8-25-2008

Learning about Learning in Games through Experimental Control of Strategic Interdependence

Jason Shachat
National University of Singapore

Todd Swarthout
Georgia State University

Learning about Learning in Games through Experimental Control of Strategic Interdependence

Jason Shachat[a] **and J. Todd Swarthout**[b]

[a] *Department of Business Policy, National University of Singapore, 1 Business Link, Singapore 117592*
[b] *Department of Economics, Georgia State University, Atlanta, GA, 30303, USA*

First Draft: June, 2002
This Draft: August 25, 2008

Abstract

We report experiments in which humans repeatedly play one of two games against a computer program that follows either a reinforcement learning or an Experience Weighted Attraction algorithm. Our experiments show these learning algorithms detect exploitable opportunities more sensitively than humans. Also, learning algorithms respond to detected payoff-increasing opportunities systematically; however, the responses are too weak to improve the algorithms' payoffs. Human play against various decision maker types doesn't vary significantly. These factors lead to a strong linear relationship between the humans' and algorithms' action choice proportions that is suggestive of the algorithm's best response correspondence. These properties are revealed only by our human versus computer experiments, and not by our standard human versus human experiments, nor our model simulations.

JEL classification: C72; C92; C81

Keywords: Learning; Repeated games; Experiments; Simulation

1 Introduction

Researchers have exerted considerable efforts identifying how individuals adjust behavior in repeated decision making tasks. A sizable literature has approached this issue by developing hypotheses about the learning process, embedding the hypotheses in a parametric model, and then estimating the model's parameters from experiments in which human subjects repeatedly play some stage game. The models within this literature have converged to a formulation with two main components. The first component is a rule that assigns a value to each of a player's actions conditional

upon the history of play, and the second component converts these values into a probability distribution that governs the player's action choice—in effect, a mixed strategy. Three examples of such learning models are the cautious fictitious play model introduced by Fudenberg and Levine (1995), the reinforcement model introduced by Erev and Roth (1998), and the experienced weighted attraction model introduced by Camerer and Ho (1999). Previous evaluations have concluded that these types of models have marked success at approximating human play (Feltovich, 2000; Mookherjee and Sopher, 1994, 1997), and provide an adequate "representative agent" description of how humans play in a broad class of games.[1] However, these models do not fair well in repeated games that provide an opportunity for players to benefit from out-of-equilibrium cooperation, such as the prisoner's dilemma (Erev and Roth, 2001; Hanaki et al., 2005).

We uncover two important properties of this class of models that should provide significant new directions in this literature. First, these models are more sensitive than human subjects at detecting exploitable trends in opponent play. Moreover, the models' action frequencies exhibit strikingly linear adjustments toward their best response. Second, these adjustments are much too weak to generate any significant gains in payoffs. The probabilistic choice components generate responses that are less aggressive than those made by subjects measured in other studies. In addition to these two results, we also introduce a novel research technique and experimental paradigm. It was only through this unique approach that we discovered these attributes of learning in games models.

As in previous studies, we start with experimental sessions in which fixed subject pairs repeatedly play a simple game, and use these data to estimate parameters of alternative learning models. Then, we conduct hybrid sessions in which different subjects each repeatedly play the same game against a computer implementation of one of the estimated learning models. Finally, we generate parallel simulations in which an estimated version of a model plays against itself. Through our hybrid experiments we control for the strategic interdependence of the adaptive behavior of subjects, which is a main source of the econometric difficulties when estimating these models (Salmon, 2001).

A simple ideal motivates our approach: a model is considered the "true" model if human versus

[1]But there are caveats as well: the models fail to capture the heterogeneous behavior exhibited across subjects (Cheung and Friedman, 1997); humans have an ability to detect and exploit intertemporal choice patterns by their opponents which the models don't (Sonsino and Sirota, 2003; Mukherji and Runkle, 2000); and humans are more successful at establishing reciprocal behavior in the repeated game environments (Andreoni and Miller, 1993; Cooper et al., 1996).

computer play is indistinguishable from human versus human play. When we identify differences in play (which will almost surely happen given the lofty benchmark), the differences should suggest how we can refine the model. Our hybrid procedure can also help in developing new learning models: by systematically varying the strategies against which people are required to play, we can obtain descriptions of how humans play against a variety of models. In addition to identifying how closely a model mimics human behavioral rules, we can also assess the effectiveness of a model by comparing its earnings to those of humans in the same settings.[2]

In our study we consider the Reinforcement (RE hereafter) and Experience-Weighted Attraction (EWA hereafter) learning models.[3] We also consider a pair of 2×2 normal forms games, both having a unique Nash equilibrium which is in mixed strategies. The following is a summary of our results:

1. In the human versus human experiments we find that, as is found uniformly across studies, the estimated adaptive rules have significant memory. As a consequence, the impact of recent outcomes on action values (and thus on mixed strategies) is very small. This leads to inert adjustment rules in simulations: mixed strategies, and the resultant pattern of action choices, are sluggish and exhibit little change from period to period.

2. The distribution of action choice frequencies of human players does not vary significantly depending on whether the opponent is a human or is one of the learning algorithms in the majority of situations considered.

3. In the simulations, models generate behavior that correspond to average human play but display less variation. Specifically, we find that, as is found in other studies, the average joint choice frequencies generated by the model simulations correspond to those generated by human pairs. However, the variance of these joint frequencies in the simulations is much lower than we observe in the human experiments.

4. In both simulations and human experiments, there is no correlation in the joint action frequencies across fixed pairs of opponents within the same game treatment.

5. The learning models are more sensitive than humans at calculating, and correspondingly

[2]This is similar in spirit to the notion of a model's economic value as discussed by Camerer et al. (2002, 2003).

[3]There are many other similarly structured models worthy of studying with our technique, but models in this class tend to generate similar play (Salmon, 2001).

adjusting play, when an opponent's action frequency deviates from the Nash equilibrium proportion. Specifically, a highly correlated relationship between the joint action frequencies of computer-human pairs emerges in our hybrid experiments. When a human's action frequency deviates from the Nash equilibrium proportion, the algorithm's action frequency systematically adjusts towards its pure strategy best response. These adjustments result in a strikingly linear relationship between the learning models' and humans' action choice frequencies. Moreover, the linear relationship is consistent with the computer players' best response correspondence.

6. While the adjustments by the models are remarkably systematic when pitted against humans, the magnitude of the adjustments is quite weak. Indeed, too weak to result in statistically significant gains in payoffs. In all six computer-human treatments, the average earnings of the models are not statistically significantly more than those of humans in equivalent roles. Moreover, in two of these cases the humans earn statistically significantly more. However, in both of these cases, we show that the probabilistic best response component—rather than superior human play—is the main source of models' lower earnings.

Our results explain and unify some previous findings counter to the implications of the considered learning models. First, our results revealing the viscous property of the learning models' mixed strategy sequences suggest a more appropriate model would generate more dramatic changes in period-to-period mixed strategy formulation. One such example is Nyarko and Schotter (2002) (NS hereafter). NS elicit subjects' beliefs of opponents' actions in a repeated game similar to what we use, and observe that subjects wildly revise their elicited beliefs from period to period. Then NS formulate a model in which the expected values of actions, calculated using the stated beliefs, are incorporated into a probabilistic logit choice rule. After conducting a series of goodness-of-fit exercises, NS conclude this model outperforms both the RE and EWA models. Our results show that the NS model gains its effectiveness from its ability to allow for greater period-to-period changes in the likelihoods of action choices. The implied volatile mixed strategies of the NS model are supported when subjects are asked to explicitly choose mixed strategies in studies such as Shachat (2002) and Noussair and Willinger (2003).

Second, our demonstration that probabilistic learning models accurately detect but only weakly

best respond to nonequilibrium play is in direct contrast with what humans do. In studies such as Shachat and Swarthout (2004), Fox (1972), and Lieberman (1962), subjects play against non-optimal but unknown mixed strategies in repeated zero-sum games. Subjects only detect these non-optimal strategies if they are far enough removed from the minimax strategies. However, once detected, subjects move decisively towards best response and increase their payoffs.

The remainder of this study is structured as follows. We next proceed with a more detailed discussion of several past studies that incorporate human versus computer game play. We then present the two learning models used in our study. In the fourth section we discuss the games used in our experiments and our experimental procedures. Section 5 covers our experiment results, findings, and interpretations. In conclusion, we integrate our results with other experimental results to provide a summary of human play in games and contrast this with current learning models.

2 Man Versus the Machine

Human players and computerized decision makers have interacted in a number of previous studies. This technique has been used to identify social preferences in strategic settings (Houser and Kurzban, 2002; McCabe et al., 2001), to establish experimental control over player expectations in games (Roth and Schoumaker, 1983; Winter and Zamir, 1997), and to identify how humans play against particular strategies in games (Walker et al., 1987). In this section, we summarize previous results regarding how humans play against unique minimax solutions, non-optimal stationary mixed strategies, and variants of the fictitious play dynamic rules in repeated constant-sum games with unique solutions in mixed strategies.

All of the studies we discuss used fixed human-computer pairs playing repetitions of one of the zero-sum games presented in Figure 1.[4] Studies by Lieberman (1962), Messick (1967), and Fox (1972) all contain treatments where humans played against an experimenter-implemented minimax strategy. In these studies, the human participants were not informed of the explicit mixed strategy adopted by their computerized counterparts.[5] All three studies reach the same conclusion: human

[4] In some of these studies the experimenters implemented stationary mixed strategies by using pre-selected computer generated random sequences in their non-computerized experiments.

[5] When reported, human participants were instructed something similar to: "The computer has been programmed to play so as to make as much money as possible. Its goal in the game is to minimize the amount of money you win and to maximize its own winnings." (Messick, 1967, pg. 35)

play does not correspond to the minimax prediction, and only in the Fox study does the human play adjust—albeit weakly—towards the minimax prediction. These results are not surprising: when a "computer" adopts its minimax strategy, as the expected payoffs of a human player's actions are all equal.

This indifference is not present when the computer adopts non-minimax mixed strategies. Lieberman (1962) and Fox (1972) also studied human play against non-optimal stationary mixed strategies and discovered that human players do significantly adjust their play (although not to the extent of exclusively playing the pure strategy best response) and also significantly increase, in a statistical sense, their payoffs above minimax value levels. In the relevant Lieberman treatment, subjects played against the experimenter for a total of 200 periods. In the first 100 periods, the experimenter played his minimax strategy of (.25, .75) and then in the final 100 periods the experimenter played a non-minimax strategy of (.5, .5). Human players were not informed that their opponent had adjusted his strategy. Human play adjusted from best responding approximately 20 percent of the time immediately after the experimenter began non-minimax play, to best responding approximately 70 percent of the time by the end of the session. This shift toward the best response was also a shift towards the human's minimax strategy, making it difficult to differentiate between the attractiveness of the minimax strategy and the best response.

In one of Fox's treatments, each human participant played 200 periods against a computer which played the non-minimax mixed strategy (.6, .4) for the entire session. This design placed the human's best response of (1, 0) on the opposite side of (.5, .5) from the human's minimax strategy of (.214, .786). Initial human play of their first action was slighty above 50 percent, and then slowly adjusted towards the pure strategy best response over the course of the experiment. Specifically, human players approached the best response 75 percent of the time. These experiments demonstrate that human participants will adjust their behavior (but not as much as possible) to take advantage of exploitable stationary mixed strategies. Furthermore, the human subjects in both studies statistically improved their payoffs.

In a study designed to ascertain how far a mixed strategy must deviate from the minimax strategy before humans exploit it, Shachat and Swarthout (2004) systematically vary the fixed mixed strategy across subjects.[6] We observed that when a computerized strategy deviates by more

[6]This study used the Pursue-Evade game adopted here and presented in Figure 2.

than 15 percent from the minimax strategy of two thirds, human play begins converging to best response. Furthermore, many subjects in this study adjusted to exclusive play of the best response action—a behavior which wasn't apparent in the aggregate data presented in the previous studies.

Messick (1967), Coricelli (2001), and Spiliopoulos (2008) all conducted experiments to evaluate how human players respond when playing against variations of fictitious play.[7] These experiments are notable in that the computer's strategy was responsive to the actions selected by its opponent. Messick studied human subjects matched against two fictitious play algorithms: one with unlimited memory and the other with only a five period memory. Against unlimited memory fictitious play, human players earned substantially more than their minimax payoff level. Human players earned an even greater average payoff against limited memory fictitious play. In the study by Coricelli, there are two treatments (both utilizing the game form introduced by O'Neill (1987)) in which human participants play against unlimited memory fictitious play with and without a belief bias. This bias holds that human subjects tend not to repeat their "P" action. In both treatments human participants win significantly more often against the algorithms than they do against human opponents.[8] Spiliopoulos studied human play against a variety of fictitious play algorithms, including those with varying memory, pattern detection, and aggressive probabilistic choice rules.[9] The main findings were that subjects generally could exploit the algorithms and that subjects play differently conditional on the algorithm they faced. Establishing that humans can "outgame" these algorithms is significant, though not surprising. It is well known that in games with a unique mixed strategy equilibrium, the fictitious play algorithm can generate strong positively serially correlated action choices that are easily exploited (Jordan, 1993; Gjerstad, 1996). It was this speculated vulnerability that partially motivated game theorists to propose and study adaptive learning models which incorporated probabilistic choice as a key component.[10]

To summarize, prior experiments pairing human subjects against algorithms in constant sum games with strictly mixed strategy solutions have taught us that: 1. human players do not tend to

[7] In the original formulations of fictitious play (Brown, 1951; Robinson, 1951), a player uses the empirical distribution of the entire history of his opponent's action choices as his belief of the opponent's current mixed strategy and then chooses a best response to this belief.

[8] Human versus human data for this conclusion are taken from O'Neill (1987) and Shachat (2002).

[9] The computerized decision makers followed the algorithm prescription 80 percent of the time and played the minimax strategy 20 percent of the time.

[10] For example, see cautious fictitious play proposed by Fudenberg and Levine (1995), and the two learning models we utilize in this study.

play their minimax strategy in response to opponents playing their minimax strategy; 2. human players exploit opponents who play mixed strategies significantly different from their minimax strategy; and 3. human players exploit adaptive algorithms which generate highly serially correlated action choices. These results suggest that the technique of humans playing against computer-implemented behavioral rules will provide insights on the appropriateness of more current models of learning in games.

3 Response Algorithms

In this section we describe the Reinforcement learning model of Erev and Roth (1998) and the Experience Weighted Attraction model of Camerer and Ho (1999). Our descriptions of the model formulations and estimation techniques follow the original presentations as close as possible. Nonetheless, we only consider 2×2 games and in some instances we simplify notation without changing the models.

3.1 Reinforcement Learning

Erev and Roth's model (hereafter RE) is motivated by the reinforcement hypothesis from psychology: an action's score is incremented by a greater amount when it results in a "positive" outcome rather than a "negative" outcome. More formally, let $R_{ij}(t)$ denote player i's score for his jth action prior to the game at iteration t; let $\sigma_{ij}(t)$ denote the probability that i chooses j at iteration t; and let X_i denote the set of player i's possible stage-game payoffs. The two initial conditions of the dynamic system are: 1. at the initial iteration, each of a player's actions has the same probability of being selected (in the 2×2 case each action is chosen with probability one-half); and 2. the initial score of each actions is

$$R_{ij}(1) = \sigma_{ij}(1)S(1)\overline{X_i},$$

where $S(1)$ is an unobservable strength parameter, which influences the player's sensitivity to subsequent experience, and $\overline{X_i}$ is the absolute value of player i's payoff averaged across all action profiles.

After each iteration, each action's score is updated as follows

$$R_{ij}(t+1) = (1-\phi)R_{ij}(t) + \left((1-\varepsilon)I_{(a_i(t)=j)} + \frac{\epsilon}{2}\right)(\pi_i(j, a_{-i}(t)) - \min\{X_i\}),$$

where ϕ is an unobservable parameter that discounts past scores, $I_{(a_i(t)=j)}$ is an indicator function for the event that player i selected action j in period t, ε is an unobservable parameter determining the relative impacts on the scores of the selected versus the unselected action, and $\pi_i(j, a_{-i}(t))$ is i's payoff when he plays action j against the deleted action profile $a_{-i}(t)$. Also player i's minimum possible payoff for any action profile, $\min\{X_i\}$, is subtracted from $\pi_i(j, a_{-i}(t))$ as a normalization to avoid negative scores. The second component of the model, a probabilistic choice rule, is specified as

$$\sigma_{ij}(t) = \frac{R_{ij}(t)}{\sum_k R_{ik}(t)}.$$

For each game we consider, parameters of the model are estimated along the lines suggested by Erev and Roth. We estimate the values of $S(1)$, ϕ, and ε by minimizing the mean square error of the predicted proportions of Left play in 20-period trial blocks for the human versus human treatments. More specifically, for each fixed triple of parameter values from a discrete grid we proceed as follows: we simulate the play of 500 fixed pairs engaging in 200 iterations, and then we calculate separately the frequency of Left play by the 500 Row players and by the 500 Column players in each 20-period block. These frequencies are the model's predictions for that triple of parameter values. The grid is then searched for the optimal parameters.

3.2 Experience-Weighted Attraction

We use the version of EWA developed by Camerer and Ho (1999). While the structure of the EWA formulation is similar to the RE learning model, it adopts a different parametric form of probabilistic choice and it updates actions' scores according to what actions actually earned in past play, and what actions hypothetically would have earned if they had been played.

According to EWA, subjects choose stage-game actions probabilistically according to the logistic distribution

$$\sigma_{ij}(t) = \frac{e^{\lambda R_{ij}(t)}}{\sum_k e^{\lambda R_{ik}(t)}},$$

where at stage t player i chooses action j with probability $\sigma_{ij}(t)$, where λ is the inverse precision (variance) parameter, and where $R_{ij}(t)$ is a scoring function, as in the RE model, albeit defined (i.e., updated) differently. The updating of $R_{ij}(t)$ involves a discounting factor $N(t)$, which is updated according to $N(t+1) = \rho N(t) + 1$ for $t \geqq 1$, where ρ is an unobservable discount parameter and $N(1)$ is an unobservable parameter, interpreted as the strength of experience prior to the beginning of play. The score $R_{ij}(t)$ is updated as follows:

$$R_{ij}(t+1) = \frac{N(t)\phi R_{ij}(t) + ((1-\varepsilon)I_{(a_i(t)=j)} + \frac{\varepsilon}{2})\pi_i(j, a_{-i}(t))}{N(t+1)},$$

where $\pi_i(j, a_{-i}(t))$, ϕ, and ε are interpreted as in the Erev and Roth model. Initial scores, $R_{ij}(1)$ for each i and j, are additional unobservable parameters.

Parameters of the EWA model are estimated via maximum likelihood. It is worth noting that EWA is a flexible specification that includes several other models as special cases. For example, a simple reinforcement learning model, which has a different parametric form than RE, is generated when $N(1) = 0$, $\varepsilon = 0$, and $\rho = 0$; and probabilistic fictitious play is generated when $\varepsilon = \rho = \phi = 1$.[11]

4 Experimental Procedure

There are three basic steps in our experimental methodology. First, we collect baseline data samples consisting of fixed human versus human pairs that play 100 or 200 rounds of one of two 2×2 games. Second, we estimate parameters for the two learning models separately for each of the two games. In the third step, a new sample of humans play one of the two games against an estimated learning algorithm. We proceed by describing the two games we used and then present more details on the outlined steps.

4.1 The Two Games

The first game we consider is a zero-sum asymmetric game called Pursue-Evade. This game was introduced by Rosenthal et al. (2002) (hereafter RSW). The normal form representation of the

[11] We refer the reader to Camerer and Ho (1999) for more discussion of how EWA can emulate various models and for a more complete interpretation of the parameters.

game is given in Figure 2. The minimax solution (and Nash equilibrium) of this game is symmetric with each player choosing Left with probability of two-thirds.

There are several reasons why this game is a strong candidate to use in our study. First, zero-sum games eliminate social utility concerns often found in experimental studies of games, thereby mitigating some behavioral effects that might arise if a human suspects he is playing against a computer rather than another human. Second, with some standard behavioral assumptions, the repeated game has a unique Nash equilibrium path which calls for repeated play of the stage game Nash equilibrium. This eliminates potential repeated game effects that the algorithms are not designed to address. Third, Pursue-Evade is a simple game in which the Nash equilibrium predictions differ from equiprobable choice. This provides a powerful test against the alternative hypothesis of equiprobable play.

Our second game poses a more difficult challenge to the learning algorithms. We refer to our second game, presented in Figure 3, as Gamble-Safe. Each player has a Gamble action (Left for each player) from which he receives a payoff of either two or zero, and a Safe action (Right for each player) which guarantees a payoff of one. This game has a unique mixed strategy in which each player chooses his Left action with probability one-half, and his expected Nash equilibrium payoff is one. Notice that this game is not constant-sum; therefore the minimax solution need not coincide with the Nash equilibrium. In this game, Right is a pure minimax strategy for both players that guarantees a payoff of one. A game for which minimax and Nash equilibrium solutions differ but generate the same expected payoff is called a non-profitable game.[12] The potential attraction of the minimax strategy can (and does) prove to be difficult for the learning algorithms which, loosely speaking, probabilistically best respond.

4.2 Protocols

4.2.1 Human versus Human Baselines

For the human versus human baseline play in the Pursue-Evade game we use the data generated by RSW. In their hand-run experiments, a pair of subjects were seated on the same side of a table with an opaque screen dividing them. The Evader was given an endowment of currency. Each

[12] Morgan and Sefton (2002) present an excellent study of human play in non-profitable games.

player was given two index cards: one labeled Left and the other labeled Right. At each iteration the players slid their chosen cards face down to the experimenter seated across the table. Then the experimenter simultaneously turned over the cards, executed the payoffs, and recorded the actions. Twenty pairs of human subjects played this treatment: fourteen for 100 periods and six for 200 periods.

The human versus human baseline experiments for the Gamble-Safe game were executed via computerized interaction. Each subject was seated at a separate computer terminal such that no subject could observe the screen of any other subject. Within a pair, each subject played either the Row or Column role for the entire session. Fifteen pairs of subjects participated in this treatment, with five pairs each playing 100 periods and ten pairs each playing 200 periods. At the beginning of each repetition, a subject saw a graphical representation of the game on the screen. A Column player's display of the game was transformed so that he appeared to be a Row player. Thus, each subject selected an action by clicking on a row, and then confirmed his selection. Each subject was free to change his row selection before confirmation. Once an action was confirmed, a subject waited until his opponent also confirmed an action. Then a subject saw the outcome highlighted on his game display, as well as a text message stating both players' actions and his own earnings for that repetition. Finally, at all times a history of past play was displayed to the subject. This history consisted of an ordered list with each row displaying the number of the iteration, the actions selected by both players, and the subject's earnings.

4.2.2 Human versus Algorithm Treatments

We conducted our hybrid treatments using both the experimental software and protocol used for the Gamble-Safe game baseline.[13] In these treatments, two human subjects played against each other for the first 23 repetitions of the game. Then, unbeknownst to the human pair, they stopped playing against each other and for the remainder of the experiment they each played against a computer that implemented either the EWA or RE learning algorithm.

We used an initial phase of human versus human play to minimize the impact of estimated initial score values of actions and focus our evaluation on the dynamics of the algorithm. During the first 23 repetitions, we allowed the action value scores to "prime" themselves with the play generated by the

[13] For the Pursue-Evade game, the Evader was given a currency endowment.

subjects. (Although updating of scores was determined by the parameter estimates obtained from the baseline treatments). That is, even though the response algorithms were not selecting actions during the first 23 repetitions, the scores were still being updated according to the specifications of the previous section. For example, consider the 24th repetition of a game. The human Row player now faces a computer that plays the Column role. Moreover, during the first 23 repetitions, the computer Column player updated the scores associated with Column's actions based on the observed actions of both humans.

We adopted a simple technique to make the "split" seamless from the subjects' perspectives. From period twenty-four on, the two human-computer pairs had no interaction except for the timing of how action choices were revealed. Specifically, although the computers generated their action choices instantly, the computers didn't reveal their choices until both humans had selected their actions. This protocol preserved the natural timing rhythm established by the humans in the first 23 stage games.

In summary, we have two experimental factors: the stage game and the type of opponent. The sample size of each treatment is given in Table 1.[14]

5 Experimental Results

5.1 Baseline Experiments and Model Estimation

Our experimental baselines are the human versus human play in each of the two games. Inspection of the aggregate data reveals that play in the two games departs from the Nash equilibrium and the dynamic features of the data suggest non-stationarity of play. After estimating the unobserved parameters of the learning models, we simulated large numbers of experiments based upon these estimated versions of the models. Simulations reveal that the learning models generate aggregate choice frequencies similar to the experimental data, but only weakly mimic the experimental data time series. Furthermore, the simulations do not reveal striking differences between the two learning models.

We use the data from RSW as the Pursue-Evade game baseline data set. Figure 4 shows contingency tables for the data aggregated across subject pairs and stage games. A graph of the

[14] We explain in the next section why we have no observations for the EWA Gamble-Safe treatment.

time series of the average proportion of Left play for the Row and Column players is shown below each contingency table. Each observation in a series is the average across a 20 period time block. As noted by RSW, the contingency table is distinctly different from the Nash equilibrium predictions (the numbers in parentheses) and Column subjects play Left significantly more often than the Row subjects.[15] In the block average time series, we see that the Column series almost always lies above the Row series and that both series exhibit an increasing trend.

Using this data, RSW estimated the parameters of both the RE and EWA models. As noted by RSW, both models have some success in explaining the deviation. Using the estimated models we simulated 10,000 experiments of twenty pairs playing the Pursue-Evade game for 200 iterations. Averages from the 10,000 simulated experiments were used to construct contingency tables and time series in the same format as those presented for the baseline data. These results are presented alongside the baseline results in Figure 4. Unsurprisingly, given the respective objective functions used to select model parameters, casual observation suggests that the EWA model generates an expected contingency very close to the human baseline and the RE model more accurately mimics dynamics in the times series.

We provide a corresponding analysis for the Gamble-Safe game in Figure 5. In the contingency table for the baseline data we observe that the Row subjects play Right significantly more than Left, while Column subjects played Left more often. This result partly comes from two pairs in which the Row and Column subjects' action profile sequence eventually converged to the profile (Safe, Gamble). This is evident around the midpoint of the times series for the baseline treatment, where we see the Column and Row subjects' series diverge. These two pairs provide a stark example of a noted weakness of the learning models considered. Specifically, the models—unlike some human subjects—don't generate play which converges to non-equilibrium but joint payoff-maximizing action profiles.

This convergence to minimax play by the Row subjects in these two pairs is problematic for the maximum likelihood estimation used in the EWA model. Specifically, the long strings of Left by Column leads the EWA model to assign a near zero probability to Right (Safe) by Row for any possible parameter values. However, since Row is repeatedly choosing Right in these instances there is a zero likelihood problem in estimating the EWA parameters. Rather than violate the

[15]Moreover, the Column subject plays Left more frequently than his Row counterpart in almost all pairs.

maximum likelihood criterion for parameter selection specified by Camerer and Ho (1999) we chose not to conduct a Human versus EWA treatment for this game.

Since the parameter selection of the RE model does not rely upon maximum likelihood estimation, we obtain estimates which generate the best fit for the baseline data. Interestingly we see that the RE contingency table is remarkably similar to the baseline table. However, the predicted RE dynamics are excessively smooth and do not resemble the baseline time series. We believe this failure results from the inability of the model to incorporate the heterogenous behavior that occurs when some players adopt the minimax strategy and other players use adaptive strategies.

Comparison of the experimental data to simulations based upon estimated versions of the learning models suggests that the learning models successfully capture some features of the humans' disequilibrium behavior. However, time series views of the simulation data exhibit less variable dynamics than the experiment data, which suggests that learning models are not as responsive as humans and tend to simply fit aggregate human choice frequencies. We will see that this conclusion couldn't be further from the truth. In the human-algorithm experiments the learning algorithms demonstrate an acute ability to detect exploitable and distinctly adjust play. However, these adjustments are too timid to be profitable.

5.2 Analysis of Learning Algorithm Response to Opponents' Play

Inspection of the pair-level data from human-algorithm experiments reveals that the learning algorithms generate choice frequencies that are linearized best responses to the choice frequencies of their human opponents. Each of Figures 6–8 is a 2×2 array of scatterplot panels. The rows of each panel array correspond to the decision maker type of the Row player: the top row indicates human decision maker and the bottom row indicates computer decision maker. Similarly the columns of each panel array correspond to the decision maker type of the Column player: the left column for human and the right column for computer. Hence the upper left panel is from the human-human baselines, the lower right panel is from the algorithm-algorithm simulations, and the off-diagonal panels are from the human-algorithm experiments.

The scatterplots show the proportions of Left play by the Row and Column players in each pair after the first 23 iterations. In the simulation panel we only use the data from a single simulated experiment with twenty pairs playing 200 iterations. Also, each off-diagonal scatterplot displays

a trend line, which is obtained by regressing the Computers' proportions of Left on the Humans' proportions of Left.

Inspecting the two main diagonal panels of each figure reveals that both human-human play and pure simulations of model interactions generate uncorrelated "clouds" of joint Left frequencies with the simulations' clouds exhibiting much smaller dispersion. The scatterplots of human-algorithm play are dramatically different. In most of the off-diagonal panels the joint frequencies exhibit strong linear correlations. Moreover, the linear relationship suggests that the algorithms' frequencies adjust toward best responding to the frequencies of their human opponents.

These notions are quantitatively expressed in Table 2. In this table we report for each scatterplot the correlation coefficient and a hypothesis test of whether the coefficient is different than zero. For four of the five pure human experiment or pure simulation panels we fail to reject zero correlation at the five percent level of significance. However, we get a near opposite result in the human-algorithm experiments. We do reject zero correlation for four out of six of the human-algorithm cases. Moreover, for each of the human-algorithm experiments, the sign of the correlation coefficient is consistent with the algorithm better responding.

One example really highlights the result that the algorithms "better" respond—rather than best respond—to human play. Consider the upper right scatter plot of Figure 6. In this scatterplot, Column RE players face human Row players in the Gamble-Safe game. One of the human players chose his Minimax strategy, Right, exclusively and his computer RE opponent best responded to this only about 70 percent of the time. More striking is how all the observations in this panel, including this extreme observation, very closely align with the fitted line.

We further explore the idea that the learning algorithms are better than the human subjects at detecting and adjusting to exploitable play by presenting the OLS results of regressing the learning algorithms' Left frequencies on their human counterparts' Left frequencies.[16] A learning algorithm that is highly sensitive and adjusts systematically to opponents' play should generate regressions that explain a high percentage of the variance of the algorithm's Left frequencies, and the estimated slope coefficient should be consistent with the best response correspondence. These features are found in the Table 3 regressions: the slope of each regression has the correct sign, three

[16]When running these regression we are now assuming that there is a causality which we didn't assume in the correlation analysis. Given the consistency of the correlations with the direction of best response for the computer we feel that the assumption is not too egregious.

of the regressions have exceedingly large adjusted R^2 statistics, and a fourth is still quite large considering the data is cross sectional. These adjusted R^2 results reflect the tight clustering to the fitted regression line observed in the scatterplots. Correspondingly, F-tests for these four regressions do not reject the significance of the regressions at the 5 percent level of significance. Interestingly, the two cases where F-tests reject the regressions are when the EWA and RE algorithms assume the Column role in the Pursue-Evade game. We do not see a reason for the differential performance, but do note that the mean of the computers' data is close to their minimax strategy in this case.

To summarize, we see that the frequency of Left play by the learning algorithms moves toward (but not all the way to) best response, and the magnitude of these responses by the algorithms is described by a surprisingly predictable linear relationship. So can we conclude that the learning algorithms out play humans?

5.3 Learning Algorithms' Lack of Effective Exploitation

Previous arguments established that the learning algorithms sensitively detect opponents' exploitable action choice frequencies and then the algorithms respond with a systematic but tempered reaction in the direction of their best response. However, we will now see that these statistically significant responses are too weak in magnitude to generate statistically significant payoff gains. Table 4 presents the average stage game winnings for all decision maker types when pitted against a human for each role and game. If the learning algorithms successfully exploit human decision makers we would expect the algorithms in each game and role to have greater winnings than a human when playing against a human in the competing role. The average stage game winnings in Table 4 do not exhibit this trait.

The reported average stage game payoff statistics are calculated by first taking the total session payoffs for each decision maker who plays against a human, and dividing by the number of stage games played. Then we partition these decision makers according to the game played, role played, and decision maker type. Finally, we report the average stage game payoffs across decision makers in each partition. For each game and player role we conduct t-tests with the null hypothesis that on average a non-human decision maker earns the same as a human when the opponent is a human. At a five percent level of significance we fail to reject the null hypothesis in four out of the six tests. In the two rejections, the human average exceeds the algorithm average.

These two rejections merit closer inspection because it is tempting to conclude that human subject are able to "outsmart" the algorithms even though the learning algorithms appear to be better responding to the opponents more often over the course of the experiments. We show this is not the case, as the rejections arise from the combination of two other factors: 1. the timidness of the algorithm to best respond, as dictated by the probabilistic choice rule; and 2. these are the two cases where Human play differs depending on whether the opponent is an algorithm or another human.

Let's first consider the Gamble-Safe sessions with Human Column players, where we see Human Row players on average earn more than RE Row players. In Figure 9 we first graph a family of iso-expected payoff curves for the Row player as a function of the joint frequency Left play by the Row and Column player. Notice that the Row player expects a payoff of one whenever he plays Left with probability zero (i.e., he chooses the safe action) or the Column player plays his Nash equilibrium Left frequency of fifty percent. More importantly, whenever Column plays Left more than fifty percent of the time Row's expected payoff is bounded below by one, and whenever Column plays Left less than fifty percent of the time Row's expected payoff is bounded above by one.

Next we plot the joint frequencies of Left play for Human Row versus Human Column pairs (denoted with open circle marks) and the RE Row versus Human Column pairs (denoted with solid triangle marks). When facing Human Row players, seven of the fifteen Human Column players choose Left more than half the time and thus ensuring that their opponents expected payoff is greater than one. However, the RE Row players typically face less favorable opportunities. When facing RE Row players, nine out of the twelve Human Column players choose Left less than fifty percent of the time. In these nine instances, the only way the RE Row player can expect to achieve a payoff equal to the average Human Row payoff of .99 would be to exclusively best respond and always choose right. Of course, the probabilistic choice rule prevents consistent best responding.

Next we consider the Pursue-Evade sessions with Human Row players, where we see Human Column players on average earn more than EWA Column players. In Figure 10 we first graph a family of iso-expected payoff curves for the Column player as a function of the joint frequency of Left play by the Row and Column players. We see that the Column player expects to earn their minimax payoff of -.67 whenever either the Row or Column player plays their minimax Left

frequency of two-thirds. We have already observed that the EWA Column players earn less than the minimax payoff, which is roughly what the Human Column players earn, when facing Human Row players. This is also evident in Figure 10 when we look at the scatterplot of the joint frequencies of Left play for EWA Column versus Human Row pairs (denoted with solid triangle marks). Notice that for all of these pairs the Human Row player chooses Left less than two-thirds of the time, and in all but one case the EWA opponent best responds, by choosing Left, more than half the time.

So why are EWA decision makers earning less than their minimax payoff? From the formulation of the algorithm one sees that if the opponent plays his minimax strategy then both Left and Right will tend to have similar values overtime. Then the probabilistic choice rule leads the EWA algorithm to play Left and Right with equal frequency. Now as the opponent deviates away from his minimax strategy, the EWA algorithm will deviate from equal probable play towards the best response. This is what we observe, but in this instance the magnitude of the algorithm's choice frequency adjustment is so small that it fails to approach its minimax strategy of two-thirds. As we observe from the location of the level curves, by failing to best respond at least two-thirds of the time the EWA algorithm earns less than the minimax payoff. This surprising result is due to the fact that EWA doesn't assess payoffs relative to what it can achieve via its minimax strategy and that its probabilistic choice rule leads to weak adjustments that are based relative to the equiprobable mixed strategy.

At this point, the human-algorithm experiments have provided us with the identification of two previously unknown properties of the learning in games models: the models adjust linearly towards their best responses to human play, and the size of the adjustments are are extremely weak. Now we turn our attention to the question of what the human-algorithm experiments tell us about human subject play.

5.4 Human Play Conditional On Opponent Decision Maker Type

Past studies have demonstrated that humans play differently against Nash equilibrium strategies than they do against other humans. However, we have also presented evidence that play by learning algorithms is more responsive to opponents' decisions than human play is. A natural question to ask is: do humans play differently against learning algorithms than they do against other humans? To answer this question we compare the empirical distributions of the proportions of Left play

by humans when facing the different decision-making types as presented in the scatter plots of Figures 6–8. We report a series of Kolmogorov-Smirnov two-sample goodness-of-fit tests (hereafter denoted KS) comparing the distributions of Human Left play proportions when facing human opponents to Human Left play proportions when facing the alternative algorithms. The main result is that we do not observe differences in human play except in two cases: when the human is the Row player in the Pursue-Evade game and when the human is the Column player in the Gamble-Safe game. These are the same two cases we just discussed for which subjects out-earned algorithms.

Figure 11 shows the empirical CDFs of proportion of Left play by human Row players as they face human, RE, and EWA Column decision maker types in the Pursue-Evade game. Additionally, the figure reports the results of Kolmogorov-Smirnov tests of whether the Humans' distribution of Left play frequencies differs when facing an algorithm opponent as opposed to a human opponent. Previously we have observed that the learning algorithms performed differently in the Column role of the Pursue-Evade game than in any other situation. This trend continues as the proportions of Left by humans in the Row role are significantly different when facing each learning algorithm than when facing another human.

Next we consider the CDFs generated by human Column players when playing against Human, RE, and EWA Row decision maker types in the Pursue-Evade game. We see in Figure 12 that play against human opponents is statistically indistinguishable from play against both EWA and RE opponents.

Next, we turn our attention to human play in the Gamble-Safe game. Figure 13 shows that human Row players' CDFs of proportion of Left play are not statistically different as they face Human and RE Column decision maker types. Finally, the CDFs and associated KS tests generated by human Column players in the Gamble-Safe game are shown in Figure 14. We see that play against human opponents differs from play against RE opponents at the six-percent level of significance.

6 Discussion

Through experiments in which humans play games against computer-implemented learning algorithms, we have established that humans do not detect nor exploit the non-stationary but rather

inert mixed strategy processes of the RE and EWA algorithms. Our experiments also establish that the learning models are more sensitive than humans in detecting exploitable opponent play. Furthermore, our experiments reveal that the learning algorithms' action choice frequencies respond uniformly and linearly to opponents' non-equilibrium action choice frequencies. However, the corresponding mixed strategy adjustments of the learning models to detected exploitable play are too weak to increase their payoffs.

Our results, in conjunction with those of other studies, reveal a different depiction of human learning in games than those suggested by currently proposed models of adaptive behavior. First, through the technique of pitting humans against algorithms we know that humans successfully increase their payoffs (but not as much as possible) against non-optimal but stationary mixed strategy play and against adaptive play that generates highly serially correlated action sequences. On the other hand humans do not exploit the subtle dynamic mixed strategy processes of the learning models examined in this paper.

Some sources of behavioral departure between learning models and humans are identified in experiments that elicit subjects' beliefs (Nyarko and Schotter, 2002) or subjects' mixed strategies (Shachat, 2002). Elicited beliefs are highly volatile and often times correspond to a belief that one action will be chosen with certainty. Similarly elicited mixed strategies show erratic adjustments and a significant amount of pure strategy play.

This set of stylized facts establishes benchmarks which new learning models should explain. Furthermore, the use of human-algorithm interactions can play an important role in future efforts to identify how humans adapt in strategic environments. First, the technique brings increased power in evaluating proposed models and overcomes some current econometric and numerical limitations. Second, this technique can be used to identify human learning behavior through the adoption of carefully selected algorithms and the subsequent measurement of human responses to these algorithms. For example, one could determine the extent to which humans can exploit serially correlated strategies by adjusting the level of variance incorporated in the probabilistic choice rule of a cautious fictitious play algorithm; or one could determine human ability to detect and exploit non-minimax mixed strategies by systematically varying the computer's mixed strategy across opponents in a matching pennies game. In these instances, the algorithms are not being evaluated but rather used as carefully chosen stimuli to generate informative measurements of

human behavior.

References

Andreoni, J. A., Miller, J. H., 1993. Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence. Econ. J. 103, 570–585.

Brown, G. W., 1951. Iterative Solutions of Games by Fictitious Play. In: Koopmans, T. C. (Ed.), Activity Analysis of Production and Allocation, John Wiley.

Camerer, C., Ho, T., Chong, J.-K., 2002. Sophisticated Experience-Weighted Attraction Learning and Strategic Teaching in Repeated Games. J. Econ. Theory 104, 137–188.

Camerer, C., Ho, T., Chong, J.-K., 2003. Models of Thinking, Learning, and Teaching in Games. Amer. Econ. Rev. 93, 192–195.

Camerer, C. F., Ho, T.-H., 1999. Experience-Weighted Attraction in Games. Econometrica 67, 827–874.

Cheung, Y.-W., Friedman, D., 1997. Individual Learning in Normal Form Games: Some Laboratory Results. Games Econ. Behav. 19, 46–76.

Cooper, R., DeJong, D. V., Forsythe, R., Ross, T. W., 1996. Cooperation without Reputation: Experimental Evidence from Prisoner's Dilemma Games. Games Econ. Behav. 12, 187–218.

Coricelli, G., 2001. Strategic Interaction in Iterated Zero-Sum Games. Technical Report 01–07, University of Arizona.

Erev, I., Roth, A. E., 1998. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. Amer. Econ. Rev. 88, 848–881.

Erev, I., Roth, A. E., 2001. Bounded Rationality: The Adaptive Toolbox, MIT Press, Cambridge, MA, chapter Simple reinforcement learning models and reciprocation in the prisoners dilemma game, 215–231.

Feltovich, N., 2000. Reinforcement-Based vs. Beliefs-Based Learning Models in Experimental Asymmetric-Information Games. Econometrica 68, 605–642.

Fox, J., 1972. The Learning of Strategies in a Simple, Two-Person Zero-Sum Game without Saddlepoint. Behavioral Science 17, 300–308.

Fudenberg, D., Levine, D., 1995. Consistency and Cautious Fictitious Play. The Journal of Economic Dynamics and Control 19, 1065–1089.

Gjerstad, S., 1996. The rate of convergence of continuous play. Econ. Theory 7, 161–178.

Hanaki, N., Sethi, R., Erev, I., Peterhansl, A., 2005. Learning strategies. J. Econ. Behav. Organ. 56, 523–542.

Houser, D., Kurzban, R., 2002. Revisiting Confusion in Public Good Experiments. Amer. Econ. Rev. 92, 1062–1069.

Jordan, J. S., 1993. Three Problems in Learning Mixed-Strategy Nash Equilibria. Games Econ. Behav. 5, 368–386.

Lieberman, B., 1962. Experimental Studies of Conflict in Some Two-Person and Three-Person Games. In: Criswell, J. H., Solomon, H., Suppes, P. (Eds.), Mathematical Methods in Small Group Processes, Stanford University Press, 203–220.

McCabe, K., Houser, D., Ryan, L., Smith, V., Trouard, T., 2001. A functional imaging study of cooperation in two-person reciprocal exchange. Proceedings of the National Academy of Sciences, U.S.A. 98, 11832–11835.

Messick, D. M., 1967. Interdependent Decision Strategies in Zero-Sum Games: A Computer-Controlled Study. Behavioral Science 12, 33–48.

Mookherjee, D., Sopher, B., 1994. Learning Behavior in an Experimental Matching Pennies Game. Games Econ. Behav. 7, 62–91.

Mookherjee, D., Sopher, B., 1997. Learning and Decision Costs in Experimental Constant Sum Games. Games Econ. Behav. 19, 97–132.

Morgan, J., Sefton, M., 2002. An Experimental Investigation of Unprofitable Games. Games Econ. Behav. 40, 123–146.

Mukherji, A., Runkle, D. E., 2000. Learning to Be Unpredictable: An Experimental Study. Fed. Reserve Bank Minneapolis Quart. Rev. 24, 14–20.

Noussair, C., Willinger, M., 2003. Efficient mixing and unpredictability in an experimental game. Technical report, Emory University, unpublished manuscript.

Nyarko, Y., Schotter, A., 2002. An Experimental Study of Belief Learning Using Elicited Beliefs. Econometrica 70, 971–1005.

O'Neill, B., 1987. Nonmetric Test of the Minimax Theory of Two-Person Zerosum Games. Proceedings of the National Academy of Sciences, U.S.A. 84, 2106–2109.

Robinson, J., 1951. An Iterative Method of Solving a Game. Annals of Mathematics 54, 296–301.

Rosenthal, R. W., Shachat, J., Walker, M., 2002. Hide and Seek in Arizona. Int. J. Game Theory 32, 273–293.

Roth, A. E., Schoumaker, F., 1983. Expectations and Reputations in Bargaining: An Experimental Study. Amer. Econ. Rev. 73, 362–372.

Salmon, T. C., 2001. An Evaluation of Econometric Models of Adaptive Learning. Econometrica 69, 1597–1628.

Shachat, J., 2002. Mixed Strategy Play and the Minimax Hypothesis. J. Econ. Theory 104, 189–226.

Shachat, J., Swarthout, J. T., 2004. Do we detect and exploit mixed strategy play by opponents? Math. Methods Operations Res. 59, 359–373.

Sonsino, D., Sirota, J., 2003. Strategic Pattern Recognition – Experimental Evidence. Games Econ. Behav. 44, 390–411.

Spiliopoulos, L., 2008. Humans versus computer algorithms inrepeated mixed strategy games. Technical Report 6672, Munich Personal RePEc Archive.

Walker, J. W., Smith, V. L., Cox, J. C., 1987. Bidding Behavior in First Price Sealed Bid Auctions: Use of Computerized Nash Competitors. Econ. Letters 23, 239–244.

Winter, E., Zamir, S., 1997. An Experiment with Ultimatum Bargaining in a Changing Environment. Technical Report 159, The Hebrew University of Jerusalem, center for Rationality and Interactive Decision Theory.

Table 1: Sample size for each treatment.

	Opponent type		
Game	Human	EWA	RE
Persue-Evade	40	30	30
Gamble-Safe	30	0	24

Table 2: Correlation coefficients of pairwise joint densities of proportion Left for each role combination and game.

Game	Row Player	Column Player	Correlation Coefficient	T-Statistic	Degrees of Freedom	P-Value
Gamble-Safe	Human	Human	-0.379	-1.533	14	0.148
Gamble-Safe	Human	RE	-0.982	-17.441	11	0.000
Gamble-Safe	RE	Human	0.928	8.285	11	0.000
Gamble-Safe	RE	RE	0.270	1.220	19	0.237
Pursue-Evade	Human	Human	0.383	1.805	19	0.087
Pursue-Evade	Human	RE	-0.338	-1.345	14	0.200
Pursue-Evade	RE	Human	0.930	9.450	14	0.000
Pursue-Evade	RE	RE	0.490	2.452	19	0.024
Pursue-Evade	Human	EWA	-0.314	-1.237	15	0.237
Pursue-Evade	EWA	Human	0.581	2.674	14	0.018
Pursue-Evade	EWA	EWA	0.200	0.891	19	0.384

Table 3: OLS regression results, algorithm left frequency $= \alpha + \beta \times$ human left frequency.

Game	Algorithm	Algorithm role	Human role	α (t-stat)	β (t-stat)	Adjusted R-squared	F statistic (p-value)
Gamble-Safe	RE	Row	Column	0.07	0.66	0.85	62.40
				(2.11)	(7.90)		(0.00)
Gamble-Safe	RE	Column	Row	0.75	-0.69	0.96	276.54
				(40.03)	(-16.63)		(0.00)
Persue-Evade	RE	Row	Column	-0.26	1.16	0.85	82.92
				(-2.89)	(9.11)		(0.00)
Persue-Evade	RE	Column	Row	0.72	-0.21	0.05	1.68
				(9.40)	(-1.30)		(0.22)
Persue-Evade	EWA	Row	Column	0.28	0.29	0.29	6.64
				(3.24)	(2.58)		(0.02)
Persue-Evade	EWA	Column	Row	0.69	-0.20	0.03	1.42
				(8.85)	(-1.19)		(0.25)

Table 4: Average stage game winnings for human and algorithm decision makers when facing a human opponent. Each t-test compares an algorithm with human decision makers for a given game and role.

Game	Decision maker type and role	Decision maker average earnings	T test statistic	Approx. d.o.f.	P-value
Gamble-Safe	Human Column	1.0776	***	***	***
Gamble-Safe	RE Column	1.0786	-0.012	23	0.990
Gamble-Safe	Human Row	0.9888	***	***	***
Gamble-Safe	RE Row	0.8983	2.187	25	0.038
Pursue-Evade	Human Column	-0.6709	***	***	***
Pursue-Evade	RE Column	-0.6829	0.498	32	0.622
Pursue-Evade	EWA Column	-0.7205	2.312	33	0.027
Pursue-Evade	Human Row	0.6709	***	***	***
Pursue-Evade	RE Row	0.6395	1.285	31	0.208
Pursue-Evade	EWA Row	0.6395	1.557	32	0.129

Lieberman

	E1 (.25)	E2 (.75)
S1 (.75)	3	-1
S2 (.25)	-9	3

Messick

	A (.556)	B (.244)	C (.2)
a (.400)	0	2	-1
b (.111)	-3	3	5
c (.489)	1	-2	0

Fox

	a1 (.426)	a2 (.574)
b1 (.214)	6	-5
b2 (.786)	-2	1

Coricelli (Introduced by O'Neill)

	G (.2)	R (.2)	B (.2)	P (.4)
G (.2)	-5	5	5	-5
R (.2)	5	-5	5	-5
B (.2)	5	5	-5	-5
P (.4)	-5	-5	-5	5

Figure 1: Zero-sum games used in previous studies. Humans are row player, payoffs are for row player, and minimax strategy proportions are next to action names.

		Column Player: Left	Column Player: Right
Row Player	Left	1 , -1	0 , 0
Row Player	Right	0 , 0	2 , -2

Figure 2: The Pursue-Evade game.

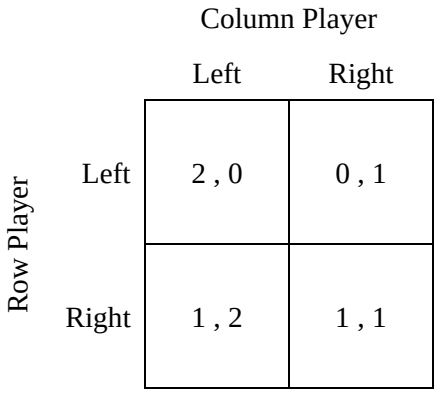


Figure 3: The Gamble-Safe game.

All Human Baseline

	Left	Right
Left	**.441** (.444)	**.169** (.222)
Right	**.276** (.222)	**.114** (.111)

Estimated RE Model

	Left	Right
Left	**.502** (.444)	**.161** (.222)
Right	**.208** (.222)	**.129** (.111)

Estimated EWA Model

	Left	Right
Left	**.442** (.444)	**.204** (.222)
Right	**.239** (.222)	**.110** (.111)

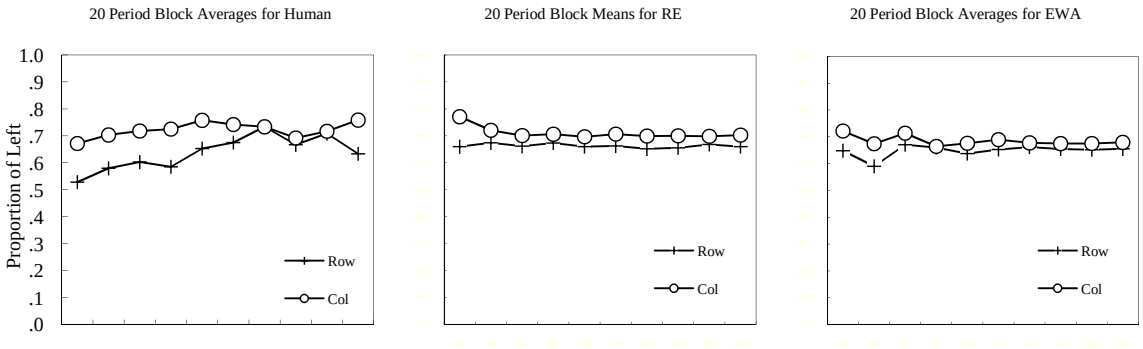


Figure 4: Baseline data and model simulation summary for the Pursue-Evade Game.

All Human Baseline

	Left	Right
Left	**.212** (.250)	**.220** (.250)
Right	**.309** (.250)	**.259** (.250)

Estimated RE Model

	Left	Right
Left	**.226** (.250)	**.215** (.250)
Right	**.297** (.250)	**.262** (.250)

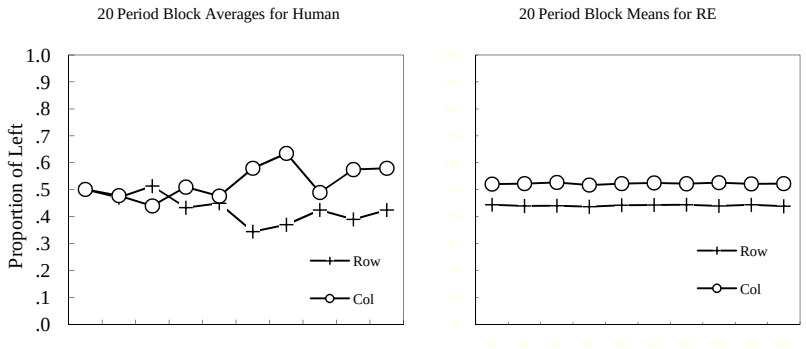


Figure 5: Baseline data and model simulation summary for the Gamble-Safe game.

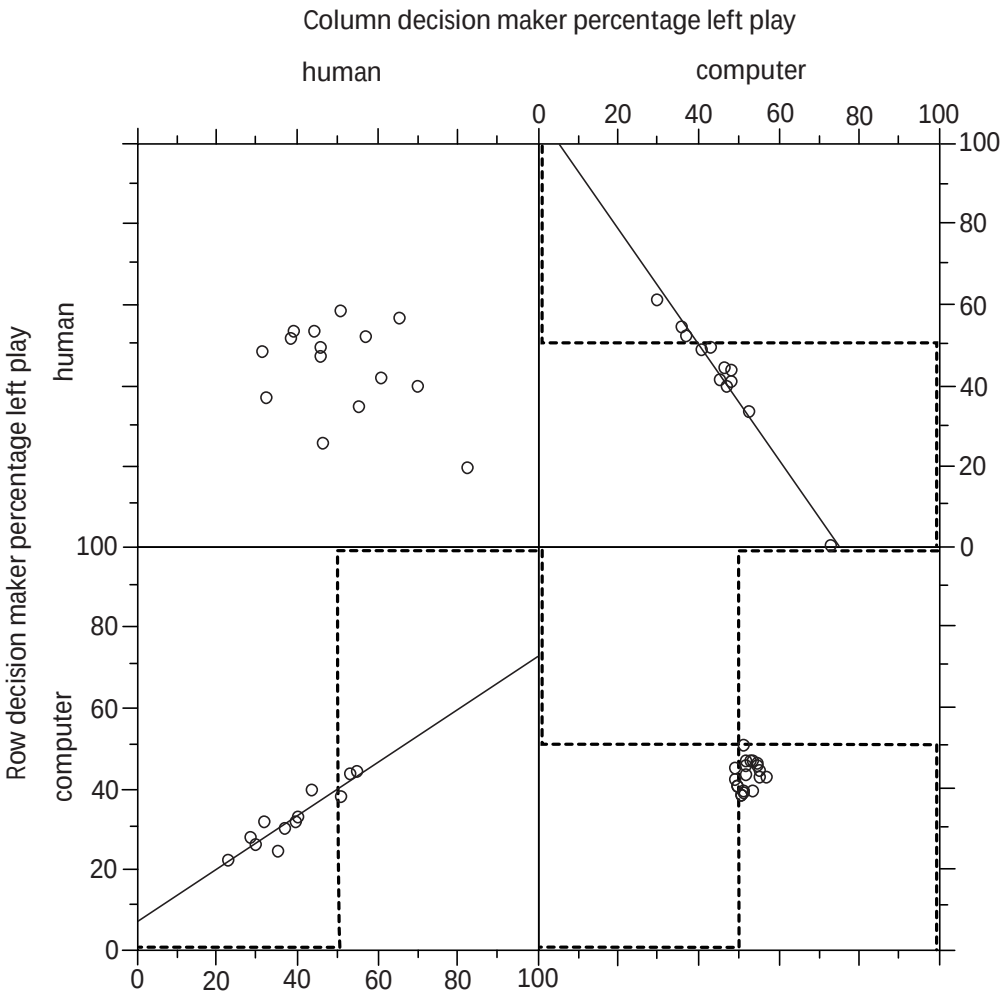


Figure 6: Pairwise joint densities of proportion Left in the Gamble-Safe game, conditioned on reinforcement algorithm interaction. Dashed lines represent computer best-response correspondences.

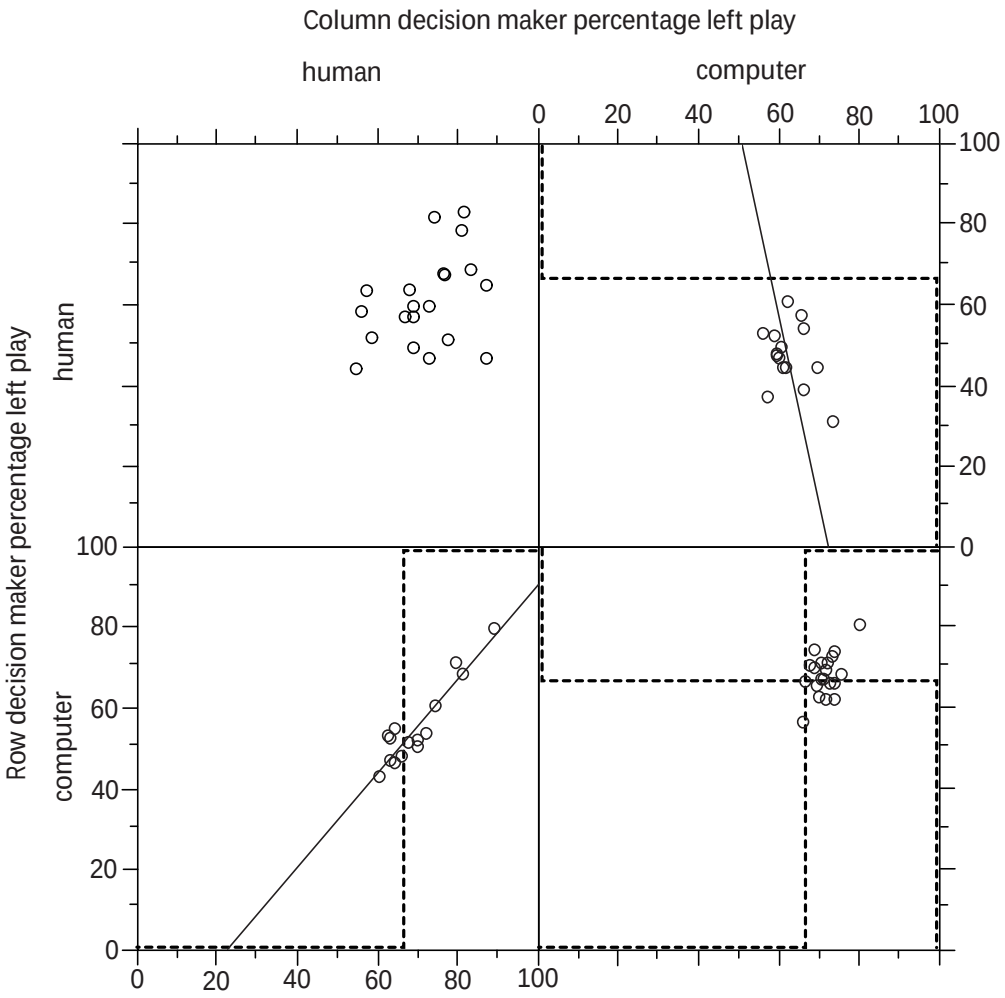


Figure 7: Pairwise joint densities of proportion Left in the Pursue-Evade game, conditioned on reinforcement algorithm interaction. Dashed lines represent computer best-response correspondences.

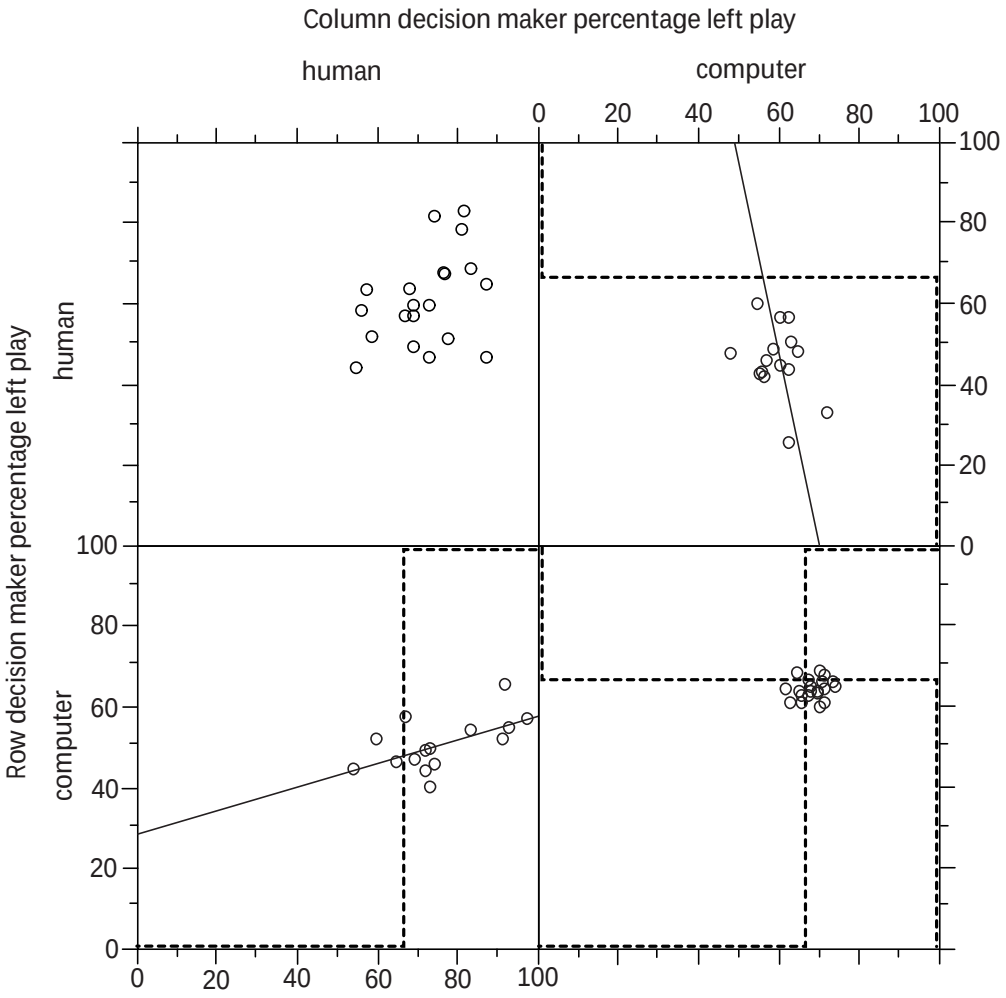


Figure 8: Pairwise joint densities of proportion Left in the Pursue-Evade game, conditioned on experience-weighted attraction algorithm interaction. Dashed lines represent computer best-response correspondences.

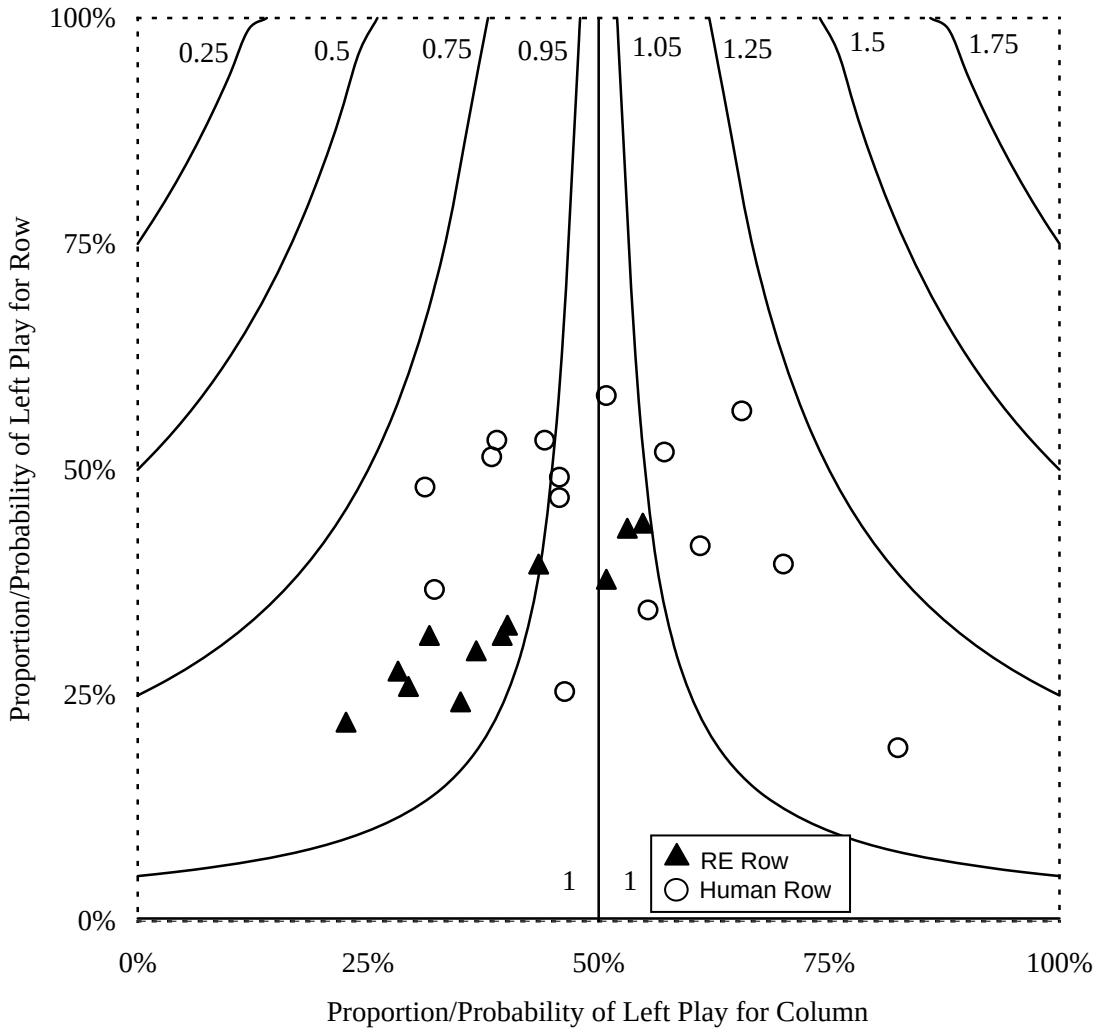


Figure 9: Row payoff contours and joint frequencies for RE and human Row players versus human Column Payers in Gamble-Safe game.

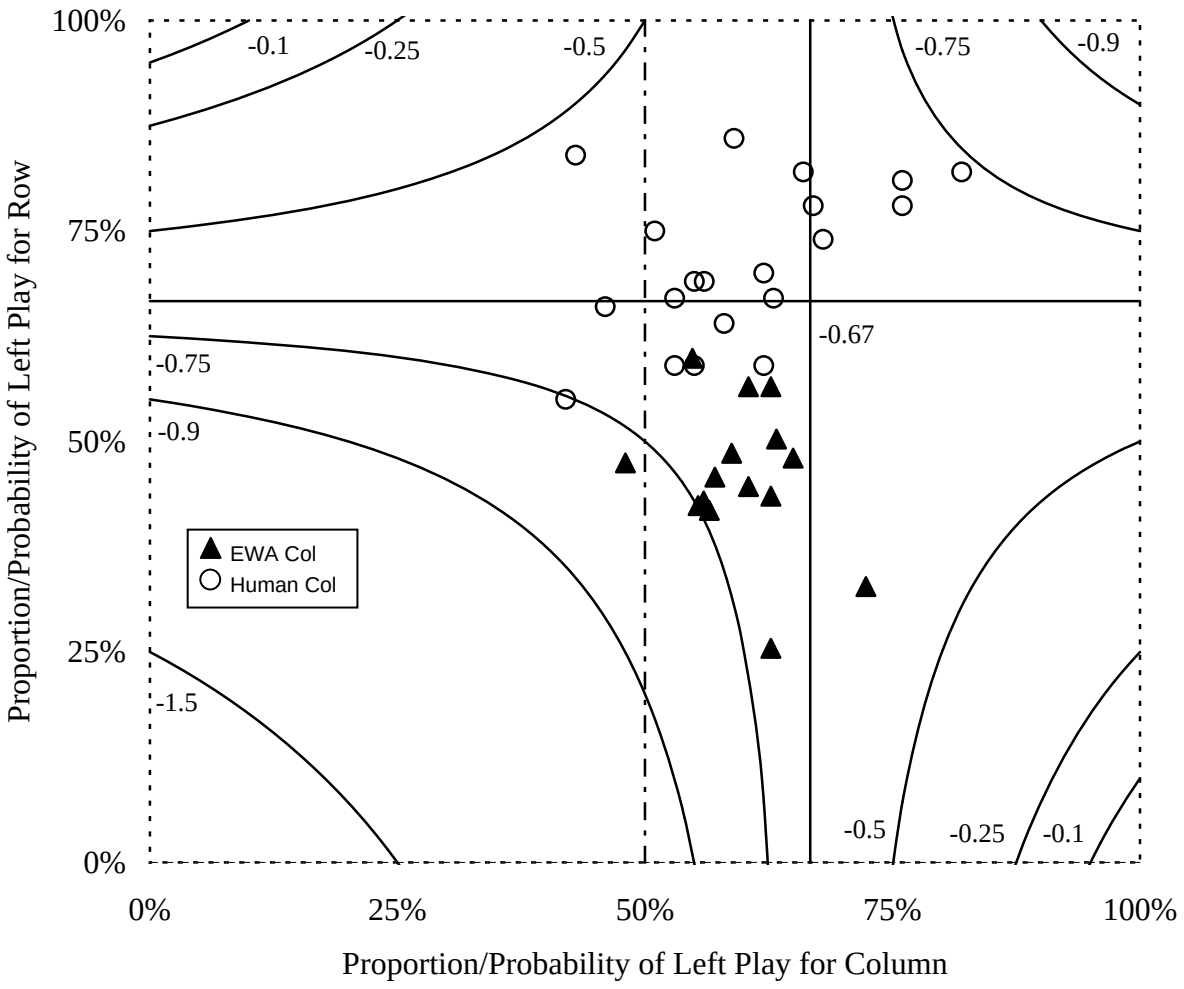
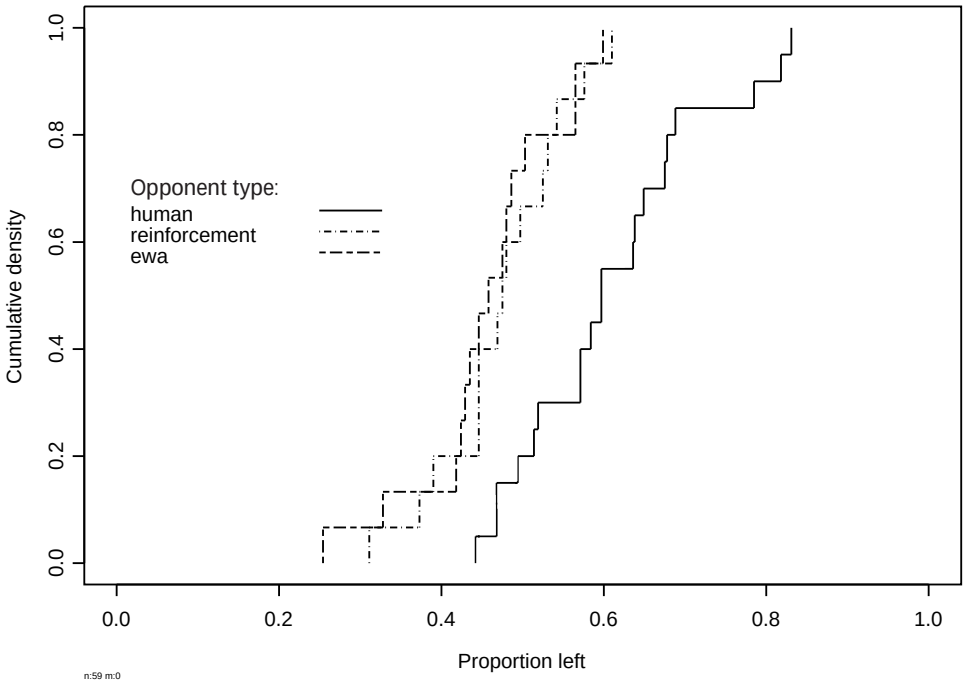
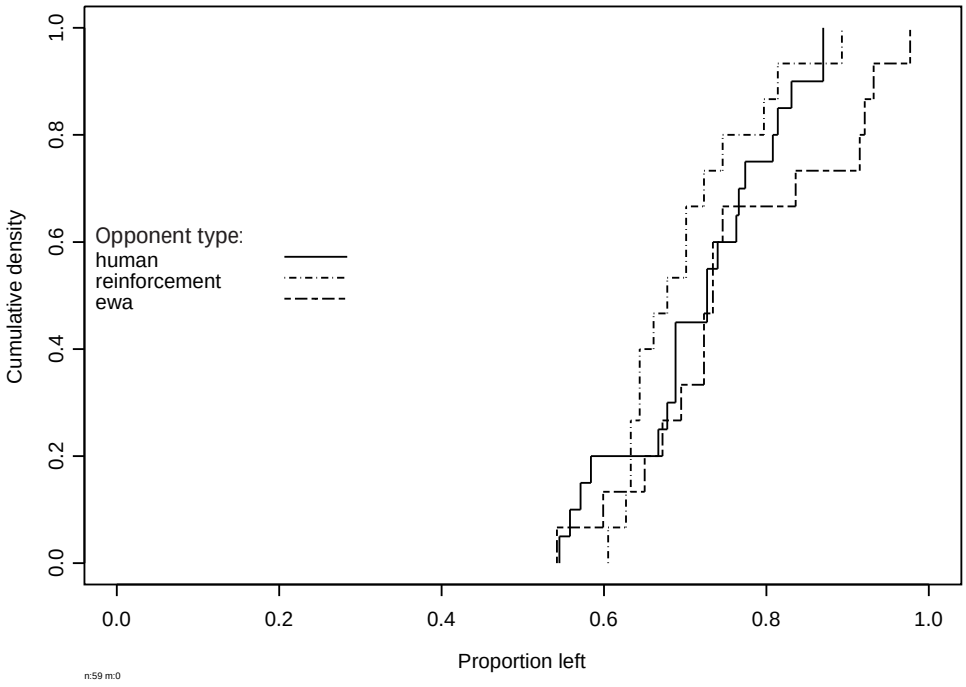


Figure 10: Column payoff contours and joint frequencies for EWA and human Column players versus human Row payers in Pursue-Evade game.



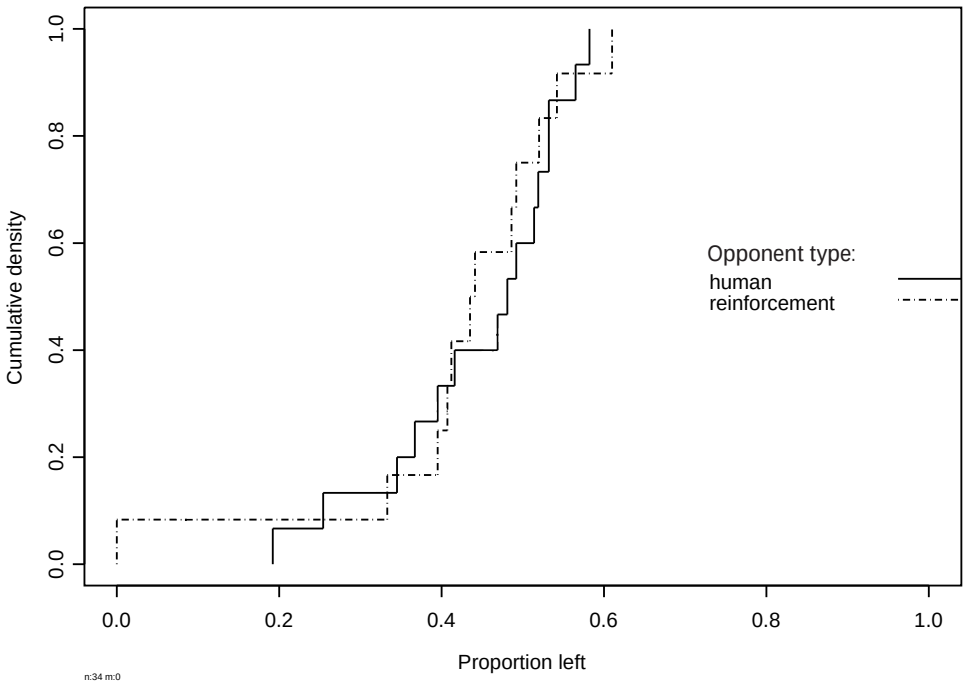
Dist. Left when facing Human tested against dist. Left when facing:	KS statistic	P-value
RE	0.567	0.005
EWA	0.633	0.001

Figure 11: Distributions of left play by human row players in the Pursue-Evade game. The Kolmogorov-Smirnov test is used to test for differences in distributions.



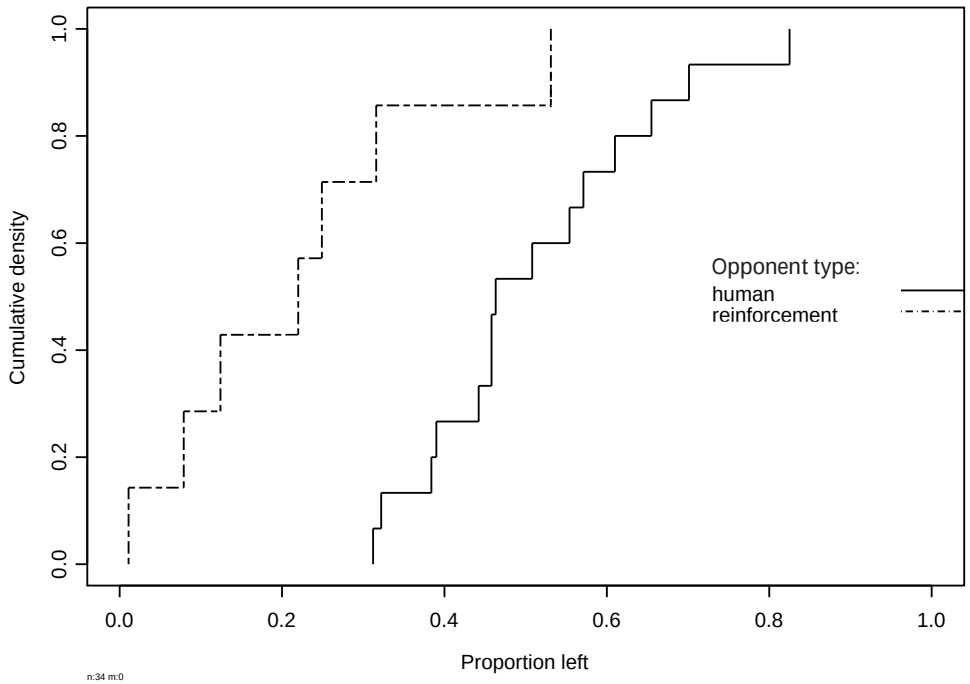
Dist. Left when facing Human tested against dist. Left when facing:	KS statistic	P-value
RE	0.283	0.435
EWA	0.267	0.507

Figure 12: Distributions of left play by human column players in the Pursue-Evade game. The Kolmogorov-Smirnov test is used to test for differences in distributions.



Dist. Left when facing Human tested against dist. Left when facing:	KS statistic	P-value
RE	0.183	0.952

Figure 13: Distributions of left play by human row players in the Gamble-Safe game. The Kolmogorov-Smirnov test is used to test for differences in distributions.



Dist. Left when facing Human tested against dist. Left when facing:	KS statistic	P-value
RE	0.483	0.061

Figure 14: Distributions of left play by human column players in the Gamble-Safe game, conditioned on opponent type. The Kolmogorov-Smirnov test is used to test for differences in distributions.