Transfer of Learned Opponent Models in Zero Sum Games

Ismail Guennouni · Maarten Speekenbrink ·

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 in games where human agents play against 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 their opponent's strategy to exploit it. Moreover, we show that participants transfer their learning to new games and that this transfer is moderated by the similarity between games as well as the level of sophistication of the computer opponent. Computational modelling shows that players start each game with a model-based learning strategy that facilitates between-game transfer of their opponent's strategy, but then switch to behaviour that is consistent with a model-free learning strategy in the latter stages of the interaction.

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

Maarten Speekenbrink
Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom
E-mail: m.speekenbrink@ucl.ac.uk

1 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 et al. 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 and 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 their 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 (e.g. risk dominance) actions in later games, having
had experience with a similar game environment previously (Teck-Hua Ho,
Camerer, and Weigelt 1998; C. Camerer and Knez 2000). As is well-known,
a Nash equilibrium means that all players in a game act such that no-one
can unilaterally improve their performance by deviating from their strategy.
When playing against an opponent with a Nash-optimal strategy, you 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,
you may be able to exploit this and also deviate from this strategy, increasing
your performance beyond what is expected at Nash equilibrium. Of course,
this comes with some risk, as your own deviation from Nash-optimal play may
leave you 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 (Spiliopoulos 2013; Shachat and Swarthout
2004). These studies used computer opponents that don't adapt their strategy
to their human opponent, mostly consisting of playing each action with a fixed
probability (a mixed strategy) or using a pre-determined sequence. Findings
showed that humans are capable of adapting to non-Nash-equilibrium play and
detect patterns in an opponent's history of play. However, the use of mixed
strategies may have limited people's ability to form accurate opponent models.

Instead of thinking of their opponents as drawing random actions 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 ...." **For example, take the game of Rock-Paper-Scissors (RPS). If I believe that you believe I will play Rock, then I would expect you to play Paper, as this beats my Rock. I would therefore choose to play Scissors, as this beats your Paper. But if I believe you apply similar reasoning, I might expect you to believe I will play Scissors, and therefore I would play Rock. But then I might expect you to play Paper in return, and so forth, ad infinitum.**

Strategies based on such iterative reasoning can explain non-equilibrium play in a range of games (C. F. Camerer 2003), such as the p-beauty contest game (Nagel 1995), dominance solvable games (C. F. Camerer, Ho, and Chong 2004), and various standard normal form games (Costa-Gomes, Crawford, and Broseta 2001).** Iterative reasoning underlies successful models in behavioural economics, such as Level-k and Cognitive Hierarchy models (C. F. Camerer, Ho, and Chong 2004), which posit that people assume their opponent applies a limited level of iterative reasoning, taking actions that are best responses to those of 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 is based on similar principles, but rather than assuming an opponent always adopts a particular level-k strategy, they are assumed to adopt each of the level-k strategies with a particular probability (i.e., the opponent's strategy is a mixture over pure level-k strategies).

In repeated simultaneous-move games such as RPS, the starting-point of iterative reasoning is often what was played on the previous round (Dyson 2019). For example, opponents are expected to repeat previous-round actions, or choosing actions that beat previous-round actions (eyler2009winning?). This leads to sequential regularities in actions. The evidence for whether these patterns are exploited is mixed, however. Brockbank and Vul (2021) found that whilst people playing the a repeated RPS game indeed exhibit sequential regularities based on their previous moves and outcomes, their opponents fail to detect most of these patterns, leading to suboptimal exploitation. STATE WHAT IS DIFFERENT IN OUR STUDY COMPARED TO THEIRS. Which patterns are detectable and therefore exploitable remains unexplored. Wang, Xu, and Zhou (2014) show that strategies using a cyclical response conditional on previous number of wins/ties/losses fits data on population level. However, this analysis was not applied to individual level strategies. THIS IS UNCLEAR, AND ALSO DOESN'T INDICATE WHETHER PATTERNS WERE EXPLOITED.

Predicting an opponent's next action can be based on learning contingencies within a game (e.g., that your opponent is likely to play Scissors if your previous action was Paper), but also on more abstract aspects, such as that the opponent's action are based on a level-2 strategy. Unlike contingencies between actions, such abstract aspects can be generalized to other games, with other actions. This can then provide an early advantage, allowing one to exploit the opponent's strategy before you have played long enough to reliably establish contingencies within a game. In what we call the Bayesian Cognitive Hierarchy model, we propose that people use Bayesian inference to determine the depth of iterative reasoning of their opponent, and use this to predict their opponent's actions and determine the best response to this.

Rather than learning to predict an opponent's next actions, and then deciding upon the best counter-move, another plausible strategy is to directly learn rewarding actions. Instead of learning that your opponent is likely to play Scissors if your previous action was Paper, and therefore responding with Rock, you could learn that it pays off to play Rock whenever you played Paper on the previous round. Such learning is implemented in model-free reinforcement learning models, which learn state-action values (the expected reward from taking an action in a given state). The Experience Weighted Attraction (EWA) model (Teck H. Ho, Camerer, and Chong 2004), popular in behavioral game theory, includes an additional mechanism to learn about the consequences of actions that were not actually taken. As both RL and EWA models learn values for actions in states, they don't allow for generalization to new games with different actions, as there no immediate way in which to map a set of actions in one game to a set of actions in another. For example, consider the game of Fire-Water-Grass (FWG), where water beats fire, grass beats water, and fire beats grass. This game is structurally the same as RPS. However, if you have learned in RPS that playing Rock after a previous play of Paper, there is no direct way to transfer this to knowing that Fire is a rewarding move after previously having played Water. Learning about the new FWG actions requires experience with the FWG game. This is in contrast to learning the level of iterative reasoning of the opponent. If you know they expect you to repeat your previous action and choose their action to beat this, then you can infer that they will play Grass after your previous choice of Water, even if you haven't played FWG before.

Iterative reasoning is likely to be more cognitively demanding than learning action contingencies. Whilst iterative reasoning allows transfer between games, once enough experience has been gained within a game, it may no longer provide an advantage over simpler reinforcement learning. Therefore, it might be more efficient to use

an RL strategy in the later stages of a game. With the use of a hidden Markov model, we allow participants to switch between strategies during the games.

In the present study, we let humans repeatedly face computer agents endowed with a limited ability for iterative reasoning based on previous-round play. The use of computer opponents allows precise experimental control over the strategy used, and whether aspects such as the depth of reasoning can be transferred to new games.** We aim to 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 repeatedly face the same opponent (Experiment 1) or two different opponents (Experiment 2) in three consecutive games: the well-known Rock-Paper-Scissors game, the 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 new games, providing them with an early advantage before gaining enough experience within a game to enable learning contingencies between previous and winning actions. This transfer is moderated by the sophistication of their opponent and the similarity between the games. Computational modelling shows that participants, when first encountering an opponent in a game, employ Bayesian inference about the level of iterative reasoning of their opponent to predict their actions and determine the best response. In later rounds of the games, they switch to a cognitively less demanding reinforcement learning strategy.

2 Experiment 1

In the first experiment, we aim to test learning transfer by making participants face the same computer opponent with a limited level of iterative reasoning in three sequential games that vary in similarity. If participants are able to learn the limitations of their opponent's iterative reasoning and generalize this to new games, their performance in (the early stages of) later games should be higher than expected if they were to completely learn a new strategy in each game.

2.1 Methods

2.1.1 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 (£1.06 on average). The experiment had 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.

2.1.2 Tasks

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 label 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

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

| 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
...	...	...	...

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 requires good knowledge of this strategy, we opted for this more deterministic formulation of a level-0 player. The level-2 computer opponent assumes in turn that the participant is a level-1 opponent, playing according to the strategy just described. To make the computer opponent's strategy not too obvious, we introduced some randomness in actions of computer opponent, 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 et al. 2016; Wang, Xu, and Zhou 2014). Table 1 shows an example of the computer opponent's actions in response to the previous round play.

2.1.3 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 (between 0.5 and 3 seconds) is shown the action chosen by the computer opponent and the outcome of that round. 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 their opponent, players were paid a bonus at the end of the experiment proportional to their final score (each point was 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 opponent. They were then debriefed and thanked for their participation.

Outsmart your opponent

Rock, Paper, Scissors

Round : 5
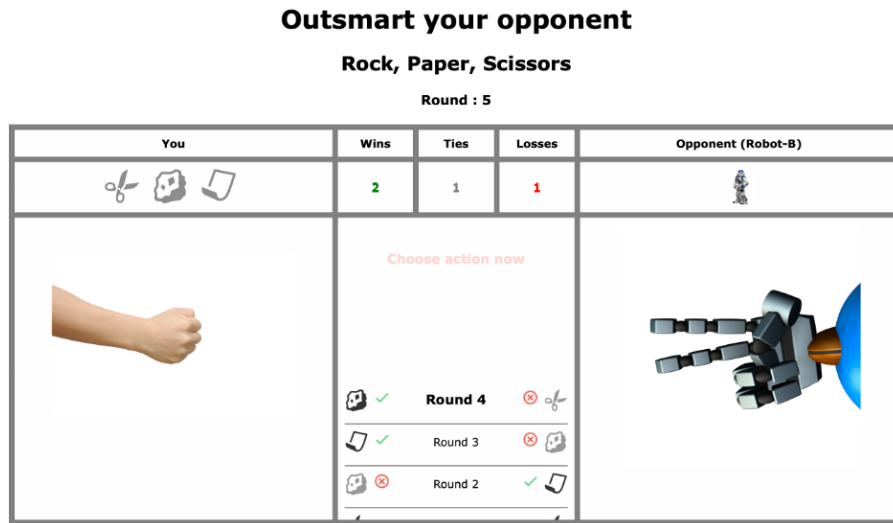
Fig. 1 Screenshot of the Rock-Paper-Scissors game in Experiment 2. Shown here is the feedback stage, after both the human (left) and computer (right) players have chosen their action. The interface was similar in Experiment 1, but excluded the history of game play in the center panel.
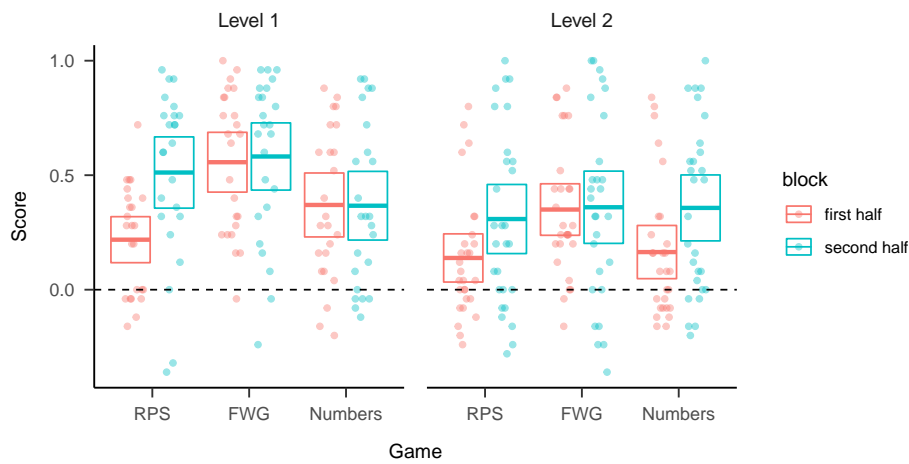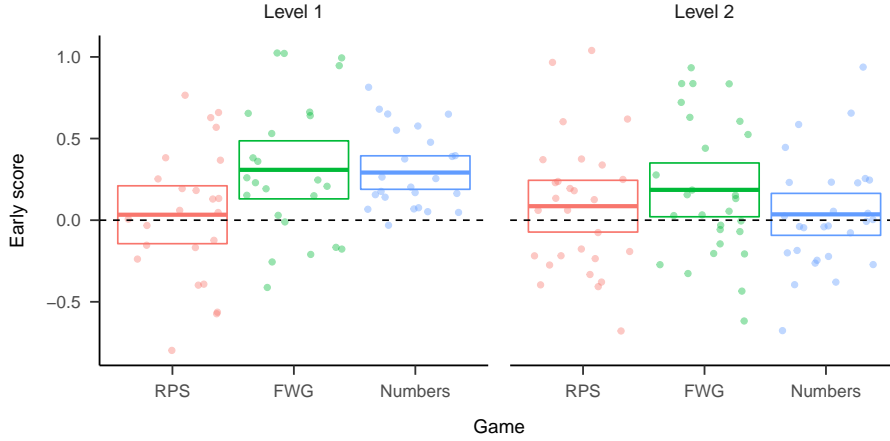

Fig. 2 Performance per game and block across conditions in Experiment 1. Points are scores of individual participants and boxes reflect the 95% confidence intervals of the mean (center line equals the mean).

2.2 Behavioral results

Participants' scores in each half of each game are depicted in Figure 2. Overall, scores in each game 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

Fig. 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).

the Nash equilibrium, this indicates successful deviation from a Nash-optimal strategy.

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 played 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 would be expected to perform at chance level in these early rounds, whilst positive scores in rounds 2-6 are consistent with generalization of prior experience. The early score score in both FWG and Numbers is significantly higher than 0 (FWG: $t(148.85) = 4.584$, $p < .001$; Numbers: $t(148.85) = 3.00$, $p = .009$). We did not expect positive early 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. We estimated a linear mixed-effects model with the early round score as the dependent variable, and Opponent (Level 1 or Level 2) and Game (RPS, FWG, Numbers), as well as their interaction, as independent variables. In addition to these fixed effects, the model included random intercepts for participants. The results showed only a significant main effect of Game ($F(2, 100) = 3.35$, $p = .04$).

3 Experiment 2

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 later games. 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 now faced both types of opponent. To perform well against both opponents, participants would need to learn distinct strategies against these opponents. 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.

3.1 Methods

3.1.1 Participants & Design

A total of 50 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 received a fixed fee of £2.5 for completing the experiment and a performance dependent bonus (£1.32 on average).

3.1.2 Tasks

Participants played three games: Rock-Paper-Scissors (RPS), Fire-Water-Grass (FWG), and the penalty Shootout game. The first two games were identical to the ones used in Experiment 1. In the Shootout game, 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 now 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
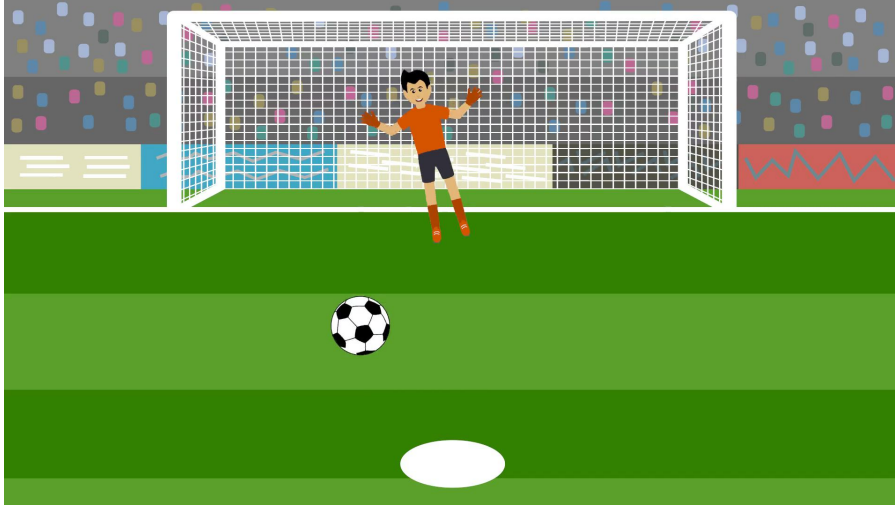
Fig. 4 Screenshot of the shootout game

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.

As in Experiment 1, 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.

3.1.3 Procedure

Participants played 3 games sequentially against both level-1 and level-2 computer opponents. As in Experiment 1, the computer opponents retained the same strategy throughout the 3 games. Participants faced each opponent twice in each game. 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 started by facing one of the opponents in stage one, then the other in stage two. This was repeated in the same order in stages 3 and 4. Which opponent they faced first was counterbalanced. All participants engaged in the three games (RPS, FWG and Shootout) in this exact order, and were aware that their opponent could not cheat and chose their action simultaneously with the player, without knowing their choices beforehand. In order to
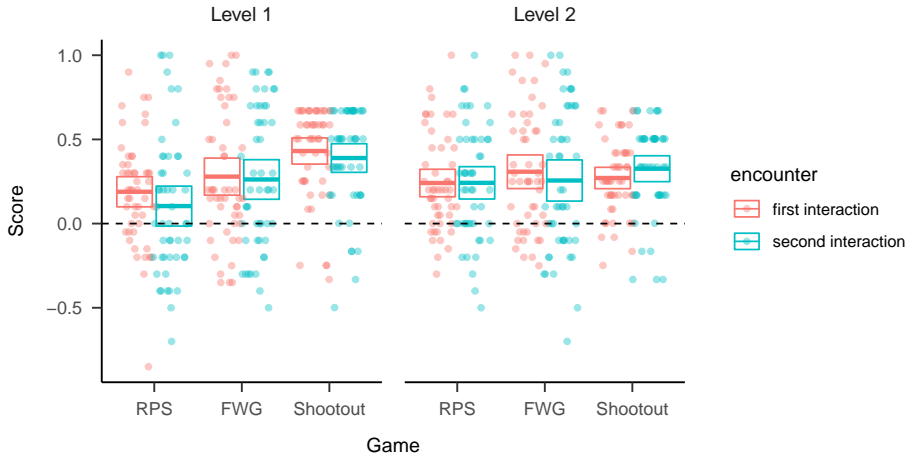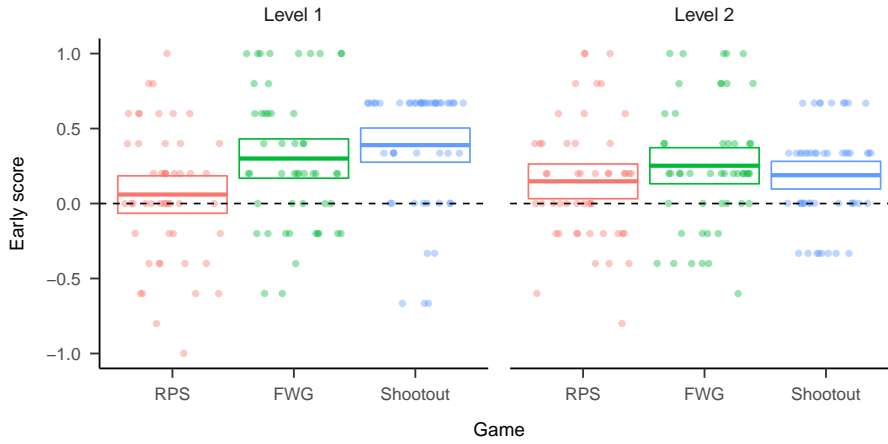
Fig. 5 Performance per game and interaction across opponents in Experiment 2. Points are scores of individual participants and boxes reflect the 95% confidence intervals of the mean (center line equals the mean).

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 random delay between 0.5 and 3 seconds was again introduced before the choice of the computer agent was revealed, as a way of simulating a real human opponent's decision 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 in each game were displayed at the top of the screen, and this scoreboard was reinitialised to zero at the onset of a new stage game.

3.2 Behavioral results

Participants' scores are depicted in Figure 5.[1] Average (adjusted) scores differed from 0 in all three games (RPS: $t(49) = 6.26$, $p < .001$; FWG: $t(49) = 7.25$, $p < .001$; Shootout: $t(49) = 13.61$, $p < .001$). As in Experiment 1, this indicates that participants successfully deviated from random (Nash-optimal) strategies.

---

[1] Scores in the Shootout game were adjusted because there were 2 out of three possible winning actions in that game, compared to one out of three in RPS and FWG. As such, the expected score of random play in the Shootout game is 1/3, whilst it is 0 in the RPS and FWG game. To make the scores comparable, we therefore subtracted 1/3 from the Shootout scores.

Fig. 6 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).

To measure transfer between games, we again focus on participants' scores in the initial 5 rounds after the first round (rounds 2-6) of each game and only during the first time they interact with an opponent in a game (see Figure 6). Early-round scores differed significantly from 0 for the both the FWG (level-1 opponent: $t(271) = 4.99$, $p < .001$, level-2 opponent: $t(271) = 4.41$, $p < .001$) and Shootout games (level-1 opponent: $t(271) = 6.66$, $p < .001$, level-2 opponent: $t(271) = 3.22$, $p = .004$). This is consistent with transfer of opponent models between games. WHAT ABOUT RPS?

We next estimated a linear mixed-effects model with the early round scores as the dependent variable, and Block (1, 2), Opponent type (level-1 or level-2) and Game (RPS, FWG, Shootout) as independent variables, including all main effects and interactions. In addition to these fixed effects, the model also included random intercepts for participants. We found a main effect of Game ($F(2, 240) = 6.57$, $p = .002$), and an interaction between Game and Opponent type ($F(2, 240) = 3.34$, $p = .037$). The latter effect is due to early-score increasing from FWG to Shootout for level-1 opponents, but decreasing for level-2 opponents. We found no effect of Block. This is important, as if early-round scores are due to general practice effects, rather than transfer of an opponent model, we would expect early-round scores to increase between block 1 (first encounter of one opponent) and 2 (first encounter of other opponent) within a game.

3.3 Discussion

The results of Experiment 2 confirm earlier findings from Experiment 1 on learning transfer in a situation where participants need to learn about two distinct opponents. As in Experiment 1, participants adapted their strategies

to both level-1 and level-2 opponents to exploit deviations from Nash-optimal play. Knowledge about both types of opponents were transferred to both the similar and dissimilar game. This transfer was moderated by the degree of similarity between games, as well as opponent type. The interaction effect showed that transfer to the dissimilar game was easier when facing the less sophisticated agent.

4 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 (self-tuning) 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 implement learning about the game and/or the opponent. Throughout, 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. The initial state is empty, and we assume that the action at a first encounter of an opponent in a game is chosen at random.

4.1 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 and Dayan 1992) for each state-action pair:

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

where $Q(s_t, a_t)$ is the value of taking action a when in state s at time t, 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 λ determines the consistency of the strategy (the higher λ, 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 space $\mathcal{S}_g$ and action space $\mathcal{A}_g$, and there is no simple way to map states and actions across games. The RL model has two free parameters: the learning rate (α) and the inverse temperature (λ).

4.2 Experience-weighted attraction (EWA) model

As discussed in the introductory section on computational models, the self-tuning Experience Weighted Attraction (EWA) model combines two seemingly different approaches, namely reinforcement learning and belief learning. The EWA model is based on updating "Attractions'' for each action over time given a particular state. The attraction of action a at time t given state s is denoted as $A_t(a, s)$ and updated as

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

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. $R(a, o_t)$ 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 actually taken by the rewards they would have provided. The parameter δ reflects the weight given to such counterfactual rewards. Setting $\delta = 0$ leads to reinforcement only of actions taken, while positive values of δ makes the update rule take into account foregone payoffs, which is similar to weighted fictitious play (Cheung and Friedman 1994). $N(t)$ 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.

In the earlier version of the EWA model (C. Camerer, Ho, and Others 1997), ϕ and δ were time-invariant free parameters. In the self-tuning EWA model (Teck H. Ho, Camerer, and Chong 2004), the values of δ and ϕ are learnt from experience. Over time, δ is updated as

$$\delta_a(t) = \begin{cases} 1 & \text{if } R(a, o_t) \geq r_t \\ 0 & \text{otherwise} \end{cases}$$

The ϕ_t parameter can be interpreted as a discount of prior experience, modelling either limited agent memory or changes in the game conditions. At the core, $\phi(t)$ depends on a surprise index $S_p(t)$:

$$\phi(t) = 1 - \frac{1}{2}S_p(t),$$

where $S_p(t)$ quantifies how the opponent deviates from past play. It is calculated in turn through the cumulative history of play, across opponent possible actions k, as $h^k(t)$:

$$h^k(t) = \frac{\sum_{\tau=1}^{t} I(o_\tau = o^k)}{t},$$

as well as an indicator vector of the most recent opponent play:

$$r^k(t) = I(o^k = o_t),$$

Where I is the indicator function as defined above. To get the surprise index, we simply sum all the squared deviations between the cumulative history vector $h^k(t)$ and the immediate history $r^k(t)$:

$$S_p(t) = \sum_{k=1}^{|\mathcal{A}_g|} (h^k(t) - r^k(t))^2$$

For more details on the self-tuning EWA model, we refer the reader to (Teck H. Ho, Camerer, and Chong 2004). As in the RL model above, actions are chosen based on a softmax decision rule:

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

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

4.3 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. For present purposes, we assume participants consider the opponent could be either a level 0, level 1, or level 2 player, and start with a prior belief that each of these types is equally likely. They then use observations of the opponent's 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 $\mathcal{D}_t = \{s_1, \ldots, s_t\}$ 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:

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

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

where $b(a,o) = 1$ if action a is a best response to opponent's action o (i.e. it leads to a win).

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 (θ), and the temperature parameter of the softmax function (λ).

4.4 Estimation

For both experiments, we fitted all models to individual participant data by maximum likelihood estimation using the DEoptim R package (Mullen et al. 2011). 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_NT), as well as a reinforcement learning (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 types of opponent model transfer. We can have transfer *within* games between the first and second interaction with the opponent. In addition, we can also have transfer *between*
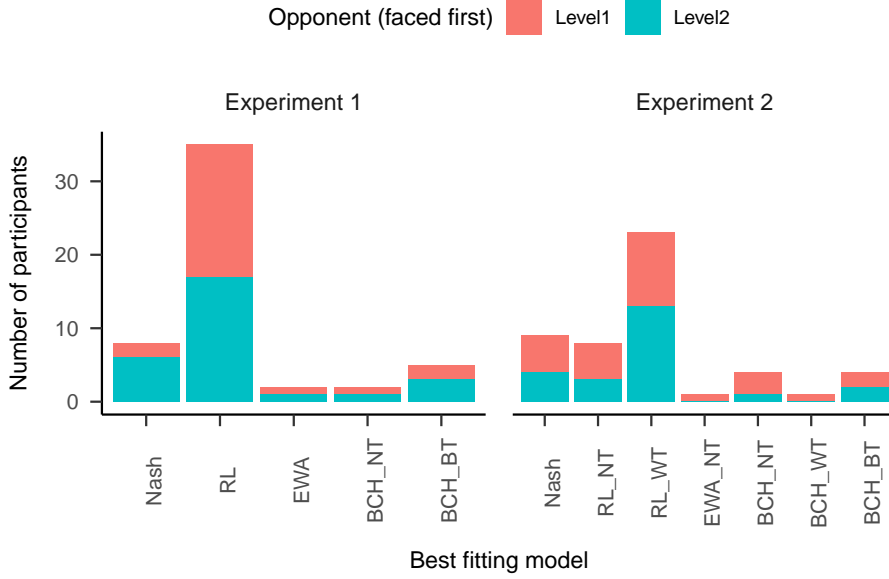
Fig. 7 Number of participants best fitted by each model in Experiment 1 and 2 according to the BIC. Results are seperated by opponent (Experiment 1) and which opponent they faced first (Experiment 2).

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_WT allows only for within-game transfer, but not between game transfer; BCH_NT allows for no transfer within or between games. As the RL model can't account for between game transfer due to a change in state and action space, we can only have models that allowing for within game transfer (RL_WT) or with no transfer within games (RL_NT). Likewise, we fit both a self tuning EWA model with (EWA_WT) and without (EWA_NT) within-game transfer. Including the baseline model of random play (Nash), we therefore fitted a total of 8 models for Experiment 2.

4.5 Model comparison

When considering which model fits best (Figure 7), we find that in both experiments, the RL model described most participants' behaviour best, followed by the random (Nash) model. Relatively few participants were best described by one of the BCH models or the EWA model.

As the RL models don't transfer learning between games, this is at odds with the behavioural findings where we found evidence for such transfer. In order to investigate this discrepancy further, we plot the likelihood by trial
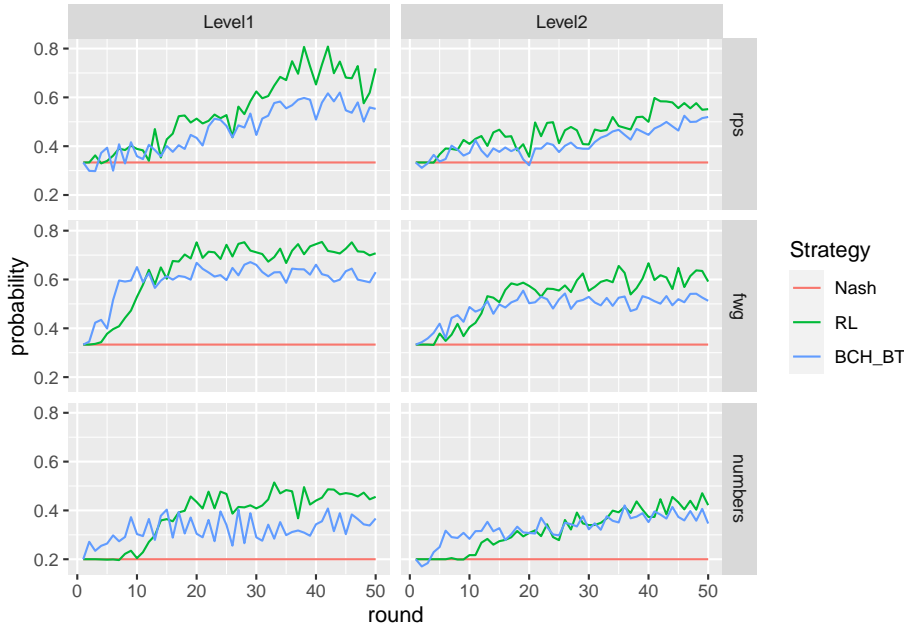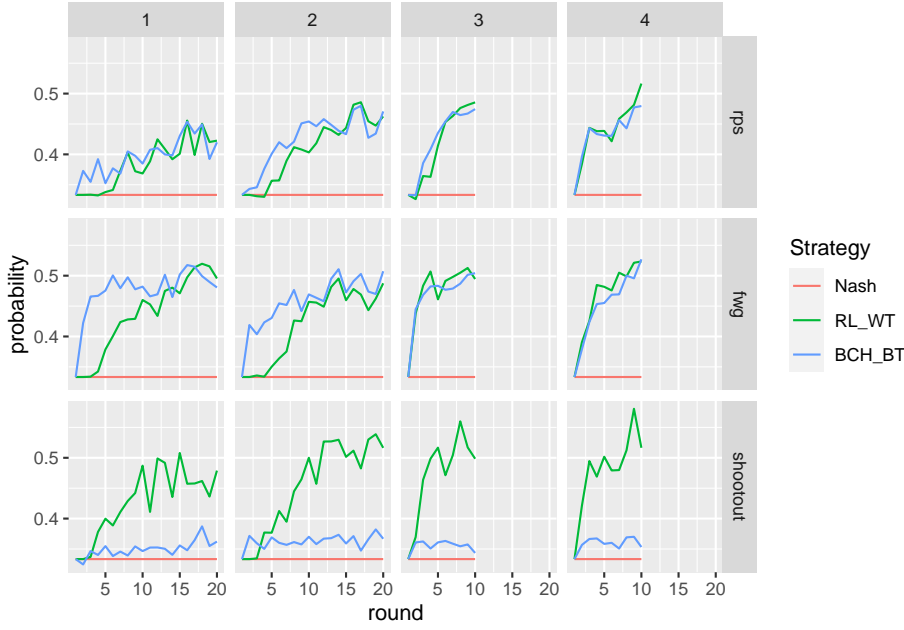
Fig. 8 Experiment 1 likelihood by trial, game and opponent faced

for each game and three strategies: reinforcement learning (RL), Bayesian Cognitive Hierarchy with between-game transfer (BCH_BT), and the random (Nash) strategy. Figure 8 shows that in the initial rounds of the the later games in Experiment 1, the likelihood for the BCH model is higher than that of the other two models. However, over time, the likelihood of the RL model increases and exceeds that of the BCH model. The same pattern holds for Experiment 2 (Figure 9). Again, the BCH_BT 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 early rounds on the learned level of their opponents iterative reasoning, switching later to learned actions through reinforcement.

## 4.6 Hidden Markov model analysis of strategy switching

We use hidden Markov models to 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 (i.e. strate-

Fig. 9 Experiment2 likelihood by trial, game and opponent faced

gies) over time. 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 and 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.

In Experiment 1, a likelihood-ratio test shows that the HMM model with switching fits significantly better than the non-switching one, $\chi^2(6) = 211.09$, $p < .001$. We should note that as the no switch model involves restricting parameters of the switching model on the bounds of the parameter space, the asymptotic p-value of this test may not be reliable. The switching model (AIC $= 1.463687 \times 10^4$, BIC $= 1.469256 \times 10^4$) fits also better than the non-switching model (AIC $= 1.483596 \times 10^4$, BIC $= 1.484988 \times 10^4$) according to the AIC and BIC. This provides further statistical evidence in favour of the hypothesis that participants switch between strategies. Figure 10 depicts the average (across participants) posterior probabilities of each state (strategy), as a function of
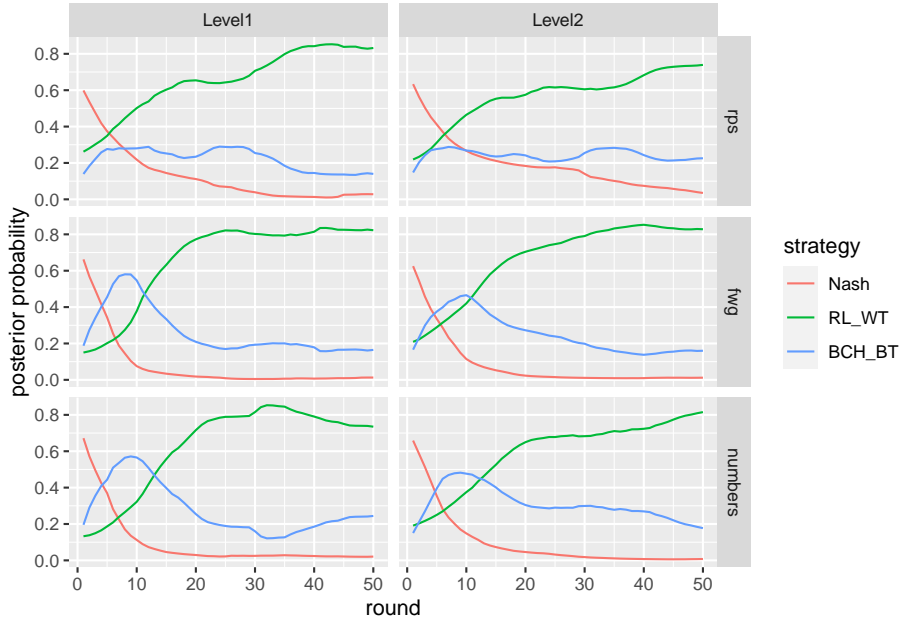
Fig. 10 Experiment 1 posterior probability of strategies by game and opponent faced

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.

In Experiment 2, the switching model (AIC $= 1.687997 \times 10^4$, BIC $= 1.693681 \times 10^4$) again fits better than the restricted non-switching model (AIC $= 1.693841 \times 10^4$, BIC $= 1.695262 \times 10^4$), $\chi^2(6) = 70.44$, $p < .001$. The posterior probabilities of the strategies (Figure 11) 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.

5 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 in 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 learnt to adapt to their opponent's strategy over multiple interactions, and generalised this learning to a similar game. Performance in early rounds indi-
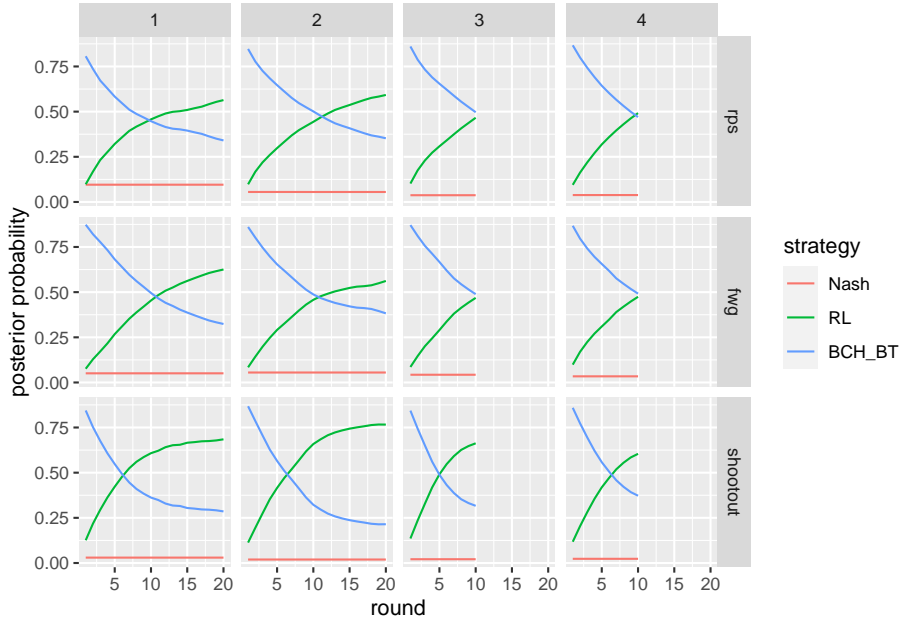
Fig. 11 Experiment 2 posterior probability of strategies by game and opponent faced

cated that learning transfer was moderated by the degree of similarity between games. In the second experiment, participants faced both types of opponents, allowing for a stronger test of opponent modelling, as participants would need to learn a different strategy for each opponent. In Experiment 1, participants could learn a single strategy for each game, making opponent modelling possibly less pertinent. In Experiment 2, we again found clear evidence of transfer in the early rounds of the later games. In addition to the moderation of transfer by the similarity between games, we also found that learning transfer was moderated by the type of opponent faced: Transfer was harder when facing the level-2 opponent as compared to the level-1 opponent. That evidence for transfer was weaker when facing the more sophisticated level-2 opponent 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 the first game (RPS) that allowed them to beat their opponent in the later games (FWG, and Numbers or 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 indicates 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 this heuristic for a level-2 player is simpler than that for a level-1 player, which is at odds with the finding that participants performed better against the level-1 opponent.

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 reason-

ing, 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 (Stahl 2000, 2003).

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. 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 (Lieder and Griffiths 2020). 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 (Daw, Niv, and Dayan 2005; Simon and Daw 2011; Wouter Kool, Joseph T. McGuire, Zev B. Rosen, Matthew M. Bovinick 2011). 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.

6 Conclusion

Our results show that people can successfully deviate from Nash equilibrium play to exploit deviations from such play by their **computer** opponents. Moreover, people can transfer knowledge about the limitations of their computer 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.

7 Declarations

7.1 Conflict of interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

7.2 Funding

7.3 Availability of Data and Material

The results are fully reproducible, and all code for the analyses (including the manuscript in R Markdown format) are available at: `https://github.com/ismailg/CBB_Learning_Transfer`

7.4 Ethics Approval

The study was conducted in line with the human participant guidelines of the 1964 Declaration of Helsinki and was approved by the local UCL Ethics Committee (EP/2014/005).

7.5 Consent to Participate and Consent for Publication

Informed consent for participation and publication of results was obtained from all participants at the start of the experiments.

References

Batzilis, Dimitris, Sonia Jaffe, Steven Levitt, John A List, and Jeffrey Picel. 2016. "How Facebook Can Deepen Our Understanding of Behavior in Strategic Settings: Evidence from a Million Rock-Paper-Scissors Games." working paper.

Brockbank, Erik, and Edward Vul. 2021. "Formalizing Opponent Modeling with the Rock, Paper, Scissors Game." *Games* 12 (3): 70.

Camerer, Colin F. 2003. "Behavioural Studies of Strategic Thinking in Games."
    Trends in Cognitive Sciences 7 (5): 225–31.

Camerer, Colin F, Teck-Hua Ho, and Juin-Kuan Chong. 2004. "A Cognitive
    Hierarchy Model of Games." *The Quarterly Journal of Economics* 119 (3):
    861–98.

Camerer, Colin, Teck-Hua Ho, and Others. 1997. "Experience-weighted at-
    traction learning in games: A unifying approach."

Camerer, Colin, and Marc Knez. 2000. "Increasing Cooperation in Prisoner's
    Dilemmas by Establishing a Precedent of Efficiency in Coordination Games."

Cheung, Yin-Wong, and Daniel Friedman. 1994. *Learning in evolutionary
    games: some laboratory results.* University of California, Santa Cruz.

Costa-Gomes, Miguel, Vincent P Crawford, and Bruno Broseta. 2001. "Cog-
    nition and Behavior in Normal-Form Games: An Experimental Study."
    Econometrica 69 (5): 1193–1235.

Daw, Nathaniel D, Yael Niv, and Peter Dayan. 2005. "Uncertainty-Based Com-
    petition Between Prefrontal and Dorsolateral Striatal Systems for Behav-
    ioral Control." *Nature Neuroscience* 8 (12): 1704–11.

Dyson, Benjamin J. 2019. "Behavioural Isomorphism, Cognitive Economy and
    Recursive Thought in Non-Transitive Game Strategy." *Games* 10 (3): 32.

Ho, Teck H, Colin F Camerer, and Juin-Kuan Chong. 2004. "The economics
    of learning models: A self-tuning theory of learning in games."

Ho, Teck-Hua, Colin Camerer, and Keith Weigelt. 1998. "Iterated dominance
    and iterated best response in experimental" p-beauty contests"." *The Amer-
    ican Economic Review* 88 (4): 947–69.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J.
    Gershman. 2017. "Building machines that learn and think like people." *Be-
    havioral and Brain Sciences* 40. `https://doi.org/10.1017/S0140525X16001837`.

Lieder, Falk, and Thomas L Griffiths. 2020. "Resource-Rational Analysis: Un-
    derstanding Human Cognition as the Optimal Use of Limited Computa-
    tional Resources." *Behavioral and Brain Sciences* 43.

Mandler, Jean Matter. 2004. *The Foundations of Mind: Origins of Conceptual
    Thought.* Oxford University Press.

Mertens, Jean-François. 1990. "Repeated games." In *Game Theory and Ap-
    plications*, 77–130. Elsevier.

Mullen, Katharine, David Ardia, David Gil, Donald Windover, and James
    Cline. 2011. "DEoptim: An R Package for Global Optimization by Dif-
    ferential Evolution." *Journal of Statistical Software* 40 (6): 1–26. `https:
    //doi.org/10.18637/jss.v040.i06`.

Nagel, Rosemarie. 1995. "Unraveling in Guessing Games: An Experimental
    Study." *The American Economic Review* 85 (5): 1313–26.

Shachat, Jason, and J Todd Swarthout. 2004. "Do we detect and exploit mixed
    strategy play by opponents?" *Mathematical Methods of Operations Re-
    search* 59 (3): 359–73.

Simon, Dylan A., and Nathaniel D. Daw. 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, Leonidas. 2013. "Strategic adaptation of humans playing computer algorithms in a repeated constant-sum game." *Autonomous Agents and Multi-Agent Systems* 27 (1): 131–60.

Stahl, Dale O. 2000. "Rule Learning in Symmetric Normal-Form Games: Theory and Evidence." *Games and Economic Behavior* 32 (1): 105–38.

———. 2003. "Sophisticated Learning and Learning Sophistication." *Available at SSRN 410921*.

Stahl, Dale O, and Paul W Wilson. 1995. "On players models of other players: Theory and experimental evidence." *Games and Economic Behavior* 10 (1): 218–54.

Visser, Ingmar, and Maarten Speekenbrink. 2010. "depmixS4: An R Package for Hidden Markov Models." *Journal of Statistical Software* 36 (7): 1–21. http://www.jstatsoft.org/v36/i07/.

Wang, Zhijian, Bin Xu, and Hai-Jun Zhou. 2014. "Social Cycling and Conditional Responses in the Rock-Paper-Scissors Game." *Scientific Reports* 4 (1): 1–7.

Watkins, Christopher J C H, and Peter Dayan. 1992. "Q-learning." *Machine Learning* 8 (3-4): 279–92.

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.