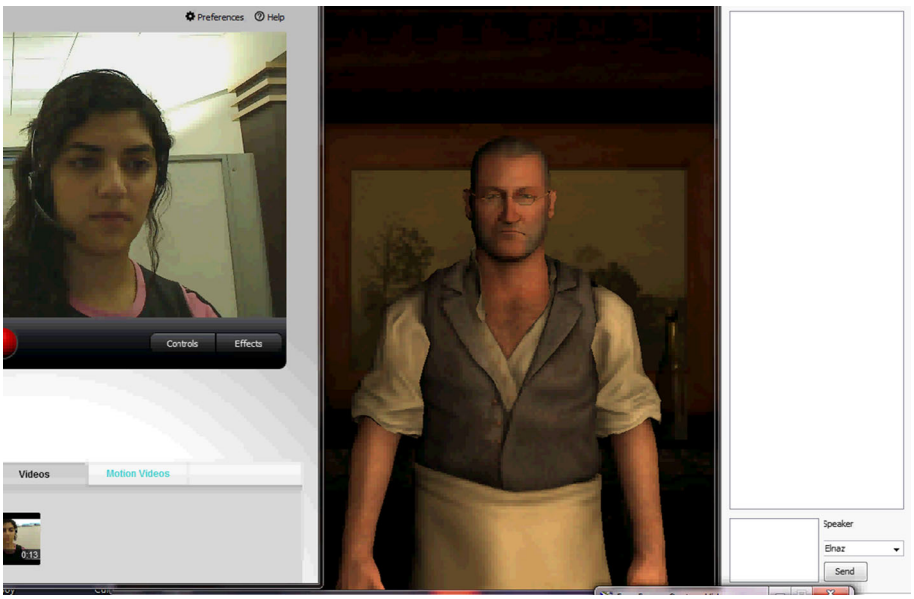

Fig. 2 Human playing the Ultimatum Game with a virtual human

4.3.2 *Experiment 2: evaluation of culture models*

We evaluated the culture models by testing the virtual humans playing against other virtual humans. We included all possible configurations (low and high) of the cultural dimensions under four different conditions considering in-group and out-group status crossed with lower and higher power. The goal was to compare cultural models with data from human players from those cultures.

In order to evaluate the result of the experiments done so far and the performance of the model, we referred to the extensive data available from the literature. Camerer (2003) provides a detailed history and data of the different ultimatum bargaining game experiments. Henrich et al. (2005) reveal more variability across cultures and along with Oosterbeek et al. (2004) provide the baseline data for our model. The average percentage of total value offered in all the experiments done so far is between 26 % (for Machiguenga in Peru) and 58 % (in Lamelara in Indonesia). Our model's virtual human proposed values also fall into this range of data. According to Camerer (2003), the rejection rate varies from 0 % (in Tsimané in Bolivia) to 67 % (in Mapuche in Chile).

A summary of our results is shown in Table 5, along with data that have been reported for human players of these cultures. We show the Hofstede values reported for the culture, as well as statistics reported for that culture's play in the Ultimatum Game, under the "Human" columns. In the virtual human (VH) columns, we show statistics for our model based on the Hofstede scores.

These initial results are encouraging. With the exceptions of Israel and Japan, the virtual human mean offers are quite close to the human means, and values are tending in the right directions (higher for US and lower for Spain). Similarly, for rejection rates, the rough orderings are similar for virtual humans and observed humans. Spain has the highest rejection rate for both humans and virtual humans. If we exclude Japan, the next group of countries with high rejection rates for both humans and virtual humans is "Israel, Sweden, USA, and Austria" followed by the group "Chile, Equador, and Germany" with lower rejection rates.

5 Second study: learning the weights from data

In our first study, the weights for different valuations were set based on intuition of the meanings of the Hofstede dimensions. Unfortunately, in some cases, it is difficult to decide on what the exact values should be. In the second study, we try to learn values that would lead to observed behavior, using a technique called IRL.

5.1 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 and 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) that 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 is 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 $\alpha_i \in A$ 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 $P(s_{i+1} \mid s_i, \alpha_i) \in P$. 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 and Barto (1998). There are several algorithms for estimating the Q-function, and we use Q-learning (Sutton and 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 dialog 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 et al. 2006), which will behave as part of the environment and will interact with the policy 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.

Table 4 Example interaction between a virtual human and a human, showing three game iterations

Player	Utterances
Virtual human proposer	Hello. We are going to play the Ultimatum Game. I am going to make offers for splitting $100, and you will accept or reject. If you accept, you'll get the offered amount and I'll get the rest. If you reject, both of us will get nothing
Human responder	Okay
Virtual human proposer	Would you accept an offer of $10?
Human responder	I reject your offer of $10
Virtual human proposer	How about $20?
Human responder	No
Virtual human proposer	I offer you $30
Human responder	I accept
Virtual human proposer	Thanks for playing with me. Good-bye
Human responder	Good-bye

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. IRL (Abbeel and 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. These data are called *expert data*. The reward function R can be expressed as follows:

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

where s is the state that the agent is in and α the action that it performs in this state, and $\hat{w}^T$ 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 and 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 are compared with the expert data, and the weights w_i

Table 5 Comparison of virtual human and human offers and rejection rates for several cultures

Culture	Hofstede: PDI, IDV, MAS, UAI	VH mean offer ($)	Human mean offer ($)	VH rejection rate (%)	Human rejection rate (%)
Austria	11, 55, 79, 70	33.13	39.21	9.1	16.1
Chile	63, 23, 28, 86	33.13	34.00	1.0	6.7
Ecuador	78, 8, 63, 67	33.13	34.50	1.0	7.5
Germany	31, 64, 61, 60	36.88	36.70	9.1	9.5
Israel	13, 54, 47, 81	21.66	41.71	25.0	17.7
Japan	54, 46, 95, 92	32.50	44.73	1.0	19.3
Spain	57, 51, 42, 86	28.75	26.66	25.0	29.2
Sweden	34, 70, 4, 26	33.13	35.33	10.2	18.2
US	40, 91, 62, 46	41.88	42.25	12.0	17.2

are calculated in an attempt to reduce their distance. RL is performed, using the updated reward function from Eq. (3) with weights w_i, to learn a new policy π_{i+1} that generates a new set of data by interacting with the SU. Then this new data are 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.

5.2 Experimental setup

We use data of the distribution of offers and acceptances or rejections for four different cultures (USA, Japan, Israel, and former Yugoslavia) reported in Roth et al. (1991). We use these four cultures as opposed to all cultures in Table 5 because these are the only cultures for which we have detailed data (i.e., distributions of offers and acceptance/rejection rates) that we can use to create SUs and compare our results with. For each culture, we generate SU-proposers and SU-responders by using probability functions that match the reported data. Roth et al. (1991) provide these data for the first and last round of the game. In our setup, the game lasts five 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. Next, we use RL, with a reward function based on these weights, 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 several features, including 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.

Table 6 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) $= 50$

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 feature ϕ_i is true in state s and action α has been performed. So to form the feature functions $\phi_i(s, \alpha)$, each feature is paired with all the available actions. Table 6 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 6 for conciseness. In fact, every possible value of AccSelf or AccOther, e.g., AccSelf = 150, AccOther = 200, etc., can form a feature together with an action. Thus, for the proposer, the feature function AccSelf = 150–offer = 20 is going 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, 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 in Sect. 6), whereas we would have to collect human data to create a SU for a new problem.

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

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

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

5.3 Experiment 1: IRL versus baseline reward functions

We perform two experiments as part of our second study. The goal of the first experiment is to show that the reasoning behind the actions of the proposer is better modeled as a complex trade-off 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 four cultures, we learn three 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 different 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, this is evidence that multiple factors are taken into consideration. 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; we report median values for all runs with the reward function.

5.3.1 Results

Figures 3 and 4 show examples of graphical representation of the human data compared to behavior generated by the three different reward functions (for Japanese proposer, and US responder, respectively).

Table 7 shows KL divergences for all reward functions and all cultures. 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 modeling negotiation. As we discussed in the introduction, IRL does not make any assumptions about the underlying utility function of the negotiators but instead relies only on the data to uncover the decision-making mechanism of the negotiators.

The next question is whether our models are really capturing performance of people from the cultures that they were trained for. We examine this question in two ways: first, by reanalyzing the data from experiment 1, and then, through experiment 2 (see below). First, we look at the KL divergences between all learned models and all original data sets (e.g., comparing human US data not just to data from the IRL policy for US, but also to the policies learned for other cultures). This is shown in Table 8. For example, the first row shows the KL divergences between the IRL-based reward data for USA and the human data for USA, Japan, Israel, and Yugoslavia for both roles (proposer and responder).

We can see that most of the time the model learned 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 modeled by the culture trained on it. However, again, 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

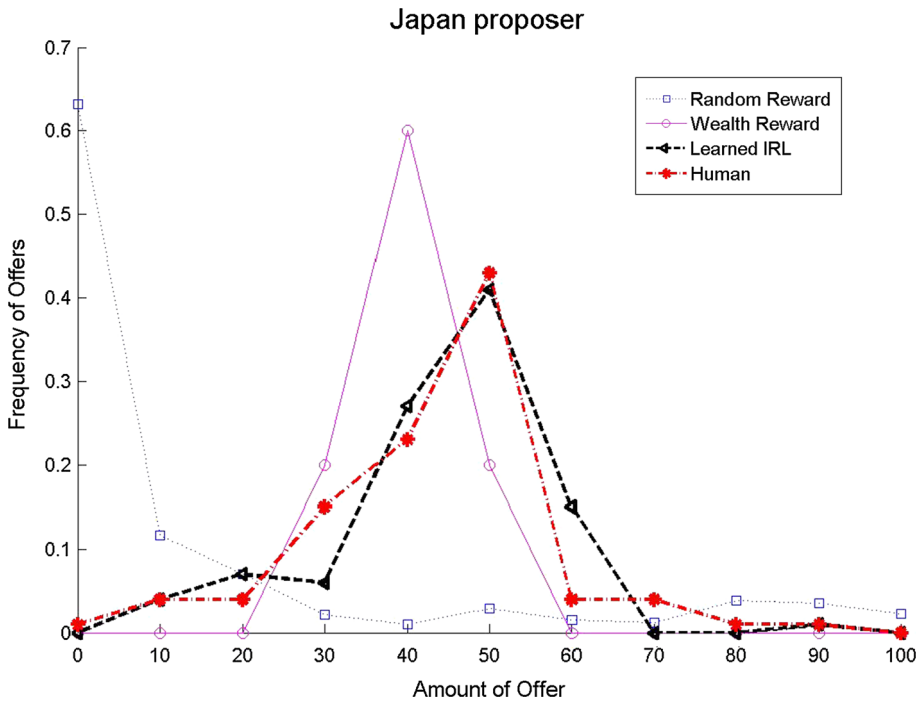

Fig. 3 Comparison of random reward, wealth reward, IRL-based reward, and human data for the Japan proposer policies tested with Japan SU-responders

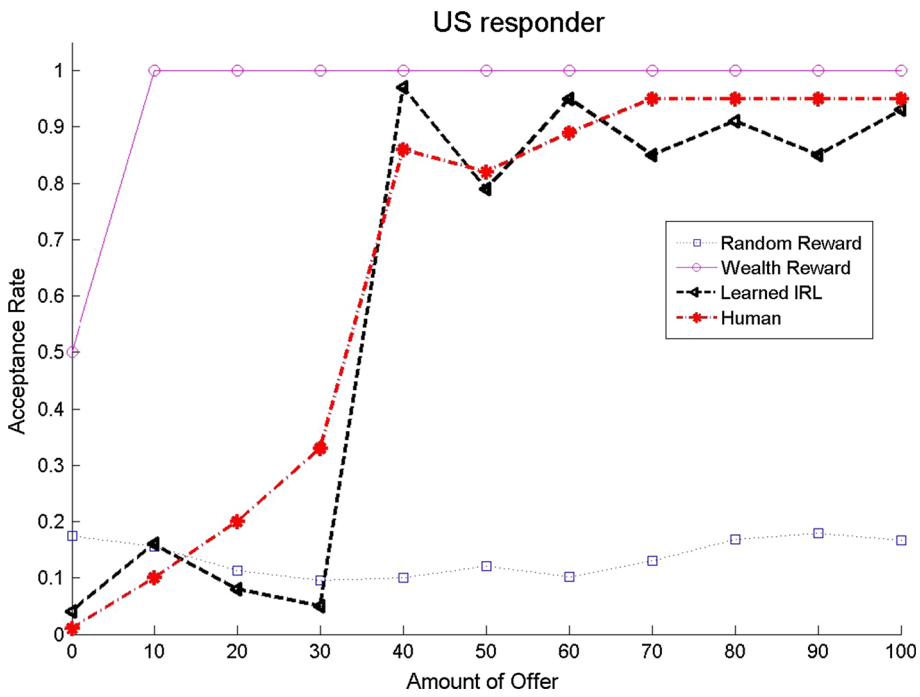

Fig. 4 Comparison of random reward, wealth reward, IRL-based reward, and human data for the US responder policies tested with US SU-proposers

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 from other cultures.

5.4 Experiment 2: reward functions from different cultures

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

Table 7 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 best values are in bold (horizontally)

Table 8 Cross-culture comparison of KL divergences between learned models (rows) and human data (columns)

	Proposer				Responder			
	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**

The best values are in bold (horizontally) and italics (vertically)

show this, we use IRL to learn a reward function for each of the four cultures and then we use these four reward functions to learn policies for each culture, yielding 16 total policies for each role. Then we test the four policies for each culture and role against SUs from the same culture that they were trained on. 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 (policy-rewardJapan-trainUS, policy-rewardYugoslavia-trainUS, and policy-rewardIsrael-trainUS).

5.4.1 Results

In Table 9 we can see the results of experiment 2. Each row shows a culture–role combination and the KL divergences for policies learned from each of the four reward functions.[1] 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.

[1] We used only some of the many reward functions learned in experiment 1 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 7, but in some cases they are not identical.

Table 9 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**

The best values are in bold (horizontally)

6 Discussion and conclusion

In this paper, we introduced the MARV model, a novel multi-attribute relational value model of decision-making in negotiation. 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 positing that members of different cultures place different weights on these factors. We applied this model to the Ultimatum Game and integrated it into a virtual human dialog system. We studied two cases: one in which the weights of the model were handcrafted and one in which the weights of the model were learned from data using IRL. The 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. The MARV model outperformed both baselines by generating behavior that was closer to the behavior of human players of the game in four different cultures. We also showed that the weights learned with the MARV 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. The second study also shows the power of IRL for uncovering the decision-making mechanism of negotiators. To our knowledge, no one has used IRL before in the Ultimatum Game or generally to learn patterns of behavior in negotiation. Turan et al. (2011) argue about the potential advantages of using IRL for learning the goals and motives of negotiation participants. 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. As Turan et al. (2011) point out, the standard practice in negotiation experiments, e.g., use scoring sheets to measure various factors that may contribute to the decision-making process of negotiation participants, has the drawback of not being able to generalize goals and motives that were not part of the experimental design. With IRL, there is no such problem because we do not make any assumptions but instead learn directly from the data, which in turn can lead to more accurate models of decision-making.

In future work, we would like to examine 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. Another idea for future work is that it may be the case that a different set of valuation functions must be considered for the Ultimatum Game—we may be missing some important elements, at least for some cultures. We also intend to refine the model of Hofstede dimensions, taking into account not just whether the culture scores low or high, but making the weight dependent on the actual values, to capture finer distinctions. Finally, it may be that the Hofstede dimensions themselves are not adequate for capturing the kind of cross-cultural variation that has been observed.

Acknowledgments This work was funded by the NSF Grant IIS-1117313 and a MURI award through ARO Grant No. W911NF-08-1-0301.

References

Abbeel P, Ng AY (2004) Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the 21st international conference on machine learning (ICML)

Allais M (1953) Le comportement de l'homme rationnel devant le risqué: critique des postulats et axioms de l'ecole americaine. Econometrica 21(4):503–546

Brett JM, Gelfand MJ (2006) A cultural analysis of the underlying assumptions of negotiation theory. In: Thompson L (ed) Frontiers of negotiation research. Psychology Press, Philadelphia, pp 173–201

Buchan NR, Croson RTA, Johnson EJ (1999) Understanding what's fair: contrasting perceptions of fairness in ultimatum bargaining in Japan and the United States. Discussion paper, University of Wisconsin

Camerer CF (2003) Behavioral game theory—experiments in strategic interaction. Princeton University Press, Princeton

Gal Y, Pfeffer A, Marzo F, Grosz BJ (2004) Learning social preferences in games. In: Proceedings of the 19th national conference on artificial intelligence

Gandhe S, DeVault D, Roque A, Martinovski B, Artstein R, Leuski A, Gerten J, Traum DR (2008) From domain specification to virtual humans: an integrated approach to authoring tactical questioning characters. In: Proceedings of the annual conference of the international speech communication association (INTERSPEECH)

Georgila K, Henderson J, Lemon O (2006) User simulation for spoken dialogue systems: learning and evaluation. In: Proceedings of the annual conference of the international speech communication association (INTERSPEECH)

Güth W, Schmittberger R, Schwarze B (1982) An experimental analysis of ultimatum bargaining. J Econ Behav Organ 3(4): 367–388

Hartholt A, Gratch J, Weiss L, Leuski A, Morency L-P, Liewer M, Thiebaux M, Marsella S, Doraiswamy P, Tsiartas A, LeMasters K, Fast E, Sadek R, Marshall A, Lee J, Pickens L (2009) At the virtual frontier: introducing Gunslinger, a multi-character, mixed-reality, story-driven experience. In: Proceedings of the 9th international conference on intelligent virtual agents (IVA)

Henrich J (2000) Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. Am Econ Rev 90:973–979

Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H, McElreath R, Alvard M, Barr A, Ensminger J, Henrich NS, Hill K, Gil-White F, Gurven M, Marlowe FW, Patton JQ, Tracer D (2005) Cross-cultural perspective: behavioral experiments in 15 small-scale societies. Behav Brain Sci 28(6):795–815

Hofstede GH (2001) Culture's consequences: comparing values, behaviors, institutions, and organizations across nations. SAGE, Thousand Oaks

Lewis DK (1969) Convention: a philosophical study. Harvard University Press, Cambridge

Mascarenhas S, Dias J, Afonso N, Enz S, Paiva A (2009) Using rituals to express cultural differences in synthetic characters. In: Proceedings of the 8th international conference on autonomous agents and multiagent systems (AAMAS)

Neumann JV, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton

Okada A, Riedl A (1999) When culture does not matter: experimental evidence from coalition ultimatum games in Austria and Japan. Discussion paper, University of Amsterdam

Oosterbeek H, Sloof R, van de Kuilen G (2004) Cultural differences in ultimatum game experiments: evidence from a meta-analysis. Exp Econ 7:171–188

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

Roth AE, Prasnikar V, Okuno-Fujiwara M, Zamir S (1991) Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. Am Econ Rev 81(5):1068–1095

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. MIT Press, Cambridge

Turan N, Dudik M, Gordon G, Weingart LR (2011) Modeling group negotiation: three computational approaches that can inform behavioral sciences. In: Mannix EA, Neale MA, Overbeck JR (eds) Negotiation and groups (research on managing groups and teams), vol 14. Emerald Group Publishing, London, pp 189–205