

## Transfer of Learned Opponent Models in Zero Sum Games

Ismail Guennouni<sup>1</sup> & Maarten Speekenbrink<sup>1</sup>

<sup>1</sup> Department of Experimental Psychology, University College London

## 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 explore opponent modelling and its transfer with the use of computer agents with human-like limited degrees of iterated reasoning. In two experiments, we find that participants deviate from Nash equilibrium play and learn to adapt to the opponent's strategy and exploit it. Moreover, we show that participants transfer their learning to new games and that this transfer is moderated by the level of sophistication of the opponent. Computational modelling shows that players start each game with a model-based learning strategy that facilitates between-game transfer of the opponent's strategy, 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).

A model of the opponent can help increase performance in a particular game, but learning more general characteristics of an opponent may also help increase performance in other games. In this paper, we are specifically interested in the latter: How do people build and use models of their opponent to facilitate learning transfer? 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 useful to address this question. The early literature on learning transfer in repeated games has mostly focused on the proportion of people who play normatively optimal (e.g. Nash equilibrium play) or use salient actions (e.g. risk dominance) in later games, having had experience with a similar game environment previously (Camerer & Knez, 2000; Teck-Hua Ho, Camerer, & Weigelt, 1998). As is well-known, a Nash equilibrium means that all players in a game play such that no-one can unilaterally improve their performance by deviating from their strategy. When playing against an opponent with a Nash-optimal strategy, I can do no better than play according to the Nash-equilibrium strategy as well. However, when faced with a player who deviates from the Nash-optimal strategy, it may be better for me to exploit this and also deviate from this strategy, increasing my performance beyond what is expected at a Nash

equilibrium. Of course, this comes with some risk, as my own deviation from Nash-optimal play may leave myself open to similar exploitation.

Studies that focused on whether people can learn to exploit deviations from Nash-equilibrium play have mostly looked at the ability of players to detect and exploit action-based learning rules (Shachat & Swarthout, 2004; Spiliopoulos, 2013). These studies used computer opponents with fixed strategies (no adaptation to their human opponent) consisting of playing each action with a fixed probability (a mixed strategy). Findings showed . . . . However, the use of mixed strategies may have limited people’s ability to form accurate opponent models. Rather than thinking of their opponents and drawing a random action from a (non-uniform) probability distribution, people may rather think of their opponents as applying a form of iterated reasoning to determine their next action. The type of reasoning we refer to takes the form of “I believe that you believe that I believe . . . .” and has been shown to correspond to people’s casual introspection (REFERENCE). It also underlies successful models in behavioural economics, such as Level- $k$  and Cognitive Hierarchy models, which posit that people assume their opponent has a limited level of iterative reasoning, with the player best responding to their (modelled) opponent.

In Level- $k$  theory, a level-0 player uses a fixed strategy without explicitly considering the strategy of their opponent. A level-1 player assumes their opponent is a level-0 player, and chooses actions to best respond to the strategy of their opponent, without considering what their opponent might believe that 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 best respond to the actions of that player. Cognitive Hierarchy theory works on similar principles, but rather than assuming an opponent always adopts a particular level- $k$  strategy, they assume their opponent adopts each of the level- $k$  strategies with a particular probability (i.e., the strategy is a mixture over pure level- $k$  strategies).

Iterative reasoning strategies can explain non-equilibrium play in a range of games, such as . . . In the present study, we endow our computer opponents with limited ability for iterative reasoning and assess whether (1) human players adapt their strategy to exploit this limited reasoning of their opponent, and (2) whether they are able to generalize a learned opponent model to other games. In two experiments, participants face the same opponent (Experiment 1) or same two opponents (Experiment 2) in three consecutive games: the well-known Rock-Paper-Scissors game, a structurally similar Fire-Water-Grass game, and a less similar Numbers (Experiment 1) or Shootout (Experiment 2) game. To foreshadow our results, we find evidence that participants transfer the learned strategy of their opponent to other games, but that this transfer is moderated by the sophistication of the opponent and the similarity of the games. Moreover, using computational modelling, we find evidence that participants switch from relying on an opponent model in the early stages of the games, to a more habitual and cognitively less demanding strategy in the later stages of the games.

## Experiment 1

In the first experiment, we . . . [intro to first exp]

## 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 name 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 level-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”, is not generalizable to this similar 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 randomly play each action with equal probability.

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, or choosing randomly from all actions. As a best response to a uniformly 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. Note that at all levels, the strategies are contingent on the actions taken in the previous round. The choice of this type of strategy is consistent with 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).

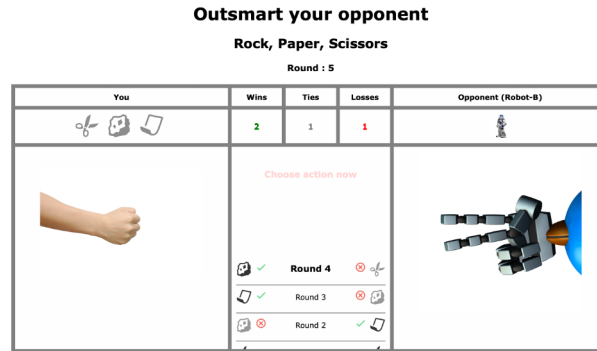
Human action $t - 1$	Computer action $t - 1$	Computer level-1 action $t$	Computer level-2 action $t$
Paper	Rock	Scissors	Scissors
Scissors	Scissors	Rock	Paper
Rock	Paper	Paper	Rock
...	...	...	...

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 with them, without prior 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. An example of the interface for the RPS game is provided in Figure 1. On each round, the human player chooses an action, and after a random delay, is shown the action chosen by the computer opponent, and the outcome of that round. A history of the chosen actions over all rounds is available in the centre panel. 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. After playing all the games, participants were asked questions about their beliefs about the computer opponent, related to whether they thought they learned their opponent's strategy, and how difficult they found playing against their particular opponent. They were then debriefed and thanked for their participation.



*Figure 1.* Screenshot of the Rock-Paper-Scissors game in Experiment 2. Shown here is the feedback stage, after both the human (left) and computer (right) players have chosen their action. The interface was similar in Experiment 1, but excluded the history of game play in the center panel.

**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



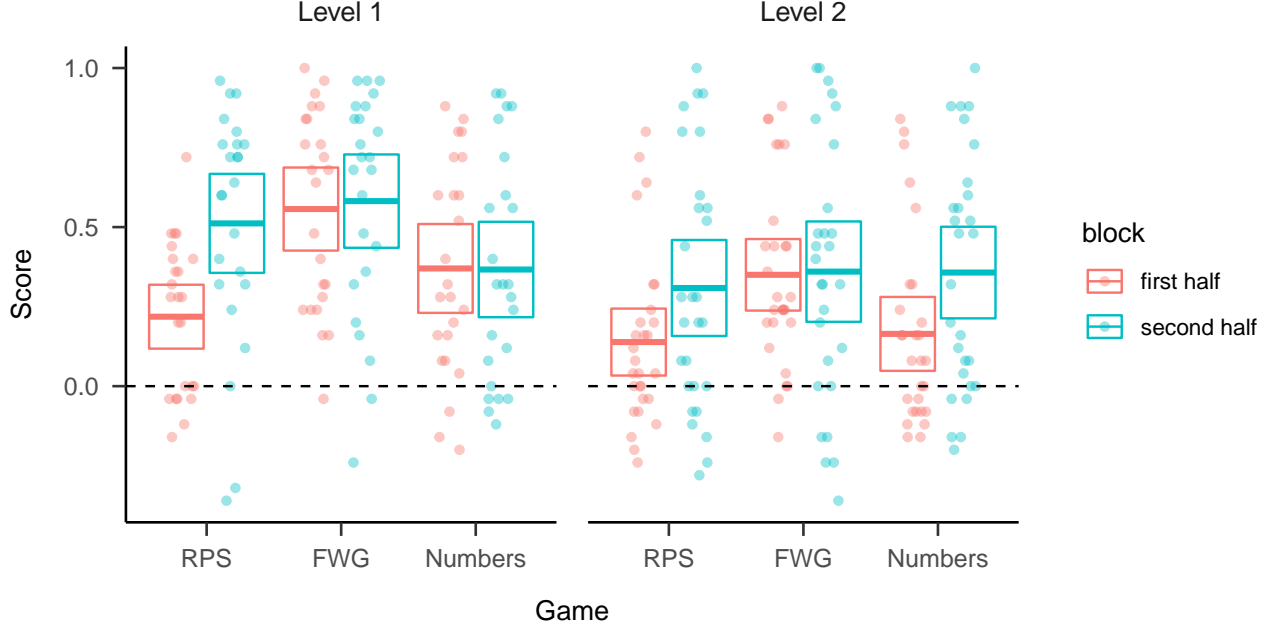


Figure 2. Performance per game and block across conditions in Experiment 1. Points are scores of individual participants and boxes reflect the 95% confidence intervals of the mean (center line equals the mean).

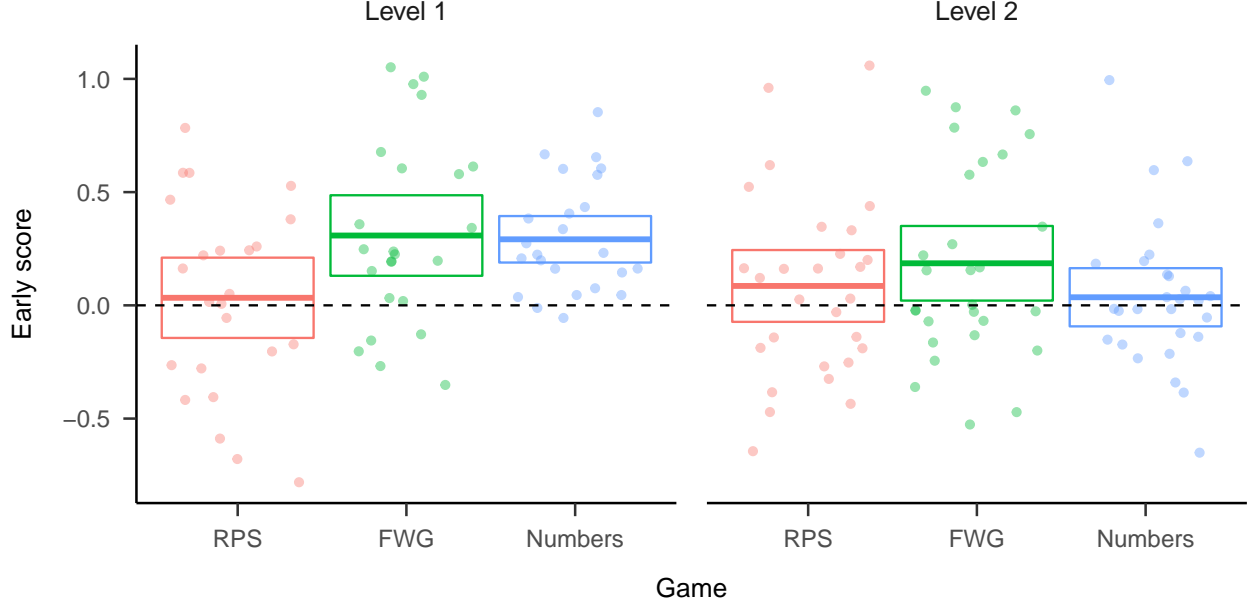
were significantly different from 0, the expected score of uniformly random play (RPS:  $t(51) = 7.26$ ,  $p < .001$ ; FWG:  $t(51) = 10.04$ ,  $p < .001$ ; Numbers:  $t(51) = 7.17$ ,  $p < .001$ ). As uniformly random play is the Nash equilibrium, this indicates successful deviation from a Nash-optimal strategy.

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 level-2 player ( $M = 0.27$ ). Thus, it appears that it was harder for participants to exploit the strategy of the more sophisticated level-2 opponent than the comparatively less sophisticated level-1 opponent. **MS: Need to check test results and report GG degrees of freedom, etc.**

**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 3). 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 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 were indeed not significantly different from 0 ( $t(148.85) = 1.04, p = .89$ ).

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 opponents. 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 ( $F(1.84, 91.77) = 3.35, \eta^2 = 0.04, p = .043$ , Greenhouse-Geisser correction to degrees of freedom applied to correct for non-sphericity). We then run statistical tests on early round



*Figure 3.* Performance in early rounds (2-6) per game and block across conditions in Experiment 1. Points are scores of individual participants and boxes reflect the 95% confidence intervals of the mean (center line equals the mean).

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.

**Discussion.** The results of Experiment 1 indicate that participants were able to learn successful strategies which exploited the deviation from Nash-optimal play of their opponents. Moreover, they were able to transfer knowledge about their opponent to other games, but this was moderated by the type of game and opponent. There was evidence of transfer from the RPS game to the similar FWG game for both opponents. However, transfer to the dissimilar Numbers game was only evident for participants facing a level-1 opponent.

## Experiment 2

In Experiment 2, we aimed to obtain a stronger test of learning transfer. Instead of facing a single level-1 or level-2 opponent throughout all games, participants face both types of opponents sequentially in this experiment. To perform well against both opponents, participants would need to learn distinct strategies in each game. To reduce effects of increased memory load due to facing distinct opponents, we provided participants access to the history of play against an opponent within each game (see Figure 1). Finally, we changed the third game to a penalty shootout game, with 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 (aim left, center, or right), it is strategically dissimilar. Unlike the Numbers game in Experiment 1, the Shootout game does not have a cyclical hierarchy between actions, making it harder to win through a heuristic based on this cyclicity.

## Methods

**Participants & Design.** A total of 48 participants (21 females, 28 males, 1 unknown) were recruited via the Prolific platform, none of which took part in Experiment 1. The average age 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 (RPS), Fire-Water-Grass (FWG), and the penalty Shootout game. The first two games were identical to the ones used in the first experiment. In the final game (Shootout) participants took the role of the a football (soccer) player in a penalty situation, with the computer opponent taking the role of the goalkeeper. Players had the choice between three actions: shooting the football to the left, right or centre of the goal. Similarly, the goalkeeper chooses between defending the left, right, or centre of the goal. If participants shoot in a different direction than where the goalkeeper defends, 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 to the other games is that there are two ways to beat the opponent: if the shooter thinks their opponent is going to choose to defend ‘right’ in the next round, they can win by either choosing to shoot ‘left’ or ‘center’. A level-1 shooter who thinks that their goalkeeper opponent will repeat their last action has thus two possible best responses. A level-1 goalkeeper, however, has only a single best response (defending where their opponent aimed in the last round). A level-2 goalkeeper, who believes their opponent is a level-1 shooter, will have two best responses however. We programmed the level-2 computer player to choose randomly between these two best responses.



*Figure 4.* Screenshot of the shootout game

**Procedure.** 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 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 action and had to wait for the timer to run out. A small random delay between 0.5 and 3 seconds was also introduced before the computer agent made their choice, as a way of simulating a real human opponent thinking time. After each round, participants were given detailed feedback about their opponent’s action and whether they won or lost the round. Further information about the history of play in previous rounds was also provided and participants could scroll down to recall the full history of each interaction against an opponent in a particular stage of a game. 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 mixed-strategy Nash equilibrium consisting of uniformly random 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 their opponents’ play.

## Results

Participants scores are depicted in Figure ???. 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 ( $M = 0.289$ ,  $SD = 0.326$ ).<sup>1</sup> Using parametric t-tests on adjusted scores, we reject the null hypothesis of random play in all three games (RPS:  $t(49) = 6.26$ ,  $p < .001$ ; FWG:  $t(49) = 7.25$ ,  $p < .001$ ; Shootout:  $t(49) = 13.61$ ,  $p < .001$ ).

To explore whether learning occurred within and between games, we performed 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 (Encounter: first or second) repeated-measures ANOVA on the average score per round, with the first factor varying between participants. This shows a main effect of Game ( $F(1.85, 88.7) = 11.81$ ,  $\eta^2 = 0.04$ ,  $p < .001$ ). Post-hoc pairwise comparisons between games (p-values adjusted using Holm method for multiple comparisons) indicate performance in the games increases steadily throughout the experiment, with FWG performance significantly higher than RPS ( $t(96) = 2.53$ ,  $p = .025$ ), and performance in the Shootout game significantly higher than in FWG ( $t(96) = 2.32$ ,  $p = .025$ ). Whilst the ANOVA shows no main effect of Condition, Opponent type, or Encounter, there was a significant interaction between Game and Opponent type ( $F(1.7, 81.82) = 5.31$ ,  $\eta^2 = 0.02$ ,  $p = .01$ ). Follow-up analysis shows that when facing level-1 agents, scores increase steadily after each game, with FWG score significantly higher than RPS ( $t(191) = 2.70$ ,  $p = .03$ ) and Shootout scores in turn significantly higher than FWG ( $t(191) = 3.05$ ,  $p = .01$ ). 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

---

<sup>1</sup> 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)

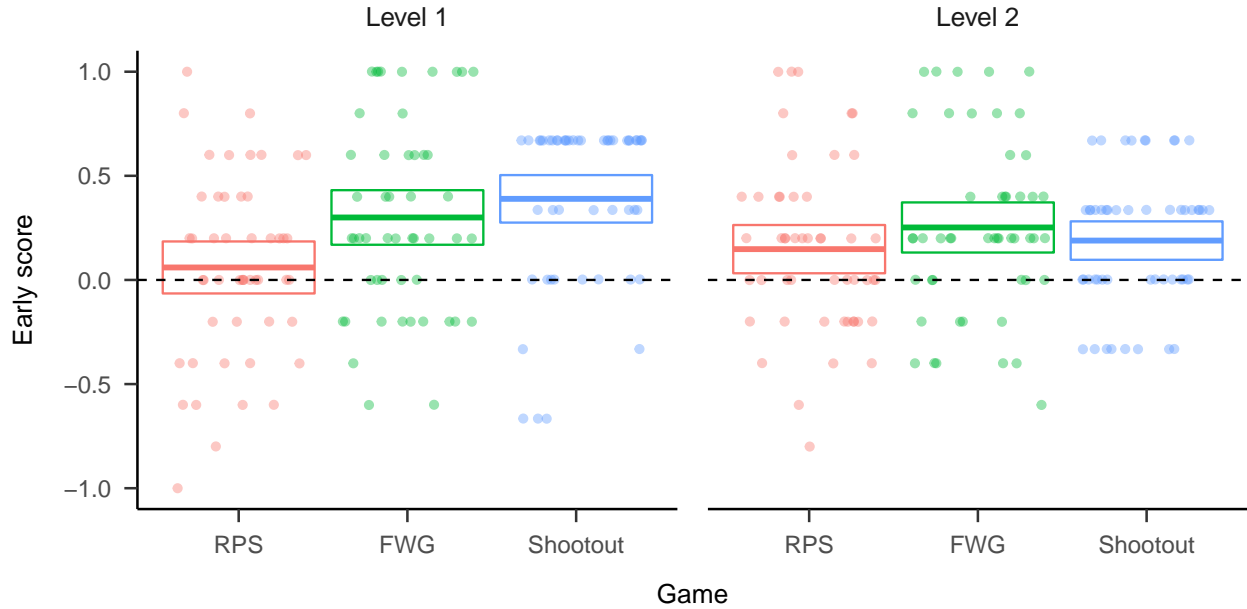


Figure 5. Performance in early rounds (2-6) per game and opponent in Experiment 2. Points are scores of individual participants and boxes reflect the 95% confidence intervals of the mean (center line equals the mean).

random (Figure ??). 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:  $t(270) = 4.99, p < .001$ ; Shootout:  $t(270) = 6.66, p < .001$ ; Level-2 opponent: FWG:  $t(270) = 4.40, p < .001$ ; Shootout:  $t(270) = 3.21, p = .004$ ). **MS: Should also report ANOVA, as for Experiment 1**

**Discussion.** The results of Experiment 2... **MS: give summary of main findings, as done for Experiment 1. You should not report new statistical tests in the discussion, which should go in the results section before.**

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) model, which is a popular model in behavioural economics.

In the following, we will describe the models in more detail, and provide some intuition into how they learn about the game and/or the opponent. We use the following notation: In each game  $g \in \{\text{RPS}, \text{FWG}, \text{Numbers}, \text{Shootout}\}$ , 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.

### 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:

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

where  $Q(s_t, a_t)$  is the value of taking action  $a$  when in state  $s$  at time  $t$ , and  $\alpha \in [0, 1]$  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:

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

where the inverse temperature parameter  $\lambda$  determines the consistency of the strategy (the higher  $\lambda$ , the more often the action with the highest  $Q$ -value is chosen. 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 state  $\mathcal{S}_g$  and action space  $\mathcal{A}_g$ , and there is no simple way to map states and actions between games.

The RL model has two free parameters: the learning rate ( $\alpha$ ) and the inverse temperature ( $\lambda$ ).

### Experience-weighted attraction (EWA) model

The self-tuning Experience Weighted Attraction (EWA) model (Teck H. 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 if 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 (Cheung & Friedman, 1994). While the assumption in expanding the update rule above is that  $\phi$  and  $\delta$  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 (Teck H. Ho, Camerer, & Chong, 2004).  $N(t)$  represents an experience weight 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. **MS: Should describe how the parameters are self-tuned with equations**

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 ( $\lambda$ ). **MS: Are there no states in the EWA? If not, it is unsurprising that this model doesn’t fit well**

## 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:

$$P(\text{level} = k | \mathcal{D}_t) \propto P(\mathcal{D}_t | \text{level} = k) \times P(\text{level} = k)$$

where  $\mathcal{D}_t = \{s_1, \dots, s_t\}$  is the data available at time  $t$ . The likelihood is defined as

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

where  $f_k(o_t | s_t) = 1$  if  $o_t$  is the action taken by a level  $k$  player when the previous round play was  $s_t = (a_{t-1}, o_{t-1})$ , 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:

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

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

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. This generalization is done simply by using the posterior  $P(\text{level} = k | \mathcal{D}_T)$  from the previous game as the prior distribution in the next game. However, the participant might also assume that the level of reasoning of their opponent

does not generalize over games. This would mean starting with a “fresh” prior  $P(\text{level} = k)$  at the start of each game. 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 ( $\theta$ ), and the temperature parameter of the softmax function ( $\lambda$ ).

### Estimation and model comparison

For both experiments, we fitted all models to individual participant data by maximum likelihood estimation. We use the Bayesian Information Criterion (BIC) to determine the best fitting model for each participant.

For Experiment 1, we fitted a total of 5 models: a baseline model assuming random play (Nash), the Bayesian Cognitive Hierarchy model allowing transfer between games (BCH\_BT) and without 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 e.g. transferring a learned opponent model from RPS to FWG. Therefore, we fitted a total of three versions of the Bayesian Cognitive Hierarchy model: BCH\_BT allows for both within and between game transfer (between game transfer without within game transfer is implausible); BCH\_NBT allows for within

but not between game transfer; BCH\_NT allows for no transfer within or between games. For Reinforcement learning models, because RL models can't account for between game transfer due to change in state and action space, we can only have models that 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 with random play (Nash) we therefore fit a total of 8 models for Experiment 2.

Figure 6 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 7. RL models have high BIC weights when they best fit the participants, and very few instances have high BIC weights for models other than RL, which fits the picture drawn by the histogram.

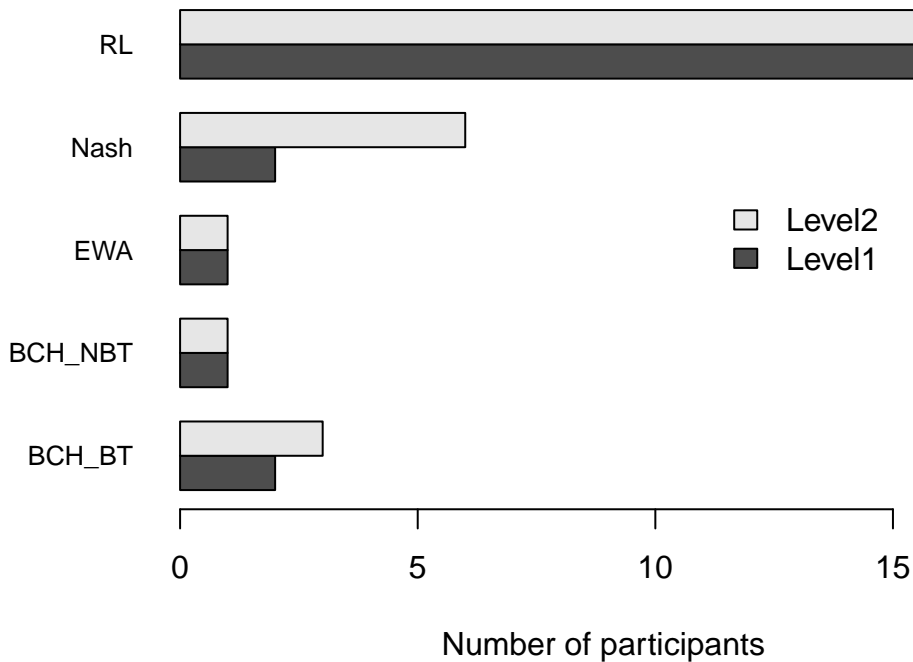


Figure 6. Experiment 1 - Histogram of best fitting computational models by condition

In Experiment 2, we can see from Figure 8 that the RL model was again more

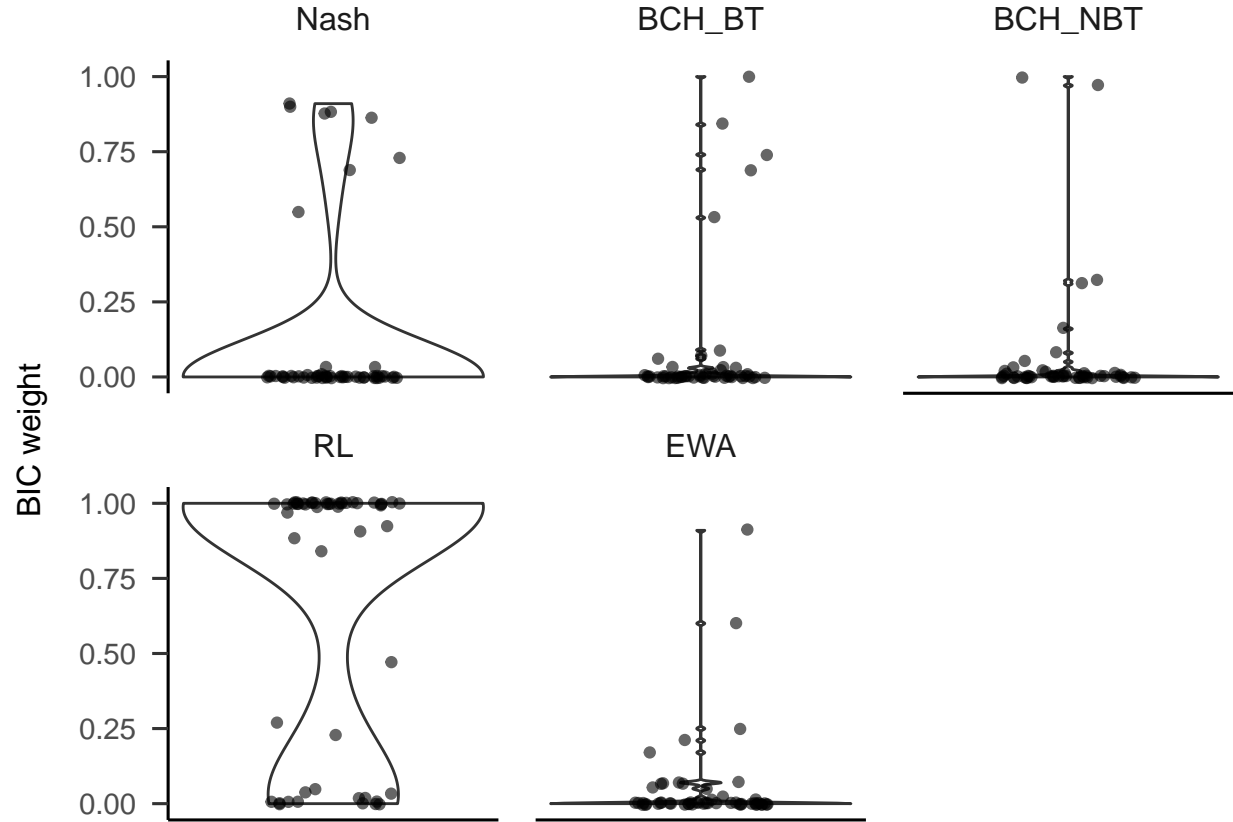


Figure 7. Model BIC weights for participants in Experiment 1.

successful than the Bayesian Cognitive Hierarchy models (with or without transfer) in fitting participants' action choices. This is also reflected in BIC weights in Figure 9. **MS: model labels in Figures should correspond to those in the text**

### Using Hidden Markov Model to explore strategy switching

The computational modelling indicates that most players are best fit by a reinforcement learning which learns good actions conditional upon the last round play. This is at odds with the behavioural findings, where we found evidence of transfer in early rounds of each game. If indeed most participants adopt a reinforcement learning strategy, they should not be able to transfer their learning to the early rounds of a new game. In order to investigate this discrepancy further, we plot the likelihood by trial for each game and three strategies: RL, Bayesian Cognitive Hierarchy models with between-game transfer, and the

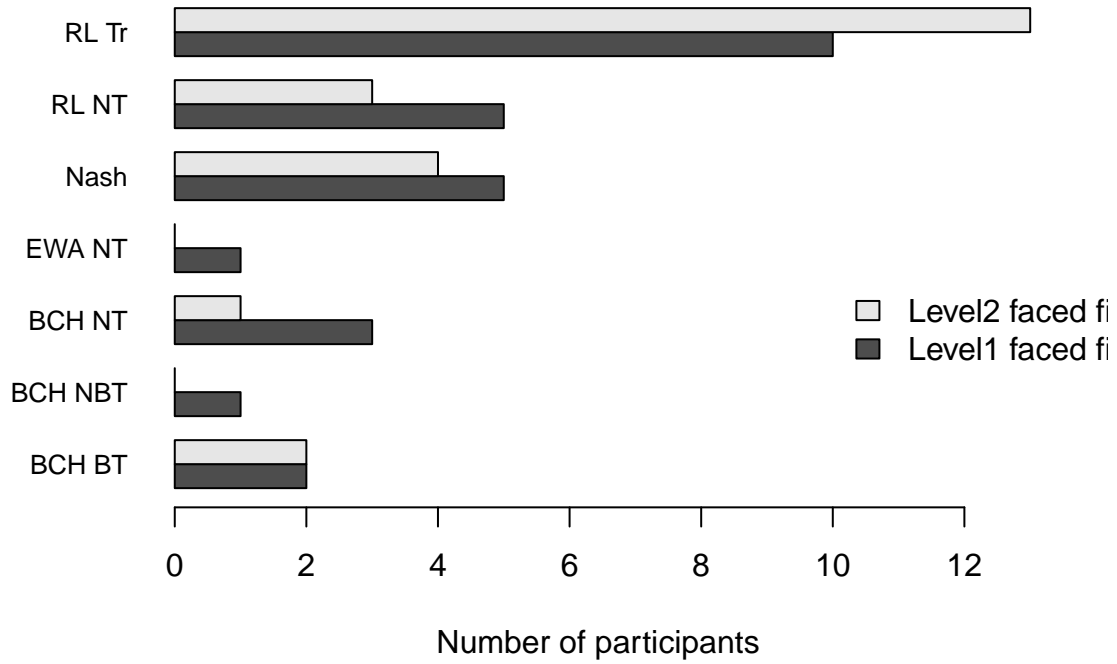


Figure 8. Experiment 2 Histogram of best fitting computational models by condition

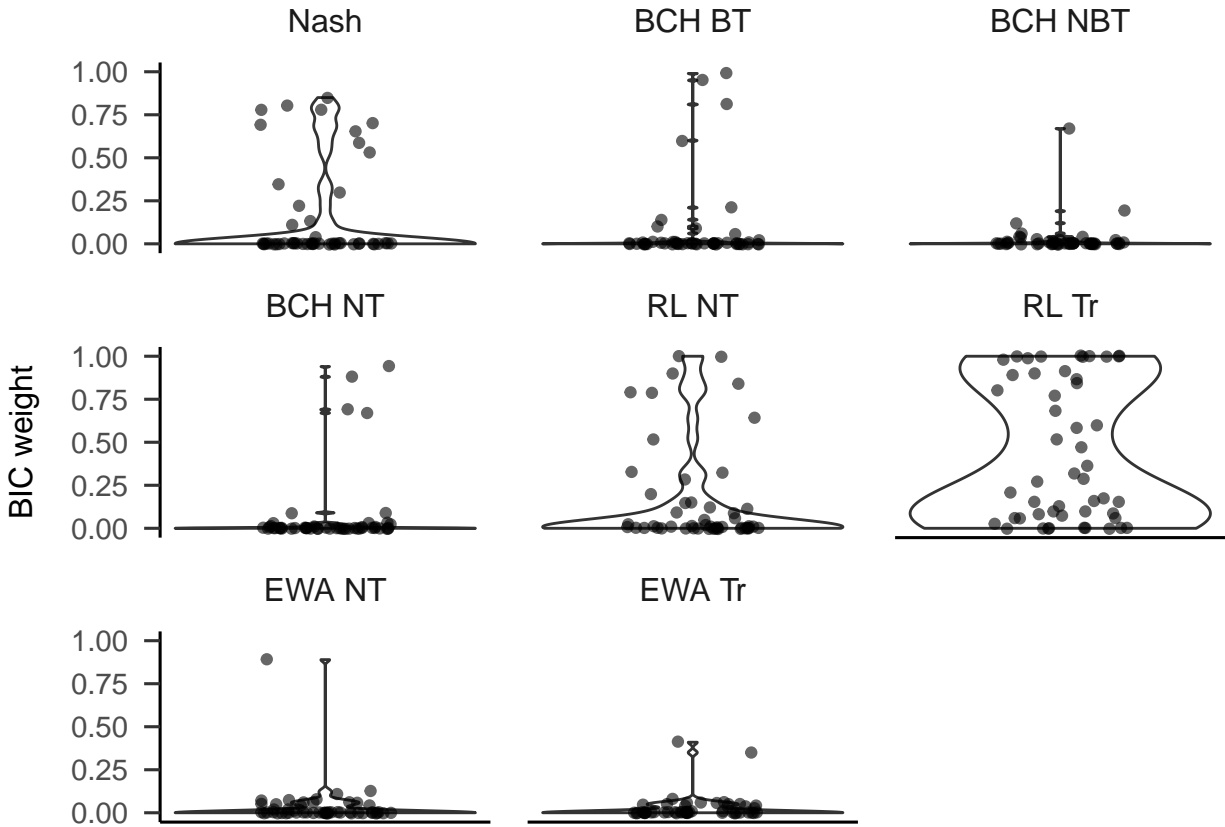


Figure 9. BIC weights for each model and participant in Experiment 2.



random (Nash) strategy. Figure 10 shows that in the later games of Experiment 1, the likelihood for the BCH models is higher than that of the other models in the initial rounds of the games. However, over time, the likelihood of the RL model increases and exceeds that of BCH model.

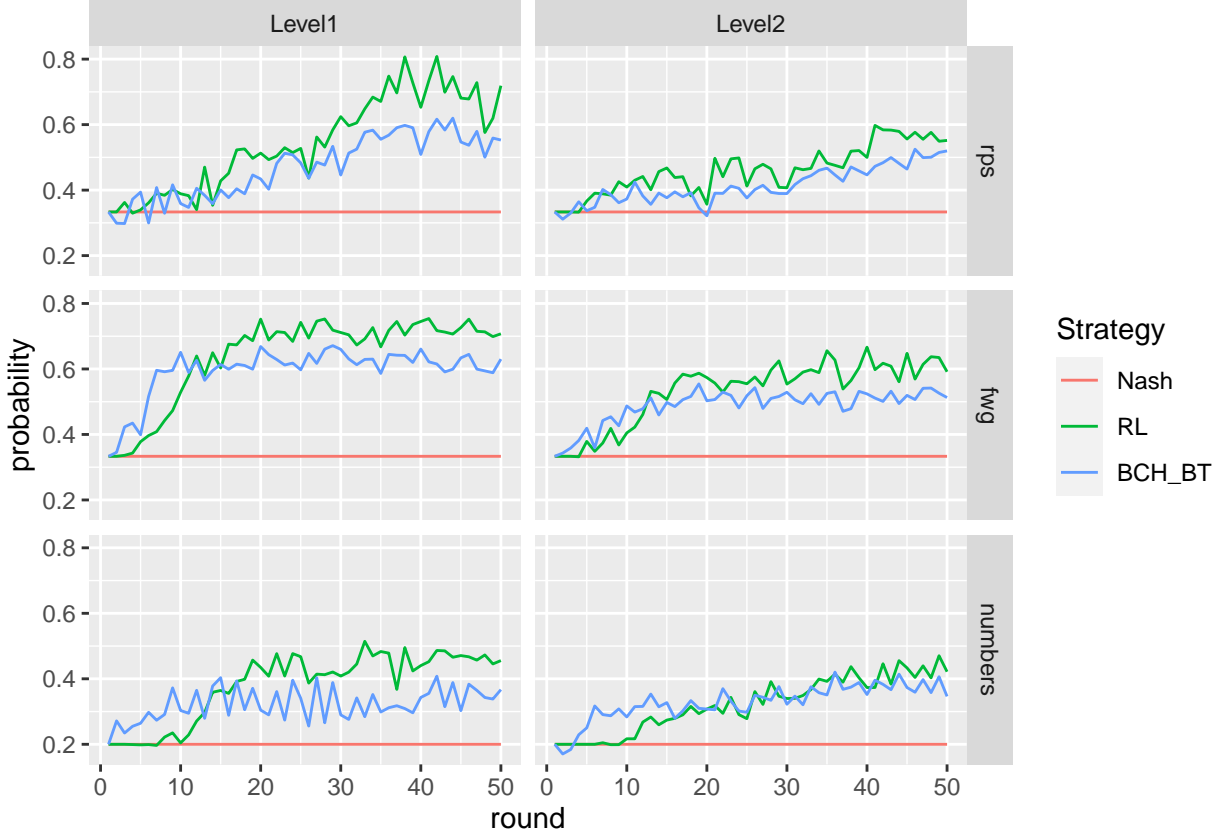


Figure 10. Experiment 1 Likelihood by trial by game and opponent faced

The same pattern holds for Experiment 2 (Figure 11). Again, the BCH model with between-game transfer has the highest likelihood in the early stages of the later games (apart from stage 1 of the shootout game, where the RL model is better). In later rounds, the likelihood of the RL model exceeds that of the BCH model.

The fact that the likelihoods of the main strategies considered cross over in both experiments could be interpreted as indicative that participants switch between strategies as the games progress. According to this interpretation, participants base their responses in

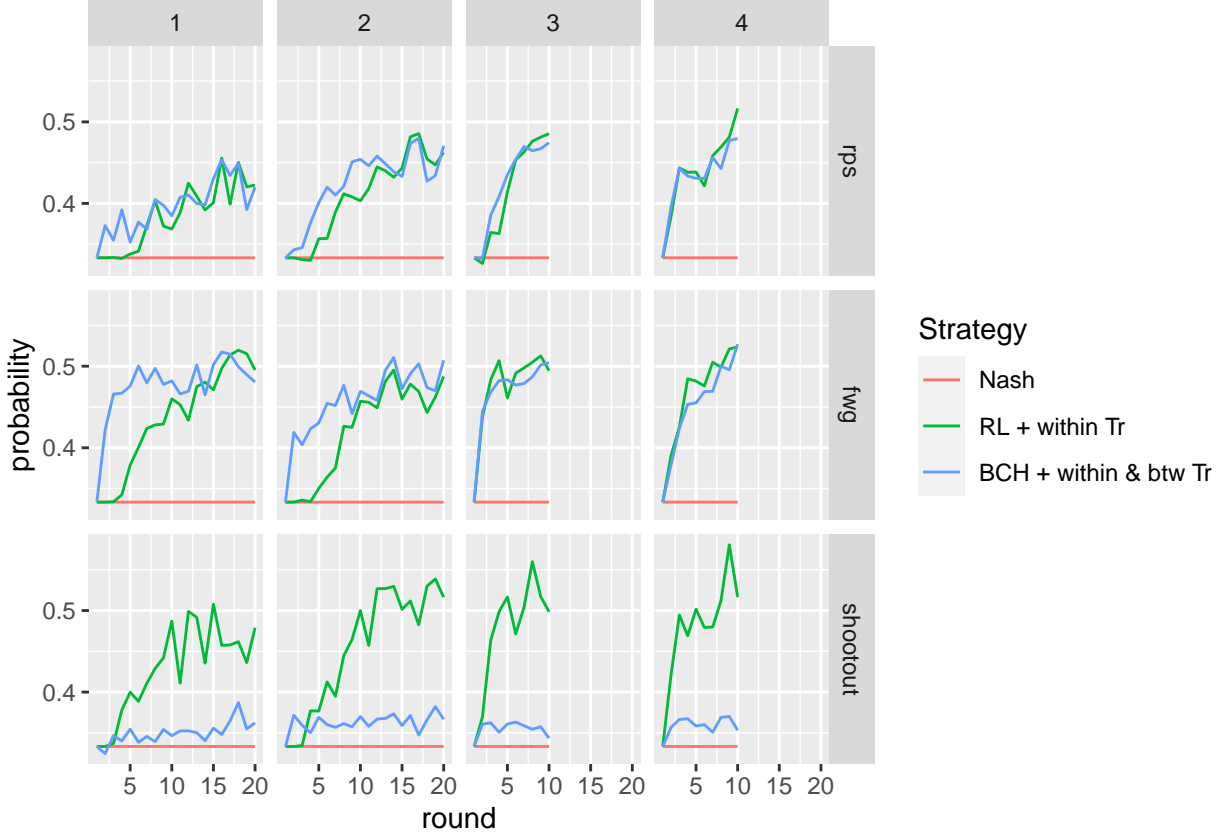


Figure 11. Experiment2 likelihood by trial by game and opponent faced

early rounds on the learned level of their opponents iterative reasoning, switching later to learned actions through reinforcement.

We use hidden Markov models to more formally test for strategy switching in participants' play. In these models, the three strategies (RL, BCH with between-game transfer, and Nash) correspond to latent states which determine the overt responses (actions chosen). The models allow for switching between the states over time, and such switches correspond to strategy switches. Hidden Markov models assume that an observable action at time  $t$  depends on a latent state at time  $t$ . Second, it is assumed that the latent state at time  $t$  depends on the latent state at the previous time  $t - 1$ . The model is specified by the state-conditional action distributions (these are provided by the likelihood of the fitted models), an initial state distribution (the distribution over the strategies at the initial round),

and the state-transition probabilities (probability of switching from one state/strategy to another). Initial state probabilities and the transition probabilities were estimated with the depmixS4 package (Visser & Speekenbrink, 2010). As a statistical test of strategy switching, we compare the hidden Markov model to a constrained version which assumes the probability of switching from one strategy to a different one is 0. This model thus assumes that when players start with a particular strategy, they continue using it throughout the experiment.

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

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

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

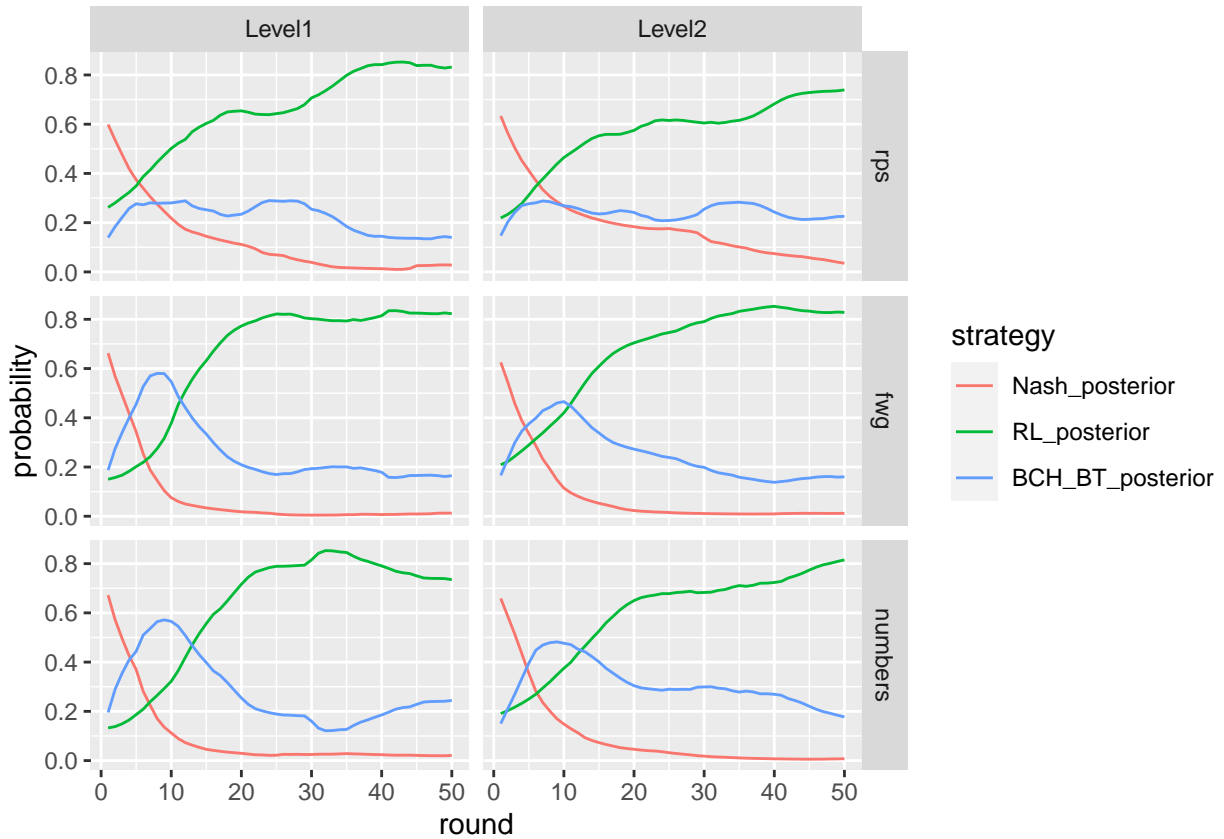


Figure 12. Experiment1 posterior probability of strategies by game and opponent faced

In Experiment 1, a likelihood-ratio test shows that the HMM model with switching fits

significantly better than the non-switching one ( $p < .001$ ). **MS: we should really use a bootstrapped Likelihood Ratio test here** This provides further statistical evidence in favour of the hypothesis that participants switch between strategies. Figure 12 depicts the average (across participants) posterior probabilities of each state (strategy), as a function of trial and opponent faced. As can be seen, there is evidence of strategy switching in the FWG and Numbers games: Initially, participants appear to use a random strategy (in the first round of a game, there is no way to predict the opponent's action), after which the BCH strategy becomes dominant. In the later rounds of the games, the RL strategy becomes dominant, however.

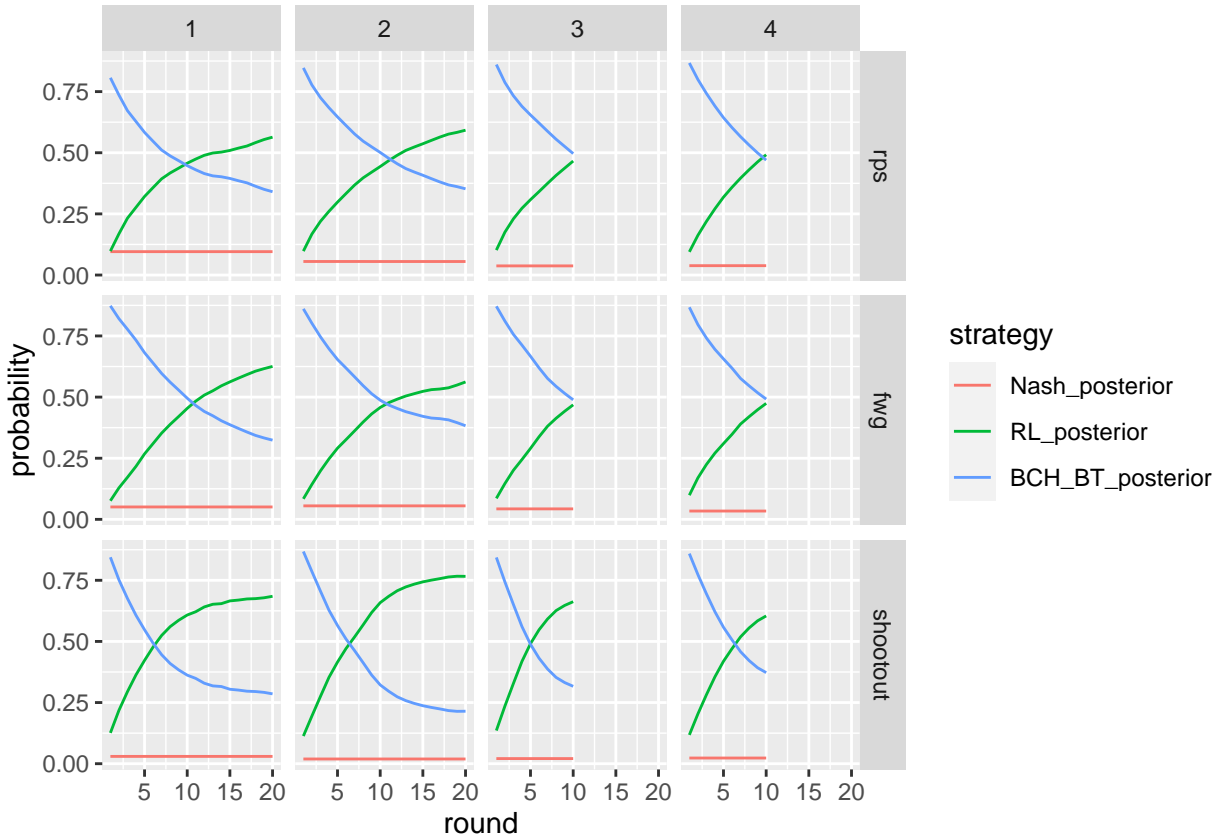


Figure 13. Experiment2 likelihood by trial by game and opponent faced

The switching model in Experiment 2 is also significantly better than the restricted non-switching model ( $p < .001$ ). The posterior probabilities of the strategies (Figure 13) show very clear evidence of strategy switching across games and stages, from using a BCH

model in the initial rounds to an RL strategy later on.

## Discussion

In this study, we investigated human learning transfer across games by making human participants play against computer agents with limited levels of iterated reasoning. We were interested 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 a similar game. Performance in early rounds indicated that transfer to the more dissimilar game was moderated by the degree of sophistication of the opponent, with evidence for transfer when players face the less sophisticated agent but not the more sophisticated one. In the second experiment, participants faced both types of opponents, which allows for a stronger test of opponent modelling, as participants would need to learn a different strategy for each opponent within a game. In Experiment 1, participants could learn a single strategy for each game, making opponent modelling possibly less pertinent. Experiment 2 also offers more opportunities to test transfer of a learnt opponent model. There were two opportunities to transfer opponent models within each game as well as two opportunities to transfer from each game to the next, giving a total of 6 transfer opportunities. **MS: we did not test transfer at all these points, or did we?** The results on learning transfer confirmed the findings from the first experiment. Again, there was clear evidence of transfer in the early rounds of the later games. We also found that learning transfer is moderated by the type of opponent faced: Evidence of transfer was weaker for the level-2 opponent as compared to the level-1 opponent. That transfer was less evident for the more sophisticated level-2 opponent in both experiments may be due to a higher difficulty of learning that opponent’s strategy. If it is more difficult to establish a model of the level-2 opponent, there is likely less knowledge

to transfer to the new game. A player cannot transfer what they have not learnt.

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 their opponent’s strategy and what form did this learning take?

One possible answer is that participants learned simple rules based on last round play. For instance, “play scissors whenever my opponent played rock in last round,” or “play paper whenever the last round play was either rock or scissors.” These are the type of strategies that are learned by the model-free reinforcement learning we used in our computational modelling. While this strategy fitted participants’ actions the best overall, there are at least two reasons why this account is not satisfactory as a complete description of participants’ behaviour. Firstly, the learned strategies are not transferable to new games. There is no simple way to map “play scissors whenever my opponent played rock in last round” in the RPS game to “play grass whenever my opponent played fire in last round.” Such a mapping may be possible by translating the rules and structure from RPS to FWG, but model-free reinforcement learning lacks the tools to do this. Model-free reinforcement learning would need to start from scratch in each new game, yet we found evidence that participants could successfully exploit their opponent’s strategy in early rounds of new games. Secondly, a reinforcement learning strategy would fare equally well against the level-1 and level-2 opponent. Whilst choosing different actions, the contingency between the state (last round play) and actions is the same for both opponents. Yet, we found that participants performed better against the level-1 opponent compared to the level-2 opponent. The difference in performance between the two types of opponent indicate that the actions of the more sophisticated level-2 opponent, or the best response to these, were somehow more difficult to predict.

We are left with two possible explanations: First, it is possible that participants discovered a heuristic that allowed them to beat their opponent without explicitly modelling

their strategy, and that this heuristic is transferable to new games. Because of the cyclicity in action choices (e.g., rock beats scissors, scissors beats paper, paper beats rock), it is possible to beat a level-2 opponent 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 transferable to other games as it does not depend on action labels. In the same vein, a heuristic that beats a level-1 player can be stated as “Choose the action that would have been beaten by my previous action.” Intuitively, it seems that the heuristic for the level-2 player is simpler than that for the level-1 player, which is at odds with the difference in performance.

A second explanation is that participants engaged in iterative reasoning, inferring their opponent’s beliefs and countering the resulting actions. For instance, this would be reasoning of the form “My opponent expects me to repeat my last action, choosing an action that would beat my last action. Hence, I will choose the action that beats their best response” or “My opponent thinks I expect them to repeat their action, hence expecting me to choose the action that beats their last action. They will therefore choose the action that beats this one, and hence I should choose the action that beats their best response.” Beating a level-1 player, in this account, requires being a level-2 player, and beating a level-2 player requires being a level-3 player. Intuitively, the additional step of iterative reasoning involved in beating a level-2 player makes the level-3 strategy more demanding and difficult to implement, which is consistent with the lower performance against the level-2 opponent.

The differences in performance between the two players, coupled with the finding of positive transfer, point to participants engaging in iterative reasoning, and learning something useful about their opponent’s limitations in this regard. This is the type of learning encapsulated by our Bayesian Cognitive Hierarchy model. 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). **MS: need**

**to check relevance of these citations. Also, they need to be in reference list** We admit that our implementation in the BCH models does not predict a performance difference between the types of opponents. Starting with an equal prior belief over the different levels of sophistication, a BCH player would perform equally well against the level-1 and level-2 opponent. There are two routes to explain the difference in performance. Firstly, prior beliefs might be biased against higher-level opponents (i.e., participants might have believed it is more likely that they would face level-1 opponent than a level-2 opponent). Secondly, if the actions of a level-2 opponent are more difficult to predict than those of a level-1 opponent, this might introduce more noise in the likelihood of the opponents actions given their level of sophistication. Either of these mechanisms would explain why learning the strategy of the level-2 opponent is more difficult and slower than learning the strategy of the level-1 opponent. **MS: This could be something to implement, actually** Using hidden Markov models, we found evidence of strategy switching between the BCH and RL strategies, and such switching seems more consistent with the latter idea. If predicting an opponent's actions through iterative reasoning is cognitively demanding and error-prone, it is resource-rational to switch to less costly yet equally successful strategies when these are available. **MS: add reference to resource-rational approach by Lieder & Griffiths** Initially, a model-free reinforcement learning strategy will be less successful than an iterative reasoning one. However, given enough experience, it will be on par with an iterative reasoning strategy. As it involves simple state-action contingencies, a model-free RL strategy may also be computationally less costly, making it overall more effective to rely on this than iterative reasoning. This is similar to the arbitration between model-free and model-based RL (Simon & Daw, 2011) **MS: add more references here.** In repeated and well-practised situations, relying on habits allows one to save cognitive resources for other demands. However, when the environment – or game – changes, it is prudent to use all available resources to reorient oneself.



## Conclusion

Our results show that people can successfully deviate from Nash equilibrium play to exploit deviations from such play by their opponents. Moreover, people can transfer knowledge about the limitations of their opponents to new situations. This transfer of a model of the opponent depends on the similarity between the prior and new game, as well as the sophistication of the opponent. Transfer is better to similar games, and for less sophisticated agents. Within games, we found evidence for a switch from a more reasoning-based strategy which allows for between-game transfer, to a more habitual strategy which does not. This is consistent with a rational trade-off between the goal of maximising performance and minimizing the cost of computing the best possible actions.

## 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*.
- Cheung, Y.-W., & Friedman, D. (1994). *Learning in evolutionary games: some laboratory results*. University of California, Santa Cruz.
- Ho, Teck H., Camerer, C. F., & Chong, J.-K. (2004). *The economics of learning models: A self-tuning theory of learning in games*.
- Ho, Teck-Hua, 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.

Visser, I., & Speekenbrink, M. (2010). depmixS4: An R package for hidden markov models. *Journal of Statistical Software*, 36(7), 1–21. Retrieved from <http://www.jstatsoft.org/v36/i07/>

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.