Modeling Human Decision Making in Cliff-Edge Environments

Ron Katz and Sarit Kraus
Department of Computer Science and The Gonda Brain Research Center
Bar-Ilan University, Ramat-Gan 52900, Israel, sarit@cs.biu.ac.il

Abstract

In this paper we propose a model for human learning and decision making in environments of repeated Cliff-Edge (CE) interactions. In CE environments, which include common daily interactions, such as sealed-bid auctions and the Ultimatum Game (UG), the probability of success decreases monotonically as the expected reward increases. Thus, CE environments are characterized by an underlying conflict between the strive to maximize profits and the fear of causing the entire deal to fall through. We focus on the behavior of people who repeatedly compete in one-shot CE interactions, with a different opponent in each interaction. Our model, which is based upon the *Deviated Virtual Reinforcement Learning* (DVRL) algorithm, integrates the Learning Direction Theory with the Reinforcement Learning algorithm. We also examined several other models, using an innovative methodology in which the decision dynamics of the models were compared with the empirical decision patterns of individuals during their interactions. An analysis of human behavior in auctions and in the UG reveals that our model fits the decision patterns of far more subjects than any other model.

Introduction

In this work we propose a model for human learning and decision making in environments of repeated Cliff-Edge (CE) interactions (Katz & Kraus 2006). CE environments, which include sealed-bid auctions, dynamic pricing and the ultimatum game (UG), are characterized by an underlying conflict for the competitor between the desire to maximize profits and the fear of causing the entire deal to fall through. Consider, for example, a proposer in the UG who needs to decide how to divide an amount of money with his opponent (Guth, Schmittberger, & Schwarz 1982): Decreasing the share offered to the opponent increases the profits accruing to the proposer, so long as the offer exceeds the acceptance threshold of the opponent. A slightly greedier proposal causes the proposer to lose the whole deal. Similarly, a bidder in a sealed-bid first-price auction (e.g. (Ockenfels & Selten 2005)) attempts to bid an amount that is only slightly higher than those put forward by opponent players. This situation is somewhat similar to that of a person standing on the edge of a cliff, trying to see the panoramic view. The more closely he approaches the cliff edge, the better the view. However, one step too many causes the viewer to fall off the cliff. Hence, interactions and games of this type are referred to as Cliff-Edge interactions.

Here we focus on one-shot CE interactions which are repeatedly competed against different opponents. Such repeated interactions occur, for example, in periodical sealed-bid auctions of the same goods, which are very popular nowadays, especially via the internet (e.g. (Zhu & Wurman 2002)). Similarly, a sequential version of the UG with different opponents is under thorough investigation in behavioral economics (Roth & Erev 1995; Brenner & Vriend ; Bourgine & Leloup 2000). Thus, understanding human behavior in such environments is important for the everyday commercial world. In addition, the current work uses concepts from Artificial Intelligence (AI) in order to explain cognitive phenomena. In so doing, we intensify the relationship between AI and cognitive science, which is likely to be of great benefit to both disciplines (Freed 2000).

Several approaches to modeling human behavior in CE environments have been previously proposed (Selten & Stoecker 1986; Roth & Erev 1995; Bourgine & Leloup 2000). Those approaches are broadly described below. In this paper, we propose a new model which is based upon the Deviated Virtual Reinforcement Learning (DVRL) algorithm (Katz & Kraus 2006), which was originally established as a mechanism for developing automated agents to compete against human opponents. We show that our approach, which integrates learning direction theory (Selten & Stoecker 1986) with the Reinforcement Learning (RL) algorithm (Roth & Erev 1995), fits the empirical decisions pattern of a considerable proportion of subjects who participated in UG and auction experiments. By examining different environments, we reduce the probability of domain-specificity influencing the findings.

The methodology by which we examine the fitness of the different models to the empirical data considers individual patterns of decision making, rather than merely the average pattern for all the subjects, as has been done elsewhere (Roth & Erev 1995; Bourgine & Leloup 2000; Brenner & Vriend)[1]. Moreover, it considers each sub-

[1]We also examined our model against the average decisions patterns of UG players from four countries, as was done by (Roth &

ject's decisions pattern during all his interactions, rather than merely the relationship between two subsequent interactions, as has been done by others (Selten & Stoecker 1986; Ockenfels & Selten 2005; Grosskopf 2003; Mitzkewitz & Nagel 1993). In contrast to all previous works, we assume that different people may have different attitudes. Therefore, we categorize subjects according to the model which best explains their decision patterns.

In the next section, we formally describe the CE environments. Then we survey previous models and the behavior patterns that they predict. In the subsequent section we detail our DVRL-based model, and afterwards we present the empirical experiment and its results. In the last section we conclude and outline directions for future work.

The CE environments' description

The general pattern of one-shot CE interactions considers a competitor required to choose an offer *i*, being an integer $0 \leq i \leq N$, where *N* is the maximum optional choice. Then a positive reward, *r*, corresponding to the offer, *i*, is determined, depending on whether the offer passed a certain threshold, τ, set by the opponent. Specifically, in the sealed-bid first-price auction, an amount, *N*, is auctioned.[2] The bidder is required to place a bid *i*, being an integer $0 \leq i \leq N$, which will gain the bidder a reward, *r*, if it exceeds the highest bid, τ, made by all other bidders in the current auction. If $\tau \leq i, r = N - i$ (the amount gained less the bid amount), otherwise r=0. The bidder is never informed as to the size of the bid offered by an opponent. In the UG, a proposer needs to divide an amount (*N*) with an opponent by offering the latter an integer amount, $i, 0 \leq i \leq N$. The reward to the proposer, *r*, is determined according to the opponent's acceptance threshold, τ. If $\tau \leq i, r = N - i$, otherwise *r*=0.

To demonstrate the challenge facing a competitor in CE environments, let *R(i)* be the reward corresponding to a successful offer *i* and let P(i) be the probability of the offer, i, succeeding (i.e. the probability that the offer will be higher than the other bids, in the case of an auction, or will be accepted by the responder, in the case of the UG).

Obviously, there is a trade-off between *R* and *P*: choosing an offer *i* which increases the expected reward, *R(i)*, decreases the probability of success, *P(i)*, and vice versa.

Related work

This section surveys previously suggested ways of modeling human behavior during repeated interactions with different human opponents in CE environments. For each model, we describe the expected pattern of decision making through the game. **Table 1** presents a representative decision pattern for each model. These patterns authentically describe the behavior of subjects who interacted repeatedly as proposers in the UG, as detailed in the Experimental Design section. Each table entry includes the offer given by a subject, and the results of that offer (S - for successful move, i.e. an acceptance of the offer by the current responder, and U - for unsuccessful move, i.e. a rejection. Similarly, in auctions we regard a winning bid as successful, and a losing bid as unsuccessful). The interaction with the first opponent is presented in the first row, with subsequent interactions following sequentially in each successive row.

LDT	RL	DVRL	Constant	Constant+DVRL
5 U	46 S	50 S	50 S	10 U
15 U	45 S	45 S	50 S	10 U
25 U	45 S	35 U	50 S	10 U
50 S	46 S	35 S	50 S	10 U
45 S	46 S	30 U	50 S	10 U
30 S	42 S	30 S	50 S	10 U
26 S	45 S	20 S	50 S	10 U
8 U	40 S	10 S	50 S	10 S
15 U	49 S	10 U	50 S	10 U
32 S	46 S	10 U	50 S	10 U
29 S	46 S	10 U	50 S	10 U
26 S	46 S	20 S	50 S	10 U
19 U	40 S	20 S	50 S	10 U
24 U	45 S	20 U	50 S	10 U
31 S	46 S	20 U	50 S	15 U
29 S	32 S	20 S	50 S	15 U
28 S	28 S	20 S	50 S	15 U
27 S	46 S	20 U	50 S	20 U
26 S	54 S	20 U	50 S	20 U
25 S	46 S	20 S	50 S	20 S
24 S	46 S	2 U	50 S	20 U
20 S	40 S	20 S	50 S	40 S
16 U	15 U	20 U	50 S	40 S
19 U	46 S	20 U	50 S	40 S
22 S	43 S	30 S	50 S	40 S
20 U	50 S	30 U	50 S	40 S
22 U	42 S	30 S	50 S	40 S
25 U	40 S	30 U	50 S	40 S
29 S	46 S	30 S	50 S	40 S
28 U	49 S	30 S	50 S	40 S
29 U	46 U	20 U	50 S	40 U
35 U	45 S	30 U	50 S	40 S

Table 1: Representative decision patterns found for each model studied. These patterns describe the empirically observed behavior of proposers in the UG during our experiments. Proposers were given 100 NIS and could choose to offer their opponent some portion of that amount, being any integer between 0 and 100. For space considerations, we present here only the first 32 interactions.

It is noteworthy that all the models considered in this paper describe the dynamics of the *medium run* (see (Gale, Binmore, & Samuelson 1995)), when subjects begin to learn and adapt their behavior to their opponents' feedback. The models do not predict a subject's decision during the first interaction. We assume that this decision is made according to individual norms that are triggered by the framing of the specific environment.

Learning Direction theory (LDT)

LDT (Selten & Stoecker 1986) is a qualitative theory about learning in repetitive decision tasks. The theory is quite simple and can best be introduced with the help of an example. Consider an archer who tries to hit a target. If the arrow misses the target on the left side, then the archer will tend to aim more to the right, and in the case of a miss to the right, the aim will be more to the left. The way in which the decision is based on experience may be described as ex-post rationality. One looks at what might have been a better choice to have made last time and adjusts the decision in this direction. Thus, in the UG, if an offer, *i*, is rejected at

Erev 1995). We found the behavior of our model to accurately reflect human behavior, and in several cases our model was even more predictive than that of Roth and Erev.

[2] In order to evade considerations of value estimations (Ockenfels & Selten 2005) the item auctioned is an amount of money.

interaction *t*, then, at interaction *t+1* the proposer will offer the opponent a higher offer, while if offer *i* is accepted at interaction *t*, then at interaction *t+1* the offer will be decreased. This pattern of decision making, which can be seen in the first column of **Table 1**, was claimed to be used by many subjects in CE environments such as in the Iterated Prisoner's Dilemma game (Selten & Stoecker 1986), at a first-price auction (Ockenfels & Selten 2005) and in the UG (Mitzkewitz & Nagel 1993).

Despite the elegance and simplicity of the LDT model, it suffers from two problems. First, it cannot explain the very common occurrence of no change in the amount offered, after both a successful and an unsuccessful interaction. (Ockenfels & Selten 2005) report that 35.5% of bids remain unchanged after a successful auction and 30.4% remain unchanged after a loss. A similar pattern was found in the current research. In the auction, 49.81% of the bids were not changed after a successful auction, the figure being 27.97% after a loss. In the UG, 43.41% of offers were not changed after an accepted offer, and 18.33% after a rejection. These findings are difficult to explain by the LDT model, nor can that model predict *when* the offers will stay unchanged. Thus, in this paper we consider as LDT-based negotiators only subjects who (almost) always *positively* obey to the LDT rules, as shown in the LDT column of **Table 1**. The second problem is that the LDT process ignores all the data experienced before the previous interaction. In contrast to the archer's case, where the target stays steady during all the trials, in our environments there is a different opponent with a different bid or acceptance threshold in each interaction. Thus, a reasonable negotiator should take into account all the experience gained from previous interactions, rather than focusing only on the result of the last interaction. Even if we assume that a typical negotiator cannot remember all previous interactions, he certainly remembers more than merely one interaction. Hence, it is difficult to accept the argument that most people obey the LDT.

Reinforcement-Learning based models

Another approach suggests that the Reinforcement-Learning (RL) method (Sutton & Barto 1998) describes the mechanism underlying human behavior when a person interacts with different opponents in CE environments such as the UG (Roth & Erev 1995; Gale, Binmore, & Samuelson 1995; Bourgine & Leloup 2000).[3] According to the RL method, the negotiator determines what offer to make from amongst all potential offers, according to her evaluation of the expected utility (EU) that each offer would yield if chosen. Expected utilities are evaluated on the basis of the results of previous interactions, and are stored in a database, termed hereafter the Q-vector. The Q-vector is updated after each interaction, according to the results of that interaction. The EU of the chosen offer is reinforced after a successful interaction, while, after an unsuccessful interaction, the EU of the chosen offer is decreased or at least remains steady (Roth & Erev 1995). The basic RL method, however, does not seem to be an appropriate human learning model for CE environments, since it does not take into consideration the inter-correlations between offers, i.e. the fact that adjacent offers have proximate successful probabilities. Moreover, once the basic RL algorithm finds a relatively rewarding offer, it hardly explores any other options.

In 1995, Roth and Erev showed that a slightly modified RL algorithm, can successfully model human players' behavior in several games, including the UG. This achievement was quite impressive, considering the fact that the UG is characterized by the existence of a significant gap between empirical human behavior and subgame perfect equilibrium predictions. The basic idea of the behavior of the proposer in their model is adaptation to the responders' acceptance thresholds. This property makes RL a possible approach for all CE environments, in which the negotiator is required to make decisions according to the opponents' actions. Roth and Erev's main modification to the basic RL algorithm was the introduction of a "generalization" parameter ϵ which prevents the probability of choosing an offer from going to zero if it is adjacent to a successful offer, as can be seen in **Algorithm 1**. Their assumption, which has general validity in CE environments, was that adjacent offers have proximate successful probabilities. In the original algorithm, where N=10, they considered as adjacent offers only the offers which are 1 above and 1 below a successful offer. Thus, if an offer of 4 is accepted by the opponent, the proposer slightly reinforces the Q-values of offers 3 and 5, as well. Here we present a generic format, which extends the range of the adjacent offers by using the γ parameter. This extension is important for environments with *N* larger than 10.[4]

Algorithm 1 ROTH & EREV'S MODEL

Notation: ϵ denotes the generalization parameter. γ is the generalization scope parameter.

1: For j=0 to N, initialize Q(j) arbitrarily
2: For each interaction, **Do**
3: Select offer i with a probability of $\frac{Q(i)}{\sum_{j=0}^{N} Q(j)}$
4: Observing opponent's move, calculate reward r
5: **If** offer i has succeeded **Then** Q(i) =Q(i) + r-ϵ
6: For j=(i-γ) to (i+γ), if j$\neq$i Q(j) =Q(j) + $\frac{\epsilon}{2\gamma}$

A negotiator who follows the RL method, as suggested by Roth and Erev, should act according to the following pattern:

1. No consistent obligation to LDT. As can be seen in the RL column in **Table 1**, an offer may be increased after a successful interaction (e.g. rows 8-9), and may be decreased

[3] Unlike Roth and Erev, (Gale, Binmore, & Samuelson 1995) proposed an evolutionary model based on replicator dynamics, and (Bourgine & Leloup 2000) proposed Gittins' index strategy in order to model the behavior of UG players. However, both of these models share the same pattern of behavior as RL, since they are based on the same ideas of learning by reinforcing profitable options, suppressing offers with lower expected payoff and selection according to their relative profitability.

[4] In their paper, Roth and Erev suggested an additional two modifications: the "cutoff parameter" and "gradual forgetting". However, these modifications start to have an effect only after many interactions, beyond the number of interactions considered here.

after an unsuccessful interaction (e.g. rows 31-32).

2. Certain offers will be chosen repeatedly throughout the whole game. Once an amount is successfully offered, the probability that this amount will be chosen again is significantly increased. As the game proceeds, it is even more difficult for other offers to be selected, since the denominator in line 3 of **Algorithm 1** increases. In addition, the small reinforcement of adjacent amounts constructs a narrow dominant range of offers which reinforce each other. In the sample from **Table 1**, it can be noticed that the amounts of 46 and 45 are regularly offered.

The DVRL-based model

The Deviated Virtual RL algorithm, as mentioned above, was originally proposed as a mechanism for computerized agents to automatically interact with people in CE environments (Katz & Kraus 2006). The algorithm, which was shown to perform impressively in empirical experiments, is briefly described here. DVRL is based upon the basic principle of RL, according to which an action is selected on the basis of its EU, which is evaluated in accordance with the results of previous interactions. One problem, however, in applying basic RL to CE environments is the disregarding of the fact that in CE the probability of an offer is gradually influenced by the size of the offer. Thus, a reasonable approach for the Q-vector update procedure in CE environments is *Virtual Learning* (VL) (Vreind 1997). According to the VL principle, the proposer in the UG, for example, treats all offers higher than an accepted offer as successful (virtual) offers, not withstanding that they were not actually proposed. Similarly, it considers all offers lower than a rejected offer as having been (virtually) unsuccessfully proposed. The rationale behind this principle is that the higher the amount proposed to the opponent, the higher the probability of the proposal being accepted. However, despite the reasonability of VL, it does not seem to be an appropriate model. The reason for this is that, while VL proceeds towards less risky offers after unsuccessful interactions, it performs no exploration of offers which are greedier than the current optimal offer, which is a deficiency it shares in common with the basic RL.

In contrast, DVRL deviates from the strict rationale underlying the VL principle, and extends the range of offers updated after each interaction. Thus, after a successful interaction, the Q-values of all the offers higher than the actual offer, as well as a few offers **below** the actual offer are increased, as described in line 8 of **Algorithm 2**. Similarly, after an offer has failed, the Q-values of all the offers lower than the actual offer, as well as a few offers **above** the actual offer are reduced, as described in line 6. Generally, this principle can be implemented in various basic algorithms besides RL (Katz & Kraus 2006). However, in this paper we will present only the extension of RL, i.e. DVRL, since the RL principle is quite intuitive, and therefore could be a suitable candidate for modeling human decision making, as in (Roth & Erev 1995). It is worth noting that the Deviated VL extension of other basic reinforcement algorithms, such as Gittins' indices strategy (see (Katz & Kraus 2006)) produces the same behavior pattern as produced by DVRL. In the DVRL version used by (Katz & Kraus 2006), the Q-values were updated by dividing the accumulated reward of each offer by the number of previous interactions where offer *i* was actually or virtually (according to the Deviated VL principle) proposed (lines 6 and 8). This was found to be more efficient than Roth and Erev's update procedure, which actually ignores unsuccessful interactions. In addition, at each interaction, the offer with the current maximum Q-value is selected (line 3).

Algorithm 2 THE DVRL ALGORITHM

Notation: α, β are two integers $0 \leq \alpha, \beta \ll$ N, (where N is the upper bound of possible offers), which denote the deviation rate. The values α and β can be gradually decreased during the learning process. The term n(j) denotes the number of previous interactions where offer j was actually or virtually proposed and r(j) is the corresponding reward for a successful offer j.

1: t=0 For j=0 to N, **Do** Q(j)=1, n(j)=0
2: For each interaction t, **Do**
3: offer i=$\arg\max_j Q(j)$
4: Observing opponent's move, calculate reward
5: **If** offer i has failed **Then**
6: For j=0 to (i + α), **Do** n(j)=n(j)+1, $Q(j) = \frac{Q(j)(n(j)-1)}{n(j)}$
7: **Else**
8: For j=(i-β) to N, **Do** n(j)=n(j)+1, $Q(j) = \frac{Q(j)(n(j)-1)+r(j)}{n(j)}$
9: t = t+1

There are two reasons for assuming that the deviation principle underlying the DVRL approach appears also in the human decision making process. Firstly, it is not necessarily a mistake to consider an offer which is slightly lower than a successful offer as successful as well. Using UG terminology, it is quite reasonable to assume that if proposal *i* had been accepted (rejected) by an opponent, that opponent would also have accepted (rejected) a slightly lower (higher) proposal. The chance that the proposal would exactly hit the acceptance threshold of the opponent is not high, especially for a large set of options. Secondly, and even more importantly, the deviation principle actually outlines a direction for optimal solution exploration, as in LDT, rather than the random trial-and-error approach that underlies other methods, such as RL. A DVRL-based proposer who successfully offered, for example, 50% of the cake (*N*) to a UG opponent in the first interaction, would offer 40% (if the configuration of β is 10) in the next interaction. The proposer would continue to decrease the offer until it is rejected. However, in contrast to the LDT, which ignores all the data experienced before the previous interaction, a DVRL-based negotiator takes into account previous interactions as well, and tries to model the distribution of the opponents' behavior, during the learning process. Thus, the DVRL algorithm provides a model which combines LDT and RL in a manner that avoids the problems of both approaches. This approach can quantitatively complete the direction marked by (Grosskopf 2003) in the conclusion of her paper on reinforcement and directional learning (LDT) in the UG:

"Therefore, a combination of the 2 approaches (reinforcement and directional learning) seems potentially worthwhile. This paper does not attempt to propose a quantitative

solution... Extending the pure reinforcement model through directional reasoning might allow for a better modeling of bounded rational but intelligent agents in different classes of similar games and serve as a crucial step towards a deeper understanding of cognition driven human behavior."

In addition to the reasonability of the DVRL as a model of human behavior, we will show here that the behavior pattern of a noticeable number of human subjects is consistent with a DVRL behavior pattern. As can be seen in the third column in **Table 1**, negotiations performed according to DVRL are characterized by the following two phases:

1. Dynamic onset - the decisions in the first interactions are made according to LDT. Since the Q-vector includes no previous information, the deviated updating of Q-values grants the maximal Q-value to an offer which is slightly lower than the last offer after a successful interaction, and to an offer which is slightly higher than the last offer after an unsuccessful interaction, as explained above.

2. Stabilization - as the game progresses, the offers become more stable and persistent. A decrease (increase) in the amount of the offer will occur only after a few successful (unsuccessful) interactions. As the negotiator gains experience, and the Q-values become higher, the system is less sensitive to the results of a single interaction, and the offer with the maximal Q-value retains its status for longer durations. This pattern is consistent with the general "Power Law of Practice", according to which learning curves become flatter over time (Blackburn 1936).[5]

The original DVRL method was developed for automated agents which have no memory restrictions. However, in order to adopt this method as a human model, we must take into account the constraints of human memory. In order to examine the influence of memory restriction on the behavior of the DVRL model, we inserted a "gradual forgetting" parameter, which gradually reduces the information retained from previous interactions, similarly to (Roth & Erev 1995). Computerized simulations revealed that increasing the forgetting parameter increases the duration of the dynamic onset, and relieves the conditions for changing offers in later interactions. However, except for extremely high forgetting parameter values, where the behavior pattern becomes identical to LDT , the general behavior pattern described above is relevant also for memory restricted negotiators.

Experimental design

In this section we empirically examine the existence of the three different patterns of behavior described above: LDT, RL-based model of Roth and Erev and DVRL-based model. For this purpose we asked people to compete iteratively against series of human opponents, in two domains: 1. In the UG, where the players had to divide 100 new Israeli Shekels (NIS, where 1 U.S. \$ $\approx$ 4.5 NIS), i.e. N=100. 2. In a first-price sealed-bid 2-bidders auction for 100 NIS. By examining different environments, we wanted to find out whether the environment influences the pattern of learning and decision making. We gave the subjects a large set of 101 optional decisions, unlike previous studies that have allowed only 10 options. This enabled their behavior to be examined with greater accuracy. However, as can be seen in **Table 1**, some subjects focused on lower resolutions of the options set (usually multiplications of 10). In our analysis, we treated those subjects according to their resolution levels, and decreased the relevant parameters in the models (such as N, α,β and γ) in order to fit their decision patterns.

The experiment included 48 participants (24 males and 24 females). All the participants were students at Bar Ilan University, aged 20-28, and were not experts in negotiation strategies nor in economic theories directly relevant to the experiment (e.g. game theory). Each of the participants was seated in an isolated room at a computer work-station.

In order to construct a series of opponents for use in our experiments, we first surveyed the behavior of 20 of the 48 subjects, as follows: In the auction environment, we collected the bid amounts of those subjects playing one game against an anonymous opponent via a computer. The bid could be any integer from 0 to 100 NIS, and the winner gained a virtual 100 NIS. In the UG environment, we extracted the minimal acceptance amounts of the 20 subjects when playing as responders, iteratively, against anonymous proposers. This method of examining responders' behaviors is widely accepted in the UG literature (e.g. (Bourgine & Leloup 2000)). The proposals were actually artificial though the participants were told that the proposals were provided by other people. We included each participant's response in the relevant series twice, in order to enlarge the size of both response series from 20 to 40. In addition, we constructed an artificial series of 40 auction bids which were randomly generated according to a normal distribution of N(71,10). In this manner, we wanted to examine the behavior of the participants against an "ideal" population which distributes normally, though there is no explicit evidence of such a distribution in any CE environment.

After extracting the series of opponents' responses for each environment, the other 28 participants were asked to interact iteratively with those series, in one-shot interactions with each opponent, functioning as proposers in the UG and as bidders in the auction (the order of the games was randomly determined for each subject). In addition, 15 of the first 20 participants competed in the auction against the artificial normally distributed population, without knowing that their "opponents" were artificial. On the whole, then, the experiment examined the decision patterns of 43 subjects who repeatedly interacted with changing opponents. After each decision, the participants were informed of the success of their offer by the "current" opponent. However, they were not informed of the bid or the acceptance threshold of the opponent. All subjects competed against the same series of opponents. At the end of the experiment, each participant was paid between 15 to 30 NIS, in proportion to her earnings in the interactions in which she participated.

[5]The original DVRL method, as described in (Katz & Kraus 2006), includes a gradual decrease in the values of the deviation parameters α and β during the learning process. Such a policy yields the same decision pattern described here, with a faster convergence to the stable phase. However, in this paper we want to emphasize the fact that the pure DVRL method induces a double-phase decision pattern even without the manipulation of decreasing the deviation parameters.

Environment (# of subjects)	LDT	RL	DVRL	Const	Const +DVRL	Unexplained
UG (28)	3.6	17.9	46.4	14.3	14.3	3.6
Auction (28)	14.3	17.9	46.4	3.6	10.7	7.1
Normally distributed Auction (15)	13.3	13.3	53.3	0.0	13.3	6.7
Average	**9.9**	**16.9**	**47.9**	**7.0**	**12.7**	**5.6**

Table 2: For each model, the percentage of model-typical observations found in each environment.

Results

Table 2 summarizes the percentage of subjects who behaved in accordance with each of the models tested, in each environment. As can be seen, about 48% of the observations can be explained by the DVRL-based model, which is considerably more than the other models.

In 7% of the observations, the participants did not change their decision throughout the game, as can be seen in the "const" column. A sample of their decision pattern is presented in the "Constant" column in **Table 1**. Another 12.7% showed an interesting pattern of decision making, which matches the DVRL-based model, except for the lack of a dynamic onset. The behavior of one such subject is presented in the "Constant+DVRL" column of **Table 1**. Subjects following this pattern maintained a persistent offer that rarely changed from the very beginning of the game. Nevertheless, changes that occurred were consistent with an LDT direction, exactly as in the later stages of the DVRL pattern. 5.6% of the observations cannot be explained by any of the models surveyed here.

As can be seen, the results from the normally distributed auction, in which players unknowingly played against artificial opponents, and from the auction environment in which the opponents were human, are quite similar, as expected. Since the former environment was the first environment the participants faced, while the latter environment was many times preceded by UG play, it can be concluded that the experience of subjects had no significant influence.

In general, the decision patterns in all the environments were very similar except for two differences between the UG and the auctions: 1. In the UG 4 players (14.3%) had a "constant" decision pattern (always offered 50%), in contrast to only 1 player out of 43 (2.3%) in the auctions. This difference may be attributable to the fact that, in the UG, the two players could perceive themselves as partners, while in the auction, they were clearly competitors. Thus, in the UG, these 4 players may have thought they should behave fairly, and always offer 50% to the other player. Alternatively, the players in the UG may have been deeply convinced that almost no responder would accept less than half of the cake, and thus they were not trying to explore other alternatives. 2. In both auction environments about 14% played according to LDT, in contrast to a figure of only 3.6% in the UG. It is possible that, in the UG, due to the possible sense of partnership between the players, subjects felt it would be too greedy to change their offer after every interaction.

Conclusion and future work

In this paper we have used an innovative methodology for modeling human iterative decision making, by examining actual decision patterns. It was shown that, in repeated CE interactions, a model based on the DVRL algorithm faithfully describes 48% of subjects' decision patterns, much more than any other model does. Interestingly, this algorithm has also been found to be the most efficient and profitable method in such environments (Katz & Kraus 2006).

In future work, we intend to examine human behavior in other CE environments, such as dynamic pricing and all-pay auctions. In addition, we intend to consider repeated CE environments, where the interaction with each opponent lasts for a number of rounds. When playing repeatedly against the same opponent, we expect players to learn the individual behavior of the specific current opponent, in addition to utilizing their own generic model of the opponent population. Moreover, in contrast to the one-shot games, each decision influences the future behavior of the current opponent, a fact that a human model must take into account.

Acknowledgements: This work is supported in part by NSF IIS0222914. Kraus is affiliated with UMIACS.

References

Blackburn, J. 1936. Acquisition of skill: An analysis of learning curves. Technical report, IHRB Report No. 73, London: His Majesty's Stationery Office.

Bourgine, P., and Leloup, B. 2000. May learning explain the ultimatum game paradox ? Technical Report GRID Working Paper No. 00-03, Ecole Polytechnique.

Brenner, T., and Vriend, N. On the behavior of proposers in ultimatum games. *J. Econ. Behav. Organ.* Forthcoming.

Freed, M. 2000. *AAAI Fall Symposium on Simulating Human Agents*. North Falmouth, MA: AAAI Press.

Gale, J.; Binmore, K.; and Samuelson, L. 1995. Learning to be imperfect: The ultimatum game. *Games Econ. Behav.* 8:56–90.

Grosskopf, B. 2003. Reinforcement and directional learning in the ultimatum game with responder competition. *Experimental Economics* 6(2):141–158.

Guth, W.; Schmittberger, R.; and Schwarz, B. 1982. An experimental analysis of ultimatum bargaining. *J. Econ. Behav. Organ.* 3:367–388.

Katz, R., and Kraus, S. 2006. Efficient agents for cliff-edge environments with a large set of decision options. In *AAMAS'06*.

Mitzkewitz, M., and Nagel, R. 1993. Experimental results on ultimatum games with incomplete information. *Int. J. Game Theory* 22(2):171–198.

Ockenfels, A., and Selten, R. 2005. Impulse balance equilibrium and feedback in first price auctions. *Games Econ. Behav.* 51(1):155–170.

Roth, A., and Erev, I. 1995. Learning in extensive form games: Experimental data and simple dynamic models in the intermediate term. *Games Econ. Behav.* 8:164–212.

Selten, R., and Stoecker, R. 1986. End behavior in sequences of finite prisoner's dilemma supergames. a learning theory approach. *J. Econ. Behav. Organ.* 7:47–70.

Sutton, R., and Barto, A. 1998. *An Introduction to Reinforcement Learning*. MIT Press.

Vreind, N. 1997. Will reasoning improve learning? *Econ. Lett.* 55(1):9–18.

Zhu, W., and Wurman, P. 2002. Structural leverage and fictitious play in sequential auctions. In *AAAI'02*.