Strategic adaptation of humans playing computer algorithms in a repeated constant-sum game

Leonidas Spiliopoulos

University of New South Wales

The Australian School of Business

School of Economics, UNSW 2052

NSW, AUSTRALIA

Tel: (612) 9385 7019

Email: l.spiliopoulos@unsw.edu.au

Abstract This paper examines strategic adaptation in participants' behavior conditional on the type of their opponent. Participants played a constant-sum game for 100 rounds against each of three pattern-detecting computer algorithms (CAs) designed to exploit regularities in human behavior such as imperfections in randomizing and the use of simple heuristics. Significant evidence is presented that human participants not only change their marginal probabilities of choosing actions, but also their conditional probabilities dependent on the recent history of play. A cognitive model incorporating pattern recognition is proposed that capture the shifts in strategic behavior of the participants better than the standard non-pattern detecting model employed in the literature, the Experience Weighted Attraction model (and by extension its nested models, reinforcement learning and fictitious play belief learning).

Keywords: Learning, Pattern detection, Computer algorithms, Constant sum games, Experience weighted attraction, Repeated games

1 Introduction

This paper presents experimental evidence that participants strategically adapt their behavior to their opponent's type. In contrast to prior research focusing on marginal choice probabilities, emphasis is placed on patterns in historical action choices, e.g., first- and second-order action choice transition probabilities. Furthermore, it is shown that existing learning models do not adequately capture strategic adaptation, due to their inability to recognize temporal patterns in action choices. Finally, an alternative behavioral model is proposed using pattern recognition that predicts the observed strategic change of participants in cross-validation data.

Pattern recognition is a relatively neglected topic in the game theory learning literature. The state of the literature includes a handful of theoretical papers on the convergence properties of simple pattern recognizing learning rules [3, 21, 44], a simulation study [4] and some empirical studies [36, 45, 46]. The proposed behavioral model will incorporate pattern recognition of varying sophistication, as defined by the temporal depth of lagged action choices used, and will be shown to significantly outperform non-pattern detecting models. Strategic adaptation of subjects' behavior to their computer opponents' behavior will be shown to arise as a consequence of pattern search.

Prior research [7, 18, 37, 49] has raised concerns over the statistical properties of the econometric techniques used in the empirical game theory learning literature. Learning models require the simultaneous estimation of a large number of heterogenous (across subjects) parameters for the belief and decision rules; consequently, these models exhibit low statistical power in rejecting hypotheses and comparing competing models.

To achieve the aforementioned goals, this study uses a within-participants experimental design, a game with high payoff curvature, and computer algorithm (CA) opponents capable of exploiting commonly observed patterns in human behavior. Replacing participants' opponents with computer algorithms, referred to as the human versus computer opponent methodology (HvC), will allow the systematic observation of strategic adaptation through the precise experimental manipulation of CA types. An optimal experimental design requires both the exact specification of a subject's opponent, and also significant variation or diversity in the behavior of these opponents. Accomplishing these two goals is difficult in a human versus human (HvH) experiment because the specification of the opponent's behavior would not be known a priori, but would have to be imprecisely estimated from the data, and variation in behavior would not be guaranteed. Given the practical restrictions of the experimental methodology, such as limited scalability,[1] significantly more observations of HvH play are required to obtain the same statistical power of a HvC experiment.

[1] The restriction on scalability is particularly important for repeated games, as the experimenter is constrained by both funding, access to participants and realistic experimental times/subject fatigue. Therefore increasing the statistical power of the relevant hypothesis tests simply by adding more participants or more rounds to an experiment has an upper limit, and also induces a tradeoff between more participants versus more rounds.

The remainder of the paper is structured in the following manner. Section 2 provides a detailed discussion of the experimental setup including the design of the computer algorithms. Section 3 analyzes the results from the experiment, documenting the degree of strategic adaptation that is present in the data. Section 4 proposes a learning model incorporating pattern recognition and Section 5 presents the estimation results, including a comparison with a standard learning model (EWA). Furthermore, this section examines how close the participants came to implementing an optimal behavioral strategy (from a set of finite state automata) against each CA. Section 6 includes an analysis of data collected from a questionnaire conducted during the experiment. Finally, Section 7 discusses the main conclusions drawn and future research directions. Appendix A presents detailed estimation results from models, Appendix B details the algorithms of the computer opponents and Appendix C describes the Experience Weighted Attraction (EWA) model in detail as it used as a baseline model. Appendix D presents general data analyses pertaining to questions such as how the type of CA opponent or the order of presentation of the CAs affected participants' payoffs, and finally, Appendix E presents the experimental instructions.

2 Methodology/Experimental design

2.1 General experimental setup

A constant-sum game was chosen as social preferences (trust, reciprocity and fairness) are irrelevant, permitting more accurate econometric estimation of pure strategic adaptation. The payoffs of the experimental game are specifically chosen to abide by two criteria: the curvature of the payoff surface should be high enough to provide significant payoff incentives to players to change strategies, thereby avoiding the flat minimum critique [26], and the mixed strategy Nash equilibrium (MSNE) of the stage game should be sufficiently far away from equiprobable play.

The experiment was run in a computer lab at Mediterranean College in Athens, Greece and participants interacted via computers running the Comlab Games Software.[2] Undergraduate students were randomly recruited through the use of fliers on campus and majored in business studies, psychology or computer science. Three sessions of eight students and one session of seven students were run for a total of 31 participants.

Every subject played the 2 × 2 constant-sum game presented in Table 1 one hundred times against each one of the three CAs. The stage game MSNE strategies are playing blue and brown 30% of the time, and yellow and white 70%. Participants were randomly assigned to one of three treatments that varied the order of presentation of the CAs—each CA was presented exactly once as the first, second and third opponent in the treatments.

[2] This software is freely available at http://www.comlabgames.com and allows the design and implementation of a wide variety of game theory or decision making experiments.

Table 1 Game payoffs

		Computer algorithm's actions	
		Brown (br)	White (w)
Participant's actions	Blue (bl)	108, -80	-32, 60
	Yellow (y)	-32, 60	28, 0

Participants were informed after each round what move their opponent made, their current payoff in that period, their average and total payoff against their opponent. Full disclosure was provided to participants by clearly stating that they were playing against three different computer opponents; they also knew when the computer opponent changed. This served to reinforce purely strategic behavioral adaptation devoid of other motivations, since human participants should not care about the outcomes of a CA.

Participants responded to an open-ended question after every 25 rounds asking what their strategy was in the last 25 rounds. After the last (100th) round against each CA, participants were asked to state how many times they thought their opponent played each action and how many times they thought they played each of their own actions.

Participants' monetary reward was dependent on the average number of points they had amassed against all three computer algorithms. They received the average number of points in euro—the average monetary reward for participants was €11.56, the minimum and maximum earnings were €4.93 and €15.80 respectively. As the experiment lasts roughly an hour this rate compares favorably to alternative sources of income for students.[3] Additionally, two prizes were offered to the participants who achieved the two highest average payoffs after the completion of all experimental runs. The subject with the highest average payoff received an additional €30, whilst the second best performer received €20. This incentive scheme, and the use of negative payoffs in the game, were used to motivate participants to keep their attention focused on the task and search for effective strategies against the different opponents—the former by providing the opportunity of a relatively high payoff, the latter by threatening payoffs that participants may have accrued in earlier rounds. This incentive structure may affect behavior through its effect on the risk associated with various strategies and the asymmetry between gains and losses. However, any such effect on behavior would be constant across all the treatments (CA opponents), and therefore comparative statics analyses of behavior across different opponents are still valid.

[3] In perspective, the minimum monthly wage for non-manual workers in Greece in 2006 was €668 [19], roughly €4.18 per hour (assuming 40 hour weeks).

2.2 Computer algorithm opponents

The studies employing a HvC methodology in the game theory literature have used a variety of types of CA players.[4] The main conclusion regarding strategic adaptation is that participants respond to the CAs and move in the direction of best response, albeit sometimes weakly, depending on the complexity and magnitude of a CA's deviation from equilibrium behavior. With the exception of a few studies [11, 29, 48], the CAs were not capable of exploiting empirically observed regularities in human behavior.

I implement more sophisticated CAs specifically designed to exploit the imperfect randomization exhibited by participants in Rapoport and Budescu [32] and Budescu and Rapoport [5]. The *fp2* and *fp3* algorithms are variants of fictitious play (*fp*), which computes the historical probability of the opponent playing each action, and predicts that future actions are drawn from the same distribution. The *fp2* algorithm computes the frequency of occurrence of two-period strategies, generating beliefs conditional on the opponent's previous action; the *fp3* algorithm tracks three-period strategies, generating beliefs conditional on the two prior actions of an opponent.

The third CA exploits a simple strategy often used by humans, the win-stay/lose-shift heuristic[5] [30], henceforth abbreviated to *ws/ls.* The single period detector *(spd)* CA observes whether participants condition on the CA's first lagged action, of which the *ws/ls* heuristic is a special case. The *spd* algorithm keeps count of the number of times a subject's action was consistent with the *ws/ls* heuristic minus the number of times it was inconsistent. Whenever this count is positive (or negative) the CA assumes the subject's next response will (not) be the one prescribed by the *ws/ls* heuristic and best responds to these assumptions.[6]

3 Evidence of strategic adaptation in the data

This section establishes that patterns of action choices are affected by the type of CA opponent. Let the subscripts i and $-i$ denote two matched players, and $a_{i,t}$ denote the action chosen by player i at time, or round, t. Table 2 presents the empirical marginal probabilities of action choices and the first-order Markov transition probabilities pooled by CA opponent type, and Table 3 the second-order Markov transition probabilities.

[4] These CAs include simple (non-Nash equilibrium) mixed strategies [20, 29, 40], fictitious play [11, 14, 30], reinforcement learning [14, 39], simple heuristic decision rules [14], a mixture of minimax and a logit model incorporating the history of play [29], a simple neural network incorporating historical play [48], a basic ACT-R cognitive model [28] and EWA [39].

[5] For 2 × 2 games, the win-stay/lose-shift heuristic dictates that a player should choose the same action in the previous period if it led to the best outcome given the opponent's action choice, and should switch to the other action if it led to the worst outcome in the previous period. Note that for this game, the win-stay/lose-shift heuristic prescribes the same action choices as the Cournot best response heuristic.

[6] This guards against the possibility that participants may learn to use the exact opposite heuristic i.e. changing strategy when winning and using the same strategy when losing.

The marginal probability of playing *bl* is significantly different for all three possible CA comparisons using two-sample tests of proportions: *fp2/fp3* $(z = -2.42, p = 0.0155)$, *fp2/spd* $(z = -6.31, p < 0.0001)$ and *fp3/spd* $(z = -3.9, p = 0.0001)$. However, the economic significance of the difference between *fp2* and *fp3* is not particularly strong, as the change in the marginal probability is only 0.03. The difference in marginal probability is greatest between *fp2* and *spd*, 0.08.

First-order Markov transition probabilities against the *fp2* and *fp3* algorithms are again similar, however there are economically significant differences between these two algorithms and *spd*. The largest change occurs for the history $(a_{i,t-1} = y, a_{-i,t-1} = w)$, as the conditional probability of playing blue increases from 0.24 against *fp2* to 0.42 against *spd*. With respect to second-order transition probabilities, again the largest strategic change appears to occur against the *spd* CA.

Table 2 Human participants' empirical marginal and first-order Markov transition probabilities

Marginal				First-order $p\left(a_{i,t} = bl \mid a_{i,t-1}, a_{-i,t-1}\right)$			
						$a_{-i,t-1}$	
	fp2	*fp3*	*spd*			*br*	*w*
bl	0.38	0.41	0.46	$a_{i,t-1}$	*bl*	0.55, 0.56, 0.46	0.39, 0.38, 0.46
y	0.62	0.59	0.54		*y*	0.62, 0.61, 0.51	0.24, 0.27, 0.42
				The conditional probability of *bl* against *fp2*, *fp3* and *spd* respectively			

Table 3 Human participants' second-order Markov transition probabilities

Second-order $p\left(a_{i,t} = bl \mid a_{i,t-1}, a_{i,t-2}, a_{-i,t-1}, a_{-i,t-2}\right)$					
		$(a_{-i,t-1}, a_{-i,t-2})$			
		(br, br)	(br, w)	(w, br)	(w, w)
$(a_{i,t-1}, a_{i,t-2})$	(bl, bl)	0.72, 0.65, 0.62	0.48, 0.55, 0.62	0.31, 0.29, 0.53	0.40, 0.45, 0.51
	(bl, y)	0.58, 0.60, 0.51	0.35, 0.48, 0.26	0.30, 0.40, 0.44	0.34, 0.35, 0.36
	(y, bl)	0.46, 0.63, 0.57	0.52, 0.65, 0.52	0.36, 0.45, 0.65	0.30, 0.32, 0.39
	(y, y)	0.57, 0.53, 0.60	0.70, 0.62, 0.46	0.39, 0.32, 0.42	0.17, 0.20, 0.26
The conditional probability of *bl* against *fp2*, *fp3* and *spd* respectively					

Strategic adaptation is verified using an alternative technique that is more parsimonious and conducive to conducting formal statistical tests of the significance of these differences. The panel mixed-effects logit model in Equation 1 is estimated, with the benefit of including learning in the action probabilities, in contrast to the Markov transition probability analysis that assumed stationarity. In the following model, A_{fp3} and A_{spd} are equal to 1 if the CA opponent is *fp3* and *spd* respectively, and zero otherwise. The variable H is a moving average of the ten prior choices of a player's opponent and therefore captures subject's sensitivity to changes in marginal probabilities of play. Individual heterogeneity is modeled using random effects for the constant $\alpha_i \sim N(\mu_\alpha, \sigma_\alpha)$.

$$\Pr(a_{i,t}|\Theta) = \Lambda\Big(\alpha_i + \sum_{j=i,-i} \beta_1 a_{j,t-1} + \beta_2 a_{j,t-1} \cdot A_{fp3} + \beta_3 a_{j,t-1} \cdot A_{spd} + \beta_4 a_{j,t-2} + \beta_5 a_{j,t-2} \cdot A_{fp3} + \beta_6 a_{j,t-2} \cdot A_{spd}] + \beta_7 H + \beta_8 H \cdot A_{fp3} + \beta_9 H \cdot A_{spd}\Big) \tag{1}$$

The results of the estimation are presented in Table 4. The average predicted marginal probabilities of playing yellow against the *fp2*, *fp3* and *spd* CAs are 0.618, 0.597 and 0.555 respectively—note that these are very close to the empirical marginal probabilities presented in Table 2. This indicates that the econometric model has captured subject behavior quite well, and therefore is a suitable tool to conduct the following hypothesis tests.

Hypothesis 1: Players' behavior is homogeneous.

$H_0 : \sigma_\alpha = 0$

The null hypothesis is strongly rejected ($p < 0.001$) by an LR test $\chi^2(1) = 54.36$, supporting the random effects specification implemented.

Hypothesis 2: Participants do not learn, or alter their choice probabilities, based on their opponent's recent action history, H (10-period moving average).

$H_0 : \beta_7 = \beta_8 = \beta_9 = 0$

Participants exhibit significant adaptation to an opponent's recent marginal probability of play, as exemplified by a LR test $\big(\chi^2(3) = 18.55, p = 0.0003\big)$. However, this is predominantly due to the play against the *spd* CA since only β_9 is statistically significant. Examining the linear time trend in H for each computer algorithm reveals a possible explanation—the strongest trend is found for the *spd* CA (0.00013), compared to *fp3* (-0.000073) and *fp2* (0.000076). If we assume that there is a non-linear sensitivity of participants to changes in H, i.e., small changes may not be detected at all or very weakly, then this would lead to the outcome observed.

Hypothesis 3: Participants' behavior is independent of the history of play (both marginal probabilities as captured by H and conditional probabilities as captured by lagged actions).

$H_0 : \beta_{\pm1} = \beta_{\pm2} = \beta_{\pm3} = \beta_{\pm4} = \beta_{\pm5} = \beta_{\pm6} = \beta_7 = \beta_8 = \beta_9 = 0$

The null hypothesis is clearly rejected by the data $\big(\chi^2(15) = 662.77, p < 0.0001\big)$, therefore there exist behavioral patterns that can be exploited by the CAs.

Hypothesis 4: Participants' behavior is independent of the prior two lags of own and opponent action choices, beyond their influence in the historical marginal probabilities given by H.

$H_0 : \beta_{\pm1} = \beta_{\pm2} = \beta_{\pm3} = \beta_{\pm4} = \beta_{\pm5} = \beta_{\pm6} = 0$

Table 4 Random effects logistic regression of action choices

Parameter	Estimate	Std. err.	z	p	95% Conf. int.	
β_1	0.098	0.083	1.190	0.235	-0.064	0.260
β_2	0.096	0.113	0.850	0.394	-0.125	0.317
β_3	-0.158	0.110	-1.440	0.150	-0.374	0.057
β_4	0.133	0.081	1.640	0.102	-0.026	0.293
β_5	0.163	0.110	1.480	0.138	-0.053	0.378
β_6	0.481	0.108	4.440	0.000	0.269	0.693
β_{-1}	1.254	0.092	13.680	0.000	1.075	1.434
β_{-2}	-0.061	0.127	-0.480	0.630	-0.310	0.188
β_{-3}	-1.028	0.129	-7.940	0.000	-1.282	-0.774
β_{-4}	0.131	0.095	1.380	0.169	-0.056	0.318
β_{-5}	0.038	0.132	0.290	0.775	-0.221	0.297
β_{-6}	0.305	0.130	2.340	0.019	0.050	0.561
β_7	0.269	0.220	1.220	0.221	-0.162	0.701
β_8	-0.090	0.283	-0.320	0.752	-0.645	0.466
β_9	0.692	0.288	2.400	0.016	0.126	1.257
μ_α	-1.112	0.067	-16.720	0.000	-1.242	-0.982
σ_α	0.226	0.038			0.163	0.314
Wald $\chi^2(15)$	662.77	$p < 0.0001$				

Participants' behavior is found to be significantly conditioned on the prior lags $\left(\chi^2(12) = 470.51, p < 0.0001\right)$ beyond what occurs from adaptation to an opponent's marginal probability of play. This provides clear evidence that behavioral patterns exist in participants' historical action profiles.

Hypothesis 5: Behavior is independent of the type of CA participants are playing against, i.e., no strategic adaptation.

$H_0 : \beta_{\pm2} = \beta_{\pm3} = \beta_{\pm5} = \beta_{\pm6} = \beta_8 = \beta_9 = 0$

There is evidence that participants strategically alter their behavior according to the type of CA $\left(\chi^2(10) = 145.10, p < 0.0001\right)$. Specifically, participants are found to behave similarly against the *fp2* and *fp3* CAs, i.e., $H_0 : \beta_{\pm2} = \beta_{\pm5} = \beta_8 = 0$ $\left(\chi^2(5) = 5.13, p = 0.399\right)$, very differently comparing the *fp2* and *spd* CAs, i.e., $H_0 : \beta_{\pm3} = \beta_{\pm6} = \beta_9 = 0$ $\left(\chi^2(5) = 112.95, p < 0.0001\right)$ and comparing *fp3* and *spd* $H_0 : \beta_{\pm2} - \beta_{\pm3} = \beta_{\pm5} - \beta_{\pm6} = \beta_8 - \beta_9 = 0$ $\left(\chi^2(5) = 89.60, p < 0.0001\right)$.

Concluding, considerable evidence has been presented that strategic adaptation does exist in this dataset, and is both economically and statistically significant. Also, the importance of modeling pattern detecting behavior by participants was corroborated by the questionnaire answers, as participants explicitly stated searching for patterns (detailed analyses can be found in Section 6). This paper now proposes and empirically validates a learning model with the capability of capturing strategic adaptation in the marginal, first- and second- order action transition probabilities.

Table 5 Table of symbols

Symbol	Explanation	Page
General		
t	Time measured by the number of rounds	5
i	Indexes players	5
$-i$	Indexes the matched opponent of player i	5
$a_{i,t}$	Action played at time t by player i	5
Memory system		
n	Depth of pattern recognition	10
ω_t	A vector of the context (history of play) at time t, elements indexed by m	10
π_i	The realized payoff of player i at time t	10
c_j	A memory chunk, indexed by j	10
A_j	Total activation of chunk j	11
B_j	The base-level activation of chunk j	11
t_q	The time (or rounds) elapsed since a chunk was observed	11
γ	The rate of chunk activation decay	11
w_m	Attention weight for element m in the context vector	11
$\varDelta_j$	Dissimilarity between the current context ω_t and a chunk's encoded context	11
Decision process		
θ	The threshold value at which an action choice is made	12
δ	Instantaneous strength of evidence, or drift of OUP process	12
τ	Time during the decision rule (OUP process)	12
$W(\tau)$	Wiener process	12
σ^2	The variance of the diffusion process	12
$X(\tau)$	The state, accumulated evidence, of the OUP process at time τ	12

4 Modeling strategic adaptation using pattern recognition

This section proposes a learning model using theories from the mathematical and cognitive psychology literature found to approximate real cognitive processes and neural behavior. The model is based on an underlying declarative memory system, the Rational Analysis of Memory (RAM) [38] used in the ACT-R cognitive framework [2], capable of similarity-based pattern recognition of arbitrary depth. The output of the memory system is transformed into action choices by a decision rule, the Ornstein-Uhlenbeck diffusion process (OUP). Evidence from memory in favor of each action race toward a threshold—the chosen action is the one that first reaches the threshold. These two systems are explained in detail in the following sections. The reader can refer to Table 5 at any time for a complete list summarizing all the parameters and symbols used throughout the paper. This model is henceforth referred to as the RAM/OUP model of decision making, stemming from the abbreviations of the two subsystems that it incorporates.

Table 6 Representation of an instance as a memory chunk

	π_j	a_j	ω_j					
Slot ID	Payoff	Choice	1	$\cdots$	$n-1$	n	$\cdots$	$M = 2(n-1)$
Chunk c_j	$\pi_i(a_{i,t}, a_{-i,t})$	$a_{i,t}$	$a_{i,t-1}$	$\cdots$	$a_{i,t-n+1}$	$a_{-i,t-1}$	$\cdots$	$a_{-i,t-n+1}$

4.1 The Rational Analysis of Memory system

The RAM declarative memory module is a symbolic system represented by chunks, each comprised of a number of slots capable of storing a single symbol or piece of information. The complete set of information stored in a memory chunk is referred to as an instance; it is comprised of a player's action choice, the associated payoff and the context (a function of the history of play) within which it occurred.[7]

The depth of pattern recognition is denoted by n and implies conditioning on the $(n-1)$ prior lags of play, e.g., $n = 2$ denotes patterns of two consecutive time periods, or equivalently conditioning on the most recent lag, $t-1$. Define the context ω_t at time $t \geq n$[8] (prior to making and observing action choices of round t) as the $2 \cdot (n-1)$-tuple $(a_{i,t-1}, \ldots a_{i,t-n+1}, a_{-i,t-1}, \ldots a_{-i,t-n+1})$ and define ω_t^m as the mth element of ω_t. The depth of pattern recognition, and by extension the context, is constrained by the size of working memory—in order to encode a chunk the current context must be available in working memory. Cowan [12] concludes from a wide survey of the relevant literature that working memory is generally constrained to 4 ± 1 items. Each chunk encoding an n-period pattern requires a working memory capacity of $2 \cdot (n-1) + 2$—given the above constraint on items in working memory $n = 2$ or possibly $n = 3$ appear to be the greatest depths that can be supported.[9]

Table 6 presents a memory chunk in tabular form as the concatenation of a player's action at time t, $a_{i,t}$, the realized payoff $\pi_i(a_{i,t}, a_{-i,t})$ and the current context ω_t. Each of the elements of the current context are encoded in one of M slots. The stored context in a chunk c_j is denoted as ω_j, and is equal to $\omega_{\bar{t}_j}$ where $\bar{t}_j$ is the time period the chunk was first created.

The declarative memory system starts at $t = 1$ with no chunks at all, and after each round t memory chunks are either created, if it is the first time that the realization of the current instance occurs, or the relevant chunk's activation is updated if it already exists (to be explained below). At any time t, the memory module will hold the set of chunks created in all the previous rounds. Note, the stored information in each chunk does not change over time.

[7] This instance-based approach has proved successful in a wide range of problems [23–25]—our model shares the same memory system with this prior work, but differentiates itself in various ways, particularly with respect to the decision rule. This prior work has employed a blending procedure as a decision rule, which basically uses a weighted average of the memory chunks according to their probabilities of retrieval, instead of the diffusion process employed in this paper.

[8] This constraint ensures that sufficient time has lapsed for the observation of at least one n-period context.

[9] In general, n-period patterns require the following items in working memory: $n-1$ prior lags of a subject's historical action profile, $n-1$ prior lags of the opponent's historical action profile, the player's current action choice $a_{i,t}$ and the realized payoff.

Each memory chunk has an activation level that reflects the likelihood that it will be needed in the future based on the history of use of the chunk. The total activation of chunk j, A_j, is comprised of two different types of activation. The first is the base-level activation denoted as B_j, specified in Equation (2), where q represents the number of times the instance of the chunk has been observed and t_q represents the time elapsed since this chunk was observed in the environment for each of q times in the past. Finally, let $\gamma = [0, \infty)$ be the rate of activation decay—an extreme value of $\gamma = 0$ implies no memory loss, whereas as $\gamma \to \infty$ only the observations in the immediately prior time period have non-zero activation.

$$B_j = \ln \sum_q t_q^{-\gamma} \tag{2}$$

The second type of activation of a chunk is the context-dependent component that depends on the degree of matching between the current context and the stored context of each memory chunk. To allow for non-exact matching, i.e., context-dependent activation of chunks that may have a similar but not identical historical context, a similarity function must be defined. Let the attention weights w_m denote the attention a subject pays to each element in the context, where $w_m = [0, 1]$ and $\sum w_m = 1$. The dissimilarity function, Δ_j between the current context ω_t and the context encoded in a memory chunk ω_j, is given by Equation 3. This is a distance function employing the city-block metric, as recommended for separable dimensions [22].

$$\Delta_j = \sum_m w_m |\omega_t^m - \omega_j^m| \tag{3}$$

The total activation of a chunk A_j is equal to the base-level activation minus the product of the dissimilarity function and the matching penalty parameter $\mu = [0, \infty)$ determining the importance of exact matching, see Equation 4. If $\mu = 0$ then the total activation of all the chunks is independent of the context, determined only by the base-level activation, whereas as $\mu \to \infty$ only chunks which are exact matches to the current context will have non-zero activation.

$$A_j = B_j - \mu \Delta_j \tag{4}$$

4.2 The Ornstein-Uhlenbeck diffusion process decision rule

With the memory system defined, a decision rule is required that describes how the information from memory is mapped into a final decision by the player. Diffusion process models of decision making have become one of the main paradigms in the mathematical/cognitive psychology literature [6, 33, 42, 43, 47]. These models assume that cognitive systems are inherently noisy and that the process of arriving at a decision occurs through the accumulation or integration of noisy samples of evidence until a decision threshold is reached. One of the main advantages of diffusion models is the clear identification of the underlying process mechanism and the simultaneous modeling of both choices and response times.

In this spirit, I propose the existence of two leaky accumulators, each of which integrates a continuous stream of evidence over time from the memory system regarding the desirability of each action. The first accumulator to reach the decision threshold, θ, triggers the corresponding action choice by a player.

The integration of evidence over time is modeled as an Ornstein-Uhlenbeck process (OUP). Define the instantaneous flow of evidence for the accumulator associated with a specific action $a = \{blue, yellow\}$ as the summation of the product of the activation of all chunks where $a = a_j$ and the encoded payoff of that chunk π_j. Therefore, denoting the instantaneous flow of evidence for each action, δ_{bl} and δ_y (dependent on the chunks in memory at time t—the subscript t is dropped for convenience):

$$\delta_{a_i} = \sum_{c_j : a_j = a} e^{A_j} \cdot \pi_j \tag{5}$$

In the case of binary choice, it is possible to model the evolution of the two accumulators as a unidimensional OUP with absorbing thresholds at $\pm\theta$ and mean drift $\delta = \delta_{bl} - \delta_y$. Let the position of this process, i.e., the accumulated evidence, at decision time τ[10] be denoted as $X(\tau)$; the starting evidence level $X(0)$ is possibly different from zero indicating an initial preference for one of the actions. Let $d\tau$ be the time step, λ be the leakage parameter of the accumulators, $W(\tau)$ be the standard Wiener process and σ^2 be the variance of the diffusion process. The OUP is then given by the following equation (see Figure 1 for an illustration):

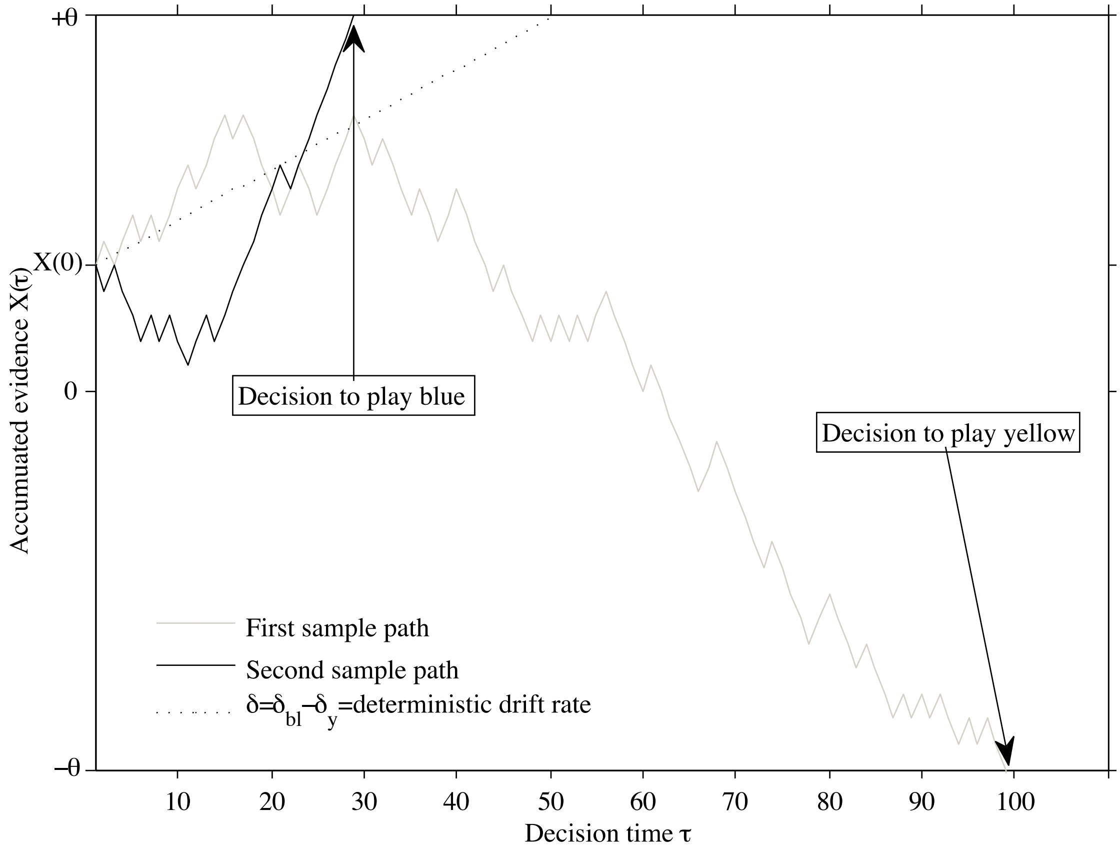
$$dX(\tau) = \delta d\tau - \lambda X(\tau) d\tau + \sigma W(\tau + d\tau) \tag{6}$$

Analytical solutions do not exist for the first passage probabilities (equivalent to the probability distribution over action choices in this application) and mean time for first passage; however, the OUP can be approximated by a birth-death Markov chain yielding simple solutions using standard linear algebra techniques. The interested reader is referred to Diederich and Busemeyer [13] for a detailed discussion of the approximation and relevant derivations.

Summarizing, the complete model is referred to as RAM/OUP(n), stemming from the abbreviation of the two systems and the chosen pattern depth n, and requires the estimation of the following parameters: γ, w_m, θ, λ, σ, and $X(0)$.

[10] Decision time τ is the time since the accumulation process of the decision rule was initiated.

Fig. 1 Discrete-time approximation of the OUP diffusion process



5 Estimation technique and results

Two variants of the proposed model are estimated, RAM/OUP(2) and RAM/OUP(3),[11] plus two baseline models—the first is simply the empirical marginal probability of the observed action choices and the second is a non-pattern detecting baseline, the Experience Weighted Attraction learning model [9]. The rationale behind the latter choice is that it has been widely studied in the literature, but also has the desirable property of nesting both reinforcement [34] and fictitious play learning models [10], thereby simplifying the model comparison procedure (see Appendix C for an introduction to EWA).

Each model allows for subject heterogeneity by estimating individual parameters for each player i given by the set Ψ_i. Let $P_i\left(t|\Psi_i\right)$ be the probability that player i plays the observed action at time t given the parameter set Ψ_i. Estimation of the model is achieved by maximizing[12] the joint log likelihood

[11] The saturated RAM/OUP(3) model has four lagged variables and attention weights. To reduce the number of parameters in the estimation process only two attention weights are estimated—the attention given to a player's own lagged variables (or opponent's) w $(1-w)$ and the attention given to the first (or second) lagged variables w_{l1} $(1-w_{l1})$.

[12] Optimization is performed using a hybrid algorithm starting with a global search over a population of 10,000 observations of randomly generated parameters to find the best candidate to serve as an initial starting point for a local search. This hybrid algorithm is used to avoid settling on a local minimum of this highly non-linear learning model.

function in equation 7 conditional on the complete parameter set $\Psi = \bigcup_{i=1}^{31} \Psi_i$ starting from $t = 11$ onwards for a total of $N = 31 \times 290 = 8990$ observations.[13]

$$\max_{\Psi} ll(\Psi) = \sum_{i=1}^{31} \sum_{t=11}^{300} P_i(t|\Psi_i) \tag{7}$$

Model comparison can be performed using the Akaike and Bayesian information criteria on the estimation data,[14] and the 10-fold cross-validation likelihood criterion, ll_{cv}, obtained using the following procedure:

1. For each player, randomly assign the rounds $t = 11$ to $t = 300$ into 10 folds (or sets) of 29 observations each, denoted by $F_{i,f}$, where f indexes the folds.
2. Estimate the model on nine of these ten sets, and then calculate the out-of-sample, or cross-validation log likelihood of the observations in the remaining fold, $ll(F_{i,f})$.
3. Repeat this procedure ten times for each player, each time withholding a different set of observations as the cross-validation set. The sum of the log-likelihood of all the cross-validation folds is the cross-validation criterion, $ll_{cv} = \sum_i \sum_f ll\left(F_{i,f}\right)$—the higher ll_{cv} is, the greater the predictive power of a model.
4. The set of $ll(F_{i,f})$ for each competing model are matched by subject i and fold f; therefore, it is possible to use paired tests to compare the distributions of $ll\left(F_{i,f}\right)$ for each model and calculate the statistical significance of performance differences.

The best performing model according to ll_{cv} and BIC is RAM/OUP(2), whereas according to AIC it is RAM/OUP(3) because it penalizes the number of parameters less—according to all criteria, the EWA model is a distant third. Tests of the differences between the models' cross-validation performance are performed using two paired tests: the non-parametric Wilcoxon signed-rank test and sign test, henceforth reported in this order. The difference between the RAM/OUP(3) and RAM/OUP(2) models is not statistically significant $(p = 0.793, p = 0.691)$; since the former has the lowest ll_{cv} and fewer degrees of freedom, the RAM/OUP(2) is regarded as the best performing model for the rest of the analysis. The difference between the RAM/OUP(2) and EWA model is statistically significant $(p < 0.0001, p < 0.0001)$—72.9% of the $ll\left(F_{i,f}\right)$ criteria were lower in the RAM/OUP(2) model compared to the EWA model.

The RAM/OUP models were also estimated with the following difference—at the end of every round, instead of only encoding a memory chunk relating to the realized payoff given the action a subject chose and the context, another memory chunk was encoded with the foregone payoff that a subject could have

[13] Each parameter set Ψ_i estimates the behavior of a single subject against all three CAs. The first ten observations for each subject are not included in the objective function as pattern detection requires an initial number of observations; however, these observations are used in the accumulation and encoding of memory chunks.

[14] These are computed using the following equations: $AIC = -2 \cdot ll(\Psi) + 2k$ and $\text{BIC} = -2 \cdot ll(\Psi) + k \cdot log(N)$

Table 7 Model estimation results and comparison

		Calibration	Information criteria		Validation
Models	*df*	ll	AIC	BIC	ll_{cv}
Baseline	31	-6016.73	12095.47	12315.69	-6053.28
EWA	186	-5736.932	11845.86	13167.18	-5870.70
RAM/OUP(2)	217	-5322.09	11078.18	12619.72	-5475.05
RAM/OUP(3)	248	-5268.22	11032.44	12794.20	-5488.69

achieved by choosing the other action. These models were found to perform worse and for the sake of brevity have not been discussed here in detail; for example, the ll_{cv} performance of an RAM/OUP(2) model with forgone payoff encoding decreased to -5503.4.

5.1 Comparison of EWA and RAM/OUP(2)

The performance of the EWA model has been shown to be inferior to RAM/OUP(2) in terms of cross-validation performance. Table 8 verifies that the RAM/OUP(2) model fits the marginal probabilities of play, pooled by CA opponent, better than the EWA model. This section tests the hypothesis that the poor performance of EWA is primarily due to its lack of sophisticated pattern detection. This is accomplished by estimating identical population-averaged GEEs (generalized estimating equations) on the experimental data, and the predicted probabilities of action choice derived from the estimated models[15]—results are reported in Table 9.

The average absolute difference between the parameter estimates obtained from the experimental data and from the EWA model is 0.272, whereas for the RAM/OUP(2) model it is considerably less, 0.137. More importantly, this difference is more pronounced for variables with coefficients that are statistically significant at the 5% level in the empirical data estimation—0.512 for EWA and 0.201 for RAM/OUP(2). The most important differences are found in the parameter estimates $\beta_6, \beta_{-1}, \beta_{-3}, \beta_9$. This verifies that the problem with EWA is predominantly its inflexibility in modeling general patterns, although it also fails to model adaptation to an opponents' moving average against the *spd* CA, as captured by β_9. Note, that the EWA model is still able to emulate limited pattern detection if it has a low memory parameter, leading it to make predictions using only the most recently observed history of play.[16]

Appendix A includes tables of the predicted marginal, first- and second-order transition probabilities of the RAM/OUP(2) model. Comparison of these more detailed results with the analogous empirically

[15] Papke and Wooldridge [31] show that this model is valid in estimating the average partial effects not only for binary dependent variables (for the model estimated on the experimental data) but also fractional response variables (for the models estimated from the outputs of the fitted models).

[16] An extreme example of this is if memory retention is zero, so that the EWA prediction is only dependent on the immediate prior lag.

Table 8 Marginal probabilities of play—empirical and model predicted

	Empirical	EWA	RAM/OUP(2)
fp2	0.380	0.462	0.387
fp2	0.410	0.461	0.379
spd	0.460	0.498	0.458

Table 9 GEE panel regressions on the experimental data and predictions of the estimated models

	Data	EWA	RAM/OUP(2)
β_1	0.094	0.212	0.095
β_2	0.097	-0.073	-0.029
β_3	-0.150	0.010	0.064
β_4	0.129	0.160	0.129
β_5	0.163	-0.032	0.048
β_6	0.480	0.077	0.248
β_{-1}	1.236	0.640	0.891
β_{-2}	-0.059	-0.048	-0.010
β_{-3}	-1.014	-0.052	-0.697
β_{-4}	0.130	0.164	0.382
β_{-5}	0.036	-0.022	0.015
β_{-6}	0.299	-0.043	0.031
β_7	0.280	0.244	0.432
β_8	-0.094	0.136	-0.137
β_9	0.669	0.131	0.666
constant	-1.088	-0.625	-1.036
Wald $\chi^2(15)$	662.55	357.98	578.65

estimated probabilities presented in Section 3 verify the conclusions reached by the above more concise comparison.

5.2 Further discussion of RAM/OUP(2)

Individual parameter estimates of the RAM/OUP(2) model can be found in Appendix A, whilst Figure 2 below provides a graphical representation of the distribution of individual parameter estimates and Table 10 exhibits the relevant distributional statistics.

Table 10 Descriptive statistics of the distribution of individual parameter estimates

Parameter	θ	λ	σ	Bias	γ	μ	w
Lower quartile	13.73	34.55	178.55	0.16	0.52	3.53	0.35
Median	26.9	50.9	258.7	0.3	0.84	7.56	0.81
Upper quartile	39.2	79.6	376.6	0.4	1.24	17.41	0.97

Since the value of $X(0)$ is not comparable across participants due to the value of the threshold parameter varying, a transformation of this variable is reported that gives a relative measure of the starting point, $Bias = 0.5 + X(0)/2\theta$. If $Bias = 0.5$ then an equal amount of evidence needs to be

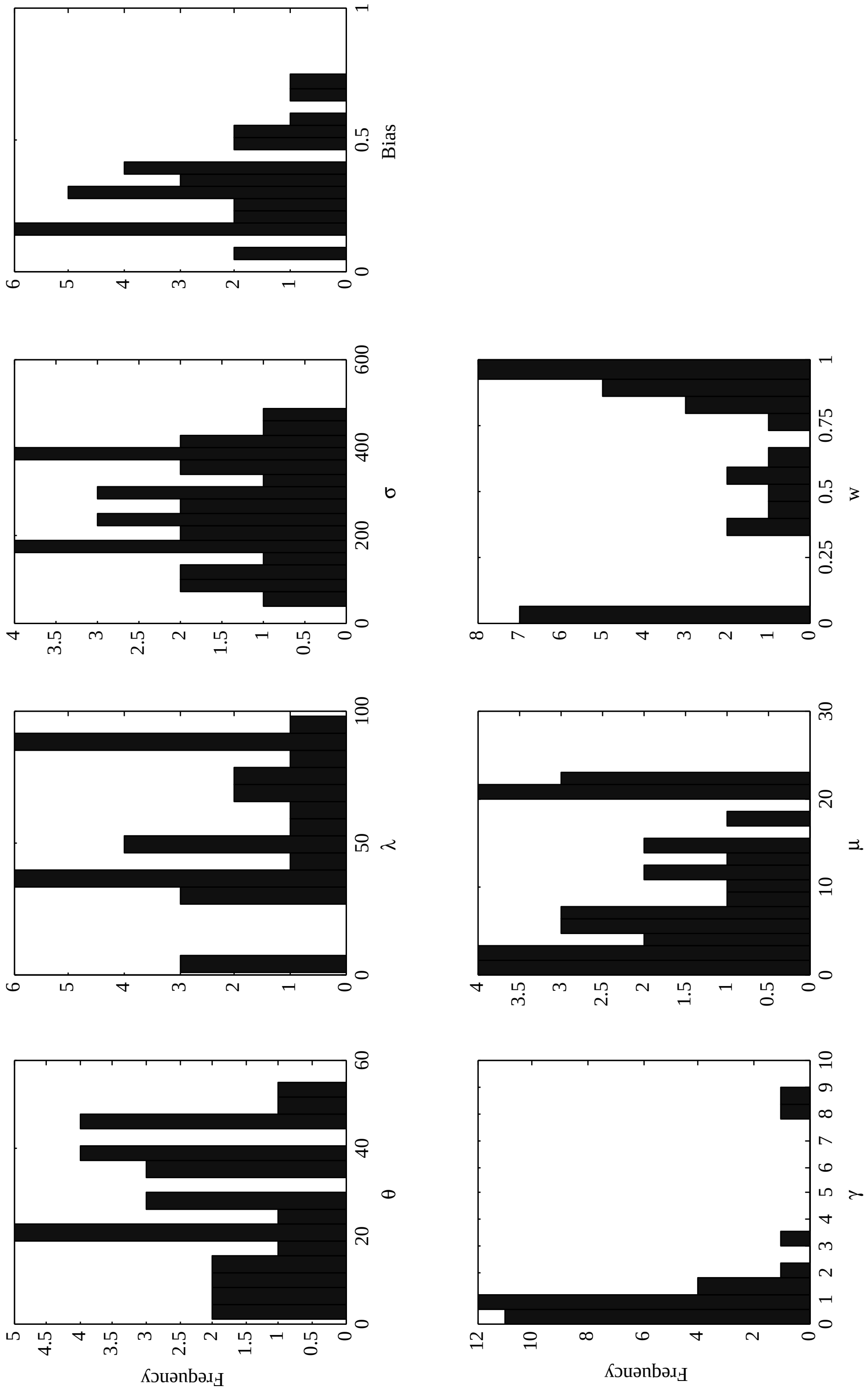


Fig. 2 Distribution of individual parameter estimates

accumulated for both action choices, whereas approaching the extreme values of zero and one implies an asymmetry in the amount of necessary evidence. The estimated bias shows a tendency for most participants to require less accumulated evidence to make a decision to play yellow than blue. Very few participants have λ estimates close to zero, defending the choice of an OUP to model the decision process instead of a Wiener process.[17]

The estimates of the memory parameter are clustered between zero and two, which is in accord with studies using ACT-R in other types of decision tasks [50]. There were only three participants whose estimated γ was high enough to imply that they almost exclusively placed weight on the most recent observations only. The matching penalty parameter estimates capture a wide range of behavior—there exist participants with high μ implying that they penalize mis-matched chunks heavily, i.e., they accumulate evidence only from chunks which are a perfect match to the current context. However, many participants are also found to use information stored in similar, but not exact chunks.

Finally, we observe there is a clustering of two groups of participants with respect to the relative weight given to own and opponents' lagged action choices, where w is the weight assigned to the latter. Only a few participants concurrently used both own and opponents' actions to encode the state, or history, of play—participants tend to fall either at one extreme or the other as two modes exist at values of zero and one.

5.3 Optimality of participants' behavior

This section examines the optimality of participants' behavior against each of the 3 CAs on the basis of payoff performance.[18] The set of permissible strategies is defined as the set of finite automata [1, 17, 35], Ξ_n, where n is the depth of pattern recognition; equivalently, the state of an automaton depends on the previous $n-1$ lagged actions. States are identical to the context as defined previously, i.e., for $n = 3$ the states are given by the context vector $\omega_t = (a_{i,t-1}, a_{i,t-2}, a_{-i,t-1}, a_{-i,t-2})$. An output function maps all possible 2^4 contexts to the action space $\{blue, yellow\}$, culminating in 2^{16} possible deterministic automata for $n = 3$. Finally, the automata are assumed to make an error in applying the

[17] As the parameter $\lambda \to 0$ the Ornstein-Uhlenbeck process collapses to the Wiener process for which accumulators are not leaky and evidence accumulation is not mean-reverting.

[18] Payoff comparisons between humans and automata are more relevant measures of the optimality of humans' learned strategies than direct comparisons of behavioral statistics (such as conditional probabilities) since they automatically accommodate for the possibility of low incentives to search for better alternatives. For example, even if a player's transition probabilities are far from optimal, if this strategy yields near optimal payoffs, the incentive to search is severely reduced.

output mapping function, leading to the wrong action with probability $\epsilon_\xi = 0.3$.[19] Simulations results are reported for the set $\varXi_3$, which implicitly includes sets of simpler automata $\varXi_1 \subset \varXi_2 \subset \varXi_3$. [20]

A comparison of the payoffs of these automata against the 3 CAs to the payoffs achieved by human participants is biased in the direction of finding human behavior to be suboptimal; human participants are learning throughout the 100 rounds of play against a CA, whereas the automata employ a stable mapping representation throughout. It is still informative however to compare the human payoffs to this ideal range to evaluate how effectively participants learnt to exploit the CAs.

Figure 3 presents Epanechnikov kernel density estimates of the probability distribution of the mean payoffs achieved over the last 50 rounds of the automata in $\varXi_3$, the first and last 50 rounds of the human participants, grouped by their CA opponent. Similar distributions imply that participants were not generating payoffs that were significantly different from those achieved by randomly choosing a strategy from the set of automata. If the distribution of human payoffs is shifted to the right of the automata distribution, this indicates that participants have strategically adapted to their opponent. Furthermore, a comparison of the human subject distributions for the first and last fifty rounds is a reflection of learning as reflected by payoff performance. Readers should keep in mind that by randomizing at the MSNE, participants could attain average payoffs of 10, regardless of the CAs' behavior.

Firstly, against all CAs, human subject payoffs for the last fifty rounds are more heavily distributed at the right tails than for the first fifty rounds, and the whole distributions are shifted towards the right. Similarly, except against the *fp3* CA, human participants payoffs are less heavily distributed at the left tails—although most participants learned to perform better against the *fp3* CA, a significant proportion of participants ended up performing worse in the last fifty rounds. Note, that the automata distribution against *fp3* is shifted significantly to the left, compared to other CA opponents, indicating that the majority of automata perform very poorly against this algorithm (indeed well below the MSNE payoffs equal to ten). Therefore, pattern search is more likely in this case to lead some participants to less desirable outcomes, from which they may not have enough experience (number of rounds) to recover from.

Comparing human payoffs in the last fifty rounds against the *fp2* and *fp3* algorithms to the automata distributions, participants played significantly better than the bulk of automata. Subject payoffs against the *spd* CA are slightly shifted towards the left, however note that the majority of observations are already at very high payoff levels, therefore the incentive to further optimize behavior is relatively small.

[19] Simulations were run for values of $\epsilon_\xi \in \{0.5, 0.4, 0.3, 0.2, 0.1, 0\}$ leading to $6 \cdot 2^{16}$ possible automata. In praxis, very low error rates for ϵ_ξ are not feasible models of human behavior given the large number of mappings–although this is reduced since the possible number of observed contexts is much smaller–and it should be expected that considerable error would enter their action choices. The results were qualitatively similar for the various ϵ_ξ values, therefore we present the results from the intermediate value $\epsilon_\xi = 0.3$.

[20] For example $\varXi_3$ includes automata whose states depend only on the previous lagged history of play $\varXi_2$, and the degenerate set of automata $\varXi_1$ that prescribe the same action regardless of the history of play; however, the error in applying the output function leads these automata to cover different marginal probabilities of play.

Fig. 3 Probability density of payoffs of automata and participants against each CA

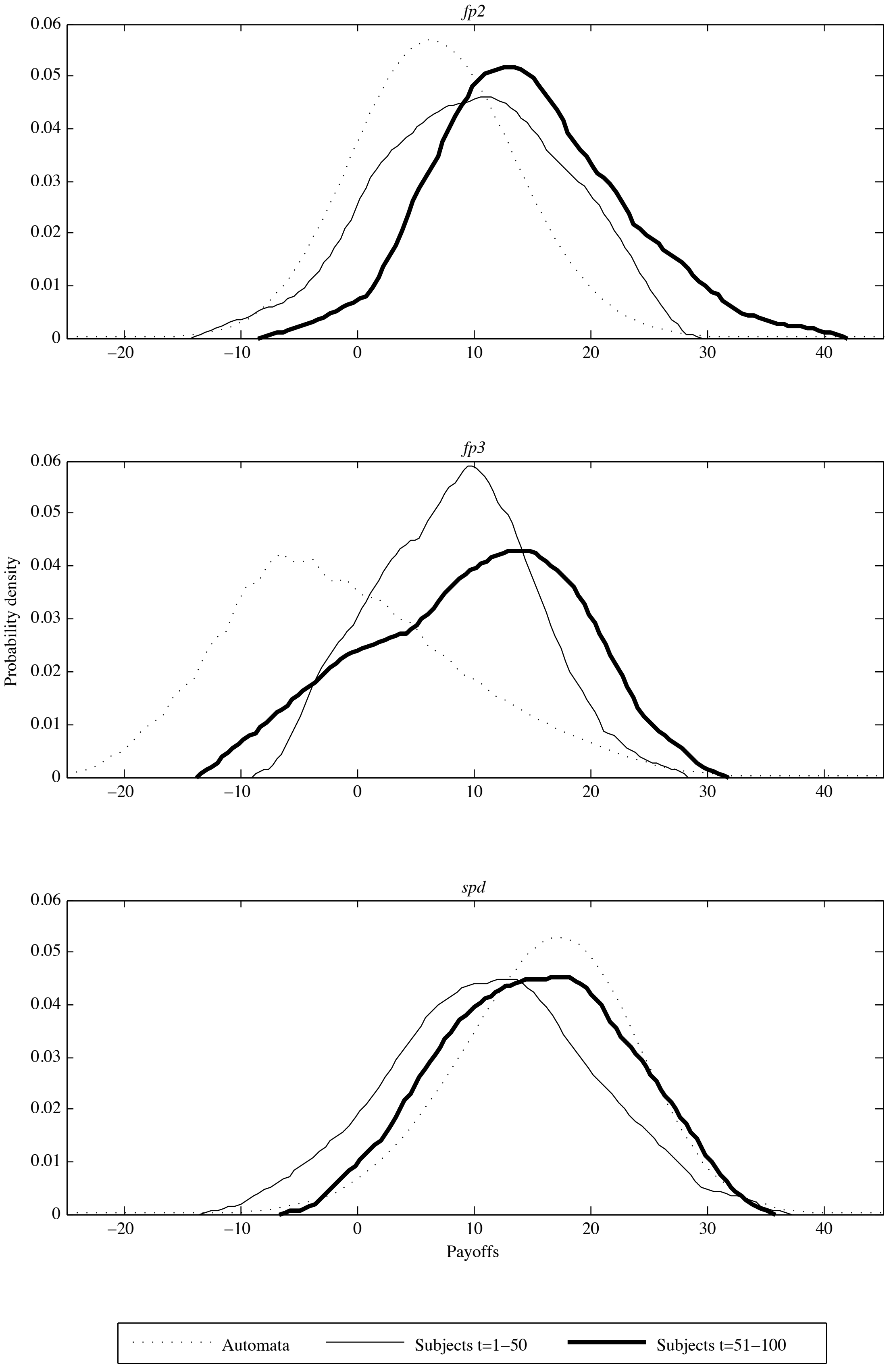


Table 11 Classification of players into types based on the questionnaire answers (%)

CA	Non-pattern detecting	Pattern detecting	*ws/ls*	Other reasoning	Random
All	18.8	21	2.2	4.6	53.5
fp2	17.8	21	2.4	5.6	53.2
fp3	16.9	22.6	1.6	3.2	55.6
spd	21.8	19.4	2.4	4.8	51.6

Most automata perform very well against the *spd* CA (as the automata distribution is shifted significantly to the right of the corresponding distributions for *fp2* and *fp3*), as long as they do not employ the *ws/ls* strategy or a close variant. For further econometric analysis of participants' payoffs and dependence on the type of CA opponent, presentation order of CAs etc., the reader is referred to Appendix D.

6 Questionnaire analysis

Earlier sections presented evidence of pattern recognition and strategic adaptation derived implicitly from observed actions; I now present explicit evidence from participants' reported descriptions of their decision strategies (collected every 25 rounds). Each answer was coded as falling into either of the following categories: non-pattern detecting strategies, pattern detecting strategies, other types of reasoning, *ws/ls* heuristic, and random.[21] To qualify for the pattern detecting category participants had to display some type of multi-period or sequential thought instead of simply looking at single-period actions. The results of this categorization are displayed in Table 11.

The proportion of participants classified under each type seems to be largely independent of the CA opponent. A large percentage of responses stated that decisions were made randomly (53.5%), followed by pattern detecting strategies (21%), non-pattern detecting strategies (19%) and a much smaller percentage, 4.6%, gave some other kind of reasoning. A striking result is that only a small percentage of responses acknowledged using the *ws/ls* heuristic. This implies that the use of the *ws/ls* heuristic is an automatic, rather than a controlled process [8], in which case players show a subconscious tendency to exhibit this type of behavior.

The correlation between the percentage of times a player stated he/she was playing randomly and that player's RAM/OUP(2) ll_{cv} was only 0.088 ($p = 0.638$). Therefore, subjects reporting random be-

[21] Examples of representative answers for each category are:

Non-pattern detecting: "The computer played white more often so I played yellow more often.", "I think it started off playing white more often but then started playing brown more often." ,"I tried to count how often I won playing each color.", "I made my choice according to the changes in my score."

Pattern detecting: "I thought the computer chose to play the same color very often.", "I am trying to figure out how many times in a row it uses each color.", "Playing three times yellow and one time blue seems to be a profitable combination.", "When I lost two consecutive times playing blue I would play yellow and vice versa."

Other reasoning: "I played yellow so as not to lose 80 points.", "Normally you should play only blue because if you win you have made up for three losses."

Win/stay, lose/shift heuristic: "Usually whenever I lost I would change color.", "If I lost playing yellow I would then choose blue and vice versa."

Random: "I was playing according to chance."

Table 12 Predictions of own and opponents' first-order play

	Opponents' actions (white action)			Own actions (blue action)		
CA	% predicted	% actual	MAD	% predicted	% actual	MAD
All	51.24%	66.15%	20.98	41.13%	41.16%	11.60
fp2	53.41%	69.59%	23.83	39.07%	37.48%	11.38
fp3	51.90%	67.81%	21.26	38.65%	40.58%	12.71
spd	48.25%	60.75%	17.71	46.00%	45.61%	10.61

havior were just as predictable as other subjects indicating that they were not really randomizing, but had difficulty explicitly reporting their strategies.

After the final round of playing a CA each subject was asked to state the number of times she thought she had played blue in the last 100 rounds and also the number of times that her opponent had played white. Table 12 presents the predicted proportions and the actual proportions of own and opponents' play pooled against all algorithms and by each computer opponent.

Predictions of opponents' play against all CAs are not calibrated well, 51.24% versus the observed value of 66.15%, in stark contrast to predictions of own play, 41.13% versus 41.16%. Individual accuracy of predictions is measured by the mean individual absolute deviations (MAD) of predictions from actual play. The difference between own and opponents' play (MAD) against all CAs, 11.6% to 20.98%, is economically and statistically significant using a Wilcoxon signed ranks test ($z = -4.308, p < 0.001$).[22]

These results imply that own actions are more easily and/or accurately retrieved from memory than opponents' actions. A possible explanation is that in the RAM model the latter are only encoded as part of the context, which does not need–and is not designed–to be directly retrievable from memory; this is in contrast to own actions which are encoded in the retrievable Slot ID "Choice". Subjects can form their probability estimate of own actions by retrieving the values of a_j from existing memory chunks and reporting a weighted average of them.

7 Discussion

This paper examines strategic adaptation and behavioral modeling of humans engaged in repeated games using a human versus computer algorithm experiment. This led to an increase in experimental control allowing for more powerful statistical testing of strategic adaptation, in response to three different types of computer opponents.

The experimental data verified that participants condition their behavior on the type of opponent, not just by modifying their marginal choice probabilities, but also conditional probabilities of play, i.e., multi-period patterns or sequences of behavior. Since standard learning models in the game theory

[22] Also, it is not possible to reject the null hypotheses that the average MAD for predictions of own and opponents' play is the same across all CAs using Friedman tests at the 5% significance level. The χ^2 (and associated p-values) of MAD predictions of own and opponents' play are 1.52 (0.468) and 2.742 (0.254) respectively.

literature do not permit sophisticated pattern recognition, this paper proposed an alternative model based on two widely used cognitive process paradigms in the mathematical psychology literature. An underlying memory system encoding the context in which memories are encoded and a similarity-based context dependent activation of memories—in this case, the context was a subset of the history of play, thereby implementing pattern recognition. Noisy samples of evidence regarding the desirability of action choices were sequentially sampled from the memory system and accumulated according to a diffusion process until a decision threshold was reached. This cognitive model significantly outperformed the widely used Experience Weighted Attraction learning model in out of sample prediction of participants' behavior.

Furthermore, the proposed model can also predict the distribution of response times, in contrast to the standard learning models in the literature that are incapable of modeling response times. Although this experimental study did not have response time data to take advantage of this, future research should attempt to jointly model decision choices and response time. The advantage of models jointly determining choice and response time is that they are more falsifiable. In light of the issues discussed early in the paper regarding the econometric estimation of learning models, the addition of response time data could also serve to increase parameter identification and lead to more powerful tests of model comparison.

Acknowledgements

The author would like to thank Kunal Sengupta, Andreas Ortmann for constructive comments, anonymous referees whose observations led to significant improvements, and acknowledge the donation of GPU-computing hardware from Nvidia. A substantial part of this research was undertaken at the University of Sydney.

References

1. Abreu, D. and A. Rubinstein (1988). The structure of Nash equilibrium in repeated games with finite automata. *Econometrica 56*(6), 1259–1281.
2. Anderson, J. (2007). *How can the human mind occur in the physical universe?*, Volume 3. Oxford University Press, USA.
3. Aoyagi, M. (1996). Evolution of Beliefs and the Nash Equilibrium of Normal Form Games. *Journal of Economic Theory 70*(2), 444–469.
4. Banerjee, D. and S. Sen (2007, April). Reaching pareto-optimality in prisoner's dilemma using conditional joint action learning. *Autonomous Agents and Multi-Agent Systems 15*(1), 91–108.

5. Budescu, D. V. and A. Rapoport (1994). Subjective randomization in one- and two-person games. *Journal of Behavioral Decision Making 7*, 261–78.

6. Busemeyer, J. (2002, July). Survey of decision field theory. *Mathematical Social Sciences 43*(3), 345–370.

7. Cabrales, A. and W. Garcia-Fontes (2000). Estimating learning models from experimental data. *University of Pompeu Fabra working paper*.

8. Camerer, C., G. Loewenstein, and D. Prelec (2005, March). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature 43*(1), 9–64.

9. Camerer, C. F. and T. Ho (1999). Experience-weighted attraction learning in normal-form games. *Econometrica 67*, 827–74.

10. Cheung, Y. W. and D. Friedman (1997). Individual learning in normal form games: Some laboratory results. *Games and Economic Behavior 19*, 46–76.

11. Coricelli, G. (2005). Strategic interaction in iterated zero-sum games. *Working paper, University of Arizona, Department of Economics*.

12. Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences 24*(1), 87–114.

13. Diederich, A. and J. Busemeyer (2003, June). Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time. *Journal of Mathematical Psychology 47*(3), 304–322.

14. Duersch, P., A. Kolb, J. Oechssler, and B. Schipper (2010). Rage against the machines: how subjects play against learning algorithms. *Economic Theory 43*(3), 407–430.

15. Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association 82*, 171–200.

16. Efron, B. and R. Tibshirani (1994). *An Introduction to the Bootstrap.* Chapman & Hall/CRC.

17. Engle-Warnick, J. and R. L. Slonim (2004, December). The evolution of strategies in a repeated trust game. *Journal of Economic Behavior & Organization 55*(4), 553–573.

18. Erev, I. and E. Haruvy (2005). On the potential uses and current limitations of data driven learning models. *J Math Psychol 49*(5), 357–371.

19. Eurostat (2006, 13 July). Minimum wages in the eu25.

20. Fox, J. (1972). The learning of strategies in a simple, two-person zero-sum game without saddlepoint. *Behavioral Science 17*, 300–308.

21. Fudenberg, D. and D. K. Levine (1998). *The Theory of Learning in Games (Economics Learning and Social Evolution).* Cambridge: MIT Press.

22. Gärdenfors, P. (2004). *Conceptual spaces.* Cambridge ; London : MIT Press.

23. Gonzalez, C. (2003, August). Instance-based learning in dynamic decision making. *Cognitive Science 27*(4), 591–635.

24. Gonzalez, C., V. Dutt, A. Healy, M. Young, and L. Bourne Jr (2009). Comparison of instance and strategy models in ACT-R. *Department of Social and Decision Sciences*, 71.

25. Gonzalez, C. and C. Lebiere (2005). Instance-based cognitive models of decision making. *Transfer of knowledge in economic decision making. New York: Palgrave McMillan*, 1–28.

26. Harrison, G. W. (1989). Theory and misbehavior of first-price auctions. *American Economic Review 79*(4), 749–62.

27. Ho, T., C. Camerer, and J. Chong (2007). Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory 133*(1), 177–198.

28. Lebiere, C. and R. L. West (1999). A dynamic ACT-R model of simple games. In *Proceedings of the Twenty-first Conference of the Cognitive Science Society*, pp. 296–301.

29. Levitt, S. D., J. A. List, and D. H. Reiley (2010). What Happens in the Field Stays in the Field: Exploring Whether Professionals Play Minimax in Laboratory Experiments. *Econometrica 78*(4), 1413–1434.

30. Messick, D. M. (1967). Interdependent decision strategies in zero-sum games: A computer-controlled study. *Behavioral Science 12*, 33–48.

31. Papke, L. and J. Wooldridge (2008, July). Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics 145*(1-2), 121–133.

32. Rapoport, A. and D. Budescu (1997). Randomization in individual choice behavior. *Psychological Review 104*(603-617).

33. Ratcliff, R. and P. L. Smith (2004, April). A comparison of sequential sampling models for two-choice reaction time. *Psychological review 111*(2), 333–67.

34. Roth, A. E. and I. Erev (1995). Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term. *Games and Economic Behavior 8*(1), 164–212.

35. Rubinstein, A. (1986). Finite automata play the repeated prisoner's dilemma. *Journal of Economic Theory 39*, 83–96.

36. Rutström, E. E. and N. T. Wilcox (2009, November). Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test. *Games and Economic Behavior 67*(2), 616–632.

37. Salmon, T. C. (2001). An Evaluation of Econometric Models of Adaptive Learning. *Econometrica 69*(6), 1597–1628.

38. Schooler, L. and J. Anderson (1997). The role of process in the rational analysis of memory. *Cognitive Psychology 32*, 219–250.

39. Shachat, J. and J. Swarthout (2011). Learning about learning in games through experimental control of strategic interdependence. *Journal of Economic Dynamics and Control*.

40. Shachat, J. and T. J. Swarthout (2004). Do we detect and exploit mixed strategy play by opponents? *Mathematical Methods of Operations Research 59*(3), 359–373.

41. Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association 62*, 626–633.

42. Smith, P. (2000, September). Stochastic Dynamic Models of Response Time and Accuracy: A Foundational Primer. *Journal of mathematical psychology 44*(3), 408–463.

43. Smith, P. L. and R. Ratcliff (2004, March). Psychology and neurobiology of simple decisions. *Trends in neurosciences 27*(3), 161–8.

44. Sonsino, D. (1997). Learning to Learn, Pattern Recognition, and Nash Equilibrium. *Games and Economic Behavior 18*(2), 286–331.

45. Sonsino, D. and J. Sirota (2003). Strategic pattern recognition—experimental evidence. *Games and Economic Behavior 44*(2), 390–411.

46. Spiliopoulos, L. (2012). Pattern Recognition and Subjective Belief Learning in a Repeated Constant-Sum Game. *Games and Economic Behavior, http://dx.doi.org/10.1016/j.geb.2012.01.005*.

47. Usher, M. and J. McClelland (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological review 108*(3), 550.

48. West, R. L. and C. Lebiere (2001). Simple games as dynamic, coupled systems: randomness and other emergent properties. *Cognitive Systems Research 1*(4), 221–239.

49. Wilcox, N. (2006). Theories of learning in games and heterogeneity bias. *Econometrica 74*(5), 1271–1292.

50. Wong, T., E. Cokely, and L. J. Schooler (2010). An Online Database of ACT-R Parameters : Towards a Transparent Community-based Approach to Model Development. In *Proceedings of ICCM*, Volume 1, pp. 282–286.

A Detailed results

Table 13 Human participants' empirical marginal and first-order Markov transition probabilities

	Marginal			First-order $p\left(a_{i,t} = bl \mid a_{i,t-1}, a_{-i,t-1}\right)$			
						$a_{-i,t-1}$	
	fp2	*fp3*	*spd*			*br*	*w*
bl	0.38	0.41	0.46	$a_{i,t-1}$	*bl*	0.55, 0.56, 0.46	0.39, 0.38, 0.46
y	0.62	0.59	0.54		*y*	0.62, 0.61, 0.51	0.24, 0.27, 0.42
				The conditional probability of *bl* against *fp2*, *fp3* and *spd* respectively			

Table 14 RAM/OUP(2) empirical marginal and first-order Markov transition probabilities

	Marginal			First-order $p\left(a_{i,t} = bl \mid a_{i,t-1}, a_{-i,t-1}\right)$			
						$a_{-i,t-1}$	
	fp2	*fp3*	*spd*			*br*	*w*
bl	0.39	0.38	0.46	$a_{i,t-1}$	*bl*	0.58, 0.56, 0.53	0.35, 0.35,0.47
y	0.61	0.61	0.54		*y*	0.53, 0.51, 0.46	0.30, 0.28, 0.41
				The conditional probability of *bl* against *fp2*, *fp3* and *spd* respectively			

Table 15 Human participants' second-order Markov transition probabilities

		Second-order $p\left(a_{i,t} = bl \mid a_{i,t-1}, a_{i,t-2}, a_{-i,t-1}, a_{-i,t-2}\right)$			
		$(a_{-i,t-1}, a_{-i,t-2})$			
		(br, br)	(br, w)	(w, br)	(w, w)
$(a_{i,t-1}, a_{i,t-2})$	(bl, bl)	0.72, 0.65, 0.62	0.48, 0.55, 0.62	0.31, 0.29, 0.53	0.40, 0.45, 0.51
	(bl, y)	0.58, 0.60, 0.51	0.35, 0.48, 0.26	0.30, 0.40, 0.44	0.34, 0.35, 0.36
	(y, bl)	0.46, 0.63, 0.57	0.52, 0.65, 0.52	0.36, 0.45, 0.65	0.30, 0.32, 0.39
	(y, y)	0.57, 0.53, 0.60	0.70, 0.62, 0.46	0.39, 0.32, 0.42	0.17, 0.20, 0.26
The conditional probability of *bl* against *fp2*, *fp3* and *spd* respectively					

Table 16 RAM/OUP(2) second-order Markov transition probabilities

		Second-order $p\left(a_{i,t} = bl \mid a_{i,t-1}, a_{i,t-2}, a_{-i,t-1}, a_{-i,t-2}\right)$			
		$(a_{-i,t-1}, a_{-i,t-2})$			
		(br, br)	(br, w)	(w, br)	(w, w)
$(a_{i,t-1}, a_{i,t-2})$	(bl, bl)	0.78, 0.75, 0.80	0.52, 0.50, 0.62	0.35, 0.40, 0.59	0.30, 0.31, 0.45
	(bl, y)	0.60, 0.59, 0.52	0.44, 0.43, 0.39	0.39, 0.37, 0.47	0.39, 0.35, 0.39
	(y, bl)	0.53, 0.59, 0.52	0.48, 0.50, 0.46	0.44, 0.44, 0.59	0.35, 0.33, 0.39
	(y, y)	0.57, 0.53, 0.56	0.52, 0.51, 0.42	0.41, 0.34, 0.44	0.23, 0.21, 0.29
The conditional probability of *bl* against *fp2*, *fp3* and *spd* respectively					

Table 17 Individual parameter estimates and confidence intervals of the RAM/OUP model

Pl.	θ			λ			σ			Bias			γ			μ			w		
1	55.3	42.6	58.4	30.0	29.4	31.7	164.8	163.7	169.6	0.35	0.35	0.35	3.31	2.33	9.27	2.42	1.19	3.02	0.58	0.44	0.71
2	26.9	13.7	42.7	34.4	34.4	68.5	316.3	250.9	440.6	0.40	0.40	0.50	0.85	0.60	0.88	22.20	3.50	22.20	0.01	0.00	0.03
3	20.1	20.1	20.1	34.2	34.2	34.2	129.3	129.3	129.3	0.30	0.30	0.30	1.00	0.96	1.03	7.56	7.56	7.56	0.91	0.89	0.93
4	35.6	17.2	38.7	62.4	48.1	64.1	195.3	127.8	195.3	0.05	0.05	0.10	0.86	0.81	0.99	7.53	7.53	17.05	0.92	0.90	0.94
5	8.3	7.5	35.8	89.0	1.0	95.0	177.3	177.2	439.1	0.15	0.15	0.35	0.30	0.06	0.41	14.80	5.21	23.90	0.06	0.02	0.96
6	34.3	16.3	49.6	36.2	16.3	48.3	368.4	148.8	378.4	0.75	0.55	0.75	0.18	0.17	0.43	1.09	0.99	20.58	0.36	0.03	0.39
7	27.6	8.9	44.1	37.1	23.0	76.6	297.2	180.8	406.2	0.60	0.25	0.65	0.49	0.12	1.56	12.19	7.15	31.39	0.05	0.01	0.07
8	11.3	3.0	44.8	77.5	1.4	77.5	407.0	143.6	453.5	0.40	0.30	0.45	0.52	0.46	9.88	6.87	3.99	22.14	1.00	0.01	1.00
9	21.0	21.0	21.1	1.0	1.0	1.0	89.8	89.7	89.8	0.20	0.20	0.20	0.93	0.89	0.98	8.58	8.57	8.59	0.84	0.80	0.87
10	1.7	1.0	15.4	80.3	7.0	98.8	426.1	228.4	498.9	0.55	0.50	0.55	1.73	0.00	3.99	5.61	1.80	49.94	1.00	1.00	1.00
11	15.9	16.0	43.2	68.6	23.0	68.7	129.1	129.1	188.1	0.05	0.05	0.10	0.54	0.51	0.63	12.56	12.55	22.69	0.87	0.55	0.93
12	47.2	24.3	47.2	68.2	68.2	99.2	369.1	292.9	398.3	0.35	0.30	0.35	0.78	0.68	0.84	23.29	12.28	23.33	0.98	0.97	0.99
13	22.7	20.1	22.7	36.8	36.8	66.2	396.6	289.3	396.6	0.40	0.30	0.40	0.16	0.14	0.33	10.83	8.01	10.83	0.99	0.98	1.00
14	13.0	2.5	41.7	35.0	34.6	97.8	99.2	41.7	228.6	0.25	0.10	0.25	1.10	0.97	1.22	2.38	2.10	8.08	0.57	0.49	0.89
15	48.4	32.2	50.0	73.7	12.1	73.7	381.9	309.3	395.2	0.15	0.15	0.40	0.64	0.54	0.67	18.28	17.41	24.59	0.34	0.34	0.51
16	45.1	39.1	45.1	31.8	31.8	61.7	183.1	155.1	188.0	0.70	0.60	0.90	1.79	1.70	1.96	1.01	0.87	1.25	0.61	0.43	0.79
17	19.3	17.0	34.6	87.0	29.5	87.4	217.8	197.4	244.7	0.15	0.15	0.30	0.84	0.73	0.98	3.33	1.79	23.79	0.87	0.67	1.00
18	46.7	46.7	46.7	1.1	1.1	1.1	234.1	234.1	234.1	0.35	0.35	0.35	0.76	0.69	0.79	22.14	22.14	22.14	1.00	1.00	1.00
19	34.0	31.8	56.6	50.2	20.8	56.6	153.8	144.9	209.7	0.30	0.30	0.35	9.02	4.48	10.00	1.46	1.18	1.64	0.53	0.44	0.62
20	1.2	1.0	15.7	48.6	28.5	98.8	243.0	228.4	331.0	0.55	0.50	0.55	0.00	0.00	3.62	21.40	2.71	21.40	0.06	0.02	1.00
21	39.2	18.7	39.2	33.0	33.0	86.7	285.6	217.8	285.6	0.20	0.15	0.20	0.84	0.80	0.97	21.67	1.53	21.67	0.99	0.80	0.99
22	37.3	14.5	43.8	44.0	43.2	97.3	227.2	135.7	295.4	0.50	0.30	0.65	1.29	1.17	1.42	21.71	4.02	21.71	0.93	0.52	0.94
23	39.2	38.3	44.1	92.2	91.8	93.9	390.9	390.1	391.0	0.15	0.15	0.15	2.29	1.65	10.00	6.05	5.99	6.71	0.00	0.00	0.00
24	27.9	16.1	47.8	89.6	13.7	91.1	447.0	272.0	478.9	0.40	0.35	0.45	0.29	0.11	0.51	14.13	14.13	41.40	0.87	0.67	0.96
25	7.8	7.2	39.6	4.4	4.4	41.6	44.3	44.2	160.0	0.30	0.10	0.35	0.96	0.93	1.04	5.15	3.35	6.79	0.75	0.65	0.77
26	24.5	23.8	39.5	53.8	25.8	86.3	182.3	181.6	226.3	0.15	0.05	0.30	1.70	1.18	10.00	3.11	2.02	18.69	0.00	0.00	0.00
27	39.3	9.0	46.1	51.9	1.0	86.8	269.4	96.3	306.8	0.15	0.05	0.35	0.83	0.65	0.93	12.11	1.31	22.39	0.42	0.19	0.95
28	20.2	4.1	43.0	92.2	24.9	94.0	300.7	108.2	478.0	0.30	0.25	0.45	0.56	0.52	0.60	4.33	1.62	18.18	0.81	0.72	0.93
29	12.4	10.7	13.4	87.9	87.7	88.1	258.7	258.6	258.9	0.25	0.25	0.25	0.52	0.40	0.58	4.14	3.82	4.60	1.00	1.00	1.00
30	9.5	9.5	26.0	99.0	25.8	99.0	492.5	429.6	492.5	0.50	0.40	0.50	0.04	0.02	0.45	21.61	21.61	50.00	0.00	0.00	0.00
31	45.3	22.2	45.7	50.9	29.9	51.2	379.1	379.1	429.0	0.30	0.30	0.45	7.92	6.55	10.00	0.26	0.00	0.51	1.00	0.51	1.00
25%	13.73			34.55			178.55			0.16			0.52			3.53			0.35		
50%	26.90			50.90			258.70			0.30			0.84			7.56			0.81		
75%	39.20			79.60			376.60			0.40			1.24			17.41			0.97		

Values for each parameter represent the estimated coefficient, lower and upper 95% jackknifed confidence intervals respectively.
Last three rows are the lower quartile, median and upper quartile of parameter estimates across individuals.

B Details of computer algorithms

The details of the belief generating equations and their updating are outlined in the following four steps. The discrete strategy set of player i is denoted S_i and the actions chosen by player i at times $t, t-1, t-2$ are denoted by a_i, a_i', a_i'' respectively.

Step 1

fp2 Starting from round 3, and for every round henceforth, after observing the action of the human subject update equations 8 and 9. Let $I_t(a_{-i}|a'_{-i})$ is an indicator function that takes a value of one if a_{-i} was the action played at time t and a'_{-i} was the action played at time $t-1$ and takes a value of zero otherwise[23]. Define for player i, the count of a_{-i} at time t given action a'_{-i} as:

$$C_i(a_{-i}|a'_{-i}, t) = \frac{I_{t-1}(a_{-i}|a'_{-i}) + \sum_{u=1}^{t-2} I_{t-u-1}(a_{-i}|a'_{-i})}{t-1} \tag{8}$$

The *fp2* beliefs of player i of action a_{-i} given action a'_{-i} are then given as:[24]

$$fp2_i(a_{-i}|a'_{-i}, t) = \frac{C_i(a_{-i}|a'_{-i}, t)}{\sum_{a_{-i} \in S_{-i}} C_i(a_{-i}|a'_{-i})} \tag{9}$$

fp3 Starting from round 4, after observing the action of the human subject update the following equations. Given actions a_{-i}, a'_{-i} a''_{-i}, $I_t(a_{-i}|a'_{-i}, a''_{-i})$ is an indicator function that takes a value of one if a_{-i} was the action played at time t and a'_{-i} and a''_{-i} were the actions played at time $t-1$ and $t-2$ respectively, and takes a value of zero otherwise:

$$C_i(a_{-i}|(a'_{-i}, a''_{-i}), t) = \frac{I_{t-1}(a_{-i}|(a'_{-i}, a''_{-i})) + \sum_{u=1}^{t-2} I_{t-u-1}(a_{-i}|(a'_{-i}, a''_{-i}))}{t-1} \tag{10}$$

The *fp3* beliefs of player i of action a_{-i} given actions a'_{-i} and a''_{-i} are then given as:

$$fp3_i(a_{-i}|(a'_{-i}, a''_{-i}), t) = \frac{C_i(a_{-i}|(a'_{-i}, a''_{-i}), t)}{\sum_{a_{-i} \in S_{-i}} C_i(a_{-i}|(a'_{-i}, a''_{-i}), t)} \tag{11}$$

spd Starting from round 4, after observing the human subject's action update equation 12 where $W_i(3)$ is initialized to be zero. The indicator function, $I_{t-1}(a'_{-i}|a''_i, a''_{-i})$ takes the value one if the subject's action a'_{-i} was consistent with the *ws/ls* heuristic and the value of -1 otherwise:

$$W_i(t) = W_i(t-1) + I_{t-1}(a'_{-i}|a''_i, a''_{-i}) \tag{12}$$

The variable $W_i(t)$ is just the net sum of the number of times the subject has exhibited *ws/ls* behavior - a count of zero means that the *ws/ls* action was played 50% of the time.

Step 2 If $t < 5$ or average payoffs to the human subject are higher than 20 proceed to step 3, otherwise proceed to step 4 with probability 0.8 or to step 3 with probability 0.2.

Step 3 Choose the computer action probabilistically according to the stage game's MSNE and proceed to step 1.

[23] At $t = 1$ the indicator function takes the value of zero for all actions a_{-i}, since a time period $t = 0$ does not exist and therefore actions are not observable.

[24] This definition assumes that the denominator is not zero i.e. that the action a'_{-i} has been played at least once in the past. In cases where a'_{-i} has not been observed beliefs are assumed to be given by a uniform distribution over $a_{-i} \in S_{-i}$.

Step 4 For the *fp2* and *fp3* algorithms, play a best response to the calculated beliefs. For the *spd* algorithm, if $W_i(t) > 0$, then assume the subject will play the *ws/ls* prescribed action and best respond to that, if $W_i(t) < 0$ assume the subject will not play the *ws/ls* prescribed action and best respond to that, and if $W_i(t) = 0$ then play the stage game MSNE. Start over from step 1.

Equations 8 and 10 implicitly assume that there is no memory decay at all so that all past actions are remembered perfectly. This was consciously chosen because a pilot study showed that increasing memory decay made the computer algorithms much more predictable and open to exploitation, a result that is also corroborated by Messick [30]. Step 2 injects noise into 20% of the decisions made to mask the underlying belief generation process.[25] The commonly used logit decision rule, by injecting noise into every single decision is overly defensive—this may be the reason why in some studies CAs using this rule have not able to gain statistically significant increases in payoffs against human participants. We believe that the approach employed in this paper provides a better tradeoff between the desire to mask the inner workings of the CAs (and avoid deterministic repetitive behavior) whilst still retaining the ability to aggressively exploit non-optimal human behavior.

C Introduction to EWA

The experience weighted attraction (EWA) learning model has the desirable property of nesting both reinforcement and fictitious play learning models as demonstrated in Camerer and Ho [9]—note, the mathematical symbols used in this chapter follow those in Camerer and Ho [9], but may clash with the use of some symbols in the main text. Equation 13 is the EWA updating formula for the attractions of each available action. These attractions are then normalized so that they sum up to one, thereby representing probabilities of playing each action, by implementing the logit decision rule in equation 14. Players are indexed by i (all other players by $-i$), individual strategies by j for a total of M_i strategies per player, $\pi_i(s_i^j, s_{-i}(t))$ are the payoffs to player i given strategies s_i^j at time t, and $I(s_i^j, s_i(t))$ is an indicator function which is equal to one if strategy j was played at time t by player i (i.e. if $s_i^j = s_i(t)$) and zero otherwise. The free parameters in the model are ϕ, which represents a decay rate on the previous period attraction and can be thought of as strength of memory, δ controls how much forgone payoffs affect attractions and κ controls how attractions grow over time - whether attractions are weighted averages of previous attractions or a cumulative sum of previous attractions. The parameter $N(t)$ is an experience weight and is modified according to equation 15, and the λ parameter in the logit decision rule[26] determines the sensitivity of agents to differences in the attractions of the available actions. Before the game starts, at $t = 0$, the experience weight $N(t)$, is assigned an initial value, $N(0)$ to be estimated, as are all attractions $A_i^j(0)$. Following Ho et al. [27] we set $N(0) = 1$ rather than estimating it, as experimental evidence finds that participants do not exhibit strong priors and the importance of this parameter rapidly decays over time.

[25] A pilot study found 20% to be an appropriate values in terms of the CAs performing well against human participants.

[26] The parametric form of the logit decision rule is such that adding a constant to all attractions does not change probabilities and therefore given that there are two possible actions it is only necessary to estimate $A_i^j(0)$ for one action and normalize the other value to some constant, in this case zero.

$$A_i^j(t) = \frac{\phi \cdot N(t-1) \cdot A_i^j(t-1) + \left[\delta + (1-\delta) \cdot I(s_i^j, s_i(t))\right] \cdot \pi_i(s_i^j, s_{-i}(t))}{N(t)} \tag{13}$$

$$P_i^j(t+1) = \frac{e^{\lambda \cdot A_i^j(t)}}{\sum_{k=1}^{M_i} e^{\lambda \cdot A_i^k(t)}} \tag{14}$$

$$N(t) = \phi \cdot (1-\kappa) \cdot N(t-1) + 1 \tag{15}$$

Reinforcement learning requires $\delta = 0$, and is in cumulative form if $\kappa = 1$ or weighted average form if $\kappa = 0$. Weighted fictitious play is attained if $\delta = 1$ and $\kappa = 0$, and the special case of Cournot best response dynamics[27] if $\phi = 0$.

D Data analysis

This section models how variables such as payoffs and randomizing efficiency are affected by the type and the order of presentation of the opponents the participants faced. The estimated models are maximum-likelihood estimated linear mixed-effects models, capturing both within- and between-participants variance, with participants modeled as random effects. As normality assumptions were clearly violated, confidence intervals were created using the bias-corrected, accelerated percentile intervals (BC_a) method proposed by Efron [15] for 2,500 replications.[28] The effects of multiple comparisons are accounted for by controlling the family-wise error rate (FWER) rather than the per-comparison error rate (PCER), through the use of the Sidak [41] correction.

The general equation setup for the following analyses will be:

$$\textit{response=random subject effects + fixed treatment effects + error} \tag{16}$$

The treatments are fixed-effects and are modeled by the inclusion of appropriate dummy variables. There are two treatments each consisting of three levels. The algorithm treatment consists of three levels, namely the three CAs that the participants faced, *fp2*, *fp3* and *spd*. Let A_{alg} denote a dummy variable equal to one whenever alg was the computer algorithm faced by the subject. Likewise, the position (or order) treatment consists of three levels i.e. whether the CA was the first, second or third opponent that a subject played against. Let T_t be a dummy variable equal to one where t is equal to the position (or order of presentation) of the game. Adhering to this convention let the estimated coefficients of these treatment effects be denoted by the lower-case Greek equivalents, α_{alg} and τ_t. Let the individual observations, i, of any particular dependent variable under investigation, x, at time t, be denoted by $x_{i,t}$. Also, x_{alg}^t is the estimate of x for players facing the CA alg, in position t.

The reference category for this model is the value of the dependent variable when participants played against the *fp2* algorithm in the 1st position, denoted by x_{fp2}^1. Note that $x_{fp3}^t = x_{fp2}^t + \alpha_{fp3}$, and similarly $x_{spd}^t = x_{fp2}^t + \alpha_{spd}$. The estimated model is given in equation 17, where $\epsilon_{i,t} \sim N(0, \sigma_\epsilon^2)$ thereby assuming homoskedasticity of errors, and $\mu_i \sim N(0, \sigma_\mu^2)$ with variance-covariance matrix $\Omega_{3n \times 3n} = I_n \otimes \Sigma_{3\times3}$, where $n = i \times t$ is the total number of observations, I_n is the identity matrix and $\Sigma_{3\times3}$ is given in equation 18.

[27] For arbitrary payoffs another necessary condition is that $\delta = 1$ so that the model weights both foregone and realized payoffs equally. It is important however to note that given the payoffs of the game used in this experiment $\delta = 1$ is not a necessary condition.

[28] Efron and Tibshirani [16] recommend 2,000 replications for the BC_a method.

Table 18 Payoff performance and its dependence on position and algorithm effects

	Coef.	Bootstrap s.e.	$lower^{2-tail}_{98.3\%}$	$upper^{2-tail}_{98.3\%}$	$lower^{1-tail}_{98.3\%}$
π^1_{fp2}	10.563	1.078	7.964	13.147	
π^1_{fp3}	7.299	1.074	4.503	9.726	
π^1_{spd}	11.662	1.217	8.629	14.612	
π^3_{fp2}	13.109	1.238	9.864	15.922	
π^3_{fp3}	9.844	1.271	6.224	12.567	
π^3_{spd}	14.207	1.205	11.352	17.232	
α_{fp3}	-3.265	1.293	-6.459	-0.243	
α_{spd}	1.099	1.356	-2.189	4.266	
$\alpha_{spd} - \alpha_{fp3}$	4.363	1.352	1.256	8.059	
τ_2	2.609	1.257			0.037
$\tau_3 - \tau_2$	-0.063	1.403			-2.774
τ_3	2.545	1.313			-0.084

Likelihood	$LR\,\chi^2(4)$	p-value
-285.30503	15.58	0.0036

$$x_{i,t} = x^1_{fp2} + \alpha_{fp3}A_{fp3} + \alpha_{spd}A_{spd} + \tau_2 T_2 + \tau_3 T_3 + \mu_i + \epsilon_{i,t} \tag{17}$$

$$\Sigma_{3\times 3} = \begin{bmatrix} \sigma^2_\epsilon + \sigma^2_\mu & \sigma^2_\mu & \sigma^2_\mu \\ \sigma^2_\mu & \sigma^2_\epsilon + \sigma^2_\mu & \sigma^2_\mu \\ \sigma^2_\mu & \sigma^2_\mu & \sigma^2_\epsilon + \sigma^2_\mu \end{bmatrix} \tag{18}$$

D.1 Are some algorithms better than others in exploiting human behavior?

The relevant hypotheses to test whether payoffs to participants, π, differed significantly across CA opponents are:

$H_0 : \alpha_{fp3} = 0,\ \alpha_{spd} = 0,\ \alpha_{spd} - \alpha_{fp3} = 0$

$H_1 : \sim H_0$

From Table 18, participants' payoffs where highest against *spd*, followed by *fp2* and finally by *fp3*. Payoffs against *fp3* are less than the payoffs from both the *fp2* and *spd* CAs, by 3.265 and 4.363 respectively, results that are not only statistically significant but also economically significant in terms of magnitude.

D.2 Do payoffs depend on the order of presentation of each algorithm?

Game specific learning implies that payoffs to participants should be increasing for higher CA positions. This will be tested by comparing the increase in payoffs from the first to second position, and from the second to third position.[29]

$H_0 : \tau_2 > 0, \tau_3 - \tau_2 > 0$

$H_1 : \sim H_0$

[29] The statistical tests in this case are one-tailed as game-specific learning necessarily implies an increase in payoffs over time. Learning from the first to second position necessarily implies that learning from the first to third period will be at least as much or more, depending on the strength of learning from the second to third position and therefore a test for learning from the first to third position is redundant. Hence, all statistical tests will be one-tailed and adjusted for two pairwise comparisons, for a strictly controlled FWER of 5%.

The results in Table 18 show that τ_2 is significantly different from zero and economically significant as payoffs rise by 2.6 points. However, $\tau_3 - \tau_2$ is not statistically significant different from zero. It can be concluded that although there is significant transfer of learning from the first period to the latter periods in terms of an increase in payoffs, there is no additional transfer of new game specific learning from the second to third time periods.

E Experimental instructions

The instructions provided to participants are presented below. Note, the game payoff matrix in the sample information screen given in the instructions is different from that used during actual game play—apart from this, the sample information screen is identical to what participants see while playing the real game.

You are about to participate in an economics experiment involving decision-making. Depending on your performance you will earn real money which will be paid to you at the end of the experiment in private. Throughout the experiment please do not talk to other participants, if you have any questions please raise your hand and the administrator will come over and answer your question.

During the experiment you will be facing three different opponents, in all cases they will be computer programs whose aim will be to maximize their own payoffs. You will play for 100 consecutive rounds against each of these three computer opponents i.e. 300 rounds during the whole experiment. Nobody will be paid the computer programs' payoffs. Your payoff after each round of the game will depend both on your actions/decisions and the computer program's actions.

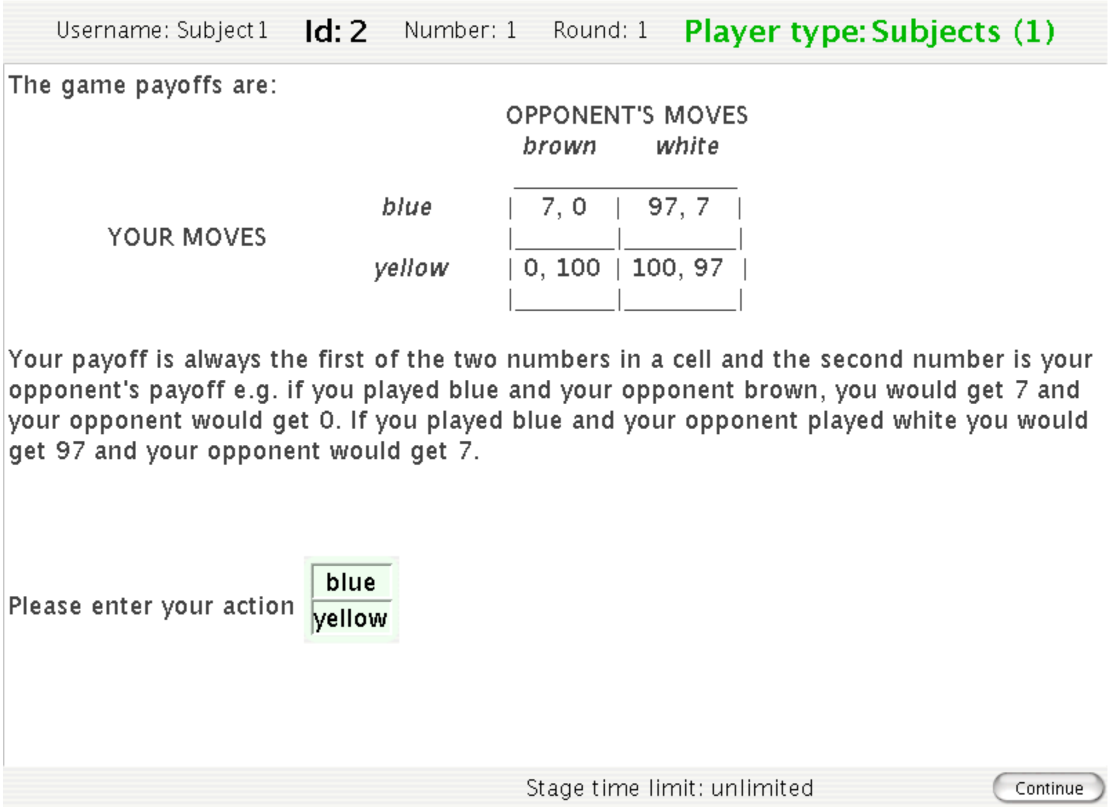
The real money that you will earn will be calculated as follows. You will receive in euro the average of your payoffs from ALL 300 rounds. For example, if your average payoff for the 100 rounds against the first opponent was 10, for the 100 rounds against the second opponent was 5 and for 100 rounds against the third opponent was 15, then you would earn (10+5+15)/3=10 euro.

You are guaranteed to receive a minimum of 5 euro for your participation i.e. if your average payoff points at the end of the experiment are less than 5 your payment will be exactly 5 euro. Furthermore, after all the sessions of the experiment are concluded, the player who achieved the highest average payoff will be paid a further 30 euro, whilst the player with the second highest payoff will receive 20 euro.

During each decision making round you will have to decide between two possible actions (or moves) referred to as blue or yellow. Your computer opponent will simultaneously make its own decision whether to play its brown or white action - neither you nor your opponent will have any knowledge regarding the other's choice in that round before deciding.

EXAMPLE:

Before every round you will be shown a screen similar to this one:



Username: Subject1 **Id: 2** Number: 1 Round: 1 **Player type: Subjects (1)**

The game payoffs are:

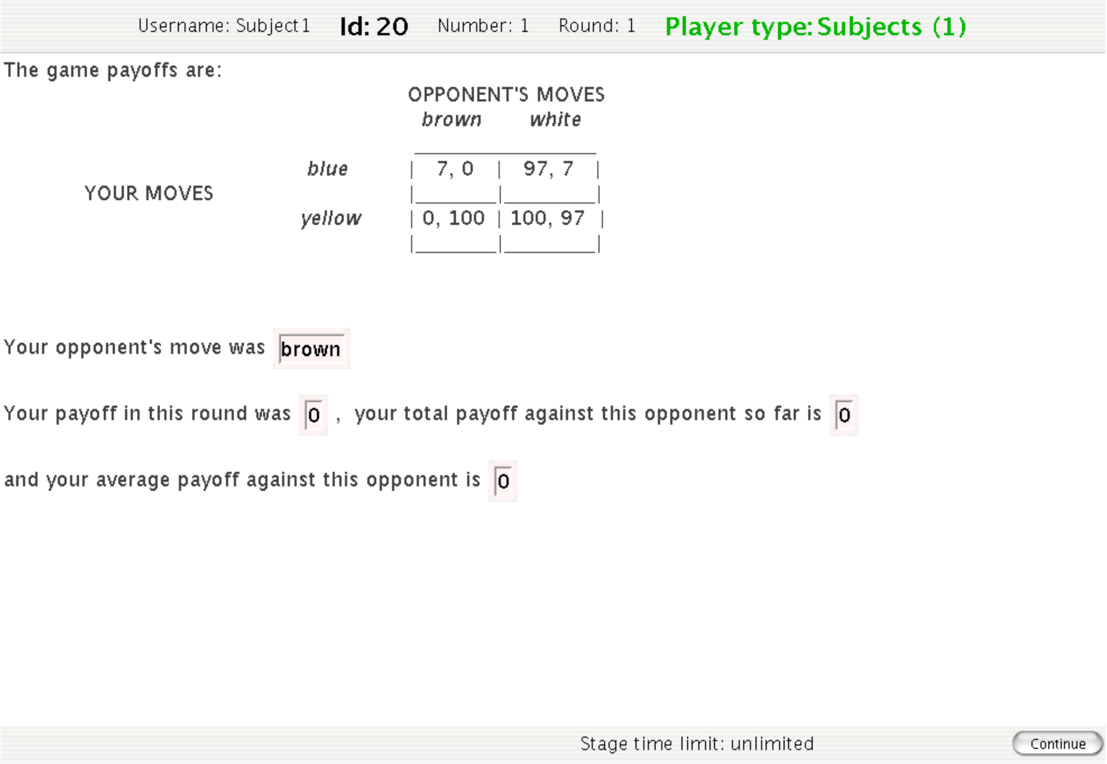
		OPPONENT'S MOVES	
		brown	*white*
YOUR MOVES	*blue*	7, 0	97, 7
	yellow	0, 100	100, 97

Your payoff is always the first of the two numbers in a cell and the second number is your opponent's payoff e.g. if you played blue and your opponent brown, you would get 7 and your opponent would get 0. If you played blue and your opponent played white you would get 97 and your opponent would get 7.

Please enter your action

Your payoff is always the first of the two numbers in each cell, and the second number is your opponent's payoff. In this example, if you played blue and your opponent brown, you would get 7 and your opponent would get 0. If you played blue and your opponent played white you would get 97 and your opponent would get 7. Please make sure that you understand how this works, if you in any way uncertain please raise your hand now. You will be playing exactly the same game (with the same payoffs) for all the rounds and against each computer program. However you do not need to memorize the payoff table as it will be presented to you before every decision round.

After you have decided which move you want to make you can enter it simply by clicking on either blue or yellow (next to the text "Please enter your action"), and then confirming it by clicking on Continue. You will then be presented with an information screen regarding the outcome of that round. You will be informed what action your opponent chose, your payoff from that round, your total payoffs and average payoffs against that particular opponent. This is a sample information screen:



Username: Subject1 Id: 20 Number: 1 Round: 1 Player type: Subjects (1)

The game payoffs are:

OPPONENT'S MOVES

		brown	white
YOUR MOVES	blue	7, 0	97, 7
	yellow	0, 100	100, 97

Your opponent's move was brown

Your payoff in this round was 0 , your total payoff against this opponent so far is 0

and your average payoff against this opponent is 0

Stage time limit: unlimited

At certain points during the experiment you will be presented with some questions, please follow the on-screen instructions and answer them as best you can. There is no “right” or “wrong” answer, and your answers do not affect how your computer opponent plays, your payoffs or the experiment in general in any way.

If you have any queries please raise your hand now, otherwise please wait for the administrator to begin the experiment.