Transfer of Learned Opponent Models in Zero Sum Games

Ismail Guennouni[1] & Maarten Speekenbrink[1]

[1] Department of Experimental Psychology, University College London

Author Note

Correspondence concerning this article should be addressed to Ismail Guennouni, Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom. E-mail: i.guennouni.17@ucl.ac.uk

Abstract

Human learning transfer abilities take advantage of important cognitive building blocks such as an abstract representation of concepts underlying tasks and causal models of the environment. One way to build abstract representations of the environment when the task involves interactions with others is to build a model of the opponent that may inform what actions they are likely to take next. In this study, we propose to explore opponent modelling and its transfer with the use of computer agents possessing human-like theory of mind abilities with limited degrees of iterated reasoning. Through two differentiated experiments, we find that participants can deviate from Nash equilibrium play and learn to adapt to the opponent strategy and exploit it. Moreover, we showed that participants transfer their learning to new games and that the transfer is moderated by the level of sophistication of the opponent. Computational modelling shows that it is likely that players start each game using a model-based learning strategy that facilitates generalisation and opponent model transfer, but then switch to behaviour that is consistent with a model-free learning strategy in the latter stages of the interaction.

Transfer of Learned Opponent Models in Zero Sum Games

Introduction

Being able to transfer previously acquired knowledge to a new domain is one of the hallmarks of human intelligence. This ability relies on important cognitive building blocks, such as an abstract representation of concepts underlying tasks (Lake, Ullman, Tenenbaum, & Gershman, 2017). One way to form these representations when the task involves interactions with others, is to build a model of the person we are interacting with that offers predictions of the actions they are likely to take next. There is evidence that people learn such models of their opponents when playing repeated economic games (Stahl & Wilson, 1995).

In this paper, we are specifically interested in the way in which people build and use models of their opponent to facilitate learning transfer, when engaged in situations involving an interaction with strategic considerations. Repeated games, in which players interact repeatedly with the same opponent and have the ability to learn about the opponent's strategies and preferences (Mertens, 1990) are particularly adapted to this task. The early literature on learning transfer in repeated games has mostly focused on measuring the proportion of people who play normatively optimal (Nash Equilibria) or salient actions (e.g Risk Dominance) in later games, having had experience with a similar game environment previously (Camerer & Knez, 2000; Ho, Camerer, & Weigelt, 1998). This doesn't allow for the possibility of learning about the opponent's strategy and potentially exploiting it.

When studies have specifically explored this aspect, they have used computer opponents that were generally programmed not to change their strategies over the course of the task, allowing better experimental control. However, they have mostly looked at the ability of players to detect and exploit action-based learning rules (Shachat & Swarthout, 2004; Spiliopoulos, 2013). The strategies implemented by the computer opponents had a style of play that was not “human-like” in the sense that humans are not very good at playing specific mixed strategies with precision, or at detecting patterns from long sequences of past play. Thus, in this study, we aim to explore opponent modelling and its transfer with the use of computer agents endowed with human-like limited degrees of iterated reasoning. The agents are either a level-1 or level-2 player, mimicking “I know that you know that I know” type reasoning, and the limited recursion depth they exhibit (Goodie, Doshi, & Young, 2012). A level 1 player adapts their play to what they believe their opponent will play, without considering what their opponent might believe they will play. A level 2 player, on the other hand, takes their opponent’s belief about their actions into account, assuming they face a level 1 player, and choosing actions to beat the actions of that player. The choice of this type of strategy is also motivated by evidence that humans strategically use information from last round play of their opponents in zero sum games (Batzilis, Jaffe, Levitt, List, & Picel, 2016; Wang, Xu, & Zhou, 2014).

We have conducted two experiments where participants interact with a computer opponent. In each experiment, we measure transfer of learning about the opponent’s strategy between games with varying degrees of similarity. The first two games we use are structurally identical except for action labels. In the first experiment, the third game is strategically similar to the first two but descriptively different, while in a second experiment, we introduce a third game that is dissimilar to the first two in terms of payoff matrix and strategic structure. In the first experiment, participants face the same opponent throughout the three games, and the opponents are randomised to be either level-1 or level-2 players. In the second experiment, participants faced both level-1 and level-2 opponent sequentially, with the order in which they are faced randomised across participants.

Experiment 1

Methods

Participants and Design. A total of 52 (28 female, 24 male) participants were recruited on the Prolific Academic platform. The mean age of participants was 31.2 years. Participants were paid a fixed fee of £2.5 plus a bonus dependent on their performance which averaged £1.06. The experiment used a 2 (computer opponent: level 1 or level 2) by 3 (games: rock-paper-scissors, fire-water-grass, numbers) design, with repeated measures on the second factor. Participants were randomly assigned to one of the two levels of the first factor.

Tasks. In the first experiment, participants played the three games against their computer opponent. These games were rock-paper-scissors, fire-water-grass, and the numbers game. A typical rock-paper-scissors game (hereafter RPS) is a 3x3 zero sum game, with a cyclical hierarchy between the two player's actions: rock blunts scissors, paper wraps rock, and scissors cut paper. If one player chooses an action which dominates their opponent's action, the player wins (receives a reward of 1) and the other player loses (receives a reward of -1). Otherwise it is a draw and both players receive a reward of 0. RPS has a unique mixed-strategy Nash equilibrium, which consists of each player in each round randomly selecting from the three options with uniform probability. The Fire-Water-Grass (FWG) game is identical to RPS in all but action labels: Fire burns grass, water extinguishes fire, and grass absorbs water. We use this game as we are interested in whether learning is transferred in a fundamentally similar game where the only difference is in the description of the possible actions. This should make it relatively easy to generalize knowledge of the opponent's strategy, provided this knowledge is on a sufficiently abstract level, such as knowing the opponent is a level 1 or 2 player. Crucially, learning simple contingencies such as "If I played Rock on the previous round, playing Scissors next will likely result in a win", as might be learned by a simple reinforcement learning algorithm, will not be able to generalize to such a game, as these contingencies are tied to the labels of the actions. The numbers game is a generalization of RPS. In the variant we use, 2 participants concurrently pick a number between 1 and 5. To win in this game, a participant needs to pick a number exactly 1 higher than the number chosen by their opponent. For example, if a participant thinks their opponent will pick 3, they ought to choose 4 to win the round. To make the strategies cyclical as in RPS, the game stipulates that the lowest number (1) beats the highest number (5), so if the participant thinks the opponent will play 5, then the winning choice is to pick 1. This game has a structure similar to RPS in which every action is dominated by exactly one other action. All other possible combinations of choices are considered ties. Similar to RPS and FWG, the mixed-strategy Nash equilibrium is to play each action with equal probability in a random way.

The computer opponent was programmed to use either a level-1 or level-2 strategy in all the games. A level 1 player is defined as a player who best responds to a level 0 player. A level 0 player plays in a non-strategic way and does not consider their opponent's actions. Here, we assume a level 0 player simply repeats their previous action. There are other ways to define a level 0 player. For instance, as repeating their action if it resulted in a win, and choosing randomly from the remaining actions otherwise. As a best response to a random action is itself a random action, defining a level 0 player in such a way would make a level 1 opponent's strategy much harder to discern. Because we are mainly interested in generalization of knowledge of an opponent's strategy to other games, which rests on good knowledge of this strategy, we opted for this more deterministic formulation of a level 0 player (whilst also introducing some randomness in the computer opponent's play). A level-2 computer opponent, will assume in turn that the participant is a level-1 opponent, playing according to the strategy just described. We also introduced some noise over the actions of computer opponents making them play randomly in 10% of all trials.

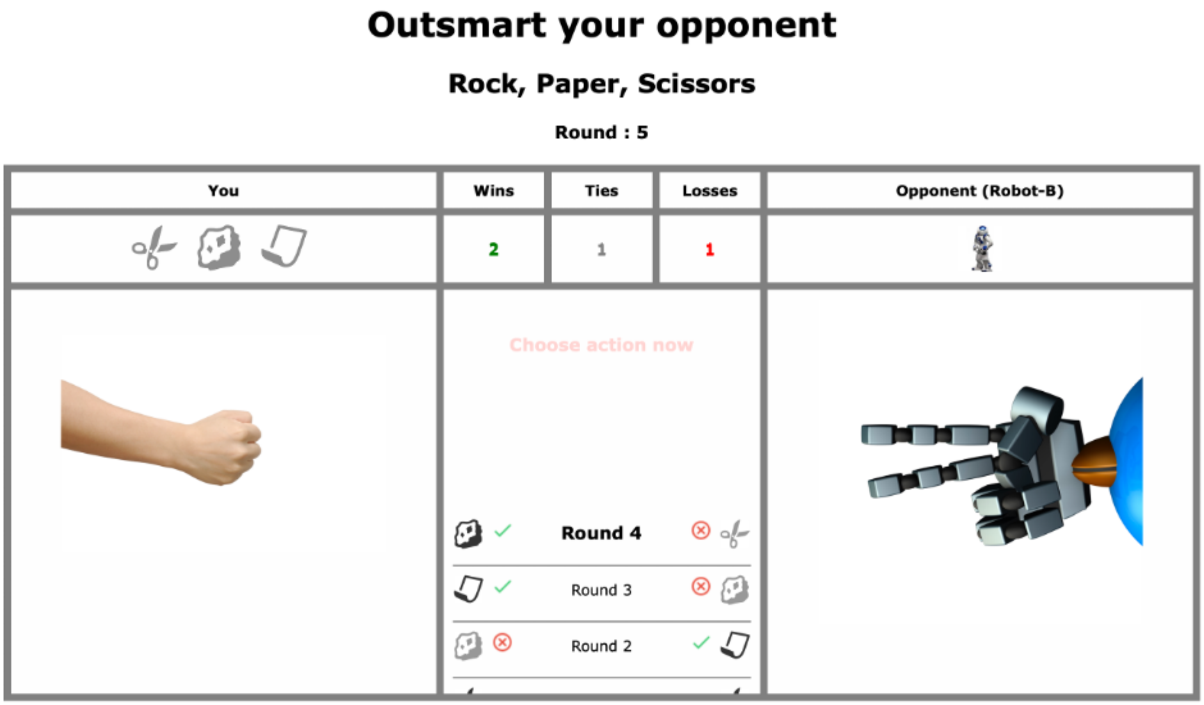
Human last	Agent last	level-1 Agent	level-2 Agent
Paper	Rock	Scissors	Scissors
Scissors	Scissors	Rock	Paper
Rock	Paper	Paper	Rock
...	...	...	...

Table 1

Example of how a level-1 and level-2 computer agent plays in response to actions taken in the previous round.

Procedure. Participants were informed they would play three different games against the same computer opponent. Participants were told that the opponent cannot cheat and will choose its actions simultaneously without knowledge of the participant's choice. After providing informed consent and reading the instructions, participants answered a number of comprehension questions. They then played the three games against their opponent in the order RPS, FGW, and NUMBERS. A total of 50 rounds of each game was played with the player's score displayed at the end of each game. The score was calculated as the number of wins minus the number of losses. Ties did not affect the score. In order to incentivise the participants to maximise the number of wins against the opponents, players were paid a bonus at the end of the experiment that was proportional to their final score. Each point is worth £0.02. An example of the interface for the RPS game is provided in Figure 1. After playing all the games, participants were asked questions about their beliefs about the computer opponent, related to whether they think they have learned their strategy and how hard they found playing against that particular opponent. They were then debriefed and thanked for their participation.

Results. On average, participants obtained the lowest score in the RPS game ($M = 0.289$, $SD = 0.348$), followed by NUMBERS ($M = 0.31$, $SD = 0.347$). Participants' performance was highest in the FWG game ($M = 0.454$, $SD = 0.354$). Scores in each game



Outsmart your opponent

Rock, Paper, Scissors

Round : 5

You	Wins	Ties	Losses	Opponent (Robot-B)
	2	1	1	

Choose action now

Figure 1. Screenshot of the feedback at the end of a round of Rock-Paper-Scissors

were significantly different from 0, the expected score of random play (RPS: $t(51) = 7.26$, $p < .001$; FWG: $t(51) = 10.04$, $p < .001$; NUMBERS: $t(51) = 7.17$, $p < .001$). To assess learning within and between games, we used a 2 (condition: level-1, level-2) by 3 (game: RPS, FWG, NUMBERS) by 2 (block: first half, second half) repeated-measures ANOVA, with the first factor varying between participants. This showed a main effect of Game ($F(2, 100) = 8.54$, $\eta^2 = 0.05$, $p < .001$), indicating that average scores varied significantly over the games. Post-hoc pairwise comparisons showed that performance in the FWG game was significantly higher than in the RPS game ($t(100) = 3.78$, $p < .001$) and the NUMBERS game ($t(100) = 3.32$, $p = .002$). The score in RPS was not significantly different from the score in NUMBERS ($t(100) = 0.45$, $p = .65$). The main effect of Block ($F(1, 50) = 22.51$, $\eta^2 = 0.03$, $p < .001$) shows that the score in the first half of each game ($M = 0.29$) was significantly lower than in the second half ($M = 0.40$), which indicates within-game learning. The main effect of Condition ($F(1, 50) = 5.44$, $\eta^2 = 0.05$, $p = .024$) indicates that scores were higher against the level-1 player ($M = 0.43$) than against the

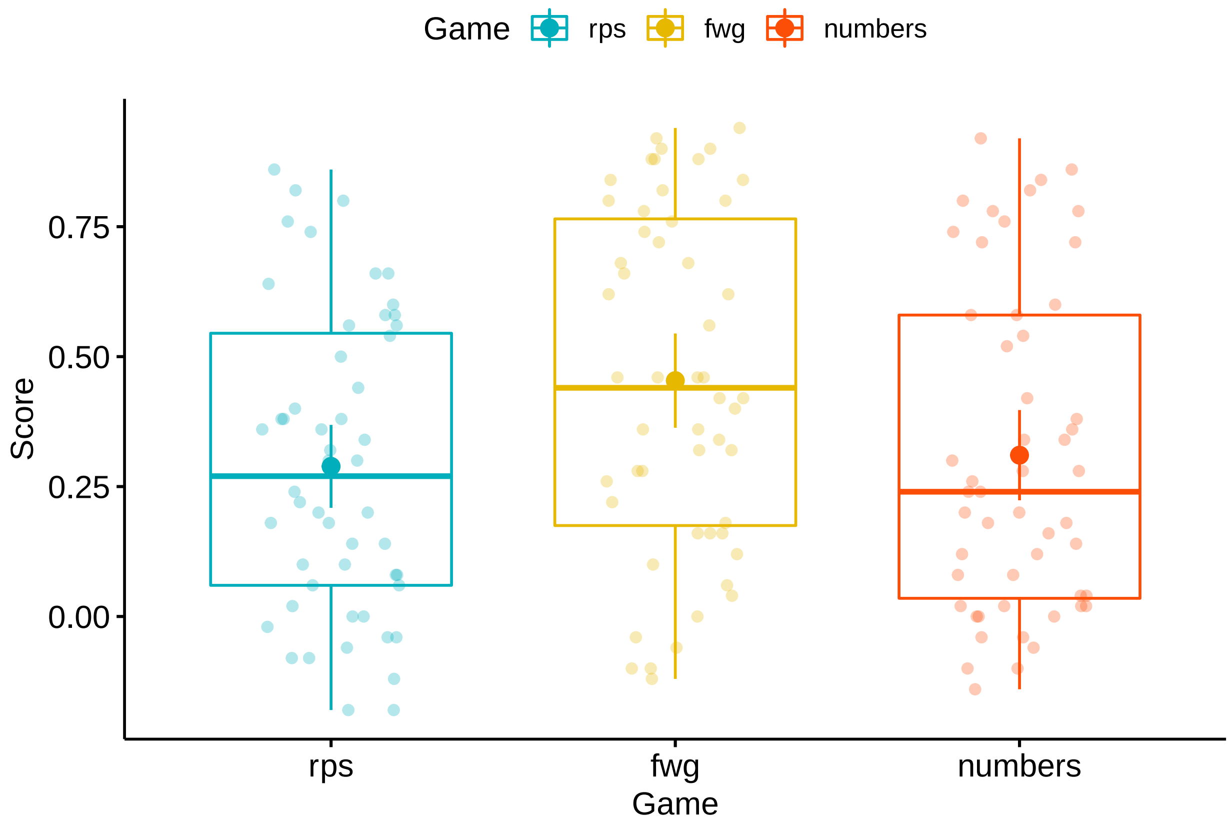


Figure 2. Boxplot of scores per game across conditions

level-2 player ($M = 0.27$). Thus, it appears that is was harder for participants to exploit the strategy of the more sophisticated level-2 opponent than the comparatively less sophisticated level-1 opponent.

Learning transfer. As a measure for learning transfer, we focus on participants' scores in the initial 5 rounds after the first round (rounds 2-6) of each game (see Figure **??**). We exclude the very first round as the computer opponent plays randomly here and there is no opportunity yet for the human player to exploit their opponent's strategy. Players with no knowledge of their opponent's strategy are expected to perform at chance level in these early rounds. Positive scores in rounds 2-6 reflect generalization of prior experience. The FWG early score score is significantly higher than 0 ($t(148.85) = 4.584$, $p < .001$). This is also the case for the NUMBERS game ($t(148.85) = 3.00$, $p = .009$). We did not expect positive scores for the RPS game, as it was the first game played and there was no opportunity for learning about the opponent's strategy. Scores in this game was indeed not significantly different from 0 ($t(148.85) = 1.04$, $p = .89$).

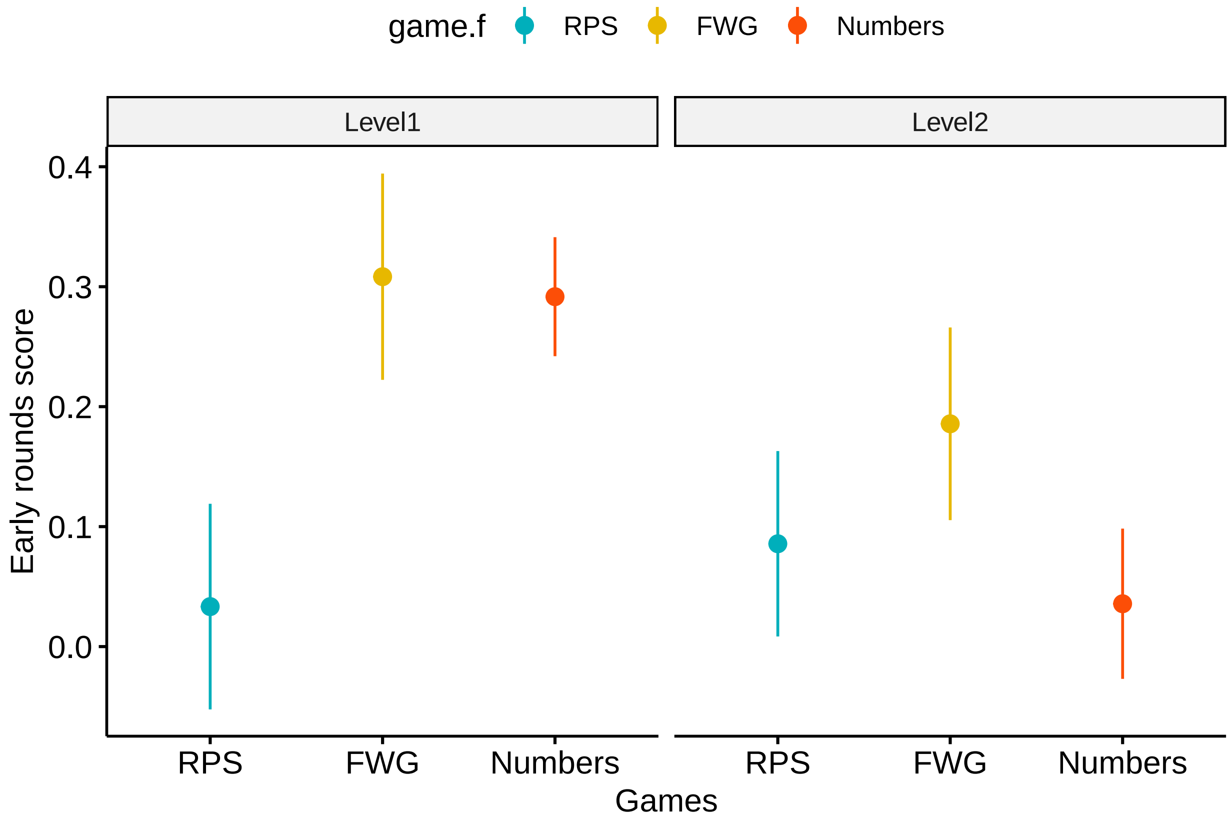


Figure 3. Mean and standard error of early scores by game and opponent

Next, we explore whether learning transfer is moderated by the type of opponent and game similarity. We expected better transfer between more similar games (i.e. better transfer from RPS to FWG than from RPS/FWG to NUMBERS), and worse transfer for the more sophisticated level 2 agent. Figure~3 indicates that the pattern over the games is indeed dissimilar between level-1 and level-2 players. To explore this, we used a 2 (condition: level-1, level-2) by 3 (game: RPS, FWG, NUMBERS) repeated measures ANOVA with the first factor varying between participants. There was a main effect of Game ($F(2, 92) = 3.35$, $\eta^2 = 0.04$, $p < .04$). We then run statistical tests on early round scores by game and opponent against the null hypothesis of 0 (no transfer). For level-1 facing players, there is evidence of learning transfer from RPS to both FWG ($t(150) = 3.96$, $p < .001$) and NUMBERS ($t(150) = 3.74$, $p < .001$). For level-2 facing players, there is evidence for transfer from RPS to the similar game FWG, albeit scores are lower than for level-1 player ($t(150) = 2.48$, $p = .01$) but not to the dissimilar game of NUMBERS.

Experiment 1 discussion: These results indicate that learning transfer to the more dissimilar game (NUMBERS) we found earlier is exclusively driven by level-1 facing players, as average early round scores in the NUMBERS game of level-2 facing players are close to 0. Therefore, both participants facing level-1 and level-2 agents can transfer learning to the similar FWG game, but only those facing the less sophisticated opponent are able to generalise to the less similar NUMBERS game.

Experiment 2

We ran a second experiment with various differentiated features to improve the opportunity to measure learning transfer. Instead of making participants face either the level-1 or level-2 player throughout, we made them face both opponents sequentially. Because there were two distinct opponents, requiring potentially holding two opponent models in memory, we also made it easier to recall the results of past rounds by providing participants with the opportunity to see the history of the game since the beginning of each interaction. Figure 1 shows an example of showing interaction history in the RPS game. Finally, we changed the third game to a penalty shootout game, whith participants aiming to score a goal and opponents playing the role of goal keepers. Whilst this game has the same number of actions as the first two, it is strategically dissimilar. Unlike the third game in the first experiment, it did not have a cyclical hierarchy between actions, making it harder to win by just following simple heuristics leveraging this cyclicality. If we see evidence for differential play against opponents, it would show participants adapting their strategies to the opponent they are facing, which is indicative of opponent modelling.

Methods

Participants & Design. A total of 48 participants (21 females, 28 males, 1 unknown) used the Prolific Academic platform to participate in the experiment. This was a new set of participants unrelated to those taking part in Experiment 1. The average age of players was 30.2 years, and the mean duration to complete the task was 39 minutes. Participants were incentivised using a two-tier payment mechanism: a fixed fee of £2.5 for completing the experiment plus a performance linked bonus that averaged £1.32.

Tasks. The three games the participants played were Rock-Paper-Scissors, Fire-Water-Grass, and the penalty shootout game. The first two games were identical to the ones used in the first experiment. In the final game (shootout) the participants took the role of the player shooting a football (soccer) penalty, with the AI opponent being the goalkeeper. Players had the choice between three actions, like in the first two games: Shooting the football to the left, right or centre of the goal. If the player shoots in a direction different from that of where the goalkeeper dives, they win the round and the goalkeeper loses. Otherwise, the goalkeeper catches the ball and the player loses the round. There is no possibility of ties in this game. Figure 4 shows a snapshot of play in the shootout game. What makes this game different however is that there are two ways to beat the opponent in each round: if we think the opponent is going to choose “’right”’ in the next round, we can win by both choosing “’left”’ and “’center”’. A level-1 player (thinks that his opponent will repeat his last action) has two ways to win the next round. As such, we have programmed the level-2 computer program to choose randomly between the two possibilities that a level-1 player may choose.

Procedure. The participants played 3 games sequentially against both level-1 and level-2 computer opponents, rather than just one like in the first experiment. Like in the first experiment, the computer opponents retained the same strategy throughout the 3

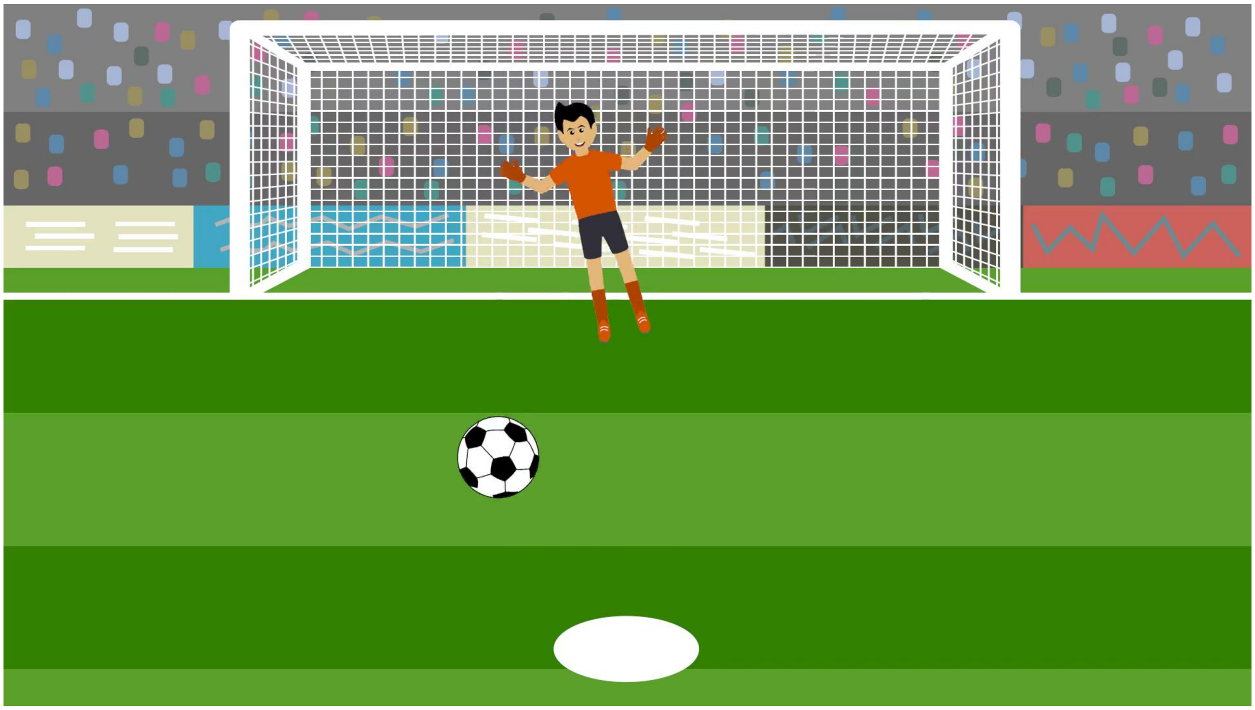


Figure 4. Screenshot of the shootout game

games, however the participants faced each opponent twice in each game. Specifically, each game was divided into 4 stages numbered 1 to 4, consisting of 20, 20, 10, and 10 rounds respectively for a total of 60 rounds per game. Participants start by facing one of the opponents in stage one, then face the other in stage two. This is repeated in the same order in stages 3 and 4. Which opponent they faced first was counterbalanced. All participants engage in the same three games (namely RPS, FWG and Shootout) in this exact order, and were aware that the opponent was not able to know their choices beforehand but was choosing simultaneously with the player. In order to encourage participants to think about their next choice, a countdown timer of 3 seconds was introduced at the beginning of each round. During those 3 seconds, participants could not choose an option and had to wait for the timer to run out. A small delay that changed randomly (between 0.5 and 3 seconds) was also introduced in the time it took the AI agent to give back their response, as a way of simulating a real human opponent thinking time. After each round, the participants were given detailed feedback about their opponent actions as well as whether they won or lost the last round. Further information about the outcome of previous rounds was also visible on the game screen below the feedback area. Throughout each stage, participants could scroll down to recall the history of interaction. The number of wins, losses and ties were clearly shown at the top of the screen for each game, and this scoreboard was reinitialised to zero at the onset of a new stage game. As in the first experiment, all the games have a unique MSNE consisting of randomising across actions. If participants follow this strategy, or simply don't engage in learning how the opponent plays, they would score 0 on average against both level-1 and level-2 players. Evidence of sustained wins would indicate that participants have learned to exploit patterns in the opponent play.

Results

The RPS game had the lowest average score per round (M = 0.194, SD = 0.345) followed by FWG (M = 0.27, SD = 0.394) and finally the Shootout game had an adjusted average score in between the two (M = 0.289, SD = 0.326).[1]. Using parametric t-tests on adjusted scores, we reject the null hypothesis of random play in all three games (RPS: t(49) = 6.26, $p < 0.0001$; FWG: t(49) = 7.25 , $p < 0.0001$; Shootout: t(49) = 13.61, $p < 0.0001$). Using the average scores obtained by participants in each game and interaction, we explore whether learning has occurred within and between games. We perform a two (condition: level-1 first, level-2 first) by two (opponent type: level-1 or level-2) by three (game: RPS, FWG, Shootout) by two (interaction: first or second) repeated measures ANOVA with the first factor varying between participants.

[1] A higher score in shootout is expected as there are 2 out of three possible winning actions, compared to one out of three in RPS and FWG. Indeed, a player not aiming to uncover the opponent's strategy and thus choosing to play randomly should be expected to have on average score per round of 0 in both RPS and FWG, and 0.33 in the Shootout game. To make the scores more comparable, and because we are interested in player's performance that is not due to chance, we will adjust all scores in the shootout game by subtracting the average score per round of a random strategy (0.33)

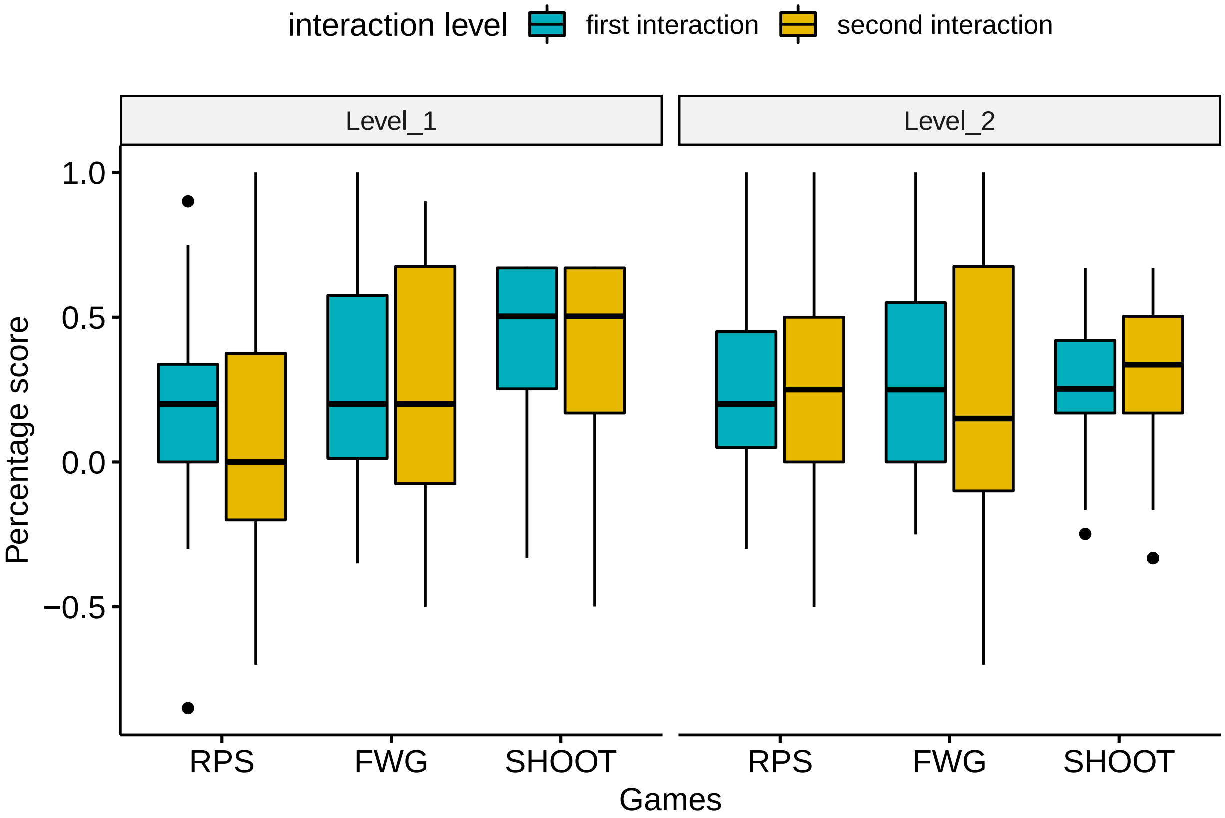


Figure 5. Boxplot of scores per game and interaction by opponent type

There is evidence for a main effect of Game on scores (F(1.85,88.7) = 11.81, η^2 = 0.04, p < .0001). To explore these differences further, we look at post-hoc analyses for pairwise comparisons between game scores (p-values adjusted using Holm method for multiple comparisons). We find the performance in the games increases steadily throughout the experiment, with FWG performance significantly higher than RPS (t(96) =2.53, p = 0.025), and performance in the Shootout game also significantly higher than in FWG (t(96) = 2.32, p = 0.025). There was no main effect of either opponent type, the interaction factor(first or second time opponent was faced) , or the condition factor (whether level-1 or level-2 opponent was faced first). There was however a significant interaction effect between Game and opponent type (F(1.7, 81.82) = 5.31,η^2 = 0.02, p = .01). Figure 5 shows boxplots of game scores, averaged across participants, by game and opponent type. We also distinguish between scores from the first time the players faced the opponent (first interaction) and the second time they did (second interaction). We see that when facing level-1 agents, scores increase steadily after each game, with FWG score significantly higher than RPS (t(191) = 2.70, p = 0.03) and Shootout scores in turn significantly higher than FWG (t(191) = 3.05, p = 0.01). There was no significant difference between average scores on any two games when facing level-2 agents however.

Learning transfer. As a measure for learning transfer we will again compare scores only on rounds 2-6 of each game, excluding the very first round where play is necessarily random.

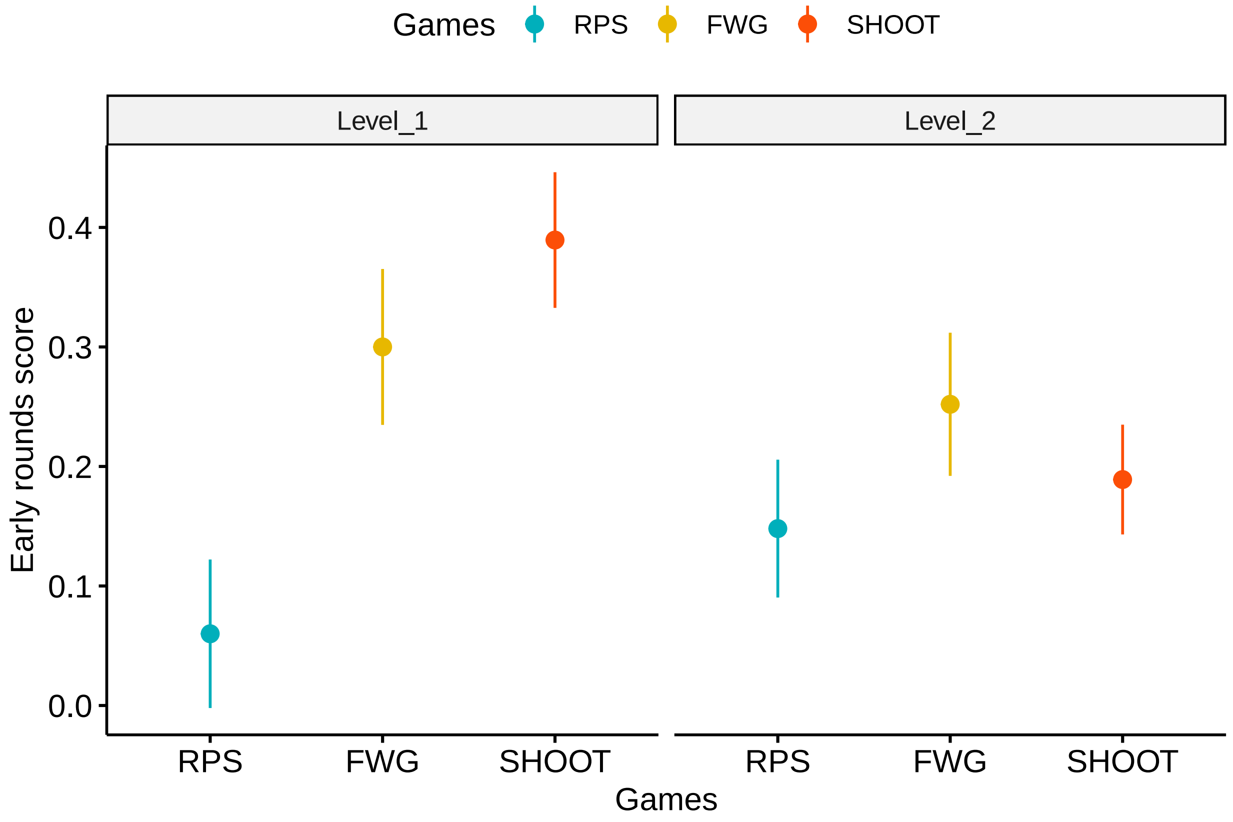


Figure 6. Mean and standard error of early scores by game and opponent for experiment 2

In Figure 6, we show the average score across participants and its 95 percent confidence interval in rounds 2-6 of the first interaction with the opponent for each game.

These scores are also averaged across the levels of condition (meaning they are irrespective of which opponent players faced first). For both the FWG and Shootout games, score in the early rounds of the first interaction are significantly higher than 0 for both opponent types. (Level-1 opponent: FWG: t(270) = 4.99, $p < 0.0001$; Shootout: t(270) = 6.66, $p < 0.0001$; Level-2 opponent: FWG: t(270) = 4.40, $p < 0.0001$; Shootout: t(270) = 3.21, p =0.004).

Experiment 2 discussion: Looking at learning transfer by type of opponent faced, we confirm the result from the first experiment that it is easier to transfer learning to the more dissimilar game (Shootout) when facing a level 1 opponent. Indeed, while the early scores of FWG for level-1 and level-2 facing players are not significantly different from each other, the score of the players facing the level-1 opponent is indeed almost 0.2 point per round higher than that of players facing level-2 opponents, and the difference is statistically significant (t(144) = 2.45 , p = 0.01). These early scores have also been adjusted to account for the fact that the shootout game has higher average scores when playing randomly, and therefore this difference is really due to better learning transfer and not due to chance.

Computational modelling

To gain more insight into participants' strategies against their computer opponents, we constructed and tested several computational models of strategy learning. The baseline model assumes play is random, and each potential action is chosen with equal probability. Note that this corresponds to the Nash equilibrium strategy. The other models adapted their play to the opponent, either by reinforcing successful actions in each game (reinforcement learning), or by determining the type of opponent through Bayesian learning (Bayesian Cognitive Hierarchy models). We also include the Expected Weighted Attraction (EWA), which is a popular model in behavioral economics.

We use the following notation. In each game $g \in \{\text{RPS}, \text{FWG}, \text{NUMBERS}\}$, on each trial t, the participant chooses an action $a_t \in \mathcal{A}_g$, and the opponent chooses action $o_t \in \mathcal{A}_g$, where $\mathcal{A}_g$ is the set of allowed actions in game g, e.g. $\mathcal{A}_{\text{RPS}} = \{R, P, S\}$. The participant then receives reward $r_t \in \{1, 0, -1\}$, and the opponent receives $-r_t$. We use the state variable $s_t = \{a_{t-1}, o_{t-1}\}$ to denote the actions taken in the previous round $t-1$ by the participant and opponent.

In the following, we will describe the models in more detail, and provide some intuition into how they they learn about the game and/or the opponent.

Reinforcement learning (RL) model

We first consider a model-free reinforcement learning algorithm, where actions that have led to positive rewards are reinforced, and the likelihood of actions that led to a negative reward is lowered. Since the computer players in this experiment based their play on the actions in the previous round, a suitable RL model for this situation is one which learns the value of actions contingent on plays in the previous round, i.e. by defining the state s_t as above. The resulting RL model learns a Q value (Watkins & Dayan, 1992) for each state-action pair:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha \left(r_t - Q(s_t, a_t) \right)$$

where $Q(s_t, a_t)$ is the value of taking action a when in state s at time t, $\alpha \in [0, 1]$ the learning rate. For instance, $Q_t(\{R, S\}, P)$ denotes the value of taking action "'Paper"' this round if the player's last action was "'Rock"' and the opponent played "'Scissors"'. Actions are taken according to a softmax rule:

$$P_t(a|s_t) = \frac{\exp\{\lambda Q_t(a, s_t)\}}{\sum_{a' \in \mathcal{A}_g} \exp\{\lambda Q_t(a', s_t)\}}$$

While this RL model allows the players to compute the values of actions conditional on past play, crucially, it will not be able to transfer learning between games, as each game has a different action space $\mathcal{A}_g$, and there is no simple way to map actions between games.

The RL model has two free parameters: the learning rate (α) and the inverse temperature parameter of the softmax decision rule (λ). We've assumed $\gamma = 0$ as changes in this parameter didn't affect the other parameter estimations or ultimately the value of the likelihoods.

Experience-weighted attraction (EWA) model

The self-tuning Experience Weighted Attraction (EWA) model (Ho, Camerer, & Chong, 2004) combines two seemingly different approaches, namely reinforcement learning and belief learning. Belief learning models are based on the assumption that players keep track of the frequency of past actions and best respond to that. By contrast, reinforcement learning does not explicitly take into account beliefs about other players, but simply increases the probability of repeating a more rewarding action. The self-tuning EWA model has been shown to perform better than either RL or belief learning alone in various repeated games and has the advantage of having only one free parameter, the inverse temperature of the softmax choice function. The EWA model is based on updating

"Attractions" for each action over time. The attraction of action a time t is written $A_t(a)$ and is updated as

$$A_{t+1}(a) = \frac{\phi \; N(t) \; A_t(a) + [\delta + (1 - \delta) \; I(a_t = a)] \; R(a, o_t)}{\phi \; N(t) + 1}$$

where $I(x)$ is an indicator function which takes the value 1 is its argument is true, and 0 otherwise, and $R(a, o_t)$ is the reward that would be obtained from playing action a against opponent action o_t, which equals the actual obtained reward when $a = a_t$, and otherwise is a counterfactual reward that would have been obtained if a different action were taken. Unlike reinforcement learning, this uses knowledge of the rules of the game to allow reinforcing actions that were not taken. We can see that setting $\delta = 0$ leads to reinforcement of past actions, while positive and high delta parameters make the update rule take into account foregone pay-offs, which is similar to weighted fictitious play @[cheung1994learning]. While the assumption in expanding the update rule above is that ϕ and δ are free parameters (Camerer, Ho, & Others, 1997), the self-tuning aspect of the model comes from the fact that these are now self-tuned using the formulas expanded in (Ho et al., 2004). $ N(t)$ represents an experience weights and can be interpreted as the number of "observation-equivalents" of past experience. We initialise it to 1 so initial attractions and reinforcement from payoffs are weighted equally.

As in the models above, actions are chosen based on a softmax decision rule:

$$P_t(a) = \frac{\exp\{\lambda A_t(a)\}}{\sum_{a' \in \mathcal{A}_t} \exp\{\lambda A_t(a')\}}$$

The self-tuning EWA has one free parameter: the inverse temperature of the softmax decision rule (λ).

Bayesian Cognitive Hierarchy (BCH) model

In what we call the Bayesian Cognitive Hierarchy (BCH) model, the participant attempts to learn the type of opponent they are facing through Bayesian learning. We assume the participant considers the opponent could be either a level 0, level 1, or level 2 player, and starts with a prior belief that each of these types is equally likely. They then use observations of the opponents actions to infer a posterior probability of each type:

$$P(\text{level} = k|D_t) \propto P(D_t|\text{level} = k) \times P(\text{level} = k)$$

where $D_t = \{a_1, o_1, \ldots, a_t, o_t\}$ is the data available at time t. The likelihood is defined as

$$P(D_t|\text{level} = k) = \prod_{j=1}^{t} \left(\theta \frac{1}{|\mathcal{A}_g|} + (1 - \theta) f_k(o_j|a_{j-1}, o_{j-1}) \right)$$

where $f_k(o_t|a_{t-1}, o_{t-1}) = 1$ if o_t is the action taken by a level k player when the previous round play was a_{t-1} and o_{t-1}, and 0 otherwise. Note that the likelihood assumes (correctly) that there is a probability $\theta \in [0, 1]$ that the opponent takes a random action. The posterior at time $t - 1$ forms the prior at time t. We assume a participant chooses an action by using the softmax function over the best response to predicted actions:

$$B_t(a) = \sum_{k=0}^{2} \sum_{o \in \mathcal{A}_g} b(a, o) P_k(o|a_{t-1}, o_{t-1}) P(\text{level} = k|\mathcal{D}_{t-1})$$

$$P_t(a) = \frac{\exp \lambda B_t(a)}{\sum_{a' \in \mathcal{A}_g} \exp \lambda B_t(a')}$$

where $b(a, o) = 1$ if action a is a best response to opponent's action o (i.e. it leads to a win), and $P_k(o|a_{t-1}, o_{t-1}) = \theta \frac{1}{|\mathcal{A}_g|} + (1 - \theta) f_k(o|a_{t-1}, o_{t-1})$ is the probability that a level k agent takes action o, as also used in the likelihood above.

Unlike the models above, the BCH model allows for between-game transfer, as knowledge of the level of the opponent can be used to generate predictions in games that have not been played before. However, the participant might also assume that the level of reasoning of their opponent does not generalize over games. We hence distinguish between two versions of the BCH model. In the No-Between-Transfer (BCH_NBT) variant, participants assume a uniform probability of the different levels at the start of each game (and hence do not transfer knowledge of their opponent between games). In the Between-Transfer model (BCH_BT), participants use the posterior probability over the levels of their opponent as the prior at the start of a new game (i.e. complete transfer of the knowledge of their opponent). Both versions of the BCH model have two free parameters: the assumed probability that the opponent chooses a random action (θ), and the temperature parameter of the softmax function (λ).

Estimation and model comparison

In both experiments, all models were fit to each participant data, with optimal parameters being estimated using maximum likelihood. Using information criteria based Bayesian model comparison (BIC), the best fitting model for each participant was chosen and we compared the number of participants whose behavior was best explained by each model.

In experiment 1, we fit a total of 5 models: A baseline model assuming random play (Nash), Bayesian Cognitive Hierarchy model allowing transfer betwen games (BCH_BT) and another with no transfer between games (BCH_NBT), as well as a Reinforcement Learning model with state space consisting of last round play (RL), and finally a self tuning EWA model with the same state space (EWA).

In experiment 2, because participants were interacting with each opponent twice within each game, we need to distinguish between two type of opponent model transfer. We can have transfer within games, between the first and second interaction with the opponent. We can also have transfer between games, as in transfering models of the opponent from RPS to FWG for instance. Therefore, we give models more flexibility to allow for these two types of transfer. As such, for Bayesian Cognitive Hierarchy, we fit a total of three models. BCH_BT allows for between game transfer. We assume that if such transfer is possible, then within game transfer should also be possible and therefore this model allows for both types of transfer. BCH_NBT allows for within but no between game transfer. Whereas BCH_NT allows for neither. For Reinforcement learning models, because RL models can't account for between game transfer due to change in ation labels, we can only have models allowing for within game transfer (RL Tr) or with no transfer within games (RL NT). Likewise, we fit both a self tuning EWA model with transfer between stages of the same game (EWA_Tr) or without transfer (EWA_NT). Counting the base model wiht random play (Nash) we therefore fit a total of 8 models in experiment 2.

Experiment 1 modelling :

Figure 7 shows the results for experiment 1: We can see that the RL model clearly described most participants' behaviour best, followed by the random (Nash) model. Only a few participants were best described by one of the BCH models, or the EWA model. Looking at BIC weights, we confirm this as seen in Figure 8. RL models hve high BIC weights when they best fit the participants, and very few instances have high BIC weights for models other then RL, whihc fits the picture drawn by the histogram.

Experiment 2

In experiment 2, we can see from Figure 9 that the RL model was again more successful than the Bayesian models in fitting player's action choices. In experiment 2 when participants faced both level-1 and level-2 agents sequentially, the Bayesian models (with or without transfer) did not fit players observed data as well. This is also reflected in BIC weights in Figure 10

Using Hidden Markov Model to explore strategy switching

The computational modelling indicates that most players are best fit by Q-learning type models with states defined by last round play. This is at odds with the findings from the section regarding learning transfer: If indeed most participants use Q-learning with states to choose their actions, then they should not be able to transfer learning to the early

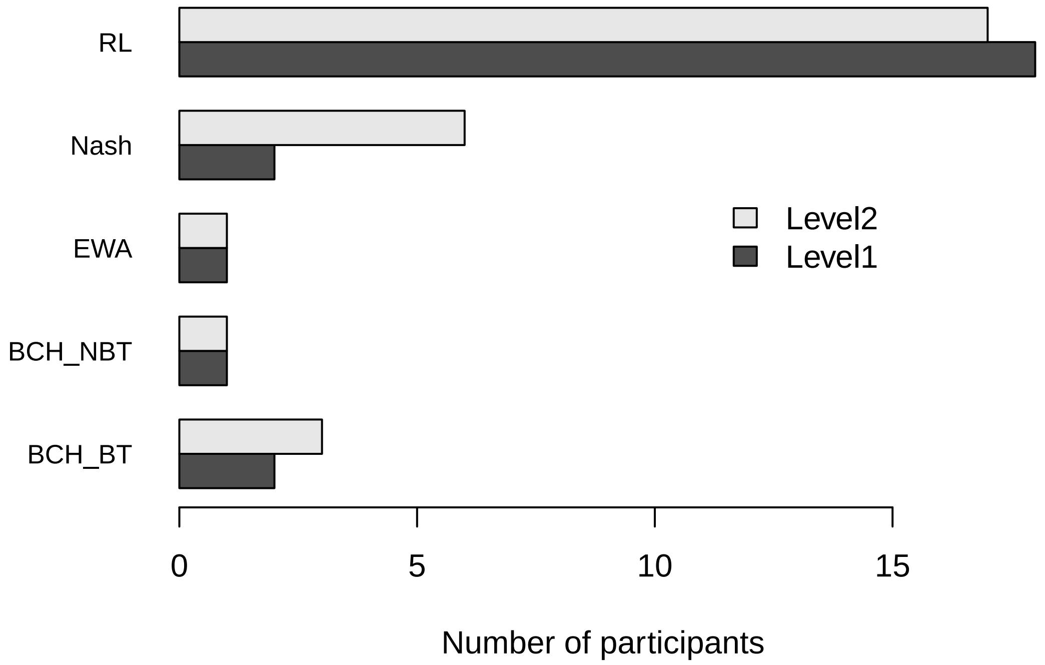


Figure 7. Experiment 1 - Histogram of best fitting computational models by condition

rounds of the new game. In order to understand better what is going on, we plot the likelihood by trial for each game and each of the three strategies: Q-learning with states, and Bayesian Cognitive Hierarchy models with and without the possibility of across game transfer.

We start with experiment 1 data. Figure 11 shows that in the later games, the likelihood for the BCH models is higher in the initial rounds in which learning transfer is measured, but that over time, the likelihood of Q-learning model becomes more important and exceeds that of BCH models.

Likewise, in experiment 2, we want to understand the dynamic of strategy choice by plotting the likelihood by trial for each strategy, using the optimal parameters found when fitting the model. Figure 12 shows that, as in experiment 1, BCH models had higher

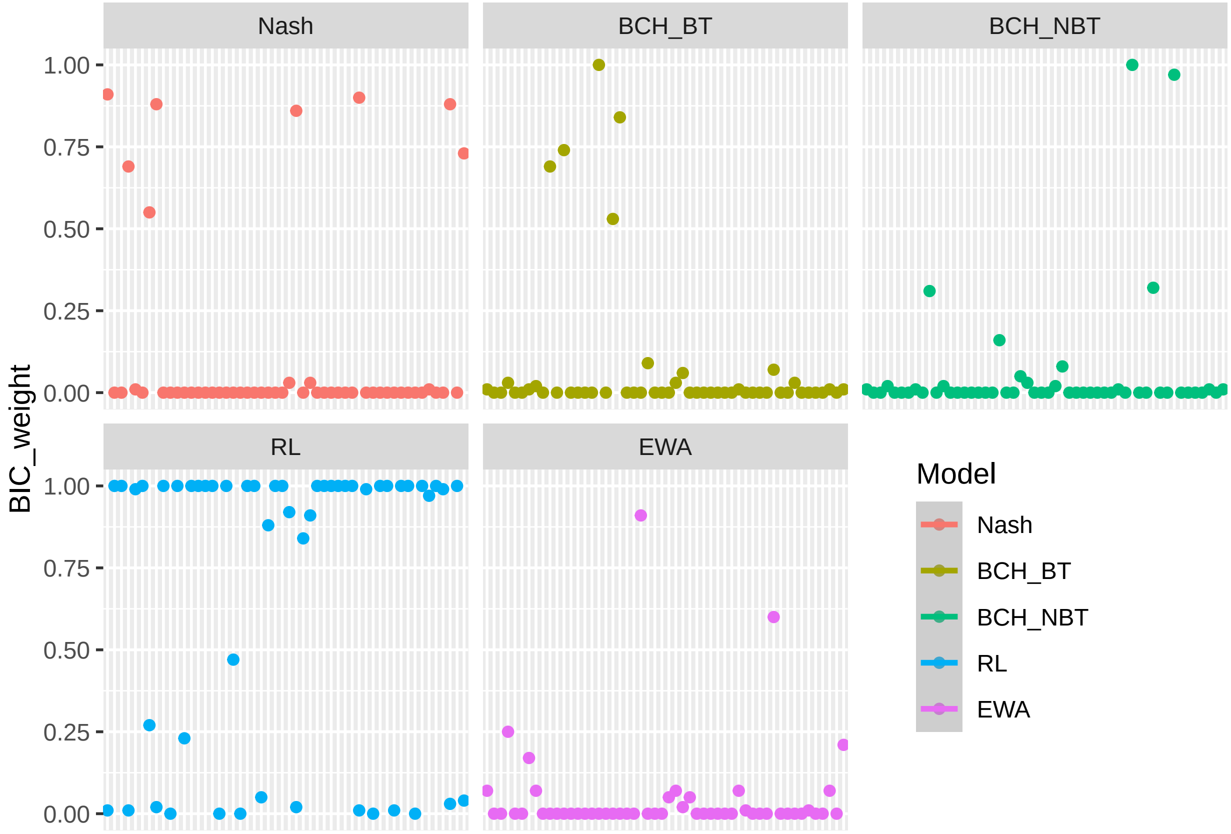


Figure 8. Model BIC weights for participants in experiment 1

likelihood in the early stages of the second (most similar) game, however the likelihood of Q-learning with states models increases steadily to be the highest in the later stages of all games. In the third and more dissimilar game, we get a result that is different from experiment 1. In this instance, the likelihoods of the BCH models stay constant and close to their initial values.

The fact that the likelihoods of the main strategies considered cross over in both experiments could be interpreted as indicative of participants switching between strategies as the games progressed. Indeed, in both experiments, following our results, it seems that in the earlier stages of the latter games, the BCH based strategies fitted observed action choices better than Q-learning based ones, with a reversal of the roles in later stages.

In order to test for the existence of strategy switching in participants' play, we fit

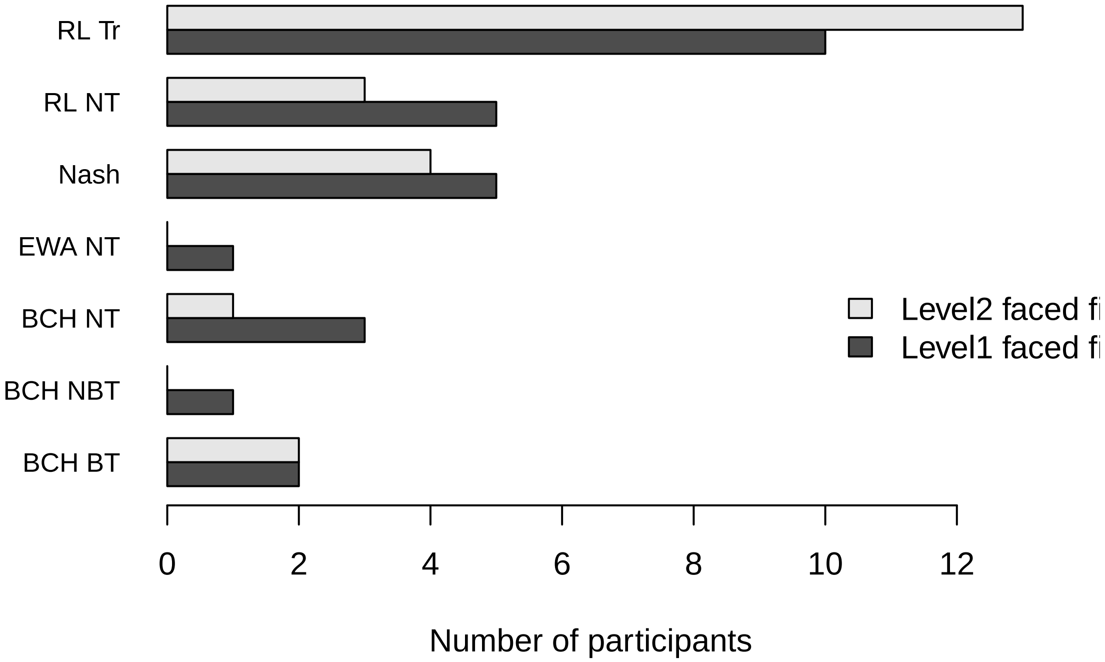


Figure 9. Experiment 2 Histogram of best fitting computational models by condition

Hidden Markov Models in which the latent states are the 3 strategies used (Q-learning with state space consisting of previous round play, BCH based model with opponent model transfer, and a base model consisting of random play consistent with a Nash equilibrium strategy). Hidden Markov models are useful tools to explore structure in observed time series. They are named as such because of two properties: First, they make the assumption that any observable action at time t results from a process whose state at time t , named S_t is "hidden" from the observer. Second, it also assumes that this hidden process has a Markov property, meaning that given state S_{t-1}, the value of S_t is independent of all states occurring before time $t-1$. We also assume that S_t has a discrete probability distribution in that it take one of K discrete values. The model is therefore specified by initial probabilities of being in each state $1, 2, ..., K$ and transition probabilities for moving from

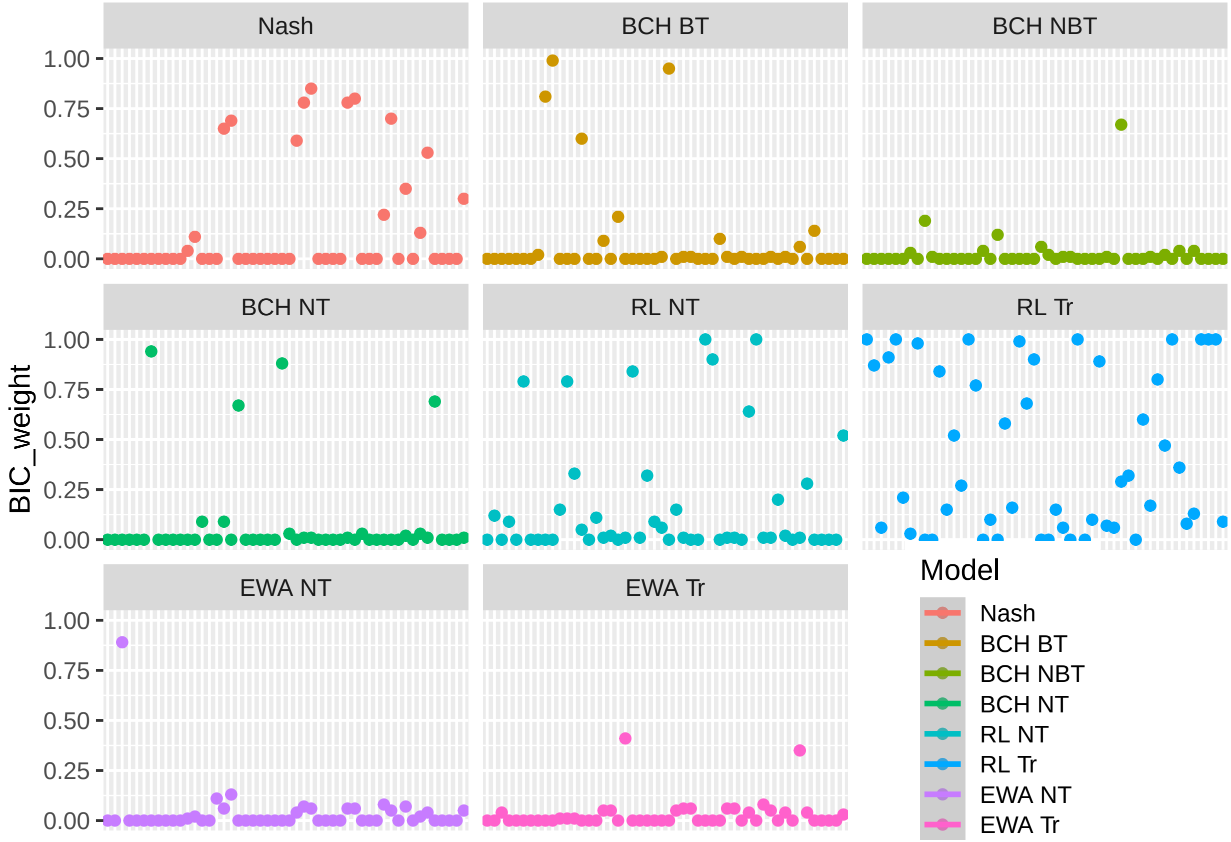


Figure 10. Model BIC weights for participants in experiment 2

state i to state j. These probabilities are fit using observed actions generated from these hidden states.

To investigate the possibility of strategy switching, we fit two different hidden Markov models with the depmixS4 R package. In the first model, we allow for a non-nil probability of players transitioning from one state (strategy) to another. In the second model, we assume that such switching does not happen, and as such assume implicitly that when players start with a particular strategy, they continue using it throughout the experiment. We then compare the likelihoods of each HMM model using a likelihood ratio test.

Experiment 1:

```
## converged at iteration 7 with logLik: -7415.978
```

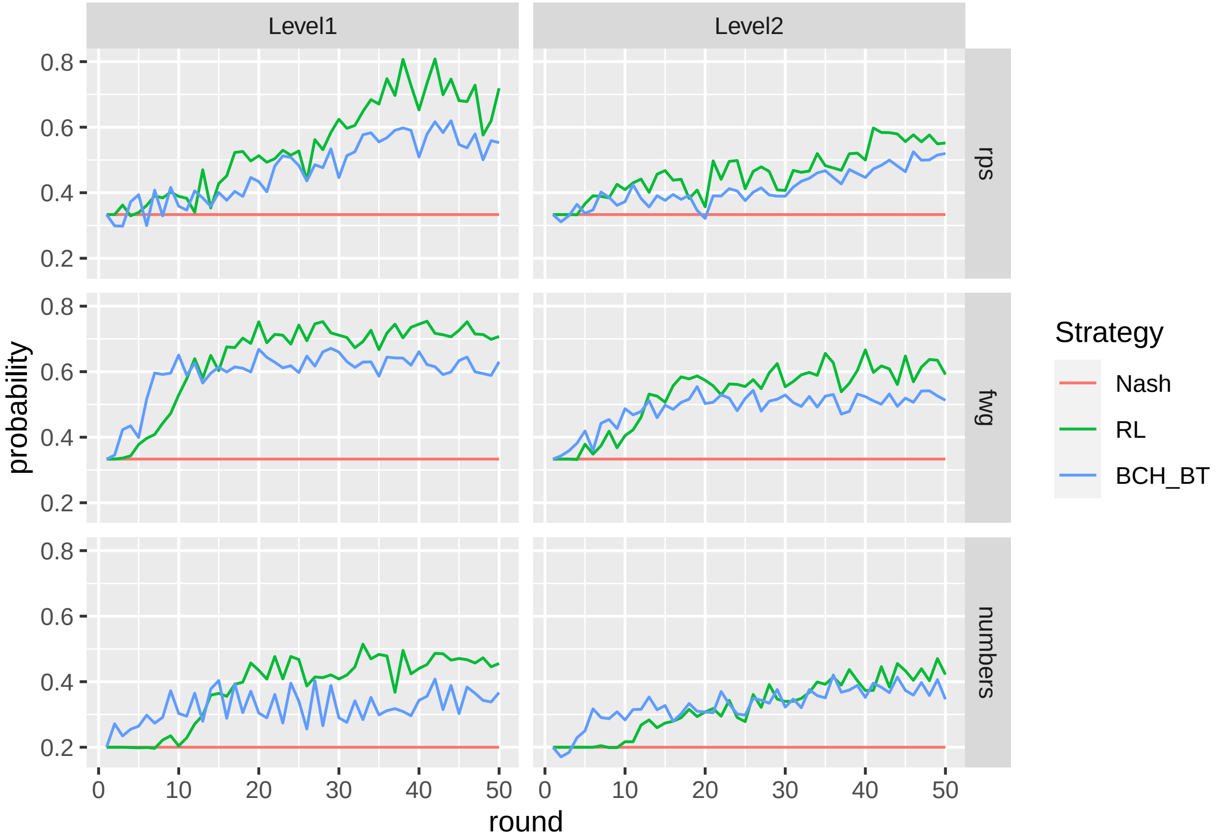


Figure 11. Experiment 1 Likelihood by trial by game and opponent faced

```
## log Likelihood ratio (chi^2): 211.087 (df=0), p=0.
```

```
## [1] 0
```

In experiment 1, the HMM model with switching fits significantly better than the non-switching one ($p < .001$). This is further statistical evidence in favour of the hypothesis that participants switch between strategies. In order to understand at which stage of the games the switching might happen, and whether there are any differences between games and type of opponents faced, we plot in Figure 13 the average (across participants) posterior probabilities of each state (strategy), as a function of trial and opponent faced. The posterior probability is the probability that an observation comes from a component distribution a posteriori, i.e. given the value of the observation. In the first experiment, we

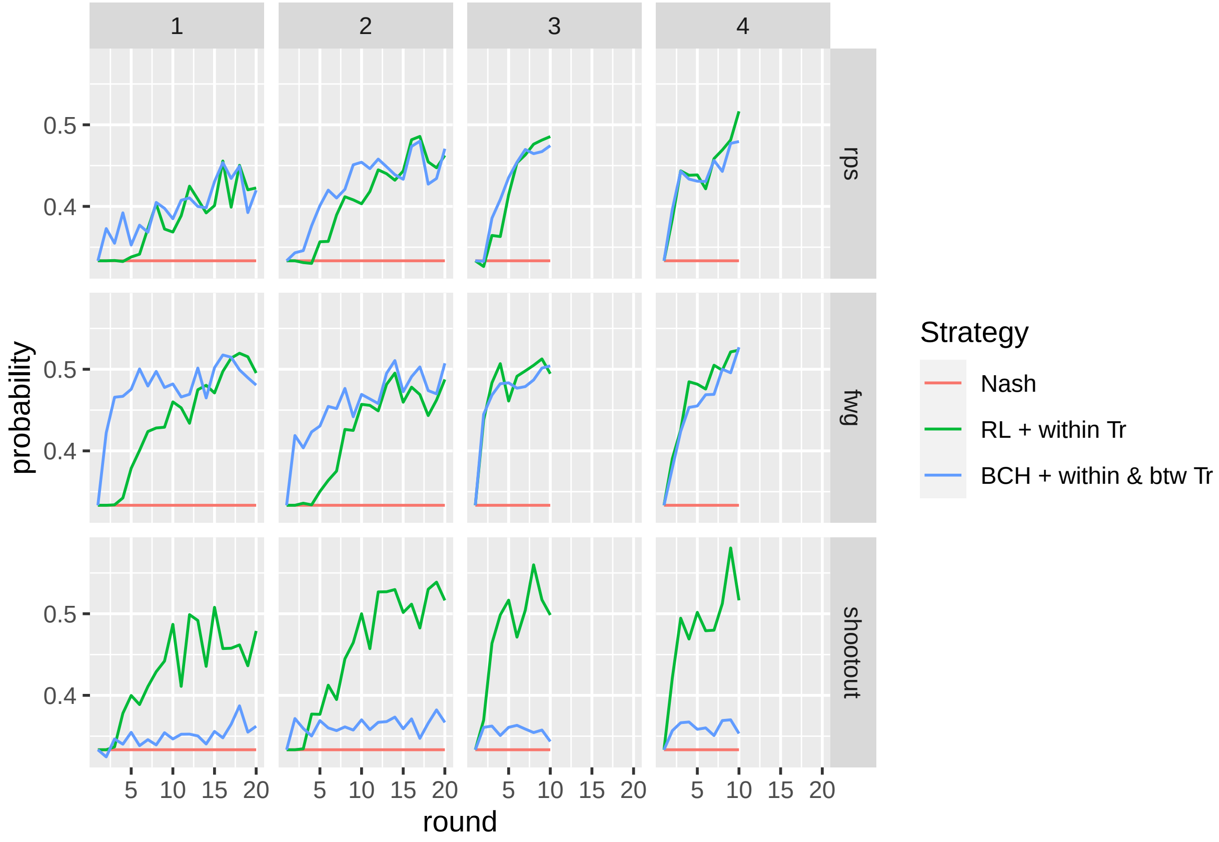


Figure 12. Experiment2 likelihood by trial by game and opponent faced

can see from the plots of FWG and numbers games for level-1 opponent that although the likelihoods are very close, the posterior probability of the Bayesian model with transfer is slightly higher than that of the RL model in the very early rounds, but decreases rapidly while the posterior probability of the QL-learning with states models keeps increasing.

Experiment 2: The switching model in experiment two has also significantly higher likelihood ($p = 0.00$). On top of indications from looking at the likelihood by trial graphs, we have therefore further evidence that participants did indeed switch their strategies as the games progressed. The posterior probability plot in Figure 14 shows switching much more clearly across games and stages. The switching also seems to happen very early on at the beginning of each game and stage, and is also consistently in the same direction: The probability of Bayesian models with transfer being initially high, then

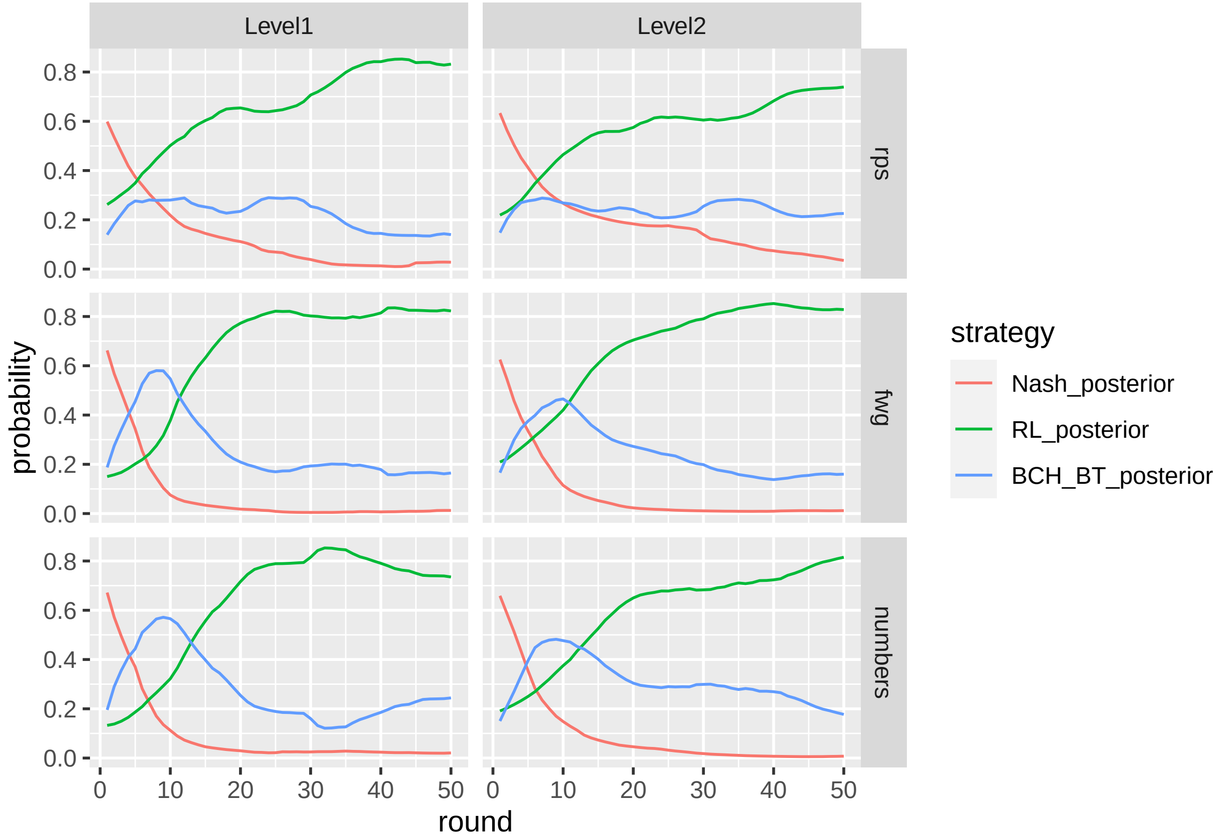


Figure 13. Experiment1 posterior probability of strategies by game and opponent faced

decreasing rapidly while the posterior probability of QL-Learning with states and within transfer learning increases rapidly.

Therefore, HMM modelling shows clear evidence in favour of strategy switching by participants, specifically after a few rounds of play. The strategy switching is consistently from BCH models towards Q-learning with states models in both experiments.

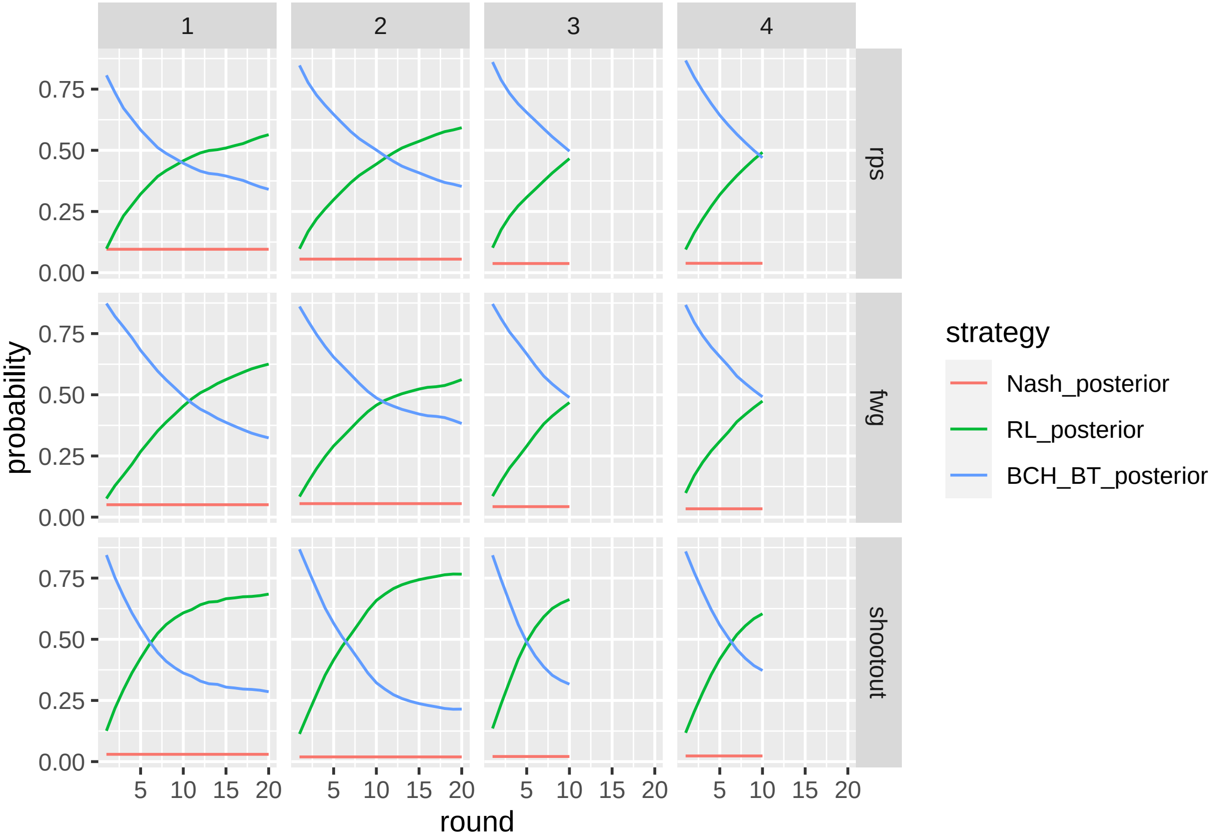


Figure 14. Experiment2 likelihood by trial by game and opponent faced

Discussion

In this study, we investigated human learning transfer across games by making human participants play against computer agents with rule-based level-k strategies. We were interested in exploring whether participants learn about the strategy of their opponent and transfer such knowledge between games, and whether this is modulated by the similarity between games and the sophistication of the agent.

The results of our first experiment show that the majority of participants learn to adapt to the opponent strategy over multiple interactions and generalise this learning to the similar game. We used results on very early rounds for measuring transfer as they are unlikely to be tainted by any within game learning. Using this approach, we showed that transfer to the more dissimilar game was modulated by the degree of sophistication of the agent, with evidence for transfer when players face the less sophisticated agent but not the more sophisticated one.

In the second online experiment, there were many more opportunities to test transfer than before: 2 opportunities to transfer opponent models within each game, and a total of three games, which means 6 opportunities to test transfer. The results on learning transfer confirmed prior findings from the first experiment. While there was no evidence of higher scores across interactions within the same game (likely due to the lower number of rounds per interaction and the higher cognitive load of facing two opponents rather than one), we found evidence for learning transfer across games as early round scores analysis confirmed. We also found that learning transfer is modulated by the type of opponent faced. When the players faced the level-1 opponent, they were able to transfer learning. However, when they faced the level-2 opponent, there was weaker evidence for transfer. The lack of transfer when facing the more sophisticated opponent might be due to the difficulty of learning that opponent strategy to start with. A player cannot transfer what they have not learnt and as such, since it might be harder to learn the strategy of the level-2 opponent, this in turn might translate into weaker evidence for transfer.

Coming back to learning transfer, we observed evidence that participants start off new games with prior knowledge as their scores are significantly higher than chance, confirmed both by early stage analysis as well as rounds 2-6 scores analysis. The question we ask ourselves therefore is: What exactly did the players learn in RPS that allowed them to beat the opponent in FWG and Shootout? what did the players learn specifically about the opponent strategy and what form did this learning take?

We will proceed by considering multiple potential answers to this question. A possible hypothesis for learning the opponent's strategy is the use of simple rules based on last round play (for instance, I play scissors whenever opponent played rock in last round, or whenever the last round play was rock/scissors, I should play paper in this round, etc. . .). Our Q-learning with states as prior-round play model is a good proxy for this type of strategies. While this approach certainly seemed to be the best fit for many player's behavior, it is unsatisfactory in explaining some of the learning transfer evidence we showed. Indeed, learning the best action in a particular state is not transferable to a new game since the state space is different and there is no single mapping between the state spaces of the initial and latter games. These rules would therefore need to be learned anew in the latter game which is inconsistent with above chance performance in very early rounds.

Likewise, assuming that players learn a complete model of the environment (for instance the transition probabilities from last round play to new play) might explain learning within games but is equally unable to account for early games transfer of learning as such models, besides being cognitively very expensive to learn, would require many rounds of practice. Another issue with these hypotheses is that they are not consistent with significant score differences between those facing level-2 and level-1 opponents. More specifically, if we assume that participants were using some type of associative learning or relying on spatial heuristics, then their scores should not depend on the degree of strategic sophistication of the opponent since their approaches would render this variable irrelevant. To be sure, if a participant learns to pick say "scissors" whenever the opponent last picked "rock", then the degree of strategic sophistication of the opponent (its level k) should not impact this learning, and we would expect in this case there would be no difference between scores when facing level-1 and level-2 opponents, which is not the case here. The fact that the degree of sophistication of the opponent matters points to the importance of opponent modelling to successful transfer of learning.

We are left with two possible explanations: First, it is possible that the players have uncovered a heuristic that allows them to beat the opponent without explicitly modelling their strategy, and is robust to transfer. Indeed, because of the cyclicality in action choices (e.g : Rock beats Scissors beats Paper beats Rock), it is possible to beat level-2 opponents most of the time by following a simple rule: Play in the next round whatever the opponent played in the last round. This is a rule that wins and is also robust to transfer as it does not depend on action labels and even works in the dissimilar game.

The second explanation of learning transfer is that it is driven by a group of participants that are able to build a mental representation of what the strategy of the opponent is. A successful mental representation would take the perspective of the opponent or endow it with intentionality in order to detect its strategy when the opponent is playing based on a level-k reasoning model. For instance, the player may think “My opponent is always trying to be one step ahead of me, therefore, I will be one step ahead of where it thinks I will be”. This mental representation would facilitate the use of theory of mind abilities and thus enable the players to learn opponent strategies when they are based on human-like reasoning models such as level-k or cognitive hierarchy. This type of learning would be deemed “explicit” in the psychology literature as a process through which knowledge consists of cognitive representations of concepts and rules, as well as the relationship between them. It involves the evaluation of explicit hypotheses and results in better problem-solving skills (Mandler, 2004). Since it is less context dependent, this type of learning is generalizable to new situations, akin to the more general framework of rule-based learning explored by Stahl (2000, 2003).

Our second experimental design allows us to test whether the first explanation holds. Since there is a simple transferable heuristic that works against level-2 players, and since as far as we know, there are no similar ones against level-1 players, if indeed participants were using this, they would perform better and transfer learning more easily when facing level-2 opponents. Because level-2 opponents use a higher level of strategic reasoning, they should in fact be harder to play against and in the absence of such a heuristic, performance and learning transfer should be worse.

Our results show that in fact, it was harder to transfer learning when facing level-2 opponents, both comparing first interactions across games and using early rounds analysis. Based on our assumptions, we conclude therefore that the most likely explanation is that participants who are able to beat the opponent and transfer learning are likely to be explicitly modelling the opponent strategy using level-k reasoning, compared to using simple learning rules they uncovered during the course of learning.

Our computational modelling allowed us to delve deeper into what might be driving the observed learning transfer. Initial modelling of observations using all available data seems to indicate that the most likely model was a Q-learning type model. However, as we argued above, that would be inconsistent with the evidence for learning transfer. Breaking down likelihoods by trial and fitting a hidden Markov model to the data with states being the various strategies that participants are assumed to be using, we showed evidence for within game switching of strategies. Participants start the early rounds of a new game acting in a way consistent with a Bayesian Bayesian Cognitive Hierarchy level, which would be accurate and generalisable but computationally expensive. However, as trials continue, participants seem to switch to a habitual type of learning (QL-models).

Why is this switching happening? We believe that participants show flexibility in their use of learning strategies. When a new game is started that is similar to a previously played game with the same opponent, participants need a way to transfer prior knowledge of the opponent and apply it to the new game in order to best respond. Adopting a Bayesian model based on BCH achieves the goal of transferring the opponent model and thus coming up with best responses in the early trials. However, Bayesian BCH models are computationally expensive and require higher order thinking (I think that you think that I think…). As such, as the games progresses, they may become burdensome and the higher amount of historical interaction in the new game allows participants to have enough data to start using the cognitively cheaper model-free learning strategies such as Q-learning. The preference for less computationally demanding strategies is well established (Wouter Kool, Joseph T. McGuire, Zev B. Rosen, Matthew M. Bovinick, 2011). Moreover, the ability to flexibly switch is also consistent with evidence from the literature on learning strategies in humans, showing that they indeed shift between model-based and model-free learning when the environment requires it (Simon & Daw, 2011).

Conclusion

Our online experiments results are consistent with behavioural game theory findings, in that human players can deviate from Nash equilibrium play and learn to adapt to the opponent strategy and exploit it when the opponent itself is deviating from Nash equilibrium. Moreover, we showed that participants transfer their learning to new games with varying degrees of similarity. The transfer is also moderated by the level of sophistication of the opponent, with participants showing more success in learning and transferring against opponents adopting a less sophisticated strategy.

Having said that, there remains a high degree of heterogeneity between players. There is a high positive association between players who learn to beat the sophisticated and less sophisticated opponents, indicating that some players are more able to detect the patterns in opponent play and learn how to exploit them. Moreover, the computational modelling shows that it is likely that players start each game using a model-based learning strategy that facilitates generalisation and opponent model transfer, but then switch to behaviour that is consistent with a model-free learning strategy as the experiment goes on. This is likely driven by a trade-off between computational complexity and accuracy between model based and model free strategies.

References

Batzilis, D., Jaffe, S., Levitt, S., List, J. A., & Picel, J. (2016). *How facebook can deepen our understanding of behavior in strategic settings: Evidence from a million rock-paper-scissors games.* working paper.

Camerer, C., Ho, T.-H., & Others. (1997). *Experience-weighted attraction learning in games: A unifying approach.*

Camerer, C., & Knez, M. (2000). *Increasing Cooperation in Prisoner's Dilemmas by Establishing a Precedent of Efficiency in Coordination Games.*

Goodie, A. S., Doshi, P., & Young, D. L. (2012). Levels of theory-of-mind reasoning in competitive games. *Journal of Behavioral Decision Making*, *25*(1), 95–108.

Ho, T. H., Camerer, C. F., & Chong, J.-K. (2004). *The economics of learning models: A self-tuning theory of learning in games.*

Ho, T.-H., Camerer, C., & Weigelt, K. (1998). Iterated dominance and iterated best response in experimental" p-beauty contests". *The American Economic Review*, *88*(4), 947–969.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*. https://doi.org/10.1017/S0140525X16001837

Mertens, J.-F. (1990). Repeated games. In *Game theory and applications* (pp. 77–130). Elsevier.

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

Simon, D. A., & Daw, N. D. (2011). Environmental statistics and the trade-off between model-based and TD learning in humans. *Advances in Neural Information*

Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011, 1–9.

Spiliopoulos, L. (2013). Strategic adaptation of humans playing computer algorithms in a repeated constant-sum game. *Autonomous Agents and Multi-Agent Systems*, *27*(1), 131–160.

Stahl, D. O., & Wilson, P. W. (1995). On players models of other players: Theory and experimental evidence. *Games and Economic Behavior*, *10*(1), 218–254.

Wang, Z., Xu, B., & Zhou, H.-J. (2014). Social cycling and conditional responses in the rock-paper-scissors game. *Scientific Reports*, *4*(1), 1–7.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3-4), 279–292.

Wouter Kool, Joseph T. McGuire, Zev B. Rosen, Matthew M. Bovinick. (2011). Decision Making and the Avoidance of Cognitive Demand. *Experimental Psychology*. https://doi.org/10.2996/kmj/1138846322

Batzilis, D., Jaffe, S., Levitt, S., List, J. A., & Picel, J. (2016). *How facebook can deepen our understanding of behavior in strategic settings: Evidence from a million rock-paper-scissors games.* working paper.

Camerer, C., Ho, T.-H., & Others. (1997). *Experience-weighted attraction learning in games: A unifying approach.*

Camerer, C., & Knez, M. (2000). *Increasing Cooperation in Prisoner's Dilemmas by Establishing a Precedent of Efficiency in Coordination Games.*

Goodie, A. S., Doshi, P., & Young, D. L. (2012). Levels of theory-of-mind reasoning in competitive games. *Journal of Behavioral Decision Making*, *25*(1), 95–108.

Ho, T. H., Camerer, C. F., & Chong, J.-K. (2004). *The economics of learning models: A self-tuning theory of learning in games.*

Ho, T.-H., Camerer, C., & Weigelt, K. (1998). Iterated dominance and iterated best response in experimental" p-beauty contests". *The American Economic Review*, *88*(4), 947–969.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*. https://doi.org/10.1017/S0140525X16001837

Mertens, J.-F. (1990). Repeated games. In *Game theory and applications* (pp. 77–130). Elsevier.

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

Simon, D. A., & Daw, N. D. (2011). Environmental statistics and the trade-off between model-based and TD learning in humans. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, 1–9.

Spiliopoulos, L. (2013). Strategic adaptation of humans playing computer algorithms in a repeated constant-sum game. *Autonomous Agents and Multi-Agent Systems*, *27*(1), 131–160.

Stahl, D. O., & Wilson, P. W. (1995). On players models of other players: Theory and experimental evidence. *Games and Economic Behavior*, *10*(1), 218–254.

Wang, Z., Xu, B., & Zhou, H.-J. (2014). Social cycling and conditional responses in the rock-paper-scissors game. *Scientific Reports*, *4*(1), 1–7.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3-4), 279–292.

Wouter Kool, Joseph T. McGuire, Zev B. Rosen, Matthew M. Bovinick. (2011). Decision Making and the Avoidance of Cognitive Demand. *Experimental Psychology.* https://doi.org/10.2996/kmj/1138846322