How to Learn to Defeat Noisy Robot in Rock-Paper-Scissors Game: An Exploratory Study[1]

Gregory Chernov

National Research University Higher School of Economics,
11, Pokrovsky blvd, Moscow, 109028, Russian Federation.
E-mail: gchernov@hse.ru

This paper studies learning in strategic environment using experimental data from the Rock-Paper-Scissors game. In a repeated game framework, we explore the response of human subjects to uncertain behavior of strategically sophisticated opponent. We model this opponent as a robot who played a stationary strategy with superimposed noise varying across four experimental treatments. Using experimental data from 85 subjects playing against such a stationary robot for 100 periods, we show that humans can decode their strategies, on average outperforming the random response to such a robot by 17%. Further, we show that human ability to recognize such strategies decreases with exogenous noise in the behavior of the robot. Further, we fit learning data to classical Reinforcement Learning (RL) and Fictitious Play (FP) models and show that the classic action-based approach to learning is inferior to the strategy-based one. Unlike the previous papers in this field, e.g. Ioannou, Romero (2014), we extend and adapt the strategy-based learning techniques to the 3×3 game. We also show, using a combination of experimental and ex-post survey data, that human participants are better at learning separate components of an opponent's strategy than in recognizing this strategy as a whole. This decomposition offers them a shorter and more intuitive way to figure out their own best response. We build a strategic extension of the classical learning models accounting for these behavioral phenomena.

[1] The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'. Author is also grateful to German Academic Exchange Service (DAAD) for their support and funding (57507441).

Author declares that there is no conflict of interest.

Author is grateful to Philipp Chapkovski, Peter Dürsch, Uri Gneezy, Heike Hennig-Schmidt, Alexander Nesterov, Jörg Oechssler, John Rust, Timothy Salmon, for helpful discussions. The author also expresses his special thanks to the anonymous referee for useful suggestions, as well as to Alexis Belianin, and Ivan Susin for their help in working on the text.

Gregory Chernov – Research Assistant: International Laboratory for Experimental and Behavioral Economics.

The article was received: 06.10.2020/The article is accepted for publication: 11.11.2020.

***Key words*:** adaptive learning; repeated games; scoring rules; simulation methods; belief learning; repeated-game strategies.

JEL Classification: D91, C53, C57, C73, D83.

DOI: 10.17323/1813-8691-2020-24-4-503-538

For citation: Chernov G. How to Learn to Defeat Noisy Robot in Rock-Paper-Scissors Game: An Exploratory Study. *HSE Economic Journal.* 2020; 24(4): 503-538.

Introduction

Decision-making in strategic environments appears in the core of many economic problems, ranging from efficient contracts [Li, 2017] to oligopolistic markets [Doraszelski, Lewis, Pakes, 2018; De Roos, 2019] and agent-based interaction [Zohreh, 2012]. However, human abilities to cope with strategic uncertainty [Brandenburger, 1996] and learn how to best respond against a rational opponent are still under-investigated at large.

The literature on learning in games [Erev, Haruvy, 2013] focuses primarily on the behavior of rational agents who base their updated decisions either on the frequency of opponents' plays (Fictitious Play (FP) type models, [Fudenberg, Levine, 1999]) or on the relative performance of own strategies (Reinforcement Learning (RL) type models [Sutton, Barto, 2017]). While these models have been remarkably successful in explaining the dynamics of learning and non-equilibrium behavior, they face a new kind of challenge when applied to learning against sophisticated opponent. Such opponents take account not only of the past actions of the opponents actions, but hypothesise over a set of behavioral rules which players can learn in the process of strategic interactions [Milgrom, Roberts, 1991; Camerer, Ho, 1999].

Optimal behavior against such players should be more than simple actions and, according, to Hanaki (2004), it may suffice to make strategies that condirion actions to th history of the previous round as first proposed by Aumann (1981). For the 2×2 games, the set of strategies constructed in that way provides a manageable learning model because the total number of distinct decision rules does not exceed $2^5 = 32$, but for larger games the number of such rules grows too quickly to be tractable by a human agent[2]. To provide a tractable model of learning in such cases, we propose and implement splitting of human decision process into simpler «elementary strategies» (see section 2.2) for a wider class of games than 2×2. Those «elementary strategies» are a very natural basis to think about boundedly rational decision rules in general.

The present paper explores human abilities to learn in a strategic setting by means of an experiment where human subjects encounter a robot with a fixed complexity of behavior. We use Rock-Paper-Scissors (RPS) game since it is a 3×3 game with no best reply in pure strategies (Rock defeats Scissors, Scissors defeat Paper, Paper defeats Rock), resulting in a unique Nash

[2] In an *N*×*N* action game, the number of possible 1-step histories is N^2. A complete «finite automaton» maps each history to one of N next actions: $A: N^2 \to N$. Two automata are distinct if they differ in any action after any history. So the total number of different automata is N^{N^2} (selecting one of *N* possible actions for each of N^2 histories). For $N = 2$ this equals 16, but even for $N = 3$ it amounts to 19683, etc.

equilibrium in mixed strategies. We introduce a robot player in order to fix the strategy and level of sophistication of an opponent to the human players, which allows us to see check whether (and how fast) she learns the optimal strategy against such a stationary opponent.

Human participants of our experiment (N = 85) were (truthfully) informed they are playing against a robot who is pre-programmed to play a particular strategy' but were not told what this strategy is. The task of human subjects was to decode this strategy in the course of 100 rounds, trying to beat the robot by the largest possible margin over the Nash equilibrium performance of 33% of gains. After the game, they were asked to complete a short questionnaire about themselves and to report the strategy they thought the robot was programmed to play. We have implemented four treatments, which differ by the level of «noise». That is, the robot played the Nash equilibrium uniform mixed strategy with fixed probabilities of 0.2, 0.4, 0.6 and 1, and behaved according to its pre-programmed strategy with complementary probability. Our experimental data shows that human players are indeed able to cope with strategic uncertainty and beat the robot on average by 17%, the most successful participant winning as many as 75 out of 100 trials (in contrast to random guesser's 33 percentage of winning). Further, efficiency of learning is inversely related to the level of noise, and the lower the noise was, the more components in the Robot strategy have the participants successfully decoded in their reports.

We extend the setup of Ioannou and Romero (2014) for the 2×2 games to the symmetric game with 3 strategies and compare two action-based learning algorithms (Weighted Fictitious Play [Cheung, Friedman, 1993) and Reinforcement Learning [Roth, Erev, 1995]) to their strategic extended counterparts. We build such extensions by incorporating «elementary strategies» of full finite automation to original learning algorithms. In our simulations, we run the various algorithms on the same playing histories that were encountered by our subjects during the experiment. We treat what models would play in those situations as their predictions. Comparing these modeled actions to those chosen by the human subjects using a prediction error metric (the Brier-score [Brier, 1951])[3]. We confirm that strategic models outperform the action-based, and Strategic RL does so by a wide margin.

The rest of the paper is organized as follows. Section 1 briefly overviews the related literature. Section 2 describes the setting of the experiment, notions of analysis, modeling and hypotheses, Section 3 offers a general overview of experimental results and relates them to the level of noise and reported strategies of the robot. Section 4 fits the action-based and strategic learning models to our experimental data and compares their predictive power. Final section summarizes our findings and concludes. Details of the experimental setup and technical results are collected in the Appendix.

1. Related literature

Current state of affairs, terminology, comparisons. Models of learning in games are now widespread in theoretical [Fudenberg, Levine, 1998; Nachbar, 1990], experimental [Roth, Erev, 1995; Camerer, Ho, 1999; Ioannou, Romero, 2014], and empirical literature [Doraszelski, Lewis, Pakes, 2018; David, De Roos, 2019]. The classical way to model strategic behavior in repeated interaction, which dates back to Cournot, is based on best response to the current statistically

[3] A companion paper [Chernov, Susin, Cheparuhin, 2020] discusses and develops this technique in detail.

prevalent action of the opponent. Here action as «units» of learning refer to particular option which could be chosen e.g. «e2-e4» in chess or «fold» in poker. Furthermore, there are about two dozen learning models that differ from the classical one either by how the *a priori* probability distribution is chosen or by the restrictions that are imposed on the rationality of the agent.

The classical rational rules based on Bayesian updates stem from the Fictitious Play (henceforth FP) [Brown, 1951] and most bounded rationality rules rely on the Reinforcement Learning (henceforth RL) [Roth, Erev, 1995]. To imitate the tendency among humans to break cycles of defeats, a weighted version of FP and RL use mixed strategies defined over *the space of actions*, i.e., moves.

Another approach to learning is to try to reproduce human behavior in difficult strategic situations [Camerer, 2018] by introducing complex cognitive strategies, e.g. pattern recognition [Spiliopoulos, 2012], conditional strategies [Hanaki at al., 2018], forward-looking beliefs [Duersch, Kolb, Oechssler, 2010]. All of these can be called «strategic» or conditional learning, in contrast to the earlier action-based learning, because the agent focuses her attention not on her own and partner's actions, but on *strategies*, i.e., rules or functions that map conditions (history) into new actions (see related extensive review on learning in [Nachbar, 2009]).

Conditional strategy-based models also relate to the experimental paper on finite automata, (first proposed by Aumann, 1981) by Hanaki (2004), who introduces conditional learning strategies into the model of learning (see also [Ioannou, Romero, 2014]).

In measuring «quality» of the learning models, we borrow from Arifovic, McKelvey, Pevnitskaya (2006), who compare learning algorithms on out-of-sample data by predictive metrics. They find that almost all mentioned action-based models perform worse than a random predictor in certain games.

Also, Mathevet and Romero (2012) conduct crosswise comparisons between the two main classes of learning models on experimental data with several 2×2 games. They also compare the typical behavior of models and humans in those games, however, at a higher level of aggregation than we do.

All this literature provides sufficient evidences that conditional strategies and predictive metrics are useful to understand learning, and we contribute to this literature by extending it to 3×3 games using new data, and exploring playing trajectories of the most successful learners.

2. Setup and notations: actions against strategies

2.1. Notions and notations: playing rules and history

We are now turning to a formal description of the experimental task. Two players $\mathcal{I} = \{1,2\}$ interact during a finite number of periods $T = 100$. Their interaction constitutes a *normal form repeated game* $G = \langle \mathcal{I}, A_{\mathcal{I}}, \{u\}, T \rangle$, where $a \in \{Rock, Paper, Scissors\} \equiv A$, $\{u\} \equiv \{u_1(a_1, a_2), u_2(a_1, a_2)\}$ are two instant payoff functions, whereas $A_{\mathcal{I}} = A \times A$ is the set of actions' profiles, $A \ni a_i = \{Rock, Paper, Scissors\} : i = \{1,2\}$. Let $x_i^{\{t-1\}}$ denote the actions of the previous round that are taken by the player i to formulate his or her decision in the current round. In case of action-base decisions, this is $x_i^{t-1} = \left(a_i^{t-1}\right) \in \{Rock, Paper, Scissors\}$,

while in the case of one-period strategies, this is one of *9 pairs* of player's and opponent's actions: $x_i^{t-1} = \left(a_i^{t-1}, a_{-\mathrm{i}}^{t-1} \right)$ where $a_i^{t-1}, a_{-\mathrm{i}}^{t-1} \in A \times A$ one of the profiles of actions taken by both players. In this paper we focus on *instant history* denoted by h^{t-1}: $h^{t-1} \equiv \left(a_i^{t-1}, a_{-i}^{t-1} \right)$, Complete history $H^{t-1} \equiv \left(h^1 \ldots h^{t-1} \right)$ consists of a sequence of all previous actions, but we limit attention to h^{t-1} or its simple aggregates like $\sum_{t=1}^{t-1} h^t$. This is consistent with the literature and the intuition about the limited information processing abilities of human subjects.

Situations in symmetric games like RPS may be further simplified thanks to the anonymity property of actions. *Anonymity* here means that re-labeling the actions will not change strategies: what matters is how players react to Win, Tie or Loss, but it does not matter whether the winning (tying, losing) action was Rock, Paper or Scissors. Hence, the «current situation» x^t can be descrybed simply as: $x^t \in (Win_i, Tie_i, Lose_i)$.

It is convenient to describe actions and strategies using modular arithmetic notation (modulo-3). It means that we enumerate the initial set A of actions as $A = \{Rock, Paper, Scissors\} = \{0,1,2\}$, then any action $a \in A$ becomes equivalent to other actions by the circular modular rule:

$$(a+3) \bmod 3 = a, (a+2) \bmod 3 = (a-1) \bmod 3, (a+1) \bmod 3 = (a-2) \bmod 3.$$

This game itself invites the player to think symmetrically in the modulo-3 arithmetic terms because Rock beats Scissors, Scissors beat Paper and Paper circularly beats Rock. Players need not care which particular action has been successful in the past. Verbal reports of the participants of our experiment after the game (see the next section) often demonstrate such *symmetric* perception of the current situation x^t.

Throughout our paper, a «conditional strategy» will be defined as the transition function from instant history to present period action[4]:

$$\xi_i : \left(a_i^{t-1}, a_{-\mathrm{i}}^{t-1} \right) \to A. \quad (1)$$

In the logic of modular arithmetic, «Stay» will denote the action repeating the previous action, «Up» – the action that beats the previous action chosen by the same player (e.g. if the

[4] Theoretically speaking, conditional strategies can be more complicated. Typically, they are formalized as finite automata, which determines how should the player change her action (state) depending on the action of the opponent. Formally the automaton is a quadruple $S_{conditional} = \left\{ q^0, Q_i, f_i, \xi_i \right\}$, where q^0 represents the initial state (unnecessary in our case, where the first transition occurs randomly), Q_i is a set of states, output function $f_i : Q_i \to A_i$ is an assignment of an action to every state, and $\xi_i : Q_i \cdot A \to Q_i$ is the transition function which maps the initial state to the new state depending on the action taken. Concrete automata used here contain only actions as states: $Q_i = A_{\mathrm{i}}^{t-1}, i = \{1,2\}$ and the output function is the same as transition function. Hence the finite automaton in the RPS game can be fully described by ξ_i and $A_{\mathrm{i}}^{t/t-1}$ only.

previous action was Rock, then the «Up» action is Paper), «Down» – the action that would be beaten by the previous one (and beats the «Up» action, i.e. Scissors in the previous example).

A very simple, unsophisticated player may expect her opponent just to repeat the previous move $a^{t+1} = a^t$. Best reply against such a strategy will be «Up». Alternatively, a Naive player may think not about the opponent, but about own success and play $Win \to Stay$, $Lose \to Shift$, where Shift may mean either $Lose \to Up$ or $Lose \to Down$. In the light of the evidence of Wang, Xu, Zhou (2014) where humans play with humans, we define the Naive-learner's strategy of switching ($\to$) as:

$$Naive: Win \to Stay,\ Tie \to Down,\ Lose \to Up. \quad (2)$$

In our experiment, one of the players is a *Robot* who is programmed to play the best reply strategy against such Naive player. Hence, the Robot playing strategically in terms of the Naive opponent's actions, uses:

$$Robot: Win \to Up,\ Tie \to Stay,\ Lose \to Down. \quad (3)$$

What should be the winning strategy of a human player against such a deterministic Robot? A moment reflection reveals that the OptAR (Optimal-Against-Robot) strategy is:

$$OptAR: Win \to Down,\ Tie \to Up,\ Lose \to Stay. \quad (4)$$

For instance, if the winning strategy of a Robot player was (without loss of generality) Rock, it is supposed to go Up to Paper, hence an optimal strategy of the human player is to stick to her previously played Scissors which will be winning now. Similarly, if the Robot has lost with Rock, he is supposed to go Down to Scissors, so the optimal strategy of a human who played Paper is to chase it Down to Rock. Finally, under Tie in Rock, the Robot is meant to stay, so it is optimal to mount Up to Paper to beat it.

2.2. Experimental setup

In our experimental settings, we ask participants to repeatedly play against the preprogrammed Robot opponent in the Rock-Paper-Scissors game.

The Rock-Paper-Scissors was chosen for our experiment because is a special type of game for learning dynamics. Only in this type of games (in stark contrast to various games from prisoners' dilemma to cooperation and chicken games) learning agent cannot guarantee some minimal payoff by just imitating the opponent instead of actually learning, as shown by Duersch, Oechssler, and Schipper (2012).

A second major design choice that separates this work from Ioannou and Romero (2014) is that in their work the focus is on how human players interact in a game, hence their learning dynamics is endogenous to the realized game path. We ask a more specific question: «when did learning happen, depending on the noise in the treatment, and how is it connected to recognition of the rule». Such inferences cannot be made when our subjects play against each other, but it is easy when one of the players is directly programmed.

Finally, for our experimental design purposes, it is important to calibrate the complexity of the rule so that the learning happens during the whole experimental session. If our rule is too

simple and our subjects learn it during the first 10% of the play – only those 10% will be «learning» data while the rest is much less interesting «best response» data. If the rule is too hard and very few subjects learn anything during the experimental session – again we are left with less useful data than if the learning complexity is just right for a feasible experimental session.

Thus, we calibrate the complexity in the following manner. Consider a Robot endowed with deterministic strategy (3) only in some random periods and plays equilibrium mixed strategy of 1/3 to each pure strategy otherwise[5]. Then we could vary the level of nonrandomness by treatment. Optimal response against such Robot remains the same as without noise because noise *per se* does not change *anything* in the deterministic component of the strategy of a Robot, hence optimal behavior against it remains the same. Also, to use OptAR our player need not have any priors about her opponent Robot: OptAR is just the maximal-expected-payoff strategy in any state.

We can look at the frequency of OptAR moves during any personal trajectory, and simply check at which moment, does it exceed 1/3. Once traced, this moment could be taken as the «learning point». However, typically the players never stop making some moves that look random, which can be attributable either to continuous experimentation or to random noise. Following for setup description, we need to introduce the extensions of classical learning models that be able to capture the behavior of OptAR Robot with noise.

2.3. Search for a «good» learning algorithm: actions, strategies and sub-strategies

2.3.1. Benchmarks and classification

To model human behavior, there exist a broad variety of algorithms, that we classify for the purpose of this paper in only two respects[6].

First, we can distinguish the *belief-based* algorithms that explore the opponent's behavior – from those *based on own wins/fails* [Fudenberg, Levine, 1998; Nachbar, 1990]. In other words, a player can either learn the opponent's actions or learn their own success per se. Among the belief-based class, we focus below on the Weighted Fictitious Play (WFP, see [Cheung, Friedman, 1993][7]. Alternatively, the Reinforcement Learning model (RL, see [Roth, Erev, 1995]) does not use any beliefs about the opponent or direct utility maximization. In response to history, RL generates a «propensity score» for each own action: the higher was the frequency of wins from a certain action, the higher will be the probability to use this action again. RL and similar learning algorithms are more sensitive to payoffs than to actions per se, unlike the belief-based schemes.

Second, both classes mentioned can be applied either to *actions* $a \in \{Rock, Paper, Scissors\}$, or to *strategies*. Here we should note that the rule of successful OptAR strategy in principle can

[5] Thus, it plays action according to Robot rule sometimes «intentionally» and sometimes by chance.

[6] See other possible classifications in Nachbar (2009).

[7] The standard FP algorithm just uses the whole current history of the opponent's actions «*frequencies*» as the predicted *probabilities* of her next action and plays the best response to these. WFP differs from FP by playing a stochastic best-response, rather than a deterministic one, to undermine the opponent's learning capacities presumably for strategic reasons. WFP better suits our purposes, being comparable with other stochastic algorithms and humans.

be decomposed into 3 «successful sub-strategies» σ_j. These are three distinct moves that can bring positive expected payoffs, namely, action a_i^t of player i should be modified as:

$$(\sigma_I): Lose \to Stay, \text{ i.e., } \left[a_{-i}^t = \left(a_R^{t-1} - 1 \right) mod\, 3 \Rightarrow a_i^{t+1} = a_i^t \right],$$

$$(\sigma_{II}): Win \to Down, \text{ i.e., } \left[a_{-i}^t = \left(a_R^{t-1} + 1 \right) mod\, 3 \Rightarrow a_i^{t+1} = \left(a_i^t - 1 \right)\ mod\, 3 \right],$$

$$(\sigma_{III}): Tie \to Up, \quad \text{i.e., } \left[a_{-i}^t = a_R^{t-1} \Rightarrow a_i^{t+1} = \left(a_i^t + 1 \right)\ mod\, 3 \right].$$

Each strategy ξ can be decomposed in a similar manner, for instance, Robot strategy is $\xi^R = \{ Lose \to Down, Tie \to Stay, Win \to Up \}$, consists of three sub-strategies. Since these three sub-strategies can be learned (or not) separately, the three sub-strategies of the Robot can be expressed in terms of non-anonymous instant history $h^{t-1} = \left(a_i^{t-1}, a_{-i}^{t-1} \right)$. In these terms, a strategy ξ is a composite of elementary strategies s – simple functions from possible «states of the world» to actions A. Such a collection of arguments and outcomes $\left(a_i^{t-1}, a_{-i}^{t-1} \right) \to a_{-i}^t$ can be perceived as vectors $\left(a_i^{t-1}, a_{-i}^{t-1}, a_{-i}^t \right)$, where the third component is the opponent's today reaction to yesterday's situation. We have denoted $0 \equiv Rock$, $1 \equiv Paper$, $2 \equiv Scissors$. Table 1 presents two particular, Robot and OptAR, *complete strategies* ξ^R, ξ^O both in anonymous (σ) or in «named» (action-specific s) elementary strategies:

Table 1.

Components of Robot's strategy ξ^R and Optimal strategy ξ^O

Anonymous ξ^O	$\sigma_I \equiv Lose \to Stay$	$\sigma_{II} \equiv Tie \to Up$	$\sigma_{III} \equiv Win \to Down$
	$s_{10} \equiv (0, 1, 0)$	$s_{13} \equiv (0, 0, 1)$	$s_{16} \equiv (0, 2, 2)$
Named ξ^O	$s_{11} \equiv (1, 2, 1)$	$s_{14} \equiv (1, 1, 2)$	$s_{17} \equiv (1, 0, 0)$
	$s_{12} \equiv (2, 0, 2)$	$s_{15} \equiv (2, 2, 0)$	$s_{18} \equiv (2, 1, 1)$
Anonymous ξ^R	$\sigma_{IV} \equiv Lose \to Down$	$\sigma_V \equiv Tie \to Stay$	$\sigma_{VI} \equiv Win \to Up$
	$s_1 \equiv (0, 1, 2)$	$s_4 \equiv (0, 0, 0)$	$s_7 \equiv (0, 2, 1)$
Named ξ^R	$s_2 \equiv (1, 2, 0)$	$s_5 \equiv (1, 1, 1)$	$s_8 \equiv (1, 0, 2)$
	$s_3 \equiv (2, 0, 1)$	$s_6 \equiv (2, 2, 2)$	$s_9 \equiv (2, 1, 0)$

Similarly to ξ^R, ξ^O (Here index O = Optimal) we can describe complete strategy ξ_N (index N = Naive). Of course, there can be many others – in principle any learning algorithm should consider *all* of them, instead of several ones, predetermined by the researcher. However, humans

cannot compare thousands of hypotheses in the course of a game. To simplify the task, we decompose the complete strategy into sub-strategies, thus reducing the space of hypotheses that the human agent needs to explore.

Indeed, looking at Table 1 and trying all possible combinations, we see that a human seeking OptAR has to explore *27 named elementary strategies* $s_k \in \{s_1,...s_{27}\}$ instead of the set of all 3^9 possible strategies – and this is exactly what we do. In principle, any complete strategy ξ can consist of various combinations of s. Therefore, the space of all named complete strategies is very (!) large: 3^9. Hardly during 100 rounds, our humans could learn the opponent trying *all* hypotheses among these 3^9 named strategies. Instead, we have programmed a version of a strategic learning algorithm that learns 27 sub-strategies (importantly, sub-strategies *are* often reported by humans as will see).

Another possible way to reduce the space of strategies would be to learn only among *anonymous* strategies. There are not too many, 9 of these. Indeed, in addition to Naive, Robot, OptAR already described, one can consider only 6 other mappings from anonymous situations $\{Win, Tie, Lose\}$ to anonymous actions $\{Up, Stay, Down\}$. Yes, as we have explained already, the symmetry of our RPS game and the symmetry of our Robot suggests that the anonymous strategic approach can be quite relevant in our specific setting of the RPS game. Moreover, the further effort-economizing possibility is to model learning as separate *learning of sub-strategies* responding to 3 separate anonymous «states of the world» $\{Win, Tie, Lose\}$. Yes, we have seen that enough players did report their finding in such terms, as anonymous sub-strategies.

The reason why we decided to define algorithms on named ones instead of unnamed ones is the possibility of generalization of such algorithms. Let us consider any deviation from the RPS game, e.g. an asymmetric payoff. We argue that the *relevance of this approach would disappear if we make payoffs asymmetric* among Rock-Paper-Scissors actions, or our players would expect asymmetric love for certain actions from Robot. Our learning algorithm should be universal, not programmed to target and find the specific answer «symmetric Robot» hidden by the researchers in this specific setting! This argument works for any deviation from the basic game.

The distinctions in approaches are represented in Table 2 as a taxonomy.

Table 2.

Taxonomy of the learning models

Models	Belief-based:	Reinforcement-based:
3-action-based:	Weighted Fictitious Play	Reinforcement Learning
strategy-based:	Strategic Weighted Fictitious Play	Strategic Reinforcement Learning

Now we should further explain our specific «strategy-based versions» of algorithms mentioned in Table 2. Following Ioannou and Romero (2014), we extend WFP and RL from the small set of actions $A = \{Rock, Paper, Scissors\}$ – to a larger set, containing *all named elementary strategies*:

$$S = \{s_1,...,s_{27}\}. \quad (5)$$

Such modifications, called «strategic versions» of FP or RL algorithms [Hanaki, 2004].

We modify the original «strategic» version by dividing ξ (a compound from 3 sub-strategies) into elementary strategies s. Among a enormous variety of possible learning algorithms, we present here four ones, already known in the literature and simple enough to implement and explain. In particular, the simplification is that our «strategies» do not use depth-2 or deeper history: they are based on instant histories only.

2.3.2. Models and extensions

For introducing the internal mechanic of learning models let's describe the rules ξ more formally. Since $u\left(a_i^{t-1}, a_{-i}^{t-1}\right) \equiv \{Win_i, Tie_i, Lose_i\}$ we quantify payoffs as $u(Lose) = 0;$ $u(Tie) = 1, u(Win) = 2.$ We will also use those payoffs as rule modifiers in our further modular calculations. Summation of actions and payoffs may look odd to a game theorist but it simplifies notation and implementation in code. Next, the «Naive» rule (here index N = Naive) ξ_i^N in these terms is defined through the transition

$$\xi_i^N\left(h^{t-1}\right) = \left(a_i^{t-1} + u\left(a_i^{t-1}, a_{-i}^{t-1}\right) + 1\right) \mathrm{mod}\, 3. \tag{6}$$

To explain, let us interpret «$Win \to Stay$» sub-strategy of the Naive player defined in eq. (2). For example, in the $t-1$ the player has chosen $a_i^{t-1} = Rock = 0$ and got $u\left(a_i^{t-1}, a_{-i}^{t-1}\right) = Win = 2$ (by our notation). Now, rule ξ_i^N produces 0 + 2 + 1 = 3, whereas $3 = 0 \ \mathrm{mod}\, 3$, thus the rule prescribes to Stay (zero shift means «repeat the previous action»). Similarly, one can check that this formula ξ_i^N also reflects two other sub-strategies $Tie \to Down$, $Lose \to Up$ in all situations Rock, Paper, Scissors.

Another strategy ($Win \to Up$, $Tie \to Stay$, $Lose \to Down$), defined in eq. (3) (henceforth «Robot», it is the best response to the previous one and describes our Robot) can be written with the modular arithmetic as:

$$\xi_i^R\left(h^{t-1}\right) = \left(a_i^{t-1} + u\left(a_i^{t-1}, a_{-i}^{t-1}\right) - 1\right) \mathrm{mod}\, 3. \tag{7}$$

Finally, the «Optimal» against such Robot strategy ($Win \to Down$, $Tie \to Up$, $Lose \to Stay$) from eq. (4) becomes

$$\xi_i^O\left(h^{t-1}\right) = \left(a_i^{t-1} + u\left(a_i^{t-1}, a_{-i}^{t-1}\right)\right) \ \mathrm{mod}\, 3. \tag{8}$$

Comparing these three strategies ξ_i^j, one can see that they differ by 1, exhausting all 3 possibilities Up, Stay, Down. Now when we have an intuition of transition mechanics of strategy we could write down two classic learning models and our extension of them (the code is available from the author).

Weighted Fictitious Play (WFP). In the WFP model [Cheung, Friedman, 1993], each player i has three counters κ_{ik}^t, one for each opponent's action $k \in \{0,1,2\}$. Starting with

$\kappa_{ik}^0 = 0$, at each period time $t \in \{1,2,\ldots T\}$, these three counters are updated (enlarged or not). Namely, based on yesterday history h^{t-1}, we enlarge the k^{th} counter κ_{ik}^t by 1 when the opponent's observed action a_{-i}^{t-1} is equal to this k:

$$\kappa_{ik}^t = \kappa_{ik}^t\left(h^{t-1}\right) := \kappa_{ik}^{t-1} + \begin{cases} 1, \ if \ a_{-i}^{t-1} = k \\ 0, \ if \ a_{-i}^{t-1} \neq k \end{cases} \forall k \in \{0,1,2\}.$$

Belief γ_{ik}^t of player i that his/her opponent $(-i)$ will choose action k at period t is defined as the relative weight, i.e., the empirical frequency of this action, aggregating the opponent's complete observed history H^{t-1}:

$$\gamma_{ik}^t = \gamma_{ik}^t\left(H^{t-1}\right) := \frac{\kappa_{ik}^t}{\sum_{j=0}^{2} \kappa_{ij}^t} \forall k \in \{0,1,2\}.$$

Unlike deterministic FP that reacts only to the most probable action of the opponent, the WFP algorithm reacts to random actions, appearing with probabilities γ_{ik}^t. It always chooses the best response, which is an offset +1 to each action. Consequently, based on history h^{t-1}, our WFP generates a random action a_i, where probability $p_{i(k+1 \bmod 3)}$ to choose an action number $(k+1 \bmod 3)$ is:

$$p_{i(k+1 \bmod 3)}^t := \gamma_{ik}^t\left(H^{t-1}\right) \forall k \in \{0,1,2\}.$$

Weighted Strategic Fictitious Play (WSFP). The transition from the WFP to the WSFP, i.e., from actions to strategies is achieved by modifying the counter for the opponent's strategy. Instead of three counters κ_{ik}^t for the opponent's actions, now we update 27 counters η_{im}^t, one for each named elementary strategy $s_{-im}^t \in \{s_1,\ldots s_{27}\}$ of the opponent (see Table 2). In other respects, WSFP algorithm is programmed alike WFP; it updates these 27 beliefs about the opponent. Namely, it adds 1 to m^{th} current counter η_{im}^t when the yesterday observed opponent's conditional strategy takes the elementary value s_m, otherwise the counter remains the same:

$$\eta_{im}^t = \eta_{im}^t\left(H^{t-1}\right) := \eta_{im}^{t-1} + \begin{cases} 1, \ \ if \ \ \left(a_i^{t-2}, a_{-i}^{t-2}, a_{-i}^{t-1}\right) = s_m \\ 0, \ \ if \ \ \left(a_i^{t-2}, a_{-i}^{t-2}, a_{-i}^{t-1}\right) \neq s_m \end{cases} \forall m \in \{1,\ldots,27\}.$$

Frequencies of opponent's elementary strategies are beliefs $\gamma_{ik}^t\left(s_{-i}^t\right)$ (that the opponent will play the named elementary strategy). The beliefs are based on 27 counters η_{im}^t:

$$\gamma_{im}^t = \gamma_{im}^t\left(H^{t-1}\right) = \frac{\eta_{im}^t}{\sum_{j=1}^{27}\left(\eta_{ij}^t\right)} \ \ (m = 1,\ldots,27).$$

Based on history H^{t-1}, our WSFP generates a player's random action a_i alike WFP. Again, probability $p_i(k+1 \bmod 3)$ to choose an action $a_i = (k+1 \bmod 3) \in \{0,1,2\}$ with number $(k+1 \bmod 3)$ is based on the beliefs, with an offset +1 to the opponent's expected action, which is the third component of m^{th} elementary strategy vector s_m:

$$p_i^t(k+1 \bmod 3) := \frac{\gamma_{ik}^t}{\sum_{j|(a_i^{t-1}, a_{-i}^{t-1}, j) \in s_j} (\gamma_{ij}^t)} (\forall m \mid (a_i^{t-1}, a_{-i}^{t-1}, k) \in s_m).$$

Reinforcement Learning. Reinforcement Learning models [Roth, Erev, 1995] assume that players adjust their strategies based on their past performance. Similar to beliefs γ_{ik}^t in the WFP model, now we update 3 counters, which are «propensities» π_{ik}^t, to generate probabilities to play each action $a \in \{0,1,2\}$. Each player i at each period t updates her propensity π_{ik}^t to play action k (starting from equal initial $\pi_{ik}^0 = 1$). The main difference from WFP is that now 3-component vector π_i^t is reinforced based on own strategies, not the opponent's ones. Reinforcement also exploits own payoffs instead of opponent's frequencies, and propensities are updated as follows:

$$\pi_{ik}^t := \pi_{ik}^{t-1} - u_{min} + \begin{cases} u(a_i^{t-1}), & if \ a_i^{t-1} = k \\ 0, & if \ a_i^{t-1} \neq k \end{cases} \forall k \in \{0,1,2\},$$

where $(u(a_i^{t-1}) - u_{min}) \in \{0,1,2\}$ is the excess payoff from the player's yesterday action a_i^{t-1} over the minimal payoff $u_{min} = \min\{u(0), u(1), u(2)\} = -1$. If some action has not been selected in this round, then its propensity remains unchanged. Similarly to FP, moves are random and the player's probability p_{ik}^t to choose any strategy k in the next round is assigned as k^{th} relative propensity:

$$p_{ik}^t := \frac{\pi_{ik}^t}{\sum_{j=1}^{J} (\pi_{ij}^t)} \ \forall k \in \{0,1,2\}.$$

Strategic Reinforcement Learning (SRL). The transition from the RL to its strategic version SRL, alike transition from the WFP to the WSFP, is achieved by modifying the counters. Instead of three counters π_{ik}^t for the player's actions, now we update 27 counters – propensities ζ_{im}^t for using each elementary strategy s_m:

$$\zeta_{im}^t = \zeta_{im}^t(H^{t-1}) := \zeta_{im}^{t-1} - u_{min} + \begin{cases} u(s_{m3}^{t-1}), & if \ (a_i^{t-2}, a_{-i}^{t-2}, a_i^{t-1}) = s_m, \\ 0, & if \ (a_i^{t-2}, a_{-i}^{t-2}, a_i^{t-1}) \neq s_m. \end{cases}$$

For $\forall m \in \{1,\ldots,27\}$. If a strategy has not been selected in this round, then its propensity remains unchanged. Again, the player's actions are random, and the probability of choosing a strategy s_m with action $a_i^t = k = s_{m3}$ in a situation $h^{t-1} = \left(a_i^{t-1}, a_{-i}^{t-1}\right)$ in the next round is similar to RL version but normalizes all propensities related to such situations (instead of beliefs):

$$p_i^t(k) := \frac{\zeta_{ik}^t}{\sum_{j|\left(a_i^{t-1}, a_{-i}^{t-1}, j\right) \in s_j} \left(\zeta_{ij}^t\right)} \left(\forall m \mid \left(a_i^{t-1}, a_{-i}^{t-1}, k\right) \in s_m\right).$$

2.4. Hypotheses

Our apriori expectation is that complete automaton structure cannot be learned during a relatively short period of 100 rounds. As we have mentioned in the introduction and section 2.2, there are too many (around sixty thousands if considering the initial starting state) automata ξ_R in a 3×3 game, so either people cannot recognize them at all or learn it only by parts through sub-strategies σ.

Almost all previous literature adopts the default theoretical assumption that any learning process can be decently approximated by action-based dynamics. Our Robot rules are constructed to outperform the simple action-based reinforcement. Therefore, we have two groups of hypotheses. The first two focus on the outcome of the experiment for human participants:

Hypothesis 1: Proportion of wins of human subjects does not exceed the Nash equilibrium share of 1/3. Essentially, this means that the difficulty of Robot strategy is too high to be learned by human subjects. Consequently, humans are not expected to be able to perform better or worse depending on the level of noise in the Robot's strategy. Hence our next hypothesis is.

Hypothesis 2: The level of noise in the Robot strategy does not affect the proportion of wins. Section 3 is devoted to testing of these hypotheses.

The remaining hypotheses concern the quality of algorithmic predictions of human subjects' behavior. Three alternative metrics are considered: (1) similarity of the outcomes of a learning algorithm to decisions of human players in terms of average payoffs, (2) similarity of algorithmic predictions to the frequency of optimal moves of human subjects, or (3) similarity of learning trajectories of the algorithm and humans (dynamic predictive quality). The last metric is the most interesting but also the most difficult computationally.

Any rational player could be expected to perform at least as well as the Nash equilibrium player choosing each strategy with equal probability. We use this player as a benchmark for learning, and compare each algorithm to this benchmark.

To compare our algorithms in three metrics with the benchmark player, we formulate the following specific hypotheses. **Hypothesis 3:** Standard action-based models outperform the benchmark player in terms of average payoff in the RPS game. **Hypothesis 4:** Action-based models playing against Robot perform at least as well as strategic-based ones in terms of average payoff in the RPS game. **Hypothesis 5:** Belief-based algorithms predict human behavior better than reinforcement-based ones. **Hypothesis 6:** Algorithms that play better against the Robot predict human behavior better.

2.5. Experiments description

Experiments were conducted in 2018 and 2020 at the Laboratory for Experimental and behavioral Economics at the Higher School of Economics (HSE), Moscow[8].

Subjects were given a very general hint: «You play against some Robot programmed to play a particular stationary strategy». Nothing was told about the Robot being stochastic or not, being maximizing some payoff or not, using any particular depth of history or not. Subjects in all treatments were given an assignment to try to guess the strategy of this Robot and over-perform it. Summaries of all sessions are presented in Table 3.

Remuneration consisted of three parts: 50 tokens[9] (150 rubles) show-up fee, per-game utility payment (0 tokens for loss, 1 for a tie and 2 for a win), and a bonus of 15 tokens for every 5 wins after the thirtieth win to stimulate extra-performance against «the correct» algorithm. We used a linear reward scheme for the number of wins exceeding what is expected by random guessing (see Appendix for the details). Average payoffs were equal to 5.95 EUR, except treatments with high school students, incentivized by prizes like books, increasing in quantity and quality according to performance (table 3).

For Instructions and details see Appendix. Our several treatments exploited a non-balanced sample of subjects (players) with the following descriptive statistics:

Table 3.

Sample description and payoffs

Noise level	Number of plrs	Percent of male	Participants – students	Average age	Min wins	Average wins	Max wins
1	8	38	university	22.8	26	**34**	39
0.6	9	33	university	23.8	30	**37**	50
0.4	39	44	high school	16.4	25	**40**	63
0.4	14	21	university	21.2	31	**43**	59
0.2	15	66	university	20.9	26	**54**	75
Total	**85**	42	–	**19.4**	**25**	**40**	63

3. Description of the experimental data: everybody tries to learn and some subjects succeed

This section expounds on how our subjects act against the Robot with fixed noise level that differs across treatments but implies the same best response for humans. We consider first the general observations and answers about whether people learn at all (hypothesis 1). Next, we proceed for the survey data analysis and show what aspects of Robot strategy that are explicitly formulated by subjects. We contrast this data with the data on the actual play history to comple-

[8] It used *oTree* [Chen, Schonger, Wickens, 2016], an open-source platform for running laboratory, online and field experiments.

[9] 1 token = 3 Russian rubles $\approx$ 0.05 euro at the time of the experiment.

ment the conclusions about the performance based on payoffs, and check whether noise affects this recognition (hypothesis 2). Finally, we explore individual learning trajectories, and try to evaluate some counterfactuals.

3.1. General observations

Table 3 shows that the average number of wins increases inversely to the level of noise: starting from 34 with noise level 1 (consistent with Nash equilibrium), the average performance of the humans monotonically goes up to 54 instances on average when the noise level drops to 0.2. The frequency of wins is significantly different from the theoretical frequency of 1/3 for noise levels of 0.2 (Wilcoxon sign rank test statistic $z = 3.81, p < 0.0015$) and 0.4 (Wilcoxon $z = 5.13$, $p < 0.000$), and are not significantly different for higher noise levels. Hence, we state that:

Result 1: Average number of wins exceeds what would be expected at Nash equilibrium in cases when the level of noise in the strategy played by the Robot is sufficiently low (0.4 or lower). Overall, participants play statistically better than Nash equilibrium player and do so more often the lower is noise. Consequently, our hypothesis 1 is rejected.

Many subjects did recognize some regularities and therefore did learn how to beat the Robot in our RPS game. Statistical z tests of this hypothesis based on individual data are provided in the App5.1.

We turn now from average payoffs to the *distribution* of personal achievements, grouping the subjects by treatments. Fig.1 compares four density plots for four treatments with noise levels 1, 0.6, 0.4, 0.2 (the less noise the darker is the tone of the bars). Decrease in the level of noise apparently provides better possibilities for learning, leading to smaller differences between individuals.

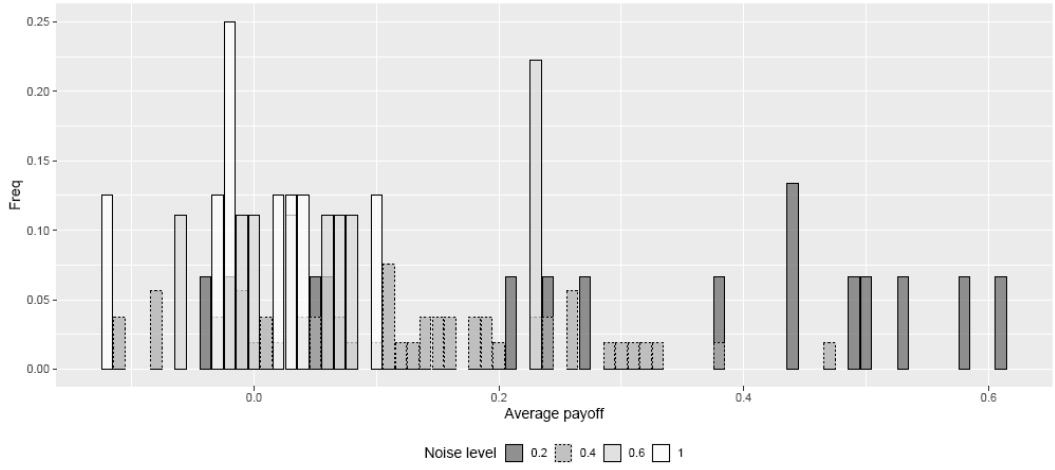


Fig. 1. Frequency of each level of final payoff under different noise levels

Observation 1. When noise decreases, both individual payoffs and their spread increase.

Histograms in Fig. 1 show the distributions of individual payoffs. Coordinate *X* denotes the per-round averaged payoff of each subject, normalized from $(0,1,2)$ matrix to $(-1,0,1)$ to center random guessing around zero. A negative value of coordinate X_j means that this *j*-s interval of subjects was beaten by the Robot. Positive X_j values correspond to the winners, an empty interval means that nobody got payoffs in this range. Coordinate *Y* denotes the share (percentage) of a related interval of payoffs within each of the four treatments and doesn't directly comparable across them.

For instance, under noise 0.2, we observe that the 15 dark bars are equally high because each payoff happened only once without coincidence.

Figure 1 visualizes a rightward shift in the distribution of payoffs when noise decreases. Indeed, under noise 1.0, the 8 payoffs (light bars) are more or less symmetrically distributed around 0, with the negligible average payoff –0.375. By contrast, under somewhat weaker noise 0.6 (gray bars with solid border), subjects have won, on average, as much as –0.111, because here winners became more numerous than losers. When noise decreases further to 0.4 (53 observations (grey bars with dotted border)), the average payoff rises up to +0.09 and we interpret that as a result of better learning.

Finally, under the weakest noise (15 observations (dark bars)) the distribution is shifted even more rightwards, with the average payoff 0.4. Here the best performer won 75 times. Hence, human subjects can cope with finding the best reply against a sophisticated opponent, provided its strategy is stationary. Further, results from our four treatments are on average remarkably consistent: the lower the noise, the better human subjects are able to beat the strategy of the Robot. Let's now turn to the other question: can subjects rationalize the reasons why they could beat the Robot, could they decode its strategy? To proceed, we use verbal descriptions of the Robot strategies solicited from the participants at the very end of the experiment.

Several verbal *descriptions* were reported in the questionnaire after the experiment[10]. These descriptions were not always complete but some examples (numbered in an alphabetic-numerical way) were suitable for interpretation and associated with a particular component of the player's strategy. Overall we found six different descriptions of strategies – Lose, Tie, Win, Last, Noise, Cycling which was mentioned by participants. In a nutshell, Lose, Tie, Win denote parts of the OptAR strategy, Last – a recognition that the Robot reacts to the last round Noise – that the Robot is somewhat irregular and randomizes, Cycling – that there are cycles in Robot's behavior. Hence, we wrote down several reports below and our interpretation to illustrate our approach of textual decomposition.

Player $v2$ (22 years old, female): «After Tie, Robot typically repeated its move. After losing to me, it most often took a move that beats my current move»[11]. Our interpretation: Player $v2$ is mistaken about the situation Tie, but correctly reveals the useful sub-strategy $Win \rightarrow Down$. Here, as everywhere below, we assume that the player says that she has in mind to play

[10] The full information of reports is available in the Appendix.

[11] In Russian: «После ничьих алгоритм чаще всего повторял свой вариант. После проигрыша алгоритм чаще всего выбирал вариант, позволяющий "побить" мой текущий вариант».

$$Tie: \left(a_{-i}^{t} = a_{R}^{t-1}\right) \Rightarrow \left(a_{i}^{t+1} = \left(a_{i}^{t} + 1\right) \bmod 3\right),$$
$$Win: \left(a_{-i}^{t} = \left(a_{R}^{t-1} + 1\right), \bmod 3\right) \Rightarrow \left(a_{i}^{t+1} = \left(a_{i}^{t} - 1\right) \bmod 3\right),$$

keeping silence about situation lose. Finally player $v2$, used term: typically' which implies non typical reaction leading Noise, thus Win, Tie, Noise parts of Robot were recognized.

Player *h*1 (16 years old, female): «Yes, there were regularities. Say, Robot has shown Rock. It expects that I will show Paper, and chooses Scissors, that is why now I should show Rock. Sometimes, this algorithm was switching to something else, and I tried to figure out how Robot behave in previous similar situations»[12]. Our interpretation: player $h1$ supposes Robot to play the best response to some Naive strategy and therefore suggests strategy $\left(a_{h1}^{t}, a_{R}^{t}\right) \to \left(a_{h1}^{t+1} = a_{R}^{t}\right)$. Though not complete, this suggestion is useful in Win situations, because it follows the best-response sub-strategy $Win \to Up$. So, this wrong hypothesis still allows $h1$ to win on average more than random. Hence only Win, part of the Robot strategy was recognized.

Player *p*2 (22 years old, male): «Robot was playing against my previous move but sometimes altered this strategy»[13]. Player $p2$ suggests from Robot strategy $\left(a_{i}^{t}, a_{R}^{t}\right) \Rightarrow \left(a_{R}^{t+1} = a_{i}^{t} + 1\right)$. Therefore, the logical $p2^{\text{th}}$ strategy should be $\left(a_{i}^{t}, a_{R}^{t}\right) \Rightarrow \left(a_{i}^{t+1} = a_{i}^{t} - 1\right)$, which turns out to be actual best-response in situation Win as well. Also Player $p2$ mentioned that the Robot alternated between some moves. Thus we interpreted the latter strategy as Win, Noise. Again, this incomplete hypothesis allows winning on average more than 0.

Finally, we present two hypotheses which could be somehow related to successful or unsuccessful strategies:

Player *l*1 (20 years old, male): «Robot was reacting to the last-move situation» Here, the player recognized the condition on the previous move but forgot or deliberately missed a detailed description. Consequently just Last are marked by us in this case.

Player *o*1 (17 years old, female): «Robot was cycling, but sometimes it repeated its previous move».

Here we must note that the latter mention of «cycling» may be actually equivalent to the player's best-response sub-strategy $Win \to Down$. Indeed, suppose that in situation Win this player somehow did use the best-response Down, and Robot responded with repetitions of (Up). Then the outcomes will be Win, Win, Win... , and the player may exploit this successful Up, Up, Up... sequence again and again, wrongly supposing that Robot simply intends to cycle Down, Down, Down.... This lucky mistake repeats until the Robot plays randomly and breaks the cycle. Afterward, whenever situation Win occurs again, the cycle can be exploited again. Hence the player reports about Cycling pattern.

---

[12] In Russian: «Да, некоторые действия подчинялись правилу. Например, Робот в последнем раунде показал камень. Он ожидает, что я покажу бумагу, показывает ножницы, значит, нужно показать камень. Иногда алгоритм нарушался, и можно было смотреть, что делал он в схожих ситуациях в прошлых ходах».

[13] In Russian: "Робот играл против моего предыдущего хода, но иногда отклонялся от стратегии».

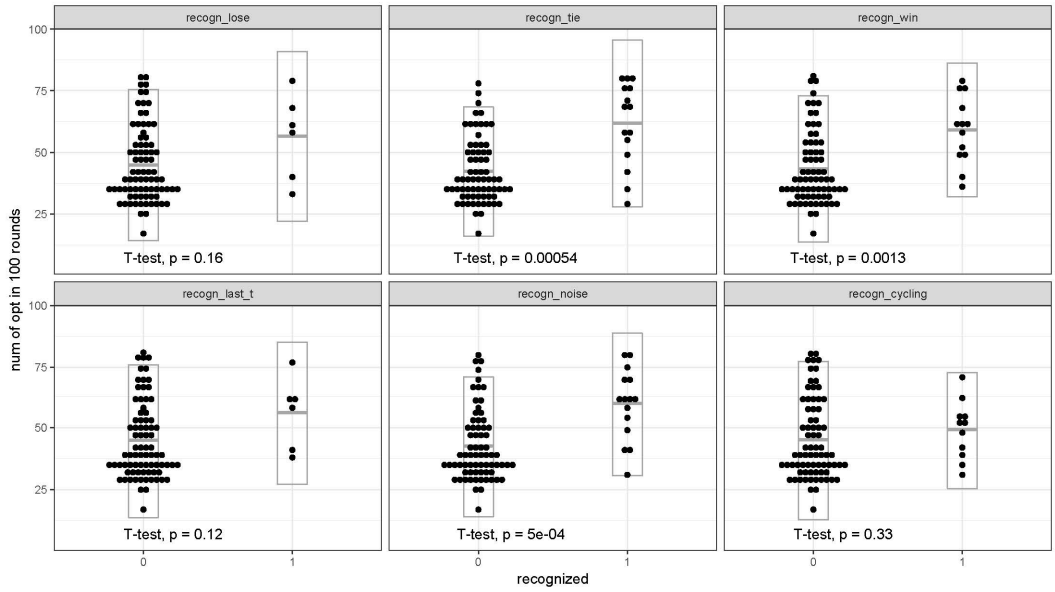


Fig. 2. The frequency of OptAR strategy among reporting and non-reporting players

Many subjects have recognized more than one of these patterns, so their separate contribution to performance is unfortunately not always possible. Figure 2 presents the performance of players who reported the six most popular patterns mentioned above against the rest of the strategies, aggregated over all four noise treatments. Strategies Lose, Tie, Win represent parts of the OptAR strategy, Last refers to recognition that the Robot plays conditional reaction to the last round, Noise – that the Robot is somewhat irregular and randomizes, Cycling – that there are cycles in Robot's behavior. Specifically Lose was mentioned by 6 players, Tie by 15, Win:12, Last:7, Noise:16, Cycling:11; these six relate to 6 pictures.

In Figure 2, each of the six plots compares the performance of those who have mentioned features of the Robot in their post-game survey with the rest of the sample. The vertical axis *Y* shows how often this player did use OptAR strategy. Of course, the payoffs will follow the OptAR frequency with some noise; that is why this frequency is a more accurate measure of successful recognition of the Robot strategy than payoffs.

For instance, the top player demonstrates as high as 77% usage of OptAR strategy, reflected similarly in all 6 pictures. This implies a very early understanding of all 3 parts of OptAR strategy, though it is not reflected in the survey. By contrast, the third-best player reports understanding these 3 best-response sub-strategies (Lose, Tie, Win) and reports Noise, but keeps silent about understanding Robot's Last-move reaction or Robot's Cycling. That is why this third-best player appears in the right column of Fig. 2 four times: in lose, Tie, Win, Noise pictures, and appears on the left in the remaining two figures (Last and Cycle). Only those five who have reached 0.8 also appear in each of the 6 pictures at this level.

Finally, the boxes show the mean (in the middle) and standard deviations of the particular groups. The statistically significant (t-test on the figure[14]) recognition of Win, Tie, Noise is not

[13] Results the same for Mann – Whitney test.

surprising. When the subject has recognized them, Lose situation is hard to encounter, so fewer subjects learned Lose strategy, therefore it is statistically insignificant. Finally, the Noise is seldom recognized alone, but usually along with the other rules.

3.2. Optimal strategy learning and its survey descriptions

Observation 2. Optimal play correlates with the recognition of opponent's strategy.

We observe that those reporting best-response sub-strategies $Tie \rightarrow Up$ (15 players) and especially $Win \rightarrow Down$ (12 players), played OptAR strategy *significantly more often* than those who didn't report it. Sub-strategy $Lose \rightarrow Stay$ was equally successful, but it was recognized only by 6 players, probably because it contradicts the Naive intuitions, forbidding to «repeat a losing action» or because successful learning of $Win \rightarrow Down$ cycle makes learning a plan for after loss unnecessarily. Players' hypotheses Last, Noise, Cycle about the Robot's behavior do not directly correspond to OptAR strategy, but still reflect some features of it.

To focus our attention on best learners, Fig. 3 modifies Fig. 2 by sorting reporting learners according to the completeness of their reports in two ways. The left panel shows only those who report that they learned 0, 1, 2, or 3 out of three OptAR sub-strategies (0, 0.33, 0.66 or 1 of maximal understanding). The right one demonstrates those who report 0, 1, 2, 3, or 4 out of 6 patterns mentioned (though nobody reported more than 4). The vertical axis again shows the percentage frequency of OptAR, which coincides with the number of related moves over the 100 rounds.

Both panels of Fig. 3 show a ***positive correlation*** (spearman coefficient 0.41 and 0.46 for the left and right panels, respectively; p<0.001 in both cases). Thus the better the players understand the strategy of the Robot, the more often their own play is optimal, and the higher is their expected payoff. This is especially well visible in the right panel, which shows that most of the points lie on an upward-rising regression line whose slope shows that articulation of one more characteristic component of the optimal strategy increases the number of wins by about 10 percentage points. Another sizeable cluster of participants makes a substantive proportion of wins without any guesses. There may be multiple reasons for that: lack of incentives to report the truth, subconscious actions, or simply higher perceived noise in the data.

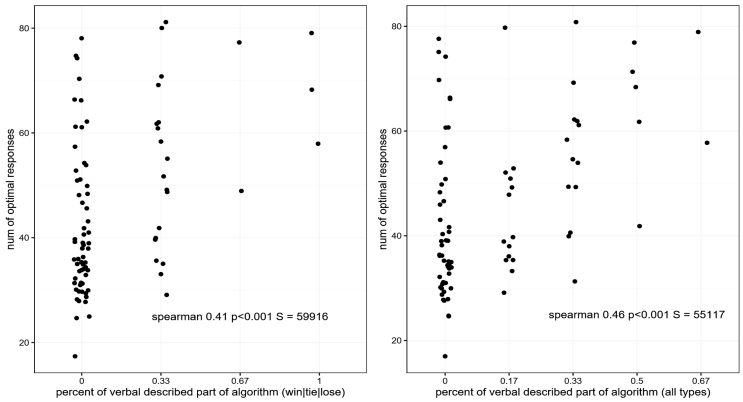


Fig. 3. Scatter plot between the number of optimal play and number of reported parts of robot

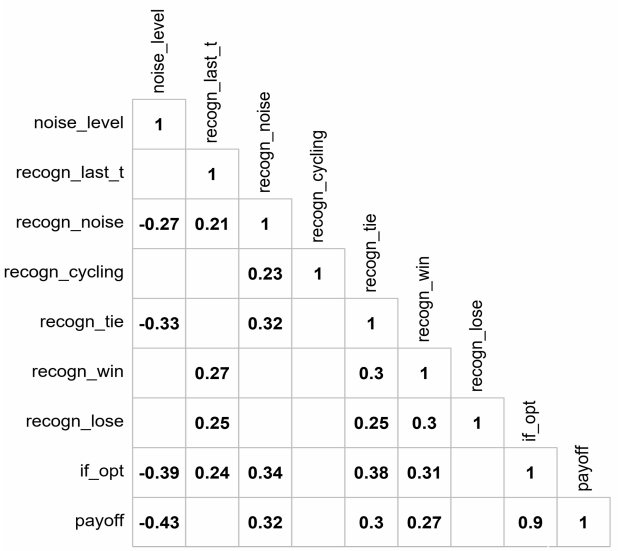


Fig. 4. Correlations between performance, noise and reported strategies

The correlogram (Fig. 4) shows the Spearman rank correlation coefficients significant at $p < 0.05$ or more (insignificant correlations are dropped). As one would expect, payoffs highly correlate with the use of the optimal strategy. Naturally, higher noise level makes learning harder – it negatively correlates with learning OptAR, its parts and payoffs. Reported learning of any OptAR sub-strategy highly correlates with two other sub-strategies, and usually also correlates with two other ideas – Noise and Last, but not with Cycling.

Table 4.

Linear regressions on sum of playing OptAR strategy

	Dependent variable					
	sum of playing OptAR strategy					
	(1)	(2)	(3)	(4)	(5)	(6)
Noise level	−18.944**	−19.205**	−21.158***	−21.216***	−26.569***	−22.027***
	(7.495)	(7.372)	(7.335)	(7.297)	(7.421)	(7.350)
recognition of last *t*	4.621					
	(5.794)					
recognition of noise	6.316	6.711				
	(4.468)	(4.249)				
recognition of cycling	−1.291					
	(4.872)					
recognition of tie	11.228**	10.790**	12.528***	12.841***	14.718***	
	(4.506)	(4.430)	(4.332)	(4.251)	(4.127)	

Continues

	Dependent variable					
	sum of playing OptAR strategy					
	(1)	(2)	(3)	(4)	(5)	(6)
recognition of win	4.827	5.773	6.828	7.342		11.133**
	(4.935)	(4.741)	(4.737)	(4.568)		(4.605)
recognition of lose	2.709	3.880	2.687			
	(6.285)	(6.092)	(6.102)			
Constant	50.116***	50.262***	51.893***	51.990***	55.983***	52.990***
	(4.092)	(3.939)	(3.836)	(3.811)	(3.747)	(3.796)
AIC	693.2	690.1	690.7	688.9	689.6	696
BIC	715.2	707.2	705.4	701.1	699.4	705.8
Observations	85	85	85	85	85	85
R2	0.319	0.312	0.291	0.289	0.266	0.209
Adjusted R2	0.257	0.269	0.255	0.263	0.248	0.190
Residual Std. Error	13.496	13.391	13.515	13.448	13.577	14.098
F Statistic	5.156***	7.177***	8.195***	10.971***	14.877***	10.823***

Note: * p < 0.1; ** p < 0.05; *** p < 0.01.

While we have a lot of observations of rounds, when we try to analyze individuals, we have only 85 observations, which precludes detailed econometric analysis. However, some important correspondences can be traced down. Table 4 provides estimates of several regressions of various recognition factors (such as reports on parts of the Robot and noise level) on the frequency of the optimal play. Noise level and recognition tie-stay (recognition of tie) part of the Robot are statistically significant across all specifications. Other recognized part of the Robot strategy, while also important from an analytic perspective, apparently are highly correlated with the noise and σ_V (Tie-Stay part of OptAR), therefore in the linear model, those influences are not detected. This ultimately confirms the next result.

Result 2: Noise level negatively influences performance: The lower the noise the higher is the number of optimal plays, rejecting Hypothesis 2.

3.3. Dynamics of learning

Our final two Observations (2 and 3), describe the observed *dynamics* of learning.

Observation 3. There is significant heterogeneity in performance both within and between the subjects and across treatments.

Fig. 5 compares three treatments: 0.2, 0.4, 0.6 noise levels. Again, the horizontal thick black line indicates the 33 correct moves out of 100 (which is 1/3), expected from Random player, and the dotted line shows the 0.05 confidence interval.

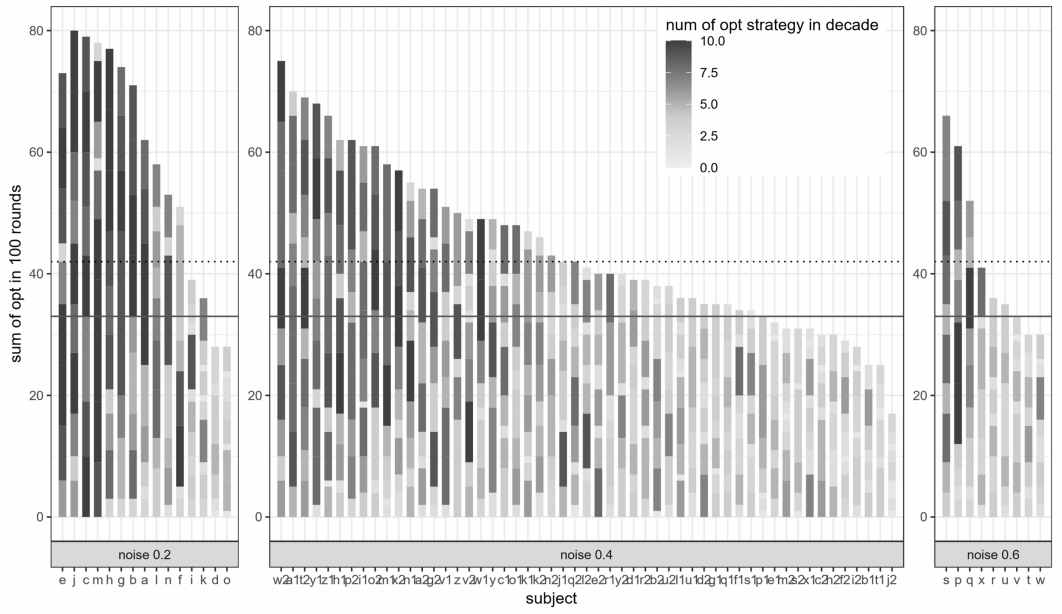


Fig. 5. Frequencies of optimal actions for each player

The players within each treatment are put in decreasing order by OptAR strategies played during 100 rounds (Y axis). Under low noise 0.2, as much as 13 out of 15 players are better than random (score 33 of optimal moves), and, moreover, 11 of them exceed this barrier on a statistically significant level!

Under higher noise 0.4, about 2/3 of players perform better than random, and about 1/3 of this cohort – significantly so. Even under high noise 0.6, 2/3 (6 out of 9) players show playing better than Random, but only 2 cases can be called statistically significant «learning.»

Each column represents the playing trajectory of a certain player, divided into 10 unequal segments. Each segment's height represents the number of OptAR strategies played during 10 rounds. They are ordered from the earlier ones at the bottom to the latest at the top. Color intensity also reflects the frequency of playing OPtAR, ranging from dark to light colors. The darkest and longest segments present the most frequent use of OptAR strategy during the next 10 rounds.

One could expect that typically the columns would be lighter (shorter) at the beginning of the game and darker in the end. Alas, such a tendency is more or less pronounced only for a subset of agents. Probably (as often reported in learning literature), the agents «fluctuate between the *exploitation* of the ideas grasped already and *exploration* of additional ideas possible. This trade-off may cause the fastest learners to often deviate from their optimal strategy and thereby perform similarly to slower learners.

However, on average the whole population becomes «darker» closer to the end of the game. Again it is evident that the lower the noise – the greater the learning on average, but we can clearly see that the best players in high-noise groups outperform the worst learners in low-noise groups.

Observation 4. Best 12 learners show rather monotone learning trajectories, their number of gains learning becoming systematically above the Nash equilibrium proportion around 60th round.

Fig.6 shows the dynamics of play for 12 selected learners among all 3 treatments who described «Win-Down» part of the Robot correctly. The *X* axis denotes rounds, and the Y axis the normalized (subtracting the expected gain of 0.33) frequency of playing OptAR strategy. Vertical lines with triangles at the top end show the average result this cohort, while squares on the same lines show the point when frequency of cumulative OptAR strategies that would imply significant outperformance of the random strategy according to z-test with $p < 0.05$. When the triangle-ended «overtakes» the squared one, it can be taken as «learning». We observe that this significant «learning moment» for this group occurs roughly after the *50th round.*

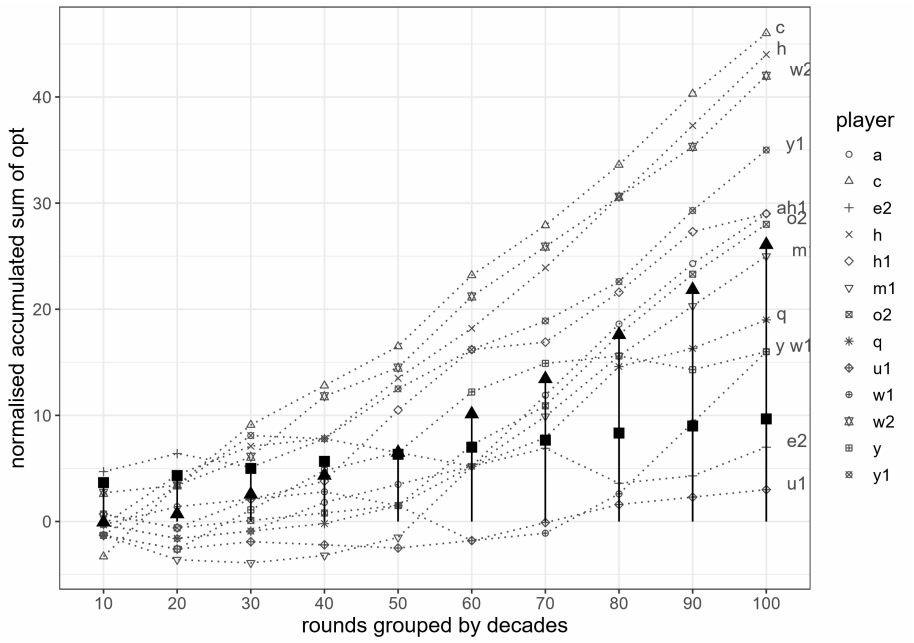


Fig. 6. Individual dynamics of learning for several players who described «win–down» part of the Robot correctly

We see that two players, c and h, seem to start determining exploration very early, about the 10th or the 20th round, which explain their outperformance over the random strategy by over 40 towards the end of the game. By contrast, a group of eight «rather good learners» have fluctuating payoffs, with 15 to 35 optimal plays by the end of the game. Finally, a group of 2 players just fluctuates near-zero (below squares at 100th round), ending up with an insignificant result. There are two possible explanations for their level of performance: either they have discovered this pattern too late for it to have an effect on their success or they were too interested in exploration at the expense of exploiting the pattern that they have discovered.

4. Comparing the models to human learning process

4.1. Simulations: Comparison of models' success in payoffs

We test our hypotheses 3 and 4 via comparing the performance of learning models against the Robot rule by simulations. In our simulations, we set each of the four algorithms mentioned in Section 2 playing as humans played in our treatments. Namely, one «simulation treatment» included 1000 attempts of an algorithm to play 100 rounds of RPS game against the Robot. Each algorithm played under 40% noise, at which level approximately half of the subjects have outperformed the random guesser (see Fig. 5)[15]. The summary of the results is in Appendix A1.

We compare the performance of the algorithms to that of humans using the proportions *z* test. The null hypothesis H_0 here is that the proportion of wins in 100 rounds should be approximately 0.33 (equal shares of Win, Tie, Lose), which corresponds to Nash equilibrium (random) play. The opposite two-sided hypothesis H_1 is that the proportion of wins is not equal 0.33. Results of related tests are presented in Table 5, which compares several algorithms' and human experimental results. Proportions *z* test of number of wins (win ratio).

Table 5.

Proportions z test of number of wins (win ratio), noise level = 0.4

Data source	Number of playing pairs	z-stat.	p-value	Test type	Mean
Lab. subjects	53	7.854	0.0	larger	0.417
SRL	1000	39.385	0.0	larger	0.391
SWFP	1000	61.298	0.0	larger	0.426
WFP	1000	–0.201	0.84	two-sided	0.33
RL	1000	–13.855	0.0	smaller	0.31

Table 5 sums up the most illustrative *z*-test results (see the complete table in Appendix A1). For human subjects, the H_0 hypothesis is rejected, they outperform the random player, as well as SRL and SWFP algorithms.

On the contrary, for action-based algorithms hypothesis H_0 (about their equivalence to a random player) is not rejected. WFP algorithm plays as good as a random player. Moreover, RL plays slightly (statistically insignificantly) worse than random, probably, because it is adaptive, like the Naive player whom our Robot is programmed to defeat. Noticing the advantage of the strategic algorithms over the random player and over the action-based ones, we make the following conclusions about our hypotheses 3 and 4.

Result 3. Standard action-based models WFP and RL do not outperform the random predictor when playing against Robot. Hypothesis 3 is rejected.

Result 4. Action-based models do not outperform their strategic analogs when playing against the Robot. Hypothesis 4 is rejected.

[15] Simulations with noise levels of 60% and 20% are contained in the Appendix.

4.2. Comparison of models' predictive performance

Now consider Fig. 7 which illustrates the difference among the learning algorithms in dynamics. It presents the cumulative sums of optimal strategy play over every 20 rounds among 100 rounds. The X axis shows intervals 0–20, 21–40, 41–60, 61–80, 81–100, including five treatments in each: action-based WFP, action-based RL, humans, SRL and SWFP. For each treatment/interval the Y axis shows the fraction (%) of Optimal moves short of the fraction explained by chance (33%), presenting the average (middle of each box), Q1–Q3 interquartile range (box), and the whole range of interval of observations (whiskers). For instance, we see that in the last period SWFP on average outperforms the random predictor for as much as 23% of moves, whereas RL performs worse than random. SRL dynamic behavior appears to be the closest to human behavior here.

Generally, as it can be seen from z-test and Fig. 7, the share of optimal (probably winning) moves among our action-based algorithms WFP, RL is close to random share or lower. On the contrary, we see that both strategic algorithms and human subjects do learn. They recognize the opponent's behavior and therefore defeat the Robot. Naturally, the shares of optimal moves differ among algorithms in the same direction as the shares of wins differ (all actual tests' results is also provided in Appendix A1).

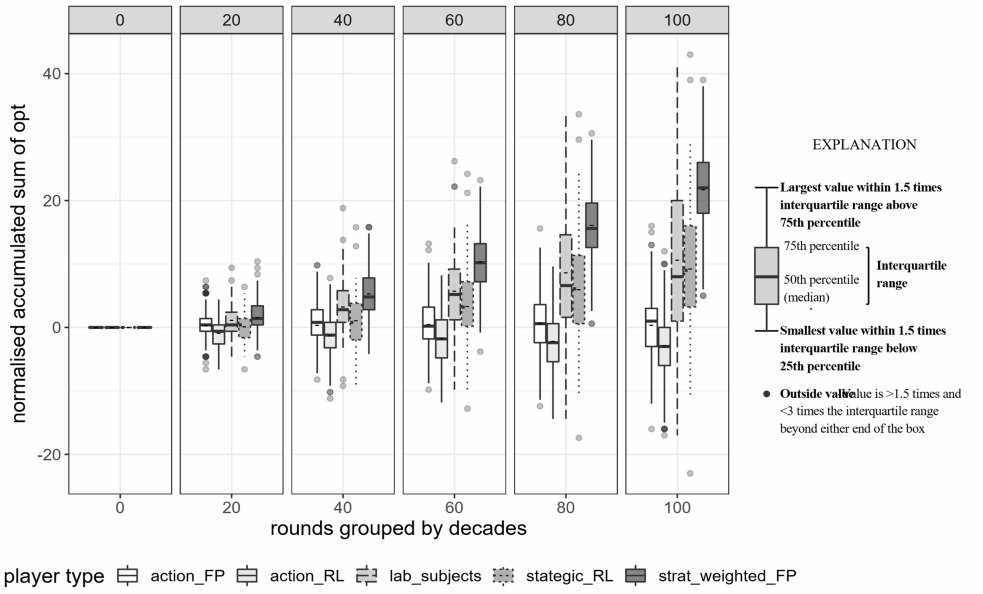


Fig. 7. Learning dynamics, noise 0.4
(subject sample includes University student + High school students)

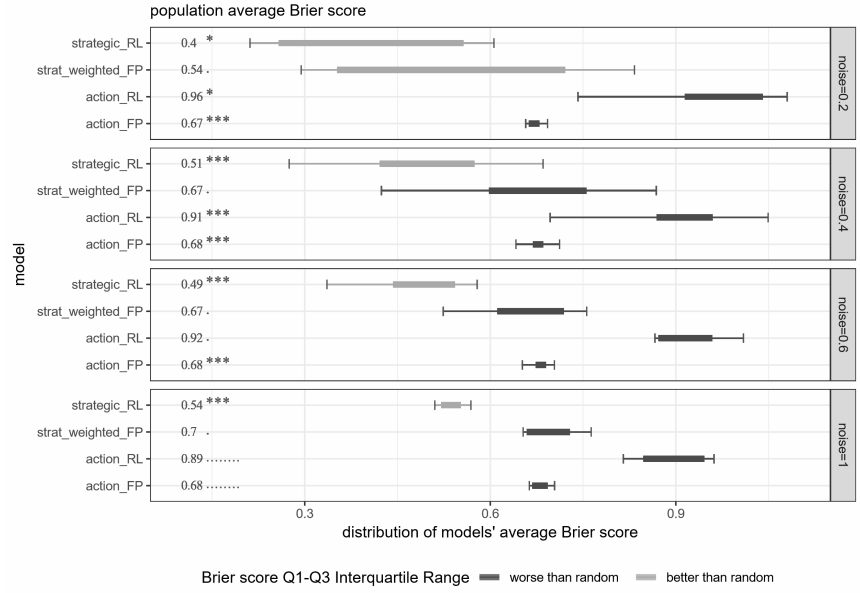
Further, to test the hypothesis 5, we analyze the predictive power of each algorithm using Brier/quadratic. These scoring rules are widely used in computer science [Gneiting, Raftery, 2007], forecasting [Satopaa et al., 2014] and experimental economics [Mathevet, Romero, 2012].

Our results appear similar for all three rules. However, since Brier takes into account the probabilities of all actions (chosen actions and also potential actions), it is more suitable for learning models. The Brier rules is a quadratic deviation, squared difference between the probability of an action and its realized outcome:

$$Brier - score = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{J}(\,Probability_{ti} - Outcome_{ti}\,)^2. \quad (9)$$

In our case, N is the number of rounds, J is the number of actions[16].

To build the Brier-score, the predictive ability of each model in the role of «Probability» was compared to human experimental data in the role of «Outcome». To perform this, complete history of each couple Human-Robot is taken from the experimental data. Further, using this personal history, we set the alternative player (one of the models) to formulate its forecast *for each move* in 100 rounds series. The algorithmic player perceives the data of the first player (human subject) as its own historical moves, and the history of moves of the opponent (Robot) as supplementary data. The Brier-score is counted for the history of each quadruple (Algorithm-Robot, Human-Robot) over the entire series of their 100 rounds. This procedure was carried out crosswise for each type of alternative algorithm and each experimental pair Human-Robot. Results are presented in Fig. 8.



Kruskal-Wallis test for two group with closest mean Brier *** < 0.001; ** < 0.01; * < 0.05.

Fig. 8. Prediction quality (Brier-scores) of human learning by the algorithms

[16] In the best scenario the Brier value would be equal to zero, in the worst equal to 2, and the random forecast is caught relational on the number of actions. In the case 3×3 game, a random forecaster produces Brier equal to 0.66.

Consistent with the previous (dynamic) result and supporting our Hypothesis 5, SFP and SRL have lower Brier scores, i.e., higher accuracy in predicting human moves than other algorithms under consideration. Figure 8 shows the resulting statistics of Brier-scores aggregated over each of our 85 human subjects, separated by noise treatments.

Nonparametric test (Kruskal – Wallis) confirmed the significance of the differences between predictions of the strategy-based and action-based models (see Fig. 8). Consistent with other performance criteria, we find that SRL demonstrates better results in predictive metrics for all groups with a lower noise level.

Result 5. SRL gets ahead of action-based and SFP ones by predictive metrics evaluated on experimental data. Hypothesis 6 is rejected.

Result 6. Belief-based strategic WFP predict experimental subjects worse than SRL. Hypothesis 3 is also rejected.

We can interpret these results and comparisons as follows. Action-based FP as a predictor performs equally badly at all noise levels, whereas the performance of action-based RL is somewhat lower, and that of SRL much lower at larger noise. We have seen that people verbally describe our Robot in terms of its «intentions» and play accordingly. This could suggest that FP algorithm better corresponds to their decision-making process than RL. However, our simulations show that SFP performs well only in the low-noise scenario. By contrast, SRL predicts humans well in all circumstances, in particular, it reflects some regularities in human behavior even when the Robot is totally random.

Our explanation of this contradictory evidence is that our human subjects had formulated their FP-like hypotheses (verbally pronounced at the end of the game) only *after* sufficiently long periods of RL-like behavior. Thus, SFP learns «too fast», predicting the speed of human behavior over the learning window worse than the SRL predictor does.

Conclusion

We have explored the noisy Rock-Paper-Scissors repeated game and compared «humans against Robot» settings with «learning algorithms against Robot» settings. We focused on two distinctions: action-based vs. strategy-based learning, or belief-based vs. reinforcement learning.

The general observations on humans confirm that: (a) many people are capable to defeat our simple Robot; (b) usually it happens in the span of 30–60 rounds, depending on the noise level; (c) people whose behavior shows learning the algorithm often can explain what they have learned, typically in belief-based terms, but it is not always the case; (d) among three parts of the optimal strategy, Tie $\rightarrow$ Up and Win $\rightarrow$ Down are more easily learned.

We have compared the standard action-based models, namely, WFP and RL with a more modern strategy-based approach [Hanaki, 2004; Ioannou, Romero, 2014]. We significantly extended this approach by focusing on partial, «elementary strategies» instead of complete automata. Our post-game surveys show that the actual process of learning is closer to our framework: people can learn one elementary strategy of the Robot but no others. Additionally, our free-form survey shows that people tend to describe their opponent in terms of some elementary strategies, not as a complete automaton even when they recognize all its parts. Further to that, our solution circumvents the curse of dimensionality for the automata. We do not need long series and large comprehensive sets of possible automata, we can «construct» them from basic elementary strategies as the game goes on.

Our sophisticated «strategy-based» versions learn to defeat the Robot; WSFP learn faster than SRL. However, it learns too fast: SRL better approximates human learning dynamics than WSFP. Moreover, SRL is the only model explored that predicts human behavior better than the random predictor in all our treatments, including a completely random Robot.

Both our descriptive picture of human learning and our attempts to model such learning suggest (though do not prove) that a strategic approach can better describe humans in some games, namely, those similar to RPS game in some respects. Can we directly transfer our empirical and modeling findings to other games? May be no. Rather, we present here a way to approach complex adaptive behavior in those other games. It is an open question whether a universal model to all learning situations can be constructed but if it is, our approach provides the necessary building blocks to it. This we expect from it help in investigating learning in a broader context. As to extensions, a broader study should find more precise stratification of learner's types and more clear approximation within an individual subject's play path.

Appendix.

A1. Z-test results

If p-value is less than a given significance level, then the null hypothesis is rejected in favor of the alternative. The null hypothesis here is that the single sample given by proportion of wins in 100 rounds was drawn from a distribution with the proportion equal random (0.33). Two-sided criterion postulates H_1 that proportion of wins not equal to 0.33 (larger > 0.33, smaller < 0.33).

Table 6.

Noise level	Data source	Variable	Z-stat	p-value	Test type	Mean
0.4.HSch	lab.subjects	ratio of opt play	3.8	0.0	two-sided	0.359
0.4	lab.subjects	ratio of opt play	3.742	0.0	two-sided	0.379
0.6	lab.subjects	ratio of opt play	1.127	0.26	two-sided	0.348
1	lab.subjects	ratio of opt play	0.097	0.92	two-sided	0.332
0.4	stategic.RL	ratio of opt play	64.725	0.0	two-sided	0.431
0.4	strat.weighted.FP	ratio of opt play	106.215	0.0	two-sided	0.498
0.4	WFP	ratio of opt play	1.551	0.12	two-sided	0.332
0.4	RL	ratio of opt play	–18.332	0.0	two-sided	0.303
0.4.HSch	lab.subjects	ratio of opt play	3.8	1.0	smaller	0.359
0.4	lab.subjects	ratio of opt play	3.742	1.0	smaller	0.379
0.6	lab.subjects	ratio of opt play	1.127	0.87	smaller	0.348
1	lab.subjects	ratio of opt play	0.097	0.54	smaller	0.332
0.4	stategic.RL	ratio of opt play	64.725	1.0	smaller	0.431

Continues

Noise level	Data source	Variable	Z-stat	p-value	Test type	Mean
0.4	strat.weighted.FP	ratio of opt play	106.215	1.0	smaller	0.498
0.4	WFP	ratio of opt play	1.551	0.94	smaller	0.332
0.4	Rl	ratio of opt play	–18.332	0.0	smaller	0.303
0.4.HSch	lab.subjects	ratio of opt play	3.8	0.0	larger	0.359
0.4	lab.subjects	ratio of opt play	3.742	0.0	larger	0.379
0.6	lab.subjects	ratio of opt play	1.127	0.13	larger	0.348
1	lab.subjects	ratio of opt play	0.097	0.46	larger	0.332
0.4	stategic.RL	ratio of opt play	64.725	0.0	larger	0.431
0.4	strat.weighted.FP	ratio of opt play	106.215	0.0	larger	0.498
0.4	WFP	ratio of opt play	1.551	0.06	larger	0.332
0.4	RL	ratio of opt play	–18.332	1.0	larger	0.303
0.4.HSch	lab.subjects	win ratio	8.84	0.0	two-sided	0.4
0.4	lab.subjects	win ratio	7.855	0.0	two-sided	0.435
0.6	lab.subjects	win ratio	2.574	0.01	two-sided	0.372
1	lab.subjects	win ratio	0.699	0.48	two-sided	0.342
0.4	stategic.RL	win ratio	39.386	0.0	two-sided	0.391
0.4	strat.weighted.FP	win ratio	61.299	0.0	two-sided	0.426
0.4	WFP	win ratio	–0.202	0.84	two-sided	0.33
0.4	RL	win ratio	–13.856	0.0	two-sided	0.31
0.4.HSch	lab.subjects	win ratio	8.84	1.0	smaller	0.4
0.4	lab.subjects	win ratio	7.855	1.0	smaller	0.435
0.6	lab.subjects	win ratio	2.574	0.99	smaller	0.372
1	lab.subjects	win ratio	0.699	0.76	smaller	0.342
0.4	stategic.RL	win ratio	39.386	1.0	smaller	0.391
0.4	strat.weighted.FP	win ratio	61.299	1.0	smaller	0.426
0.4	WFP	win ratio	–0.202	0.42	smaller	0.33
0.4	RL	win ratio	–13.856	0.0	smaller	0.31
0.4.HSch	lab.subjects	win ratio	8.84	0.0	larger	0.4
0.4	lab.subjects	win ratio	7.855	0.0	larger	0.435
0.6	lab.subjects	win ratio	2.574	0.01	larger	0.372
1	lab.subjects	win ratio	0.699	0.24	larger	0.342
0.4	stategic.RL	win ratio	39.386	0.0	larger	0.391

Continues

Noise level	Data source	Variable	Z-stat	p-value	Test type	Mean
0.4	strat.weighted.FP	win ratio	61.299	0.0	larger	0.426
0.4	WFP	win ratio	–0.202	0.58	larger	0.33
0.4	RL	win ratio	–13.856	1.0	larger	0.31
0.6	stategic.RL	ratio of opt play	43.931	0.0	two-sided	0.398
0.6	strat.weighted.FP	ratio of opt play	75.235	0.0	two-sided	0.448
0.6	WFP	ratio of opt play	2.762	0.01	two-sided	0.334
0.6	RL	ratio of opt play	–21.854	0.0	two-sided	0.298
0.6	stategic.RL	ratio of opt play	43.931	1.0	smaller	0.398
0.6	strat.weighted.FP	ratio of opt play	75.235	1.0	smaller	0.448
0.6	WFP	ratio of opt play	2.762	1.0	smaller	0.334
0.6	RL	ratio of opt play	–21.854	0.0	smaller	0.298
0.6	stategic.RL	ratio of opt play	43.931	0.0	larger	0.398
0.6	strat.weighted.FP	ratio of opt play	75.235	0.0	larger	0.448
0.6	WFP	ratio of opt play	2.762	0.0	larger	0.334
0.6	RL	ratio of opt play	–21.854	1.0	larger	0.298
0.6	stategic.RL	win ratio	19.357	0.0	two-sided	0.359
0.6	strat.weighted.FP	win ratio	30.56	0.0	two-sided	0.377
0.6	WFP	win ratio	–0.599	0.55	two-sided	0.329
0.6	RL	win ratio	–10.674	0.0	two-sided	0.314
0.6	stategic.RL	win ratio	19.357	1.0	smaller	0.359
0.6	strat.weighted.FP	win ratio	30.56	1.0	smaller	0.377
0.6	WFP	win ratio	–0.599	0.27	smaller	0.329
0.6	RL	win ratio	–10.674	0.0	smaller	0.314
0.6	stategic.RL	win ratio	19.357	0.0	larger	0.359
0.6	strat.weighted.FP	win ratio	30.56	0.0	larger	0.377
0.6	WFP	win ratio	–0.599	0.73	larger	0.329
0.6	RL	win ratio	–10.674	1.0	larger	0.314
0.2	lab_subjects	ratio of opt play	21.249	0.0	two-sided	0.601
0.2	lab_subjects	ratio of opt play	21.249	1.0	smaller	0.601
0.2	lab_subjects	ratio of opt play	21.249	0.0	larger	0.601
0.2	lab_subjects	win ratio	16.328	0.0	two-sided	0.542
0.2	lab_subjects	win ratio	16.328	1.0	smaller	0.542

Continues

Noise level	Data source	Variable	Z-stat	p-value	Test type	Mean
0.2	lab_subjects	win ratio	16.328	0.0	larger	0.542
0.2	stategic.RL	ratio of opt play	31.644	0.0	two-sided	0.441
0.2	strat.weighted.FP	ratio of opt play	86.445	0.0	two-sided	0.626
0.2	WFP	ratio of opt play	0.8103	0.42	two-sided	0.333
0.2	RL	ratio of opt play	–8.293	0.0	two-sided	0.303
0.2	stategic.RL	ratio of opt play	31.644	1.0	smaller	0.441
0.2	strat.weighted.FP	ratio of opt play	86.445	1.0	smaller	0.626
0.2	WFP	ratio of opt play	0.810	0.79	smaller	0.333
0.2	RL	ratio of opt play	–8.293	0.0	smaller	0.303
0.2	stategic.RL	ratio of opt play	31.644	0.0	larger	0.441
0.2	strat.weighted.FP	ratio of opt play	86.445	0.0	larger	0.626
0.2	WFP	ratio of opt play	0.810	0.21	larger	0.333
0.2	RL	ratio of opt play	–8.293	1.0	larger	0.303
0.2	stategic.RL	win ratio	25.635	0.0	two-sided	0.419
0.2	strat.weighted.FP	win ratio	65.783	0.0	two-sided	0.561
0.2	WFP	win ratio	0.390	0.7	two-sided	0.331
0.2	RL	win ratio	–7.396	0.0	two-sided	0.306
0.2	stategic.RL	win ratio	25.635	1.0	smaller	0.419
0.2	strat.weighted.FP	win ratio	65.783	1.0	smaller	0.561
0.2	WFP	win ratio	0.390	0.65	smaller	0.331
0.2	RL	win ratio	–7.396	0.0	smaller	0.306
0.2	stategic.RL	win ratio	25.635	0.0	larger	0.419
0.2	strat.weighted.FP	win ratio	65.783	0.0	larger	0.561
0.2	WFP	win ratio	0.390	0.35	larger	0.331
0.2	RL	win ratio	–7.396	1.0	larger	0.306

A2. Instructions

General. Welcome to the experimental session. You have to make some decisions, and You will get the opportunity to earn money. How much you earn will depend on your decisions, and on the decisions of your opponent. Therefore, it is very important that you carefully read these instructions. The money that is prescribed to you by the result of the experiment will be paid to you in cash at the end of the experimental session. Your decisions, as well as your results, are anonymous. We guarantee the confidentiality of your decisions and answers, and we will ana-

lyze them only in depersonalized. These instructions are for your personal use only. Throughout the experimental session, you are not allowed to communicate with other participants. Violation of this rule may lead to removing from the experiment and losing all the experiment's money. If you have any questions, please raise your hand. We will approach your workplace and answer your questions individually.

During the experiment, we will not use rubles but will use tokens (conventional monetary units of the experiment). At the end of the experimental session, this result in tokens will be converted into rubles at the rate 3 rubles to 1 token, besides, you will receive a participation fee of 150 rubles. At the end of the session, each participant will receive their money individually. This experimental session consists of 100 rounds. You have to make 1 decision in each round, in the game Rock-Paper-Scissors. For each move, you will be given 45 seconds.

Rules. In each round, you need to choose one of three possible actions: Rock, Scissors or Paper. Your opponent does the same regardless of you. The winner in each round is determined by the following rule: paper conquers rock; rock defeats scissors; scissors beat rock. If both players have chosen the same action, the outcome of the round is a tie.

The opponent. A computer program will play with you as an opponent. The actions of the program are managed by the algorithm. There are several things which are known about the algorithm: At the time of making a decision in each round, it does not know your action in the current round. Information about the actions of both participants (both yours and hiss) in previous rounds are available to its. The program algorithm does not change throughout the game. The algorithm could be determined by a certain rule (regularity), but could be not. Knowing the rule will allow you to better predict the actions of the opponent (computer program). You know nothing more about its algorithm.

Compensation. For each win in each round, 2 tokens are awarded to you, for a tie – 1 token, for losing to you – 0 tokens. The number of tokens is charged by the result is shown in table 7.

Table 7.

Payor matrix

Robot\human	Rock	Paper	Scissors
Rock	1	2	0
Paper	0	1	2
Scissors	2	0	1

In addition, a bonus of 15 tokens is awarded for every five wins starting from 30 (see table 8). For example, the bonus for 4 wins will be 0 tokens, for 30 wins – 15 tokens, for 41 wins 30 tokens. You can get maximum bonus points for a series of 100 wins, then the bonus will be 225 tokens. That is, your total gain in tokens will be calculated by the formula:

$$Win\_in\_tokens = (Number\_of\_Wins) \cdot 2 + (Number\_of\_ties) \cdot 1 + (Bonus) \cdot 15.$$

Where the bonus is an incomplete quotient of dividing the Number of Wins, starting from 30 with step 5.

Examples. With the result of the game 33 wins 33 ties 33 losses the player, his winnings will be: (33) × 2 + (33) × 1 + ([33/5]) × 15 = 66 + 33 + 15 = 114 tokens, then the gain in rubles ex-

cluding payment for participation will amount to 342 rubles, together with payment for participation 492 rubles. When the result of the game is 50 wins 25 ties 25 losses the player, his winnings will be: (50) × 2 + (25) × 1 + ([50/5]) × 15 = 100 + 25 + 75 = 200 tokens, then the gain in rubles without payment for participation will be – 600 rubles, along with payment for participation 750 rubles. With the result of the game 100 wins 0 ties 0 losses the player, his winnings will be: (100) × 2 + (0) × 1 + ([100/5]) × 15 = 200 + 0 + 225 = 425 tokens, then the gain in rubles without payment for participation will amount to 1275 rubles, together with payment for participation 1425 rubles.

Table 8.

Additional incentive scheme

Wins	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	95–99	100
Bonus	15	30	45	60	75	90	105	120	135	150	165	180	195	210	225

A3. Reports about robot play by subjects

Table 9.

Verbalized strategies reported by participants

ID	REC-OGN_LAST_T	REC-OGN_NOISE	REC-OGN_CYCLING	REC-OGN_TIE	REC-OGN_WIN	REC-OGN_LOSE
l2	1	0	0	0	0	0
p2	1	1	0	0	1	0
p2	0	1	1	0	0	0
q2	0	1	1	1	0	0
t2	0	1	0	1	0	0
u2	1	0	0	0	0	0
v2	0	1	0	0	0	0
v2	0	1	0	1	1	0
y2	0	0	0	0	0	1
y	0	0	0	0	1	0
g1	0	0	0	1	0	0
h1	0	1	0	0	1	0
i1	1	0	0	0	0	1
m1	1	0	0	1	1	1
n1	0	0	1	1	0	0

Continues

ID	REC-OGN_LAST_T	REC-OGN_NOISE	REC-OGN_CYCLING	REC-OGN_TIE	REC-OGN_WIN	REC-OGN_LOSE
o1	0	0	1	0	0	0
p1	0	0	0	0	0	1
u1	0	0	0	0	1	0
v1	0	0	1	0	0	0
w1	0	0	0	1	1	0
x1	0	1	1	0	0	0
y1	0	0	0	1	1	1
e2	0	1	0	0	1	0
f2	0	0	0	1	0	0
g2	0	1	1	0	0	0
q	0	0	0	0	1	0
u	0	0	1	0	0	0
a	1	1	0	0	1	0
b	0	1	1	1	0	0
c	0	1	0	1	1	1
e	0	1	0	1	0	0
h	1	0	0	1	1	0
i	0	0	1	0	0	0
j	0	0	0	1	0	0
l	0	1	0	1	0	0
n	0	0	1	0	0	0

* *
*

References

Arifovic J., McKelvey R.D., Pevnitskaya S. (2006) An Initial Implementation of the Turing Tournament to Learning in Repeated Two-person Games. *Games and Economic Behavior*, 57, 1, pp. 93–122.

Aumann R.J. (1981) Survey of Repeated Games. *Essays in Game Theory and Mathematical Economics in Honor of Oskar Morgenstern,* Birkhauser Verlag GmbH.

Brandenburger A. (1996) Strategic and Structural Uncertainty in Games. *Wise Choices: Games, Decisions, and Negotiations* (eds. R.J. Zeckhauser, R.L. Keeney, J.K. Sebenius), Brighton: Harvard Business School Press, pp. 221–232.

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

Byrne D.P., De Roos N. (2019) Learning to Coordinate: A Study in Retail Gasoline. *The American Economic Review*, 109, 2, pp. 591–619.

Camerer C.F. (2018) Artificial Intelligence and Behavioral Economics. *The Economics of Artificial Intelligence: An Agenda.* University of Chicago Press.

Camerer C., Ho T.H. (1999) Experienced-Weighted Attraction Learning in Normal Form Games. *Econometrica*, 67, 4, pp. 827–874.

Chernov G., Susin I., Cheparuhin S. (2020) *Evaluation of Econometric Models of Adaptive Learning by Predictive Measures.* Available at: https://ssrn.com/abstract=3658087

Cheung Y.-W., Friedman D. (1997) Individual Learning in Normal Form Games: Some Laboratory Results. *Games and Economic Behavior*, 19, 1, pp. 46–76.

Chen D.L., Schonger M., Wickens C. (2016) oTree – An Open-source Platform for Laboratory, Online, and Field Experiments. *Journal of Behavioral and Experimental Finance*, 9, pp. 88–97, ISSN 2214-6350.

Doraszelski U., Lewis G., Pakes A. (2018) Just Starting Out: Learning and Equilibrium in a New Market. *The American Economic Review,* 108, 3, pp. 565–615.

Duersch P., Kolb A., Oechssler J., Schipper B.C. (2010) Rage Against the Machines: How Subjects Play Against Learning Algorithms. *Economic Theory*, 43, 3, pp. 407–430.

Duersch P., Oechssler J., Schipper B.C. (2012) Unbeatable Imitation. *Games and Economic Behavior*, 76, 1, pp. 88–96.

Fudenberg D., Levine D.K. (1998) *The Theory of Learning in Games*. The MIT Press.

Gneiting T., Raftery A.E. (2007) Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102, 477, pp. 359–378.

Hanaki N. (2004) Action Learning versus Strategy Learning. *Complexity*, 9, 5, pp. 41–50.

Hanaki N., Kirman A., Pezanis-Christou P. (2018) Observational and Reinforcement Pattern-Learning: An Exploratory Study. *European Economic Review*, 104, pp. 1–21.

Ioannou C.A., Romero J. (2014) A Generalized Approach to Belief Learning in Repeated Games. *Games and Economic Behavior*, 87, pp. 178–203.

Li S. (2017) Obviously Strategy-Proof Mechanisms. *The American Economic Review*, 107, 11, pp. 3257–3287.

Mathevet L., Romero J. (2014) *Predictive Repeated Game Theory: Measures and Experiments.* Mimeo.

Milgrom P., Roberts J. (1991) Adaptive and Sophisticated Learning in Normal Form Games. *Games and Economic Behavior,* 3, 1, pp. 82–100.

Nachbar J. (2009) Learning in Games. *Encyclopedia of Complexity and Systems Science*. New York: Springer, pp. 5177–5188.

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

Satopää V.A., Baron J., Foste, D.P., Mellers B.A., Tetlock P.E., Ungar L.H. (2014) Combining Multiple Probability Predictions Using a Simple Logit Model. *International Journal of Forecasting*, 30, 2, pp. 344–356.

Selten R. (1991) Properties of a Measure of Predictive Success. *Mathematical Social Sciences*, 21, 2, pp. 153–167.

Spiliopoulos L. (2012) Pattern Recognition and Subjective Belief Learning in a Repeated Constant-Sum Game. *Games and Economic Behavior*, 75, 2, pp. 921–935.

Sutton R.S., Barto A.G. (2017) *Reinforcement Learning: An Introduction (in progress).* London, England, Cambridge, MA: The MIT Press.

Wang Z., Xu B., Zhou H.J. (2014) Social Cycling and Conditional Responses in the Rock-Paper-Scissors Game. *Scientific Reports*, 4.

Zohreh Emami (2012) Social Economics and Evolutionary Learning. *Review of Social Economy,* Taylor & Francis Journals, 70, 4, pp. 401–420.