Learning about Learning in Games through Experimental Control of Strategic Interdependence

by

Jason Shachat

IBM TJ Watson Research Center

and

J. Todd Swarthout

University of Arizona

May, 2002

Abstract

We conduct experiments in which humans repeatedly play one of two games against a computer decision maker that follows either a reinforcement learning algorithm or an Experience Weighted Attraction algorithm. The human/algorithm interaction provides results that can't be obtained from the analysis of pure human interactions or model simulations. These learning algorithms are more sensitive than humans to exploitable opponent play. Learning algorithms respond to these calculated opportunities systematically; however, the magnitude of these responses are too weak to improve the algorithm's payoffs. Human play against various decision maker types does not vary significantly. These results demonstrate that humans and current models of their behavior differ in that humans do not adjust payoff assessments by smooth transition functions but when humans do detect exploitable play they are more likely to choose the best response to this belief.

1 Introduction

Identifying how humans respond and adapt their behavior in repeated strategic decision making tasks has emerged as a core, but difficult to answer, question in the social sciences. To address this question, most studies formulate alternative adaptive or "learning" models and estimate model parameters from human experimental data. These estimated models are either evaluated for goodness-of-fit by various statistical criteria or used to generate simulations which are subsequently compared to human experimental play. Unfortunately, definitive conclusions are difficult to achieve using this approach because current econometric techniques generate exceedingly high rates of Type I and Type II errors when evaluating alternative adaptive models of play (Salmon [21] (2001)).

One source of this difficulty is the nature of the problem. Learning in games is a multivariate stochastic process. One component of this process is a set of latent variables, such as beliefs about opponents or values associated with alternative actions, which each individual uses to select actions and adjusts according to game history. If play doesn't coincide with an equilibrium and players condition actions on the common observed history of play, strong interdependencies are likely to exist among the adjustment rules of the players' latent variables. When a researcher seeks to identify the rules underlying the interdependent latent processes, typically the only observable information is the sequence of realized choices from discrete action sets. Unfortunately, these interdependent and dynamic latent processes, and the resultant sequences of discrete choices are substantial obstacles for current econometric methods.

In this study we adopt a technique that exercises experimental control of strategic interdependence and enables us to gain greater insight into the rules humans use to adjust their play in games. We conduct hybrid experiments in which humans play simple 2×2 games against alternative computer-implemented learning algorithms. Each game has a unique Nash equilibrium in mixed strategies. This technique lets us directly control the nature of the dependence of one of the players. In turn this allows us to more accurately assess human response to opponents adopting particular learning rules. Furthermore, we are able to better evaluate the appropriateness and properties of alternative hypothesized models of human adaptive behavior in games.

A myriad of recently proposed learning models embody two common principles: smooth adjustment rules for values of actions and probabilistic choice. First, for each of the players' actions the models ascribe a latent variable which represents the value of the action. This value is updated after each play of the game according to an adaptive rule and the stage game outcome. Second, in each stage game a player selects his action according to a probabilistic choice rule. This probabilistic choice rule assigns higher probabilities to actions with greater latent values. Typically these models include unobservable parameters whose values are estimated from experimental data. Uniformly across studies, estimated parameter values specify adaptive rules that have significant memory and thus the incremental relative stage game impact on calculated action values is small. In turn, this leads to "smooth" adjustment rules: from period to period, action choices and resultant mixed strategies do not drastically vary.

For this study we adopt two prominent learning models of this type: Erev and Roth's [6] (1998) reinforcement learning model and Camerer and Ho's [2] (1999) experienced weighted attraction model. There are many other similarly structured models worth studying with our technique, but models in this class tend to generate similar play (Salmon [21]) and consequently most of the potential insights can be gained through the evaluation of just one or two models of this type.

We believe our technique reveals properties about both human learning and learning models which cannot be discovered through either pure human experimentation or pure simulation. We present the following summary of our main results. Human play does not significantly vary depending on whether the opponent is a human or a learning algorithm. In contrast, algorithm play markedly differs when playing against a human rather than an identical algorithm in a simulation. When humans' action frequencies deviate from their Nash equilibrium proportions, the algorithms' action choice proportions respond with *systematic adjustments* towards their pure strategy best responses. Adjustments of algorithms in response to their human opponents' play result in a strikingly linear relationship between the learning models' and humans' action frequencies. Moreover, the linear relationship is suggestive of the computer players' best response correspondence. While adjustments by the algorithms are remarkably regular, their linear nature produces quite *muted response*

magnitudes. In fact, magnitudes of the adjustments are too small to result in statistically significant gains in their game payoffs.

Our experimental study is just one of several that exploit laboratory control to better measure some of the latent variables underlying human play in games.[1] The composite finding of these studies paints a significantly different picture of human learning in games than the class of models considered by this study. Specifically, experiments with unique mixed strategy Nash equilibrium games have shown that humans' beliefs about opponent play are highly volatile from period to period (Nyarko and Schotter [16](2002)), and correspondingly players' mixed strategies exhibit significant variability with significant amounts of switching between pure and mixed strategy play (Shachat [22](2002)). Furthermore, humans are also quite successful at significantly increasing their payoffs when computerized opponents play either stationary non-equilibrium fixed mixed strategies (Lieberman [12](1961) and Fox [7](1972)) or highly serially correlated action sequences (Messick [14](1967) and Coricelli [4](2001)). In summary, human play is characterized by volatile beliefs, variable mixed strategy choices, and successful exploitation of some strategies. In contrast, the learning models we evaluate generate beliefs that are smooth, make only minor mixed strategy adjustments from period to period, and have an inability to take advantage of calculated payoff-increasing opportunities.

We proceed with a discussion of several past studies incorporating human versus computer game play. Then we present the two learning models adopted 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 currently proposed learning models.

[1]For example Camerer, Johnson, Rymon, & Sen [3](1993) and Crawford, Costa-Gomes, & Broseta [5](2000) studied information look-up patterns of subjects. Also, Nyarko & Shotter [16](2002) elicited subjects' beliefs of opponents' future actions.

2 Literature Review

In a number of past studies, researchers have used the technique of humans interacting with computerized decision makers. This technique has been used in various studies to identify social preferences in strategic settings (Houser and Kurzban [10] (2000), and McCabe et al [13] (2001)), to establish experimental control over player expectations (Roth and Shoumaker [20] (1983), and Winter and Zamir [24] (1997)), and to identify how humans play against particular strategies in games (as in Walker, Smith, and Cox [23] (1987)). In this section, we discuss the last type of study and summarize established results on how humans play against unique minimax solutions, non-optimal stationary mixed strategies, and variants of the fictitious play dynamic (with deterministic choice rules) in the context of repeated constant-sum games with unique minimax solutions in mixed strategies.

All of the studies we discuss incorporated fixed human-computer pairs playing repetitions of one of the zero-sum games presented in Table 4.[2] Studies by Lieberman [12] (1961), Messick [14] (1967), and Fox [7] (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.[3] All three of these studies reach the same conclusion: human 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 because when a "computer" adopts its minimax strategy the human's expected payoffs are equal for all of his actions.

However, this indifference is not present when the computer adopts non-minimax mixed strategies. Lieberman [12] and Fox [7] both studied human play against non-optimal stationary mixed strategies and discovered that humans do adjust their play to exploit (though not fully) their opponents. In the relevant Lieberman treatment, subjects played against the experimenter for a total of 200 periods. In the first 100 periods, the experimenter played

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

[3] When reported, humans 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 [14], page 35)

his minimax strategy of (.25, .75) and then in the final 100 periods the experimenter played a non-minimax strategy of (.5, .5). Humans players were not informed that their opponent had adjusted his strategy. Human play adjusted from best responding approximately 20 percent of the time right after the experimenter began non-minimax play, to best responding approximately 70 percent of the time by the end of the session. However, this experimental design made it difficult to differentiate between the attractiveness of the minimax strategy and the best response since they both lay in the direction of this observed shift.

In one of Fox's treatment, humans 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, (1, 0), on the opposite side of (.5, .5) from the human's minimax strategy, (.214, .786). Human play started slightly above (.5, .5) and then slowly adjusted towards the pure strategy best response over the course of the experiment. Specifically, humans were best responding approximately 75 percent of the time by the latter stages of the experiment. These experiments established that humans will adjust their behavior to take advantage (but not as much as possible) of exploitable stationary mixed strategies.

Messick [14](and Coricelli [4] (2001) conducted experiments to evaluate how humans respond when playing against variations of fictitious play.[4] These experiments are notable in that the computer's strategy was responsive to the actions selected by its opponent. Messick studied humans matched against two fictitious play algorithms: one with unlimited memory and the other with only a five period memory. Against unlimited memory fictitious play, humans earned substantially more than their minimax payoff level. Humans 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 [17] (1985)) in which humans play against unlimited memory fictitious play and against the same algorithm that has a bias in the beliefs that subjects tend not to repeat their "P" action. In both treatments humans win significantly more often against the algorithms than they do against humans.[5] Establishing that humans can "outgame" these algorithms is significant, though

[4]In the original formulations of fictitious play (Brown [1](1951) and Robinson [18](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.

[5]Human vs. human data for this conclusion are taken from O'Neill [17] (1985) and Shachat [22](2002).

not surprising. It is well known that in games with unique mixed strategy equilibrium, the fictitious play algorithm can generate strong positively serially correlated action choices that are easily exploited.[6] 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.[7]

To summarize, through the use of experiments pitting humans against algorithms in constant sum games with strictly mixed strategy solutions we have learned (1) that humans do not tend to play their minimax strategy in response to opponents playing their minimax strategy, (2) humans exploit (but not fully) opponents who play mixed strategies significantly different from their minimax strategy, and (3) humans exploit adaptive algorithms which generate highly serially correlated action choices.

3 Response Algorithms

A large number of adaptive behavioral models have been recently introduced into the literature on games. Most of these models have similar frameworks with two main components. First, each player retains a latent "score" for each of his available actions, and the score of each action is adjusted after each game iteration based on the outcome. Second, each player chooses an action according to a probability distribution that places higher probability on actions with higher scores. Unfortunately, for obvious reasons we must limit the number of models we consider. We focus on two of the more popular models in the experimental games literature: Erev and Roth's [6](1998) Reinforcement model and Camerer and Ho's [2](1999) Experience Weighted Attraction model.

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

[6]See Jordan [11](1993) and Gjerstad [9](1996).

[7]For example, see cautious fictitious play proposed by Fudenberg and Levine [8] (1995), and the two learning models we utilize in this study.

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 for the dynamical system are (1) that at the initial iteration, each of a player's actions has the same probability of being selected (i.e., in our two games, $\sigma_{ij}(1) = .5$ for each player i and each action j) and (b) that

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

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 vs. the unselected action; and $\pi_i(j, a_{-i}(t))$ is i 's payoff when he plays action j against the opponent's stage-t action $a_{-i}(t)$. (Player i's minimum possible payoff for any action profile, $\min\{X_i\}$, is subtracted from $\pi_i(j, a_{-i}(t))$ for normalization purposes and 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 $S(1), \phi$, 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 as presented in Camerer & Ho [2](1999). While the structure of the EWA formulation is similar to RE learning, 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

$$N(t+1) = \rho N(t) + 1 \quad \text{for} \quad t \geqq 1,$$

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

The 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 is generated when $N(1) = 0$, $\varepsilon = 0$, and $\rho = 0$; and probabilistic fictitious play is generated when $\varepsilon = \rho = \phi = 1$. [8]

[8] We refer the reader to Camerer and Ho [2](1999) for more discussion of how EWA can emulate various models and for a more complete interpretation of the parameters.

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 playing 100 or 200 rounds with 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 sampling of humans play one of the two games against an estimated learning algorithm. We proceed by describing the two games we used and then presenting more details on the outlined steps.

4.1 The Two Games

The first game we consider is a zero-sum asymmetric matching penny game called Pursue-Evade. This game was introduced by Rosenthal, Shachat, and Walker [19](2002) (hereafter RSW). The normal form representation of the game is given in Table 1. The Nash equilibrium (and minimax solution) of this game is symmetric: each player chooses Left with a probability of two-thirds.

There are several reasons why this game is a strong candidate to use in our study. (1) The zero-sum nature eliminates any 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. (2) 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. (3) Pursue-Evade is a simple game in which the Nash equilibrium predictions differ from equiprobable choice. This generates a powerful test against the alternative hypothesis of equiprobable play.

We selected our second game to pose a more serious challenge to the learning algorithms. We refer to our second game, presented in Table 2, 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. The difference between the Nash

equilibrium and the minimax solution makes this game challenging for the learning models. 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 whose minimax and Nash equilibrium solutions differ but generate the same expected payoff is called a non-profitable game.[9] The potential attraction of the minimax strategy can (and does) prove to be difficult for the learning algorithms which, loosely speaking, have best response flavors.

4.2 Protocols

4.2.1 Human vs. Human Baselines

For the human vs. human 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 player was given two index cards: one labelled Left and the other labelled Right. At each iteration the players slid the chosen card to the experimenter seated across from them. 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. All subjects participated in either 100 or 200 repetitions of the game maintaining a constant role throughout.

The protocol for each period was as follows. At the beginning of each repetition, a subject saw a graphical representation of the game on the screen. Column players had the display of their game transformed so that they appeared to be a Row player. Thus, all subjects selected an action by clicking on a row, and then confirming their selection. Subjects were free to change row selections before confirmation. Once an action was confirmed, a subject waited until his opponent also confirmed an action. Then, a subject saw the resultant

[9]Morgan and Sefton [15, (2002)] present an excellent study investigating human play in Non profitable games.

outcome highlighted on the game display, as well as a text message stating both actions and the subject's earnings for that repetition. Finally, at all times a history of past play was displayed to the subjects. 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 vs. Algorithm Treatments

We conducted our hybrid treatments using both the computer program and protocol used for the Gamble-Safe game baseline.[10] 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 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 twenty-three stage games.

The non-human opponent treatments began with an initial stage of human versus human play in order to give the algorithms a better chance of successfully "standing in" for the human whose place it will take. Both RE and EWA rely on actions' scores to determine the chosen action in a probabilistic manner. During the first 23 repetitions, we allow these scores to "prime" themselves with the play generated by the subjects. (Although the updating of the scores is determined by the parameter estimates obtained from the baseline treatments). That is, even though the response algorithms are not selecting actions during the first 23 repetitions, the scores are still being updated according to the specifications of the previous section. For example, consider the 24th repetition of a game. The human Row player is now facing a computer that is playing the Column position. Moreover, during the first 23 repetitions, the computer Column player has been updating the scores associated with Column's actions based on the observed actions of both humans. We conjecture this will

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

de-emphasize the impact of the estimated initial score values of the actions.

In summary, we have two treatment variables, the stage game and the type of opponent. The data samples we have for each treatment cell are given in Table 3.[11]

5 Baseline Results, Model Estimation, and Model Simulation

Our experimental baselines are the human versus human play in each of the games we consider. 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. The 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 1 shows contingency tables for the data aggregated across subject pairs and stage games. A graph of the time series of the average proportion of Left play for the Row and Column players is shown below each table. Each observation in a series is the average across a twenty 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.[12] 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

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

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

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 1. 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 2. 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.

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

Since the selection of parameters for 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 remarkably 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 the occurs when some players adopt the minimax strategy and other players adopt adaptive strategies.

The 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 much smoother and less extreme 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.

6 Analysis of Human/Algorithm Interaction

In the previous section we used a common technique of comparing experimental data to simulation results to evaluate the appropriateness of alternative learning models. Now we proceed to present analysis of human/algorithm interaction which reveals a significantly different story. The action choice frequencies by the algorithms are more responsive to opponents' play than the humans' action choice frequencies. Moreover, the action frequencies by each algorithm adjusts linearly toward its best response to its opponent's non-equilibrium action frequencies. However, the magnitude of these adjustments is too small to generate payoff gains for the learning algorithms. Finally, we see that human play does not vary significantly whether the opponent is another human or a learning algorithm. Examination of the human/human experiments and the model simulations don't reveal these results.

6.1 Learning Algorithm Response to Opponents' Play

We now introduce pair-level data to better highlight differences in play across treatments. Inspection of the Row and Column players' proportions of Left play in each pair reveals surprising differences from purely human play and the simulations reported in the prior section. The learning algorithms are quite responsive to human deviations from Nash equilibrium play. Specifically, the algorithms' frequencies of Left play have a striking linear correlation to their human opponents' Left play proportions. Moreover, these linear relationships are consistent with a linear approximation of the algorithms' best response correspondences.

These results are most easily seen in Figures 3 - 5. Each of these figures is a 2×2 array of scatterplot panels. The rows in the panel array correspond to the decision maker type for the Row player: the top row indicates human decision maker and the bottom row indicates computer decision maker. Similarly the columns of the panel array correspond to the decision maker type for 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 and algorithm/human experiments.

The scatterplots are of 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 of these scatterplots displays a trend line and a dashed line for the computer's best response correspondence.

Examination of these figures reveals important common results across the two games and learning models. Comparisons between the two main diagonal panels reveal consistent differences and similarities between human/human play and pure simulations of model interaction. Both types of interactions generate uncorrelated "clouds" with the simulations' clouds exhibiting much smaller dispersion.[13] This raises the issue of whether the learning models are quite aggressive in adaptation and quickly converge to an equilibrium or instead the models are quite insensitive to opponents' play and just stubbornly mimic human aggregate frequencies. We can ask a similar question regarding human play. Do the humans' dispersed clouds result from high variance in the humans' propensities to play Left coupled with little response to the opponents' play or is it the result of differential skill in human play in which some humans more successfully exploit other humans' play?

Inspection of the human and learning algorithm interactions answers these questions. In contrast to the model simulations and human/human play, the scatter plots of human and learning algorithm interactions (found in the off-diagonal panels of Figures 3 - 5) exhibit strongly correlated interactions. This is evident by the tight clustering of the data around the plotted regression lines. Also, in each case the regression line is in the direction of the computer players' best response correspondence (the dashed correspondence given on each scatterplot). In other words, the computer "better" responds instead of best responds. This is best illustrated by an observation in the upper right scatter plot of Figure 3. In this scatterplot, Column RE players play against 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. Hence, we see that

[13]F-tests reject the significance of the presented regression lines; this gives statistical support for claims of no correlation.

(1) the frequency of Left by the learning algorithms move toward (but not all the way to) the best response to their opponents' frequencies, and (2) the magnitude of these responses by the algorithms is a surprisingly predictable linear relationship.

Table 5 gives some quantitative support for these observations by presenting the OLS results of regressing the learning algorithms' Left frequencies on their human counterparts' Left frequencies. 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 5 regressions: the slope of each regression has the correct sign, three 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 and correspondingly the detection and systematic reaction by the learning algorithms to calculated payoff-increasing opportunities. 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.

6.2 Learning Algorithms' Lack of Effective Exploitation

Previous arguments established that the learning algorithms are quite sensitive in detecting opponents' exploitable action choice frequencies and then the algorithms respond with a systematic but tempered best response. However, we will now see that these statistically significant responses are too weak in magnitude to generate statistically significant payoff gains. Table 6 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 6 do not exhibit this trait.

The reported average stage game winning statistics are calculated by first taking the total session winnings for each decision maker who plays against a human, and dividing by the number of stage games played.[14] Then we partition these decision makers according to the game played, role played, and decision maker type. Finally, we report the average stage game earnings 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 5 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.

Why don't the learning algorithms, which are sensitive and reactive to opponent play, generate higher payoffs than humans? The answer is twofold. First, the two games we consider have fairly flat payoff spaces in the mixed strategy domains presented in Figures 3 - 5. Thus a pair must be far removed from the Nash equilibrium to generate large payoff deviations from Nash equilibrium payoffs. Second, whenever the algorithm calculates a difference between its two action scores, it adjusts choice probabilities without assessing whether this difference is statistically significant. If this difference is not statistically significant, then there is no adjustment that can generate a real increase in payoff. Alternatively, an adjustment to a statistically significant score difference may also fail to generate a real increase in payoffs. Why? We have already seen that algorithms adjust in statistically significant ways, but these adjustments are relatively small in magnitude. These weak adjustments are the product of probabilistic choice rules, which were adopted to avoid generating transparent serially correlated choice patterns.

6.3 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 also have presented arguments that suggest learning algorithms' play is more responsive to opponents' decisions than human play is. A natural question to ask is, do humans play differently against learning algorithms than

[14]We normalize this way because in the baseline data for Pursue-Evade and Gamble-Safe some pairs played 100 stage games and others 200.

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 3 - 5. We report a series of Kolmogorov-Smirnov two-sample goodness-of-fit tests (hereafter denoted KS) comparing the distributions of Left play proportions against human opponents to Left play proportions against the alternative algorithms. The main result is that we can't find differences in human play except in the case when the human is the Row player in the Pursue-Evade game.

Figure 6 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 Human's 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 7 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 8 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 9. We see that play against human opponents differs from play against RE opponents at the six-percent level of significance.

7 Discussion

Through experiments in which humans play games against computer- implemented learning algorithms, we have established that humans do not detect nor exploit the estimated models'

non-stationary but rather smooth mixed strategy processes. Furthermore, our experiments provide a unique evaluation of the learning models by establishing that the models are more sensitive than humans in detecting exploitable opponent play. However, the models' corresponding mixed strategy adjustments are systematic but too weak to increase their payoffs.

Recall the common formulation of both the RE and EWA models. We see their adaptive functions generate sequences of action scores which adjust smoothly across periods because stage game outcomes weakly impact action scores. Furthermore, our experiments reveal that the learning algorithms' mixed strategies respond uniformly and linearly to opponents non-equilibrium action choice frequencies. The algorithms' uniform better responses are too weak to generate significant payoff gains.

Our study, in conjunction with other studies, reveals a different depiction of human learning in games. 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 [16]) or subjects' mixed strategies (Shachat [22]). Elicited beliefs are highly volatile and often times correspond to a belief that one action will be chosen with certainty. Similarly elicited mixed strategies will show erratic adjustments and a significant amount of pure strategy play.

This set of stylized facts should set benchmarks which new learning models need to 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. Second, the adoption of carefully selected algorithms will facilitate further identification of human learning behavior. For example, one could determine the extent of human ability to exploit serially correlated strategies by altering the variance incorporated in the probabilistic choice rule of a cautious fictitious play algorithm. In this instance, the algorithm is not being evaluated: rather it is

a carefully chosen stimulus to yield informative measurements of human behavior.

References

[1] G. W. Brown. Iterative solutions of games by fictitious play. In T. C. Koopmans, editor, *Activity Analysis of Production and Allocation*. John Wiley, 1951.

[2] Colin F. Camerer and Teck-Hau Ho. Experience-weighted attraction in games. *Econometrica*, 67:827–874, 1999.

[3] Colin F. Camerer, Eric J. Johnson, Talia Rymon, and Sankar Sen. Cognition and framing in sequential bargaining for gains and losses. In Ken Binmore, Alan Kirman, and Piero Tani, editors, *Frontiers of Game Theory*. MIT Press, 1993.

[4] Giorgio Coricelli. Strategic interaction in iterated zero-sum games. Technical Report 01–07, University of Arizona, 2001.

[5] Miguel Costa-Gomes, Vincent Crawford, and Bruno Broseta. Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69:1193–1235, 2001.

[6] Ido Erev and Alvin E. Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, 88:848–881, 1998.

[7] John Fox. The learning of strategies in a simple, two-person zero-sum game without saddlepoint. *Behavioral Science*, 17:300–308, 1972.

[8] Drew Fudenberg and David Levine. Consistency and cautious fictitious play. *The Journal of Economic Dynamics and Control*, 19:1065–1089, 1995.

[9] Steven Gjerstad. The rate of convergence of continuous play. *Economic Theory*, 7:161–178, 1996.

[10] Daniel Houser and Robert Kurzban. Revisiting confusion in public good experiments. *American Economic Review*, 2002. Forthcoming.

[11] James S. Jordan. Three problems in learning mixed-strategy nash equilibria. *Games and Economic Behavior*, 5:368–386, 1993.

[12] Bernhardt Lieberman. Experimental studies of conflict in some two-person and three-person games. In Joan H. Criswell, Herbert Solomon, and Patrick Suppes, editors, *Mathematical Methods in Small Group Processes*, pages 203–220. Stanford University Press, 1962.

[13] Kevin McCabe, Daniel Houser, Lee Ryan, Vernon Smith, and Theodore Trouard. A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences, U.S.A.*, 98(20):11832–11835, 2001.

[14] David M. Messick. Interdependent decision strategies in zero-sum games: A computer-controlled study. *Behavioral Science*, 12:33–48, 1967.

[15] John Morgan and Martin Sefton. An experimental investigation of unprofitable games. *Games and Economic Behavior*, 2002. Forthcoming.

[16] Yaw Nyarko and Andrew Schotter. An experimental study of belief learning using real beliefs. *Econometrica*, 2002. Forthcoming.

[17] Barry O'Neill. Nonmetric test of the minimax theory of two-person zerosum games. *Proceedings of the National Academy of Sciences, U.S.A.*, 84:2106–2109, 1987.

[18] J. Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54:296–301, 1951.

[19] Robert W. Rosenthal, Jason Shachat, and Mark Walker. Hide and seek in arizona. Technical report, IBM TJ Watson Research Laboratory, 2002.

[20] Alvin E. Roth and Francoise Schoumaker. Expectations and reputations in bargaining: An experimental study. *American Economic Review*, 73:362–372, 1983.

[21] Timothy C. Salmon. An evaluation of econometric models of adaptive learning. *Econometrica*, 69:1597–1628, 2001.

[22] Jason Shachat. Mixed strategy play and the minimax hypothesis. *Journal of Economic Theory*, 2002. Forthcoming.

[23] James W. Walker, Vernon L. Smith, and James C. Cox. Bidding behavior in first price sealed bid auctions use of computerized nash competitors. *Economic Letters*, 23:239–244, 1987.

[24] Eyal Winter and Shmuel Zamir. An experiment with ultimatum bargaining in a changing environment. Technical Report 159, The Hebrew University of Jerusalem, December 1997. Center for Rationality and Interactive Decision Theory.

		Column player	
		L	R
Row player	L	1, -1	0, 0
	R	0, 0	2, -2

Table 1: Pursue-Evade

		Column player: L	Column player: R
Row player	L	2, 0	0, 1
	R	1, 2	1, 1

Table 2: Gamble-Safe

	Opponent treatment		
Game treatment	Human	EWA	RE
Pursue-evade	40	30	30
Gamble-safe	34	0	24

Table 3: Number of subjects that participated in each treatment.

Zero-Sum Games Used In Previous Studies

(Humans are row player, Payoffs are for row player, minimax strategy proportions are next to action names)

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

Table 4:

Table 5:

OLS Regression Results

Computer Left Frequency = $\alpha + \beta * \text{Human Left Frequency}$

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

Table 6:

Average Stage Game Winnings For Decision Makers When Facing A Human Opponent

Game	Human Role	Human's Opponent	Decision Maker Avg. Earnings	T-test Statistic	Approx. d.o.f.	P-value
Gamble-Safe	Row	Human Column	1.0776	***	***	***
Gamble-Safe	Row	RE Column	1.0786	-0.012	23	0.990
Gamble-Safe	Column	Human Row	0.9888	***	***	***
Gamble-Safe	Column	RE Row	0.8983	2.187	25	0.038
Pursue-Evade	Row	Human Column	-0.6709	***	***	***
Pursue-Evade	Row	RE Column	-0.6829	0.498	32	0.622
Pursue-Evade	Row	EWA Column	-0.7205	2.312	33	0.027
Pursue-Evade	Column	Human Row	0.6709	***	***	***
Pursue-Evade	Column	RE Row	0.6395	1.285	31	0.208
Pursue-Evade	Column	EWA Row	0.6395	1.557	32	0.129

All Human Baseline

Row Action	Column Action: Left	Column Action: Right
Left	**.441** (.444)	**.169** (.222)
Right	**.276** (.222)	**.114** (.111)

Estimated RE Model

Row Action	Column Action: Left	Column Action: Right
Left	**.502** (.444)	**.161** (.222)
Right	**.208** (.222)	**.129** (.111)

Estimated EWA Model

Row Action	Column Action: Left	Column Action: Right
Left	**.442** (.444)	**.204** (.222)
Right	**.239** (.222)	**.110** (.111)

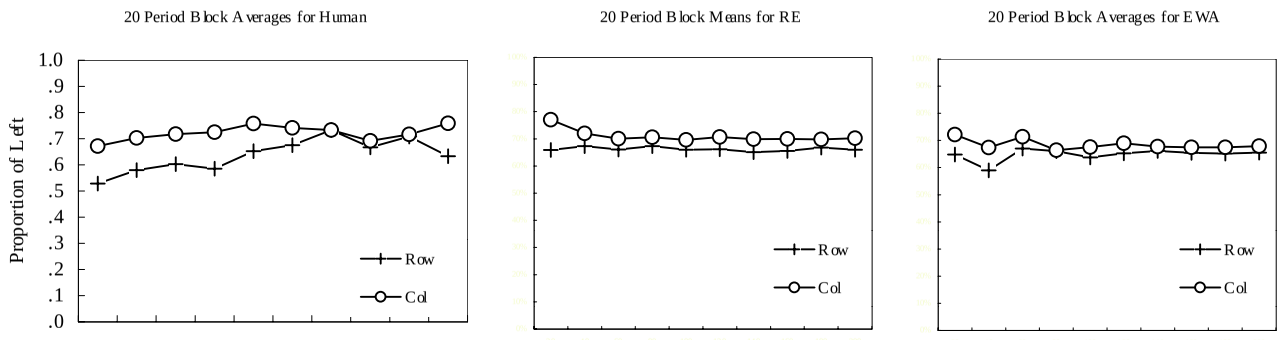


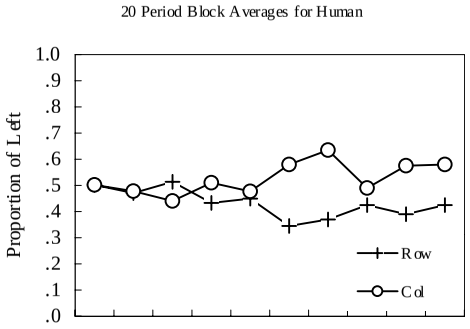
Figure 1: Baseline Data and Estimated Model Summary for Pursue-Evade Game.

All Human Baseline

Row Action	Column Action: Left	Column Action: Right
Left	**.212** (.250)	**.220** (.250)
Right	**.309** (.250)	**.259** (.250)

Estimated RE Model

Row Action	Column Action: Left	Column Action: Right
Left	**.226** (.250)	**.215** (.250)
Right	**.297** (.250)	**.262** (.250)



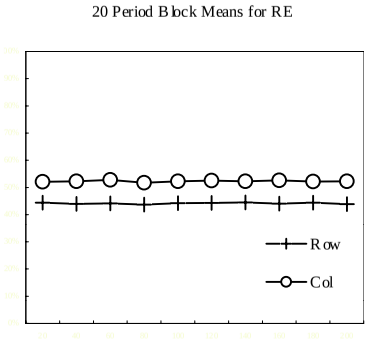


Figure 2: Baseline Data and Estimated Model Summary for Gamble-Safe Game.

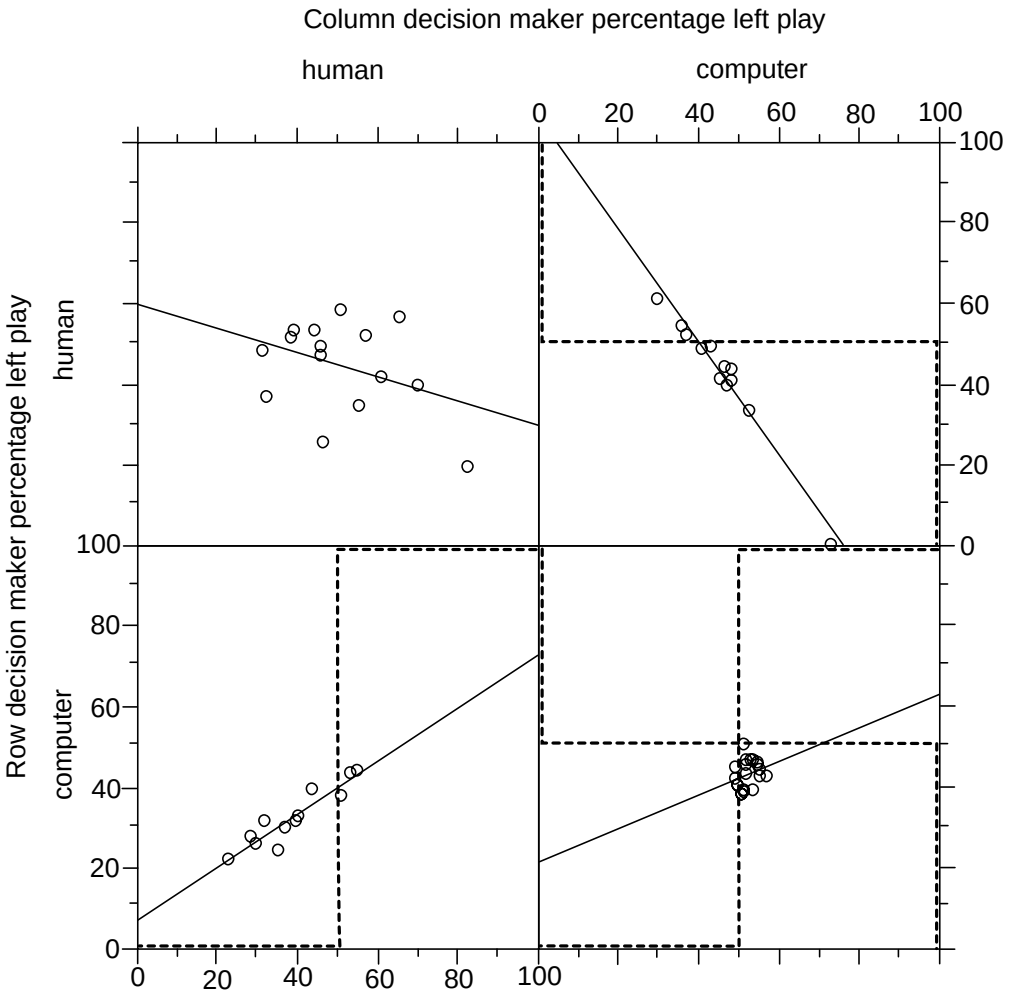


Figure 3: Gamble-Safe joint densities of proportion Left; RE interactions.

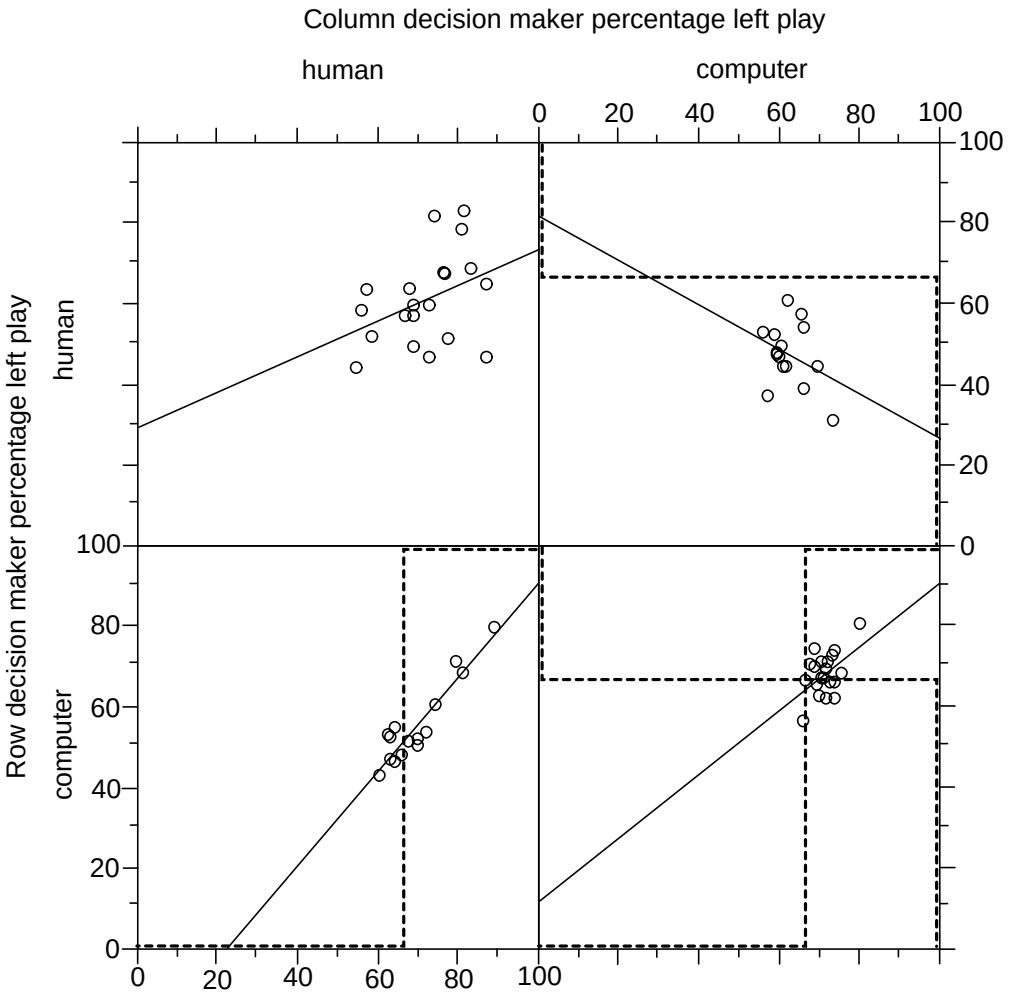


Figure 4: Pursue-Evade joint densities of proportion Left; RE interactions.

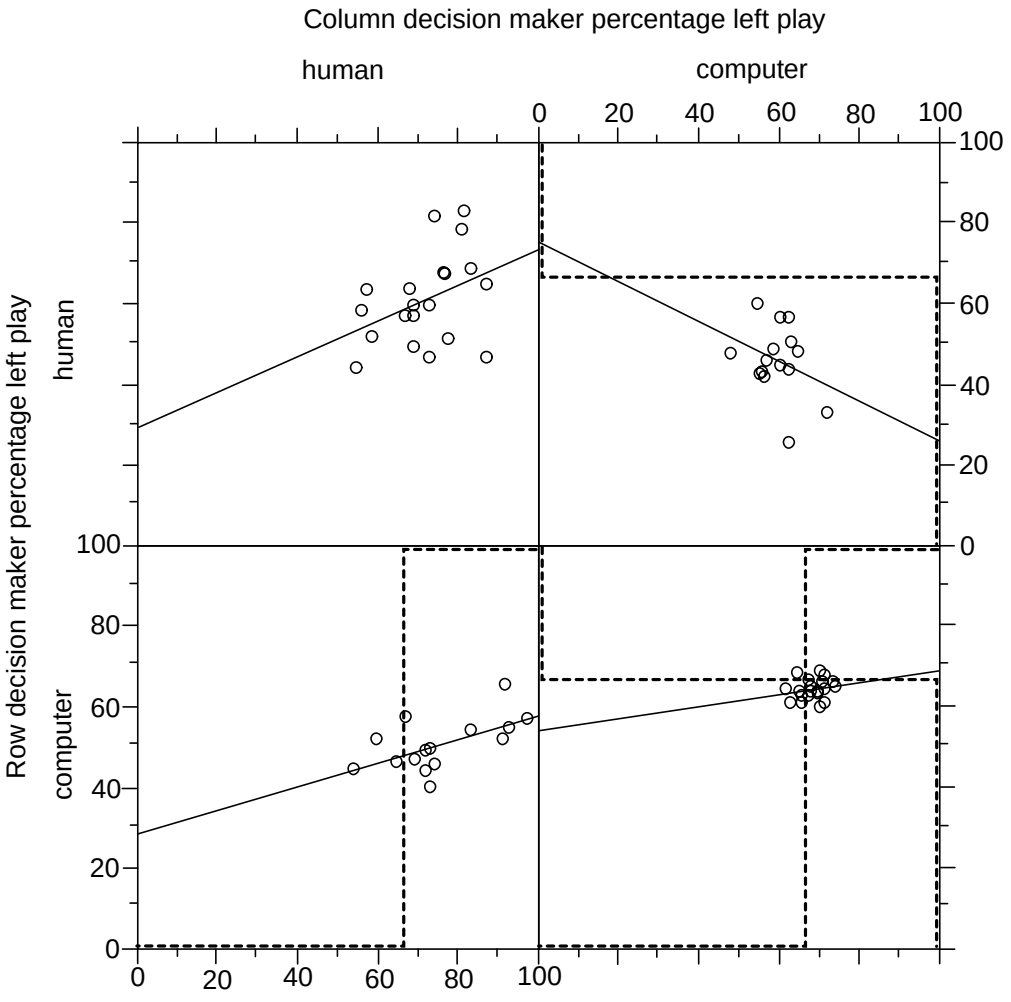
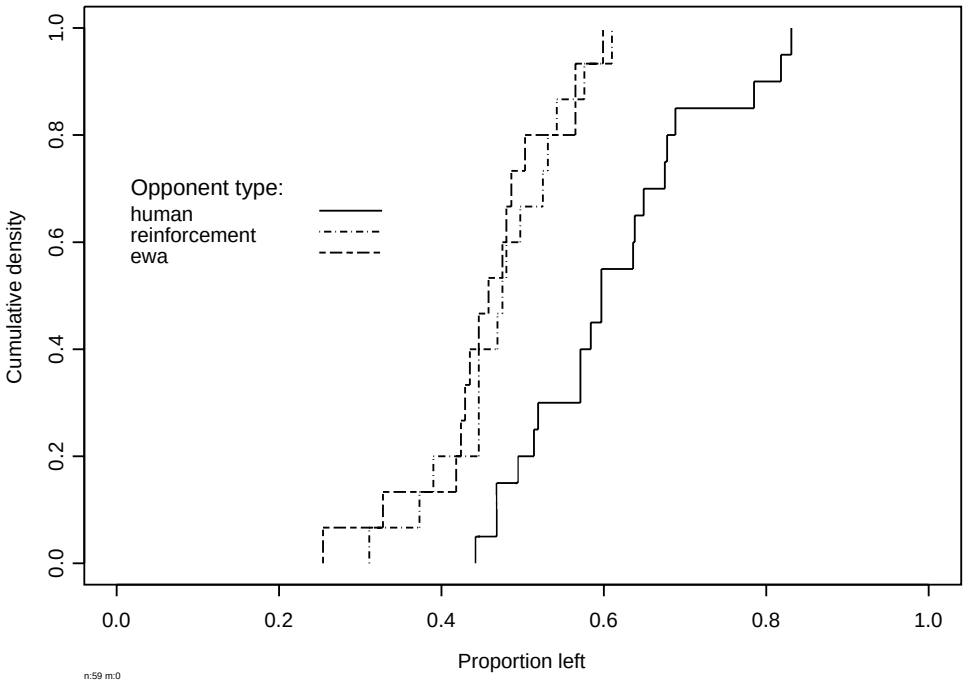
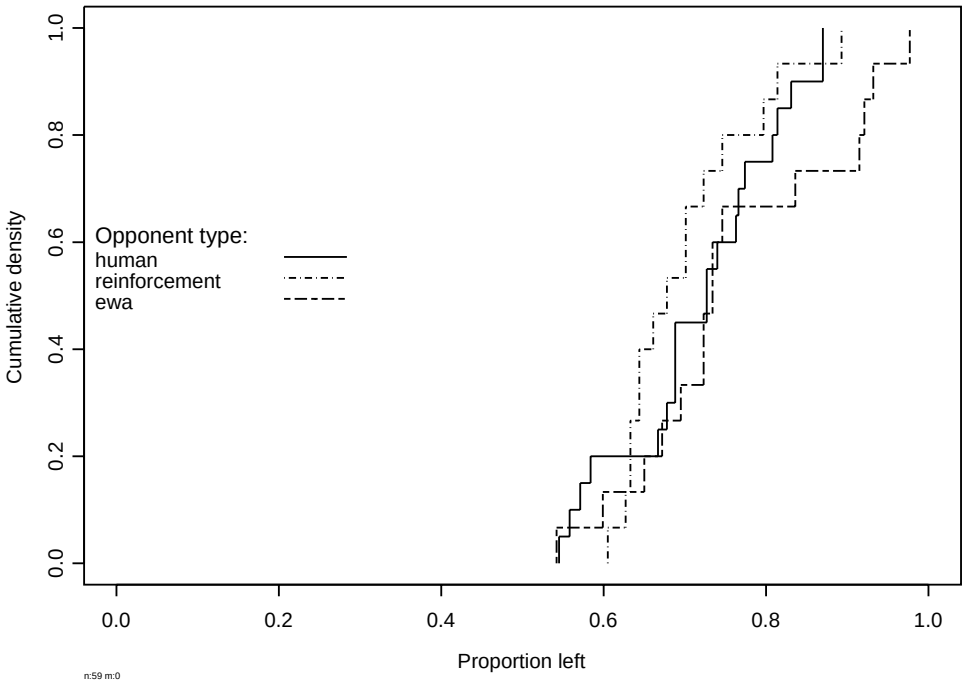


Figure 5: Pursue-Evade joint densities of proportion Left; EWA interactions.



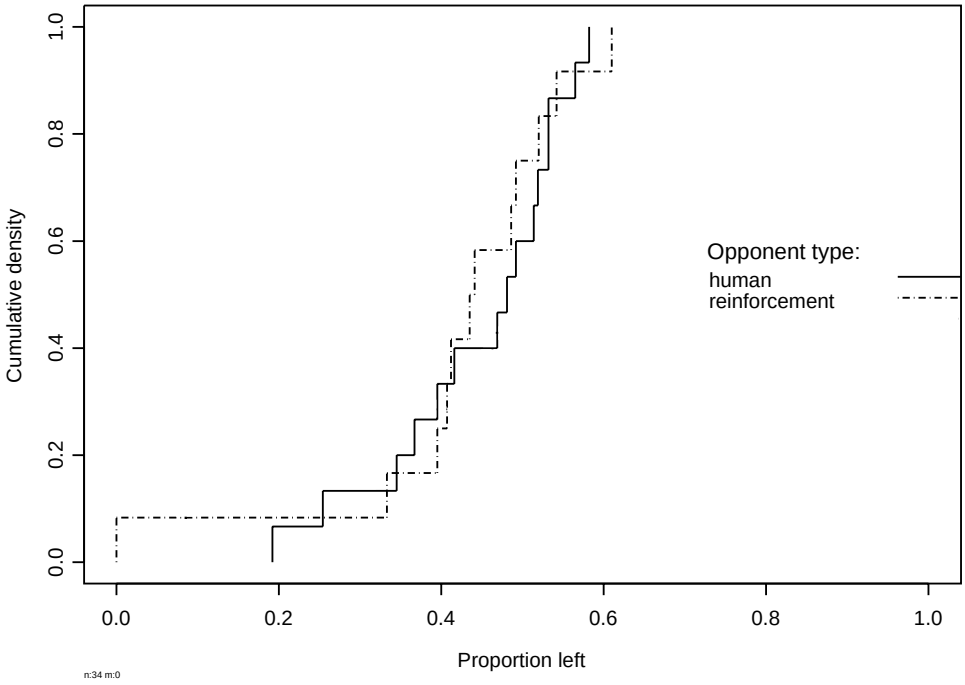
Dist. Left when facing Human tested against dist. Left when facing:	KS statistic	P-value
RE	0.567	0.005
EWA	0.633	0.001

Figure 6: Distributions of Left by Human Row players in Pursue-Evade.



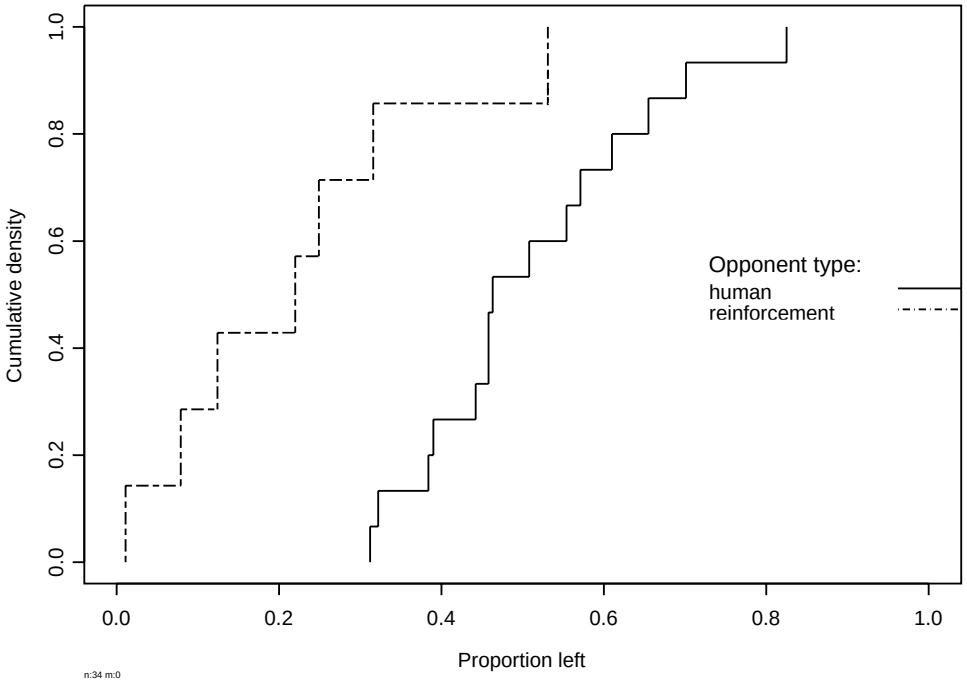
Dist. Left when facing Human tested against dist. Left when facing:	KS statistic	P-value
RE	0.283	0.435
EWA	0.267	0.507

Figure 7: Distributions of Left by Human Column players in Pursue-Evade.



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

Figure 8: Distributions of Left by Human Row players in Gamble-Safe.



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

Figure 9: Distributions of Left by Human Column players in Gamble-Safe.