₁   Transfer of Learned Opponent Models in Zero Sum Games

₂   Ismail Guennouni[1] & Maarten Speekenbrink[1]

₃   [1] Department of Experimental Psychology, University College London

₄   Author Note

₅   Enter author note here.

₆   Correspondence concerning this article should be addressed to Ismail Guennouni,

₇ Department of Experimental Psychology, University College London, 26 Bedford Way,

₈ London WC1H 0AP, United Kingdom. E-mail: i.guennouni.17@ucl.ac.uk

Abstract

Enter abstract here. Each new line herein must be indented, like this line.

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. Humans are naturally endowed with the ability to extract relevant features from a situation, identify the presence of these features in a novel setting and use previously acquired knowledge to adapt to previously unseen challenges using acquired knowledge. More formally, Perkins (1992) defines transfer of learning as the application of skills, knowledge, and/or attitudes that were learned in one situation to another learning situation. This typically human skill has so far eluded modern AI agents. Deep neural networks for instance can do very well on image recognition tasks and can even reach super-human performance levels on video and strategic board games. Yet they struggle to learn as fast or as efficiently as humans do, and more importantly they have a very limited ability to generalize and transfer knowledge to new domains. Lake, Ullman, Tenenbaum, & Gershman (2017) argue that human learning transfer abilities take advantage of important cognitive building blocks such as an abstract representation of concepts underlying tasks and compositionally structured 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 person we are interacting with that may inform what actions they are likely to take next. Once we learn something about them, we can use this knowledge to inform how to best behave in novel situations. This may lead to very efficient generalization of knowledge, even to situations that are dissimilar to the history of interaction, assuming what we have learned about others is an abstract representation that is not too dependent on the environment of the initial interaction. There is evidence that people learn models of their opponents when they play repeated economic games (Stahl & Wilson, 1995), engage in bilateral negotiations (Baarslag, Hendrikx, Hindriks, & Jonker, 2016), or simply try to exploit a non random player in chance games such as

Rock-Paper-Scissors (Weerd, Verbrugge, & Verheij, 2012). In this paper, we are specifically interested in the way in which people build and use models of their opponent to facilitate learning transfer, when engaged in situations involving an interaction with strategic considerations. These situations arise frequently such as in negotiations, auctions, strategic planning and all other domains in which theory of mind abilities (Woodruff, Sciences, Brain, & 1978, n.d.) play a role in determining human behaviour. In order to explore learning transfer in these strategic settings, it is generally useful to study simple games as a model of more complex interactions. More specifically, we need a framework that allows the study of whether and how a player takes into consideration, over time, the impact of its current and future actions on the future actions of the opponent and the future cumulative rewards. Repeated games, in which players interact repeatedly with the same opponent and have the ability to learn about the opponent's strategies and preferences (Mertens, 1990) are particularly adapted to the task of opponent modelling.

Early literature on learning transfer in games has mostly focused on measuring the proportion of people who play normatively optimal (Nash Equilibria) or salient actions (e.g Risk Dominance) in later games, having had experience with a similar game environment previously. For instance, Ho, Camerer, & Weigelt (1998) measure transfer as the proportion of players who choose the Nash Equilibrium in later p-beauty contest games, after training on similar games. They find there is no evidence of immediate transfer (Nash equilibrium play in the first round of the new game) but positive structural learning transfer as shown by the faster convergence to equilibrium play by experienced vs non experienced players. (**???**) test learning transfer in players exposed to two games with multiple equilibria sequentially and explore the ability of players to coordinate their actions to choose a particular equilibrium in subsequent games having reached it in prior ones. They distinguish between games that are similar in a purely descriptive way, meaning similar choice labels, identity of players, format and number of action choices; and games that are similar in a strategic sense, meaning similar payoffs from combination of actions, identical equilibrium properties or

significant social characteristics of payoffs such as possibility of punishment, need for fairness

and cooperative vs competitive settings. They find that transfer of learning (successful

coordination) occurs more readily in the presence of both descriptive and strategic similarity.

If the games were only strategically similar, then the transfer was much weaker.

Juvina, Saleem, Martin, Gonzalez, & Lebiere (2013) made a similar distinction between

what they deemed surface and deep similarities and find that both contribute to positive

learning transfer. However, they show that surface similarity is not necessary for deep

transfer and can either aid or block this type of transfer depending on whether it leads to

congruent or incongruent actions in later games. In a series of experiments using economic

signalling games (Cooper & Kagel, 2003, 2008) the researchers found that participants who

have learned to play according to Nash Equilibrium in one game can transfer this to

subsequent games, even though the actions consistent with Nash Equilibrium in later games

are different. They show that this transfer is driven by the emergence of sophisticated

players who are able to represent the strategic implications of their actions and reason about

the consequences of changed opponent payoffs.

Most of these studies fail to offer a formal explanation of this transfer or a modelling

framework that can explain the experimental observation of transfer between games and

generalise it to extensive classes of games. A notable exception is the effort by Haruvy &

Stahl (2012) to specify a model of learning where players learn abstract rules that they can

generalise and transfer across dissimilar games, rather than action choices that can only be

used within the same game. Participants played ten games, presented as 4x4 normal form

(matrix payoffs). Their results suggest that subjects do transfer learning over descriptively

similar but strategically dissimilar games and that this learning transfer is significant. They

also showed that players learn abstract aspects of the game that are then transferred to new

settings. Their rule-learning model, based on Stahl & Wilson (1995), was able to capture

participants' dynamic behavior and show that the propensity to select particular rules is

90 perfectly transferred across games.

91   Still, one of the commonalities in studies of how humans adapt to computerised

92 opponents is that they have mostly looked at the ability of players to detect and exploit

93 action-based learning rules. The strategies implemented by the computer opponents had a

94 style of play that was not "human-like" in the sense that humans are not very good at

95 playing specific mixed strategies with any precision of at detecting patterns from long

96 sequences of past play due to cognitive constraints. It is therefore important to have agents

97 that "play like humans", and one way of achieving that is to embed theses agents with

98 human-like theory of mind abilities based on limited steps of recursive reasoning. Simon

99 (1972) explains that humans have limited cognitive capacities and as such cannot be expected

100 to solve computationally intractable problems such as finding Nash equilibria. Instead, they

101 will try to "satisfice" by choosing a strategy that is adequate in a simplified model of the

102 environment, rather than an optimal one. This concept finds its natural application in

103 "level-k" theory, first adopted by Stahl & Wilson (1995). It posits that deviations from Nash

104 equilibrium solutions in simple games are explained by the fact that humans have a

105 heterogenous degree of strategic sophistication. At the bottom of the ladder, level-0 players

106 are non-strategic and play either randomly or use a salient strategy in the game environment

107 Arad & Rubinstein (2012). Level-1 players are next up the ladder of strategic sophistication

108 and will assume all their opponents belong to the level-0 category and as such will best

109 respond to them given this assumption. Likewise, a level-2 player will choose actions that are

110 the best response given the belief that all opponents are exactly one level below, and so on.

111   In this study, we propose to explore opponent modelling and its transfer with the use

112 of computer agents possessing human-like theory of mind abilities with limited degrees of

113 iterated reasoning. The agents will have a fixed strategy played stochastically and will either

114 be level-1 or level-2 behavioral rule, mimicking human theory of mind abilities and the

115 limited recursion depth they exhibit Goodie, Doshi, & Young (2012).

₁₁₆   Our choice of using computer oponents instead of matching groups of participants

₁₁₇ makes it easier to disentangle the process of learning about the opponent from that of

₁₁₈ learning about the game structure and payoffs. When playing against other human

₁₁₉ opponents, players are learning about the game and the strategy of the opponent

₁₂₀ simultaneously. Thus, it is harder to focus on an individual and how her strategies are

₁₂₁ changing and adapting to the opponent's play if we cannot experimentally control the

₁₂₂ behaviour of the opponent. The use of computer opponents to elicit learning behavior has

₁₂₃ been explored in the literature with encouraging results. For instance, Spiliopoulos (2013)

₁₂₄ made humans play constant sum games against 3 computer opponents programmed to take

₁₂₅ advantage of known patterns in human play such as imperfect randomization and heuristics

₁₂₆ use and found that human participants do adapt to the opponent they are facing. Shachat &

₁₂₇ Swarthout (2004) made human participants face computer opponents playing various mixed

₁₂₈ strategies in a zero-sum asymmetric matching pennies game. They found that the players

₁₂₉ changed their strategies towards exploiting the deviations from the Mixed Strategy Nash

₁₃₀ Equilibrium (MSNE), and that this exploitation was very likely if the deviation from the

₁₃₁ MSNE play was high.

₁₃₂   We chose to use a set of purely competitive zero-sum games that incentivise reasoning

₁₃₃ about the opponent without introducing social confounding factors. We avoid social dilemma

₁₃₄ and coordination games that crystallise the conflict between competing with others out of

₁₃₅ self-interest or cooperating to reach a socially optimal outcome. Playing these games

₁₃₆ repeatedly implicates important social constructs such as reputation building, trust and

₁₃₇ other individual psychological attributes such as cooperativeness and inequity aversion.

₁₃₈ Coordination games also test the ability of choosing the safe self-interested choice compared

₁₃₉ to the risky cooperative choice and may depend on similar latent factors. As such, these

₁₄₀ confounding factors may impede teasing out the process of learning an opponent model and

₁₄₁ that of whether/how these models are transferred. By contrast, purely competitive settings

₁₄₂ and zero sum games in particular would be more appropriate. They do not incentivize any

cooperation, since one player's gain is necessarily the opponent's loss, and are agnostic to trust mechanisms or reputation building. Also choosing games with no pure strategy Nash Equilibrium shift the focus away from learning a normatively optimal action to that of reasoning dynamically about the opponent which is more amenable to modelling learning. Games with unique mixed strategy Nash equilibrium leading to nil average rewards facilitate inferring whether participants have learned to exploit the non-random play of the opponent.

We measure transfer of learning about the opponent strategy between games with varying degrees of similarity. The first two games we use are identical except for action labels. In one experiment, the third game is strategically similar to the first two but descriptively different, while in a second experiment, we introduce a third game that is dissimilar to the first two in terms of payoff matrix and strategic structure while retaining, like all other games, a unique mixed strategy Nash equilibrium of random action with a fixed expected reward against a Nash player.

Hence, we ran a total of two experiments where human participants played 3 different games against either 1 (first experiment) or 2 (second experiment) computer opponents. In the following sections, we will describe each experiment's methodology and results separately, then discuss the findings before using computational modelling to shed light on some of the results.

<p style="text-align:center;">161</p>

Experiment 1

162 **Methods**

163 **Participants**

164   A total of 52 (28 female, 24 male) participants were recruited on the Prolific Academic

165 platform. The mean age of participants was 31.2 years. Participants were paid a fixed fee of

166 £2.5 plus a bonus dependent on their performance which averaged £1.06. The study used a

167 2 (computer opponent levels 1 and 2) by 3 (games of rock-paper-scissors, fire-water-grass and

168 numbers) design, with repeated measures on the second factor.

169 **Procedure**

170   In the first experiment, participants played 3 games sequentially against the same

171 computer opponent. The computer opponent either used a level-1 or level-2 strategy. The

172 three games were rock-paper-scissors, fire-water-grass and the numbers game. A typical

173 rock-paper-scissors game (hereafter RPS) is a 3x3 zero sum game, with a cyclical hierarchy

174 between possible actions: rock blunts scissors, paper wraps rock, and scissors cut paper. If

175 one player choses an action which dominates their opponent's action, the player wins

176 (receives a reward of 1) and the other player loses (receives a reward of -1). Otherwise it is a

177 draw and both players receive a reward of 0. It has a unique MSNE consisting of randomly

178 playing one of the three options each time.

179   The second game is identical to Rock-Paper-Scissors in all but action labels. We call it

180 Fire-Water-Grass (FWG): Fire burns grass, water extinguishes fire, and grass absorbs water.

181 We are interested in exploring whether learning is transferred in a fundamentally similar

182 game where the only difference is in the description of the choice actions. Finally, the

183 numbers game is a generalization of rock-paper-scissors. In the variant we use, 2 participants

184 concurrently pick a number between 1 and 5. To win in this game, a participant needs to

185 pick a number exactly 1 higher than the number chosen by the 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 choice. All other possible combination of choices that are not consecutive are considered ties. A win would add 1 point to the score of the player, while a loss deduces one point and a tie does not affect the score. Similar to RPS, the MSNE is to play each action with equal probability in a random way.

Participants were informed they would play three different games against the same computer opponent. Each participant plays all three games consecutively and in the same order described above. Participants were told that the opponent cannot cheat and will choose its actions simultaneously without knowledge of the participant's choice. 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. The overall score of the players was translated into the bonus by making each point worth £0.02. This bonus is significant as players could increase the total payoff from the experiment by up to 60% assuming they'd won all rounds against the computer opponent. An example of the interface for the rock-paper-scissors game is provided in Figure 1.

Results

Looking at the aggregate scores (See Figure 2), the RPS game had the lowest average score across participants (M = 0.289, SD = 0.348) followed by NUMBERS (M = 0.31, SD = 0.347) and finally the FWG game had the highest average score (M = 0.454, SD = 0.354). Aggregate average scores for each game were significantly different from 0 (hypothesised
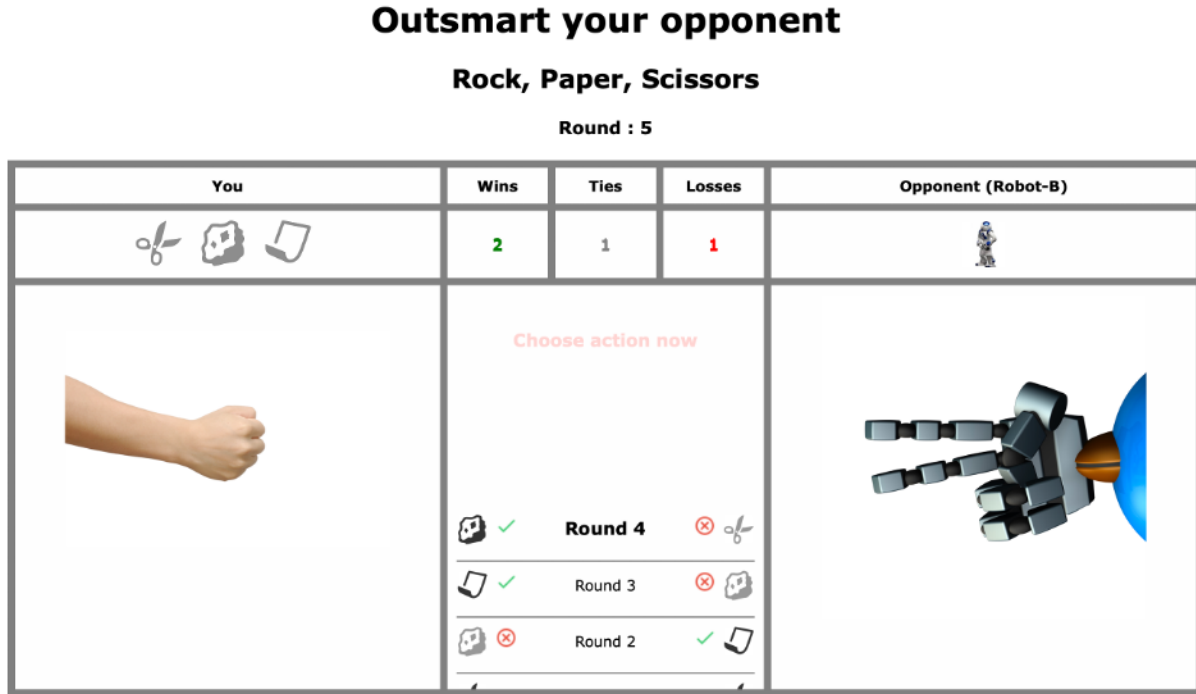
Figure 1. Screenshot of the feedback at the end of a round of Rock-Paper-Scissors

value of random play) using parametric one sample t-tests (RPS: $t(51) = 7.26$, p-value $<$ 0.001 ; FWG: $t(51) = 10.04$, p-value < 0.001 ; NUMBERS: $t(51) = 7.17$, p-value < 0.001).

To analyse within and between game learning, 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. There was a main effect of Game ($F(2,100) = 8.54$, $\eta^2 = 0.05$, $p < 0.001$), showing 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 ($t(100) = 3.78$, $p = 0.0008$), and the performance in NUMBERS was significantly lower than FWG game ($t(100) = -3.32$, $p = 0.0024$). The score in RPS was not significantly different from the score in NUMBERS however ($t(100) = 0.45$, $p = 0.65$). The main effect of Block ($F(1,50) = 22.51$, $p < .001$, $\eta^2 = 0.03$) shows that the average score in the first half of games (M = 0.29) was significantly lower than in the second half of the games played (M = 0.40), which translates to within-game learning.
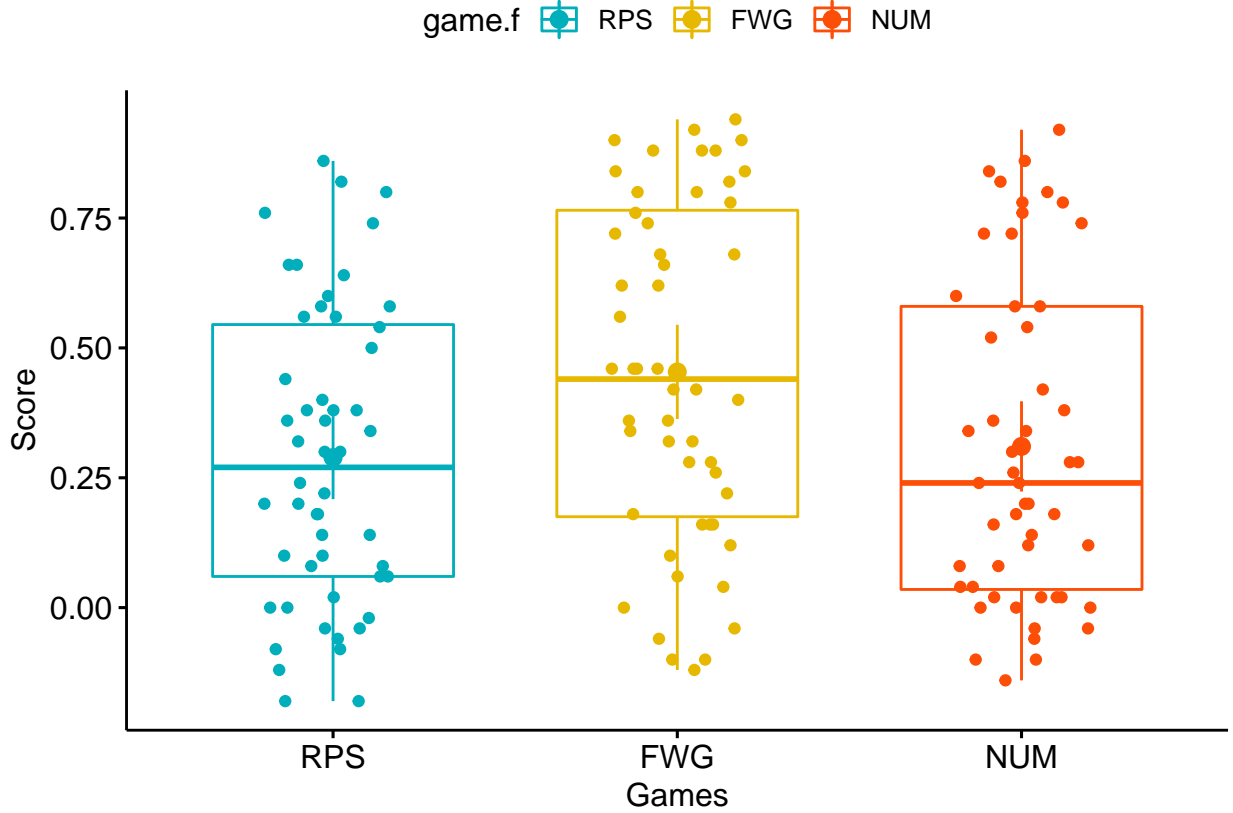
Figure 2. Average scores per game across conditions

225 As players interact more with the agent, they learn how to win more often. The main effect

226 of Condition (F(1,50) = 5.44, p = .024, $\eta^2 = 0.05$) indicates that scores were higher against

227 the level-1 player (M = 0.43) than against the level-2 player (M = 0.27). This means that it

228 was harder for participants, on average, to learn the strategy of the more sophisticated

229 opponent (level-2) compared to that of the comparatively less sophisticated agent (level-1).

230   Finally, the analysis showed a significant interaction effects of block by game (F(2,100)

231 = 6.92 , p = .002, $\eta^2 = 0.02$), indicating that within-game learning differed between the

232 games. Indeed, second half scores in RPS are significantly higher than first-half scores

233 (t(150) = 5.59, p <.0001), while there was no significant difference between block scores for

234 the other two games. This is indicative of the significant within game learning happening in

235 the first game when players have no epxerience against the opponent, as opposed to much
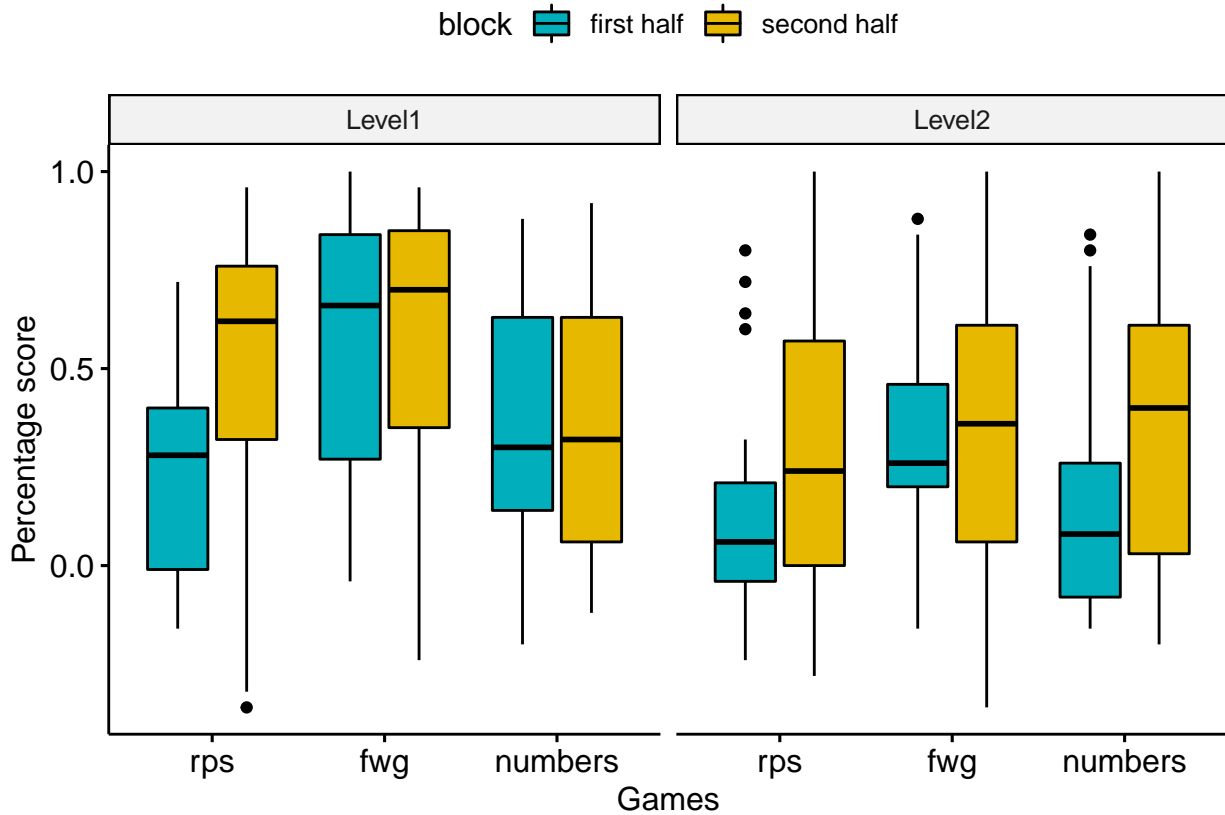
Figure 3. Average scores by game, block and condition

₂₃₆ lower within game performance improvement in the latter games when participants have had

₂₃₇ some exprience playing against the opponenet and start with higher scores indicative of

₂₃₈ transfer. There was also a three-way interaction between condition, game, and block (

₂₃₉ $F(2,100) = 3.88$, p = .023, $\eta^2 = 0.01$), which indicates, as seen in Figure 3 that within-game

₂₄₀ learning changes across games also depend on the sophistication of the opponent. For

₂₄₁ instance, there is more within game learning in the third game against level-2 opponents,

₂₄₂ since the initial scores are lower than agaisnt level-1 opponent. The explanation for this will

₂₄₃ become clearer when we discuss the factors moderating learning transfer in the next section.

₂₄₄ **Learning transfer.** As a measure for learning transfer, we focus on participants'

₂₄₅ scores in the first 5 rounds excluding the initial round (rounds 2-6). We exclude the very first

₂₄₆ round for which the agent is programmed to play randomly as it has no data on prior rounds

on which to build its response, and therefore we cannot expect the participants to exhibit any learned best responses either. A group of players with no experience of the game are expected to have scores not significantly different from 0 over the early rounds of a new game, as was the case in RPS. Any significantly positive group average scores would therefore reflect prior learning from past experiences.

We test the average scores for each game against a hypothesised value of 0 for a non-experienced player using parametric one sample t-tests. As expected for the initial RPS game, the average score is not significantly different from 0 as this is the first game and no learning is possible ($t(148.85) = 1.04$, $p = 0.89$). In FWG, the score is significantly higher than 0 ($t(148.85) = 4.58$, $p < 0.0001$). This is also the case for the more dissimilar game : NUMBERS ($t(148.85) = 3.00$, $p = 0.0092$).

Next, we explore whether learning transfer is moderated by the type of opponent and game similarity. Figure @ref(fig: exp1-score-by-opp) shows the mean scores for rounds 2-6 by game for both level-1 and level-2 facing players. Graphically we can see that the pattern is dissimilar between level-1 and level-2 players, and we suspect transfer to be positively associated with similarity and negatively with degree of sophistication of the agent. To test these hypotheses, we run statistical tests on early round scores by game and opponent against the null hypothesis of 0 (no transfer).

For level-1 facing players, there is evidence of learning transfer from RPS to both FWG ($t(150) = 3.96$, $p < 0.001$) and NUMBERS ($t(150) = 3.74$, $p < 0.001$) . 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 ($t(150) = 2.48$, $p = 0.01$) but not to the dissimilar game of NUMBERS.

Our results when averaging across conditions (previous section) showed that there was indeed evidence for transfer to the more dissimilar game (NUMBERS). We can see from
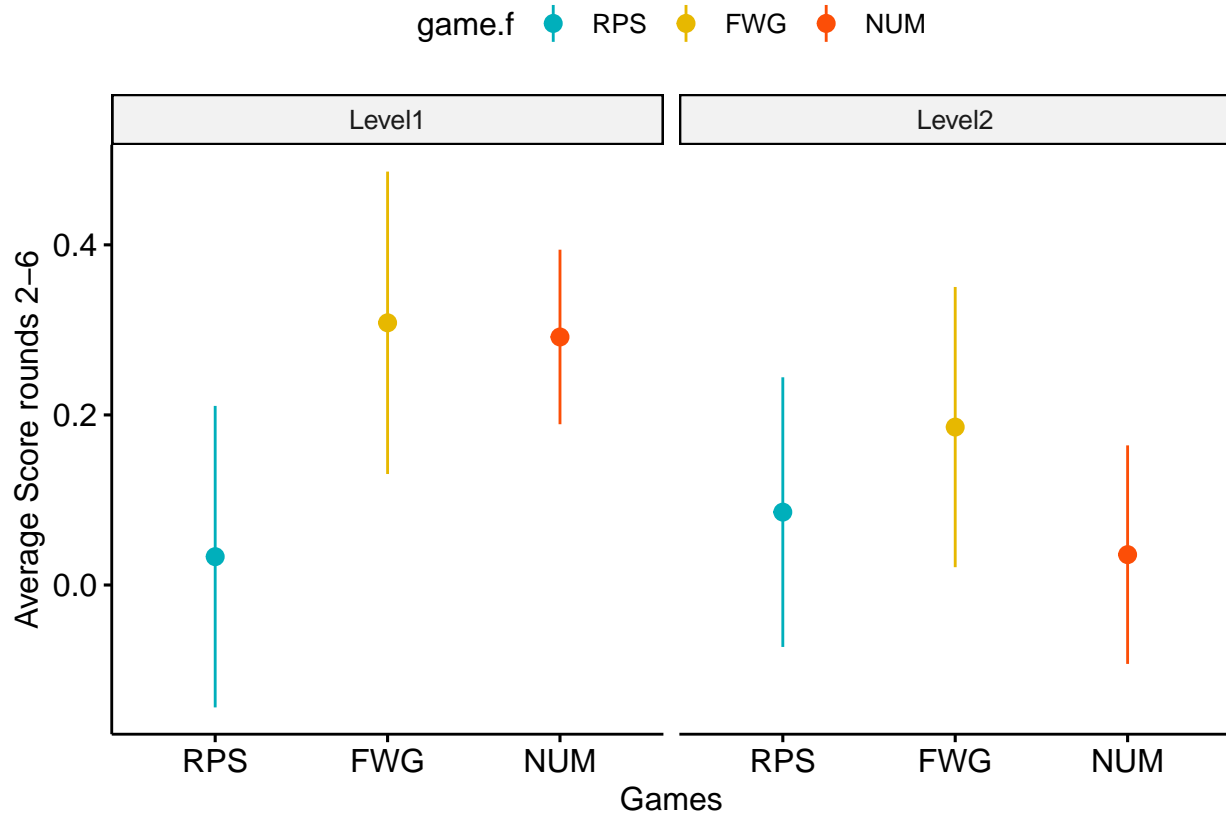
Figure 4. Average early scores by game and type of opponent faced

272 splitting the participants by opponent faced that this transfer is exclusively driven by level-1

273 facing players, as average early round scores of level-2 facing players are close to nil in the

274 NUMBERS game. Therefore, both participants facing level-1 and level-2 agents can transfer

275 learning to the similar game, but only those facing the less sophisticated opponent are able

276 to generalise to the less similar game.

₂₇₇   **Second Experiment**

₂₇₈ **Methods**

₂₇₉   **Participants.**   A total of 48 participants (21 females, 28 males, 1 unknown) used the

₂₈₀ Prolific Academic platform to participate in the experiment. This was a new set of

₂₈₁ participants unrelated to those taking part in Experiment 1. The average age of players was

₂₈₂ 30.2 years, and the mean duration to complete the task was 39 minutes. Participants were

₂₈₃ incentivised using a two-tier payment mechanism: a fixed fee of £2.5 for completing the

₂₈₄ experiment plus a performance linked bonus that averaged £1.32.

₂₈₅   **Procedure.**   The participants each played 3 games sequentially against both level-1

₂₈₆ and level-2 computer opponents, rather than just one like in the first experiment. The three

₂₈₇ games were Rock-Paper-Scissors, Fire-Water-Grass, and the penalty shootout game. The

₂₈₈ first two games were identical to the ones used in the first experiment. In the final game

₂₈₉ (shootout) the participants played the role of the player shooting a football (soccer) penalty,

₂₉₀ and the AI opponent was the goalkeeper. Players had the choice between three actions, like

₂₉₁ in the first two games: Shooting the football to the left, right or centre of the goal. If the

₂₉₂ player shoots in a direction different from that of where the goalkeeper dives, they win the

₂₉₃ round and the goalkeeper loses. Otherwise, the goalkeeper catches the ball and the player

₂₉₄ loses the round. There is no possibility of ties in this game. Figure 5 shows a snapshot of

₂₉₅ play in the shootout game. What makes this game different however is that there are two

₂₉₆ ways to beat the opponent in each round: if we think the opponent is going to choose

₂₉₇ "'right"' in the next round, we can win by both choosing "'left"' and "'center"'. A level-1

₂₉₈ player (thinks that his opponent will repeat his last action) has two ways to win the next

₂₉₉ round. As such, we have programmed the level-2 computer program to choose randomly

₃₀₀ between the two possibilities that a level-1 player may choose.

₃₀₁   Like in the first experiment, the computer opponents retained the same strategy

₃₀₂ throughout the 3 games, however the participants faced each opponent twice in each game.
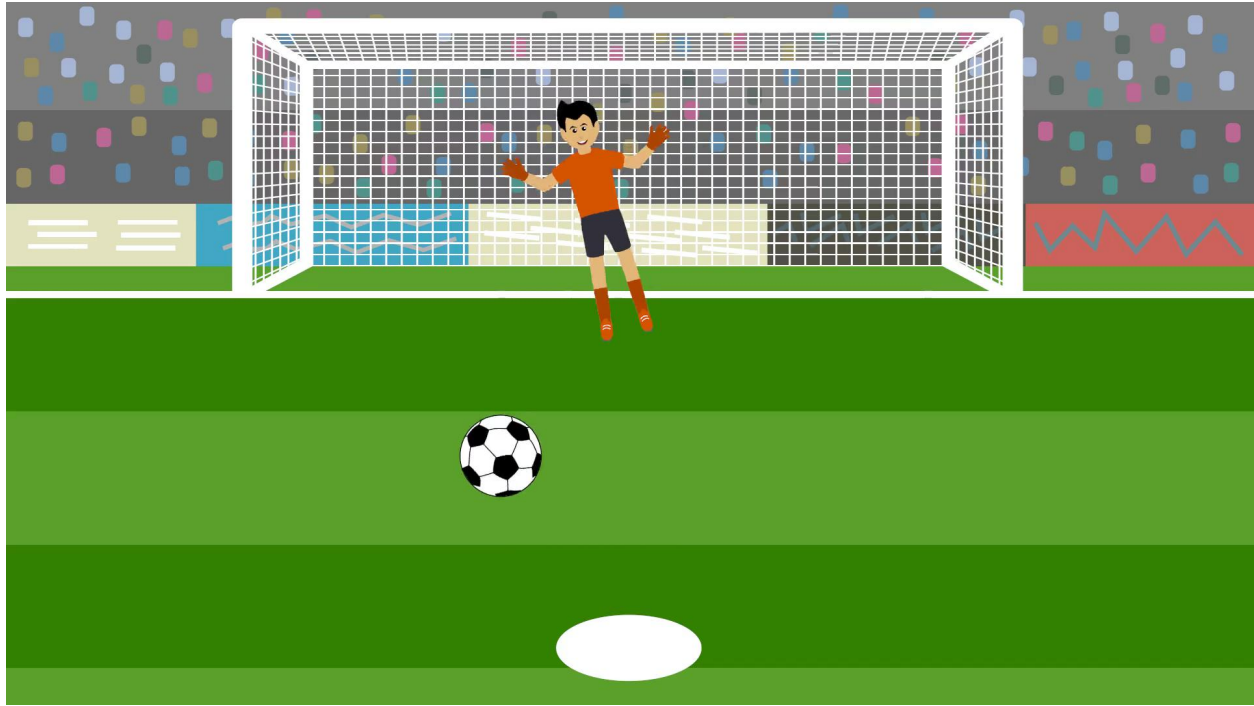
Figure 5. Screenshot of the shootout game

³⁰³ Specifically, each game was divided into 4 stages numbered 1 to 4, consisting of 20, 20, 10,

³⁰⁴ and 10 rounds respectively for a total of 60 rounds per game. Participants start by facing

³⁰⁵ one of the opponents in stage one, then face the other in stage two. This is repeated in the

³⁰⁶ same order in stages 3 and 4. Which opponent they faced first was counterbalanced. All

³⁰⁷ participants engage in the same three games (namely RPS, FWG and Shootout) in this

³⁰⁸ exact order, and were aware that the opponent was not able to know their choices

³⁰⁹ beforehand but was choosing simultaneously with the player.

³¹⁰   In order to encourage participants to think about their next choice, a countdown timer

³¹¹ of 3 seconds was introduced at the beginning of each round. During those 3 seconds,

³¹² participants could not choose an option and had to wait for the timer to run out. A small

³¹³ delay that changed randomly (between 0.5 and 3 seconds) was also introduced in the time it

³¹⁴ took the AI agent to give back their response to make it seem like it is also thinking about

³¹⁵ its last actions, as a way of simulating a real human opponent. After each round, the

316 participants were given detailed feedback about their opponent actions as well as whether

317 they won or lost the last round. Further information about the outcome of previous rounds

318 was also visible on the game screen below the feedback area, throughout each stage game

319 and opponent could go back many rounds to study the history of interaction. The number of

320 wins, losses and ties were clearly shown at the top of the screen for each game, and this

321 scoreboard was reinitialised to zero at the onset of a new stage game.

322   Like games used in the first experiment, all the games used in this seeting have a

323 unique MSNE consisting of randomising across actions. If participants follow this strategy,

324 or simply don't engage in learning how the opponent plays, they would score 0 on average

325 against both level-1 and level-2 players. Evidence of sustained wins would indicate that

326 participants have learned to exploit patterns in the opponent play.

327 **Results**

328   The RPS game had the lowest average score per round (M = 0.194, SD = 0.345)

329 followed by FWG (M = 0.27, SD = 0.394) and finally the Shootout game had the highest

330 average score (M = 0.622, SD = 0.326). The higher score in shootout is expected as there

331 are 2 out of three possible winning actions, compared to one out of three in RPS and FWG.

332 Indeed, a player not aiming to uncover the opponent's strategy and thus choosing to play

333 randomly should be expected to have on average score per round of 0 in both RPS and

334 FWG, and 0.33 in the Shootout game. To make the scores more comparable, and because we

335 are interested in player's performance that is not due to chance, we will adjust all scores in

336 the shootout game by subtracting the average score per round of a random strategy (0.33).

337 Using parametric t-tests on adjusted scores, we reject the null hypothesis of random play in

338 all three games (RPS: $t(49) = 6.26$, p-value < 0.0001 ; FWG: $t(49) = 7.25$, p-value < 0.0001

339 ; Shootout: $t(49) = 13.61$, p-value < 0.0001).

340   Using the average scores obtained by participants in each game and interaction, we

341 explore whether learning has occurred within and between games. We perform a two

342 (condition: level-1 first, level-2 first) by two (opponne type: level-1 or level-2) by three

343 (game: RPS, FWG, Shootout) by two (interaction: first or second) repeated measures

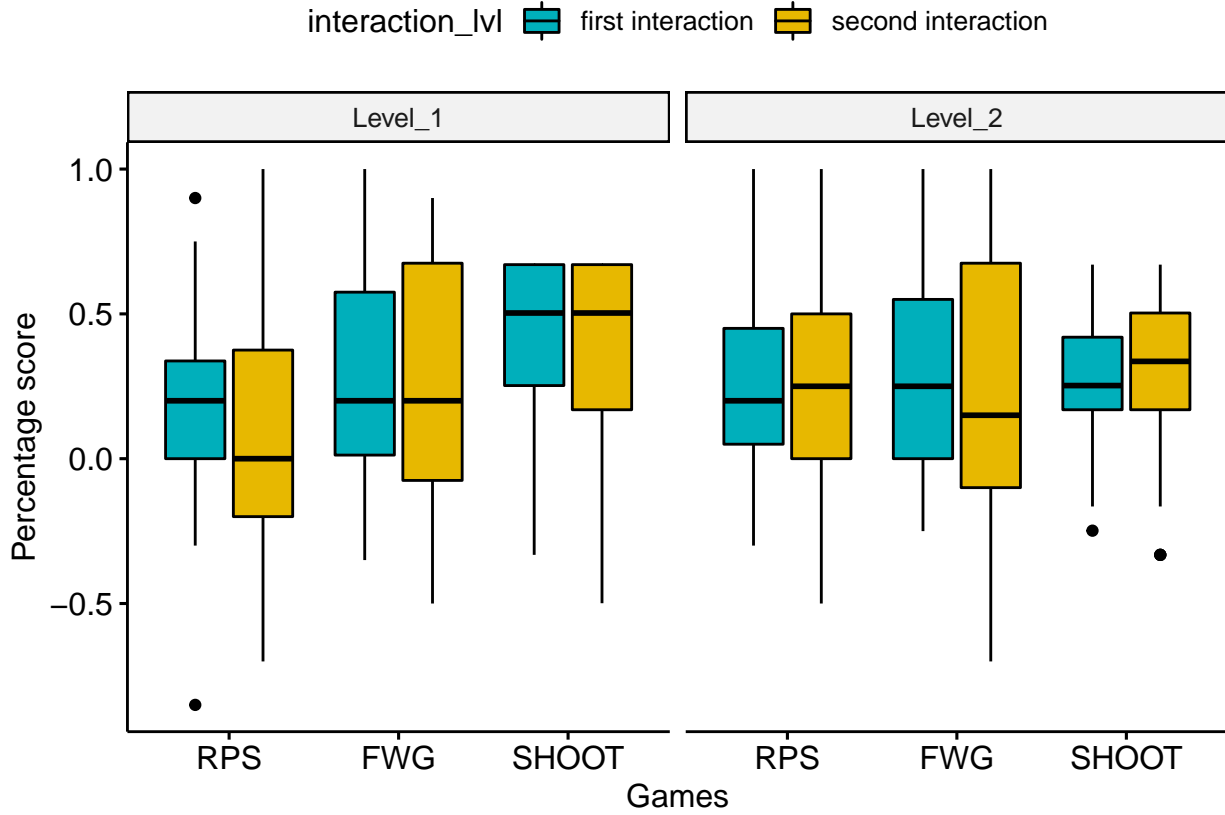344 ANOVA with the first factor varying between participants.

Figure 6. Average scores per game by opponent type

345   There is evidence for a main effect of Game on scores (F(1.85,88.7) = 11.81, $\eta^2 = 0.04$,

346 p < .0001). To explore these differences further, we look at post-hoc analyses for pairwise

347 comparisons between game scores (p-values adjusted using Holm method for multiple

348 comparisons). We find the performance in the games increases steadily throughout the

349 experiment, with FWG performance significantly higher than RPS (t(96) =2.53, p = 0.025),

350 and performance in the Shootout game also significantly higher than in FWG (t(96) = 2.32,

351 p = 0.025). There was no main effect of either opponent type, the interaction factor(first or

second time opponent was faced) , or the condition factor (whether level-1 or level-2

opponent was faced first). This means that neither the type of opponent faced, nor whether

it was the first or second interaction with the opponent nor which opponent was faced first

did have a main effect on performance. There was however a significant interaction effect

between Game and opponent type ($F(1.7, 81.82) = 5.31, \eta^2 = 0.02$, p = .01). As can be seen

in Figure 6, when facing level-1 agents, scores increase steadily after each game, with FWG

score sgnificantly higher than RPS ($t(191) = 2.70$, p = 0.03) and Shootout scores in turn

significantly higher than FWG ($t(191) = 3.05$, p = 0.01). There was no significant difference

between average scores on any two games when facing level-2 agents however.

Learning transfer. As a measure for learning transfer we will again compare scores

only on rounds 2-6 of each game, excluding the very first round where play is necessarily

random. Because the number of rounds here is very limited, there should be very little

learning within games, and we should be better measure learning transfer if it exits.

In Figure 7, we show the average score across participants in rounds 2-6 of the first

interaction with the opponent for each game. These scores are also averaged across the levels

of condition (meaning they are irrespective of which opponent players faced first). For both

the FWG and Shootout games, score in the early rounds of the first interaction are

significantly higher than 0 for both opponent types. (Level-1 opponent: FWG: $t(270) = 4.99$,

p < 0.0001; Shootout: $t(270) = 6.66$, p < 0.0001; Level-2 opponent: FWG: $t(270) = 4.40$, p

< 0.0001; Shootout: $t(270) = 3.21$, p =0.004). Looking more specifically at early scores 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

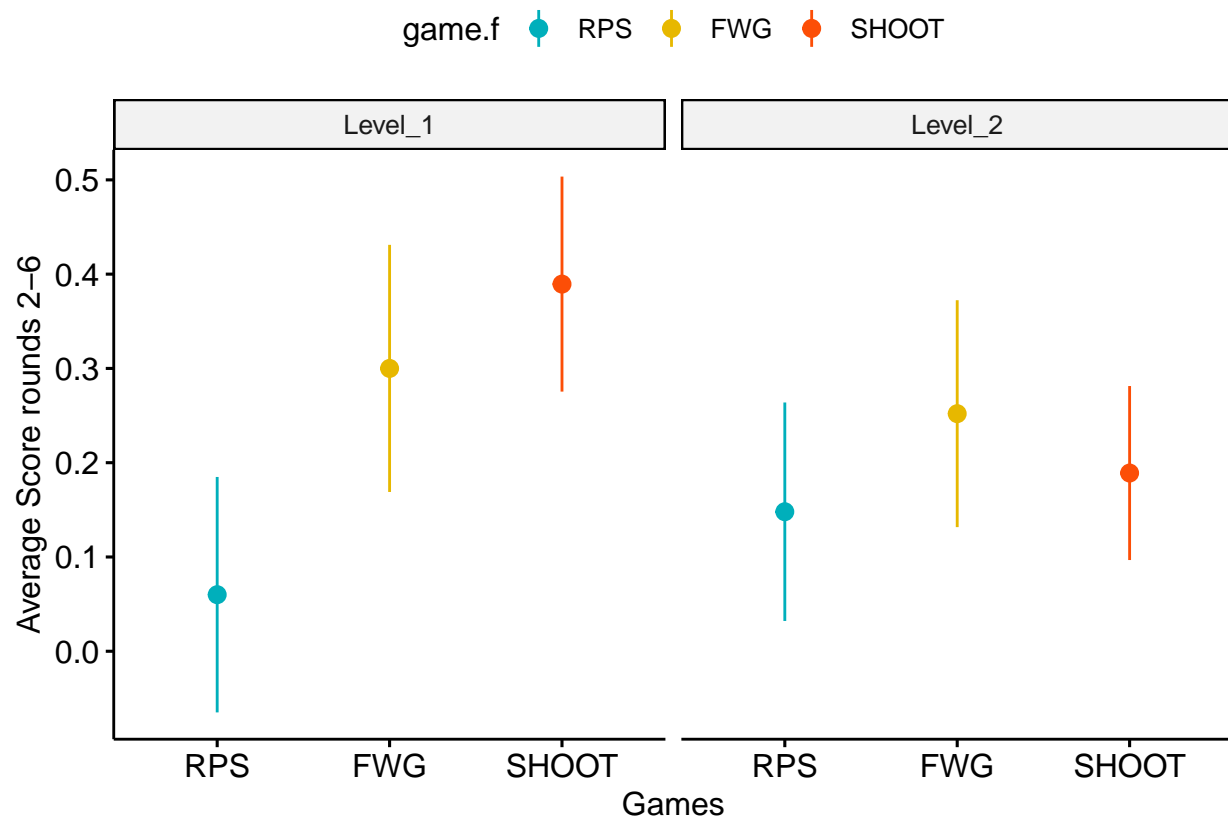also been adjusted to account for the fact that the shootout game has higher average scores

Figure 7. Average early round scores by game for experiment 2

379   when playing randomly, and therefore this difference is really due to better learning transfer

380   and not due to chance.

Computational modelling

382   To gain more insight into how participants played the games against the computer

383   opponents, we compared multiple models of strategies the players may have been using to

384   learn how to beat the opponent. As a base model, we assume play is random, and each

385   potential action is chosen with equal probability. Note that this corresponds to the Nash

386   equilibrium strategy. In this section, we will go through the various models we have used and

387   explain how they update what they learn about the game or the opponent

388   **Reinforcement Learning**

389   We include for comparison purposes a simple model-free reinforcement learning

390   algorithm, that reinforces actions that have led to positive rewards, and conversely lowers

391   the likelihood of choosing actions that led to a negative reward, irrespective of any state. We

392   will use a simple temporal difference learning update rule:

$$V_{t+1}(a) = V_t(a) + \alpha * (R_t - V_t(a))$$

393   Where $V_t(a)$ is the value associated with action a at time t, α is the learning rate and

394   R_t the reward at time t. Actions are chosen based on action values using a softmax choice

395   rule.

396   We extend this very simple model by adding a state space that consists of last round

397   human and agent play. This is akin to using a Q-learning algorithm (Watkins & Dayan,

398   1992). The update rule becomes:

$$Q^{new}(s_t, a_t) = Q(s_t, a_t) + \alpha * (R_t + \gamma * \max_a Q(s_{t+1}, a) - Q(s_t, a_t))$$

399   Where $Q(s_t, a_t)$ is the value of taking action a when in state s at time t, γ is the

discount rate applied to future rewards. For instance, $Q(RS, P)$ denotes the value of taking action "Paper" this round if the player's last action was "Rock" and the opponent played "Scissors". This is a much richer model allowing the players to compute the values of actions conditional on past play.

In experiment 2, we allow this richer model to have a memory of the interaction. In effect, we have a Q-learning model with a state space consisting of last round play, but in one instance, we fit a model (QLS) in which, when encountering an opponent, learning starts anew with no prior knowledge from previous encounters with the opponent used. In another instance, we fit a model in which what was learned about an opponent in the first interaction is transfered to the first round of the second interaction, so there is a within-game trasnfer of Q-values allowed (QLS-within). We fit both models to all participant's data in the second experiment where both opponents are faced.

EWA Models. Next, we use a self-tuning Experience Weighted attraction model (Ho, Camerer, & Chong, 2004). EWA models particularity is that thet nest two seemingly different approaches, namely reinforcement learning and belief learning. Belief based models are based on the assumption that players keep track of the frequency of past plays and best respond to that. In contrast, reinforcement learning does not take into account beliefs about other players, but is such that an action followed by a positive reward is more likely to be repeated than an action followed by a negative reward. The self-tuning EWA model has been shown to perform better than both these nested models in multiple repeated games and has the advantage of having only one free parameter, the inverse temperature in the softmax choice function.

Let's define some notation in order to write the update rule of the self-tuning EWA model. For player i, there are m_i strategies, denoted s_i^j (i.e player i's strategy number j). Strategies actually played by i in period t, are denoted $s_i(t)$, while the opponent's strategy at time t is denoted $s_{-i}(t)$. After playing s_i^j at time t, player i payoff is denoted $\pi_i(s_i^j, s_{-i}(t))$,

426 and the actual payoff the player received is $\pi_i(t)$.

427   The EWA model is based on updating "Attractions" for each action over time. For

428 instance, the attraction of strategy j to player i at time t is written $A_i^j(t)$. Future action

429 choice probabilities are based on these attractions using the softmax playing rule:

$$P_i^j(t+1) = \frac{e^{\lambda . A_i^j(t)}}{\sum_{k=1}^{m_i} e^{\lambda . A_i^k(t)}}$$

430   The attractions are updated over every time step t using the followng update rule :

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

431   Here, I(x,y) is the indicator function equal to 1 if $x = y$ and 0 otherwise. A simple

432 way to think about this update rule is that attractions are multipled by a parameter that

433 represents experince $(N(t-1))$ which is itself decaying by a wight ϕ. The result is added to

434 either the payoff received (when the indicator function is 0), or to δ times the foregone payoff

435 (when indicator function is 1). We can see that setting $\delta = 0$ leads to reinforcement of past

436 actions, while positive and high delta parameters make the update rule take into account

437 foregone payoffs, which is similar to weighted fictitous play (Cheung & Friedman, 1994).

438 While the assumption in expanding the update rule above is that ϕ and δ are free

439 parameters (Camerer, Ho, & Others, 1997), the self-tuning aspect of the model comes from

440 the fact that these are now self-tuned using the formulas expanded in (Ho et al., 2004).

441   **ToM models.** In this set of models, we assume that participants have a belief that

442 the opponent is a level-k agent, with uniform probability of the level k, and use evidence of

443 past play to update their beliefs in a Bayesian way about the true value of k. We use values

444 of k in $0, 1, 2$. We distinguish between multiple ToM models based on their ability to keep

445 what was learned about the opponent in memory and hence facilitate transfer. In a

No-Between-Transfer (NBT) model, players have no memory of what was learned about the opponent and start every new game assuming each level-k has equal probability. In the context of Experiment 2 where players face both opponents, this model assumes that participants transfer learning within the same game, from the first to the second interaction with the opponent, but are not able to transfer that learning to new games (So within but no between trasnfer)

Conversely, In a Between-Transfer model (BT), players are assumed to keep in memory what was learned about the type of opponent faced (vector of probabilities of level-k) and use that at the beginning of each new game. In the context of experiment 2, we still assume that if between transfer is present, then within game transfer is also present (from first to second interaction).

We fit the above two models to experiment 1 data. In experiment two, on top of these two models, we fit two other models: One in which we assume that players don't distinguish between the two opponents faced, so the opponent model vector is never updated (Naive). And finally, we also fit a model in which all stages of the game and all new games start with a fresh uniform probability of level-k opponent (NT), so no within or between opponent model learning transfer.

In summary, we fit two ToM models for experiment 1-play (NBT and BT), and 4 ToM models for experiment 2 (NBT, BT, NT and Naive). In both experiments, all models were fit to each participant data, with optimal parameters being estimated using maximum likelihood. Using information criteria based Bayesian model comparison (BIC), the best fitting model for each participant was chosen and we compared the number of participants whose behavior was best explained by each model.

Experiment 1 modelling :

Figure 8 shows the results for experiment 1: we can see that while some participant's learning behavior was either random or explained by some of the base models, a significant number of participants in experiment one had learning most consistent with Q-learning conditional on last round play.
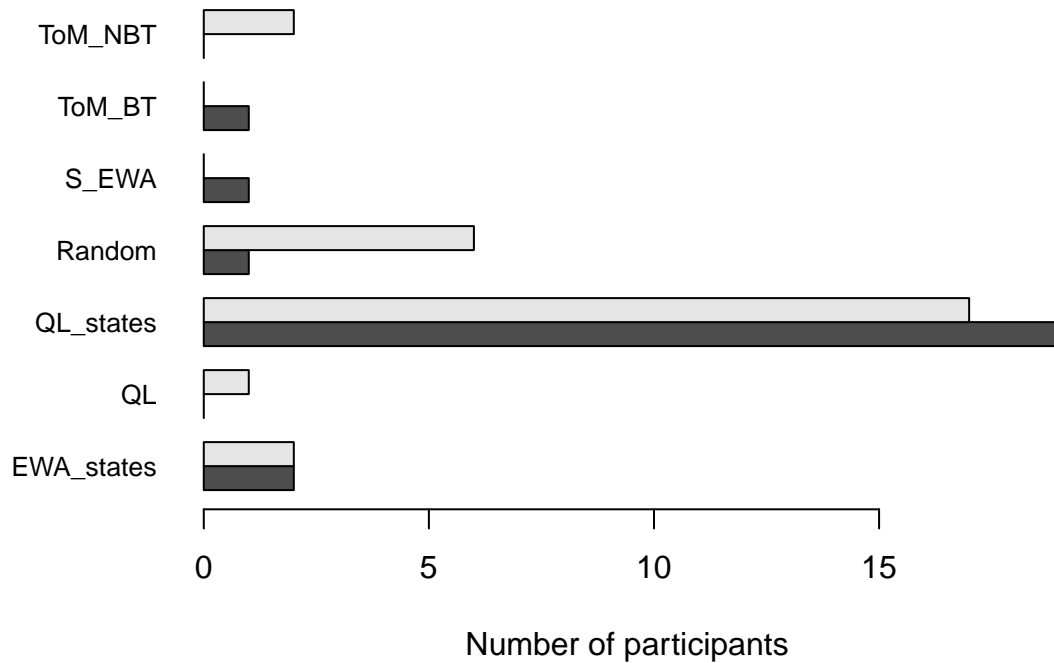

Figure 8. Experiment 1 - Histogram of best fitting computational models by condition

Next we compared the performance of players whose actions are best fit by each of our hypothesized models. Figure 9 shows the average cumulative performance of players across games, for participants grouped by which model best fits their behavior in experiment 1. We can see that participants whose actions are most consistent with learning a ToM opponent model in a Bayesian way had the best overall performance (witout transfer), followed by Q-learning conditional on last round play. EWA, QL and random players had,
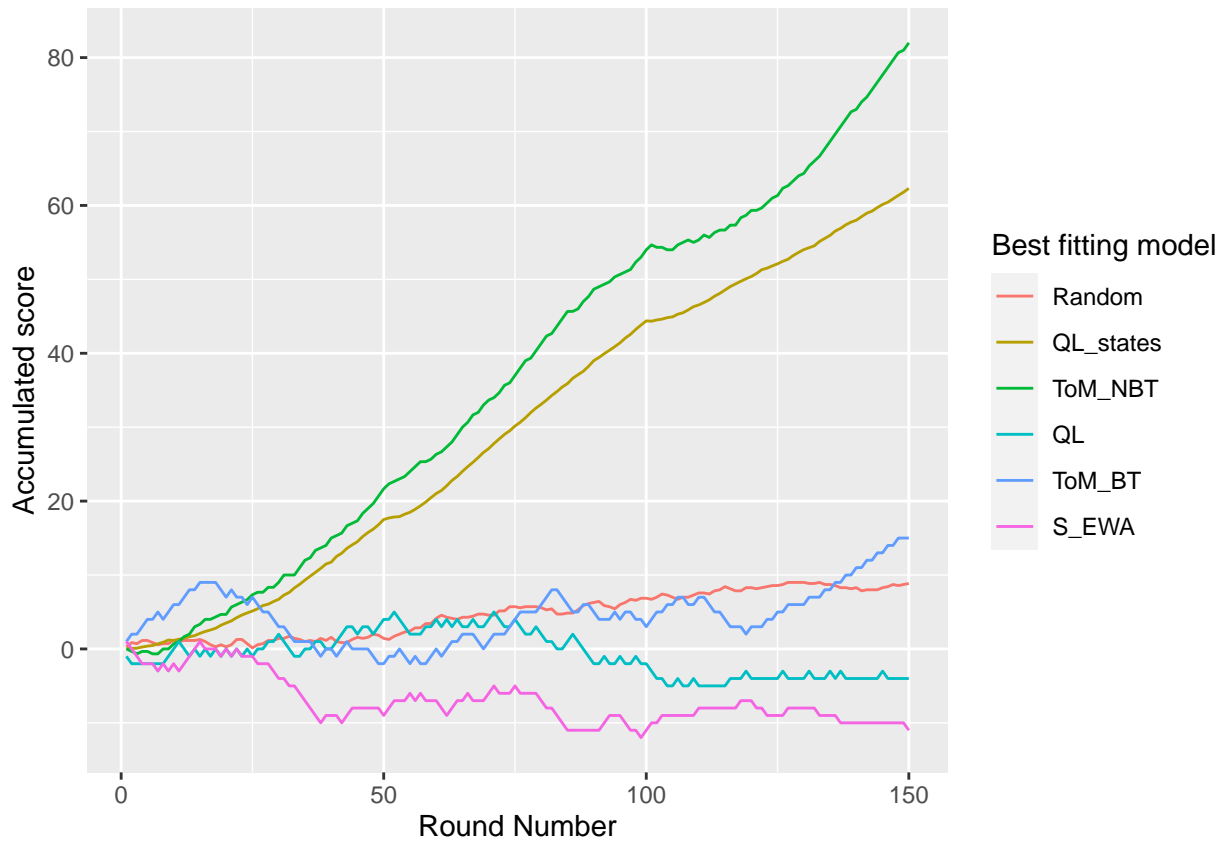
480 understandably the lowest performance.

Figure 9. Experiment 1 - Average cumulative scores of participants by best fitting model

481 However, this is at odds with the findings from the previous section regarding learning

482 transfer: If indeed most participants use Q-learning with states to choose their actions, then

483 they should not be able to transfer learning to the early rounds of the new game. In order to

484 understand better what is going on, we plot the likelihood by trial for each game and each of

485 the three strategies: Q-learning with states, and Theory of Mind models with and without

486 the possibility of across game transfer.

487 Figure 10 shows that in the later games, the likelihood for the ToM models is higher in

488 the initial rounds in which learning transfer is measured, but that over time, the likelihood of

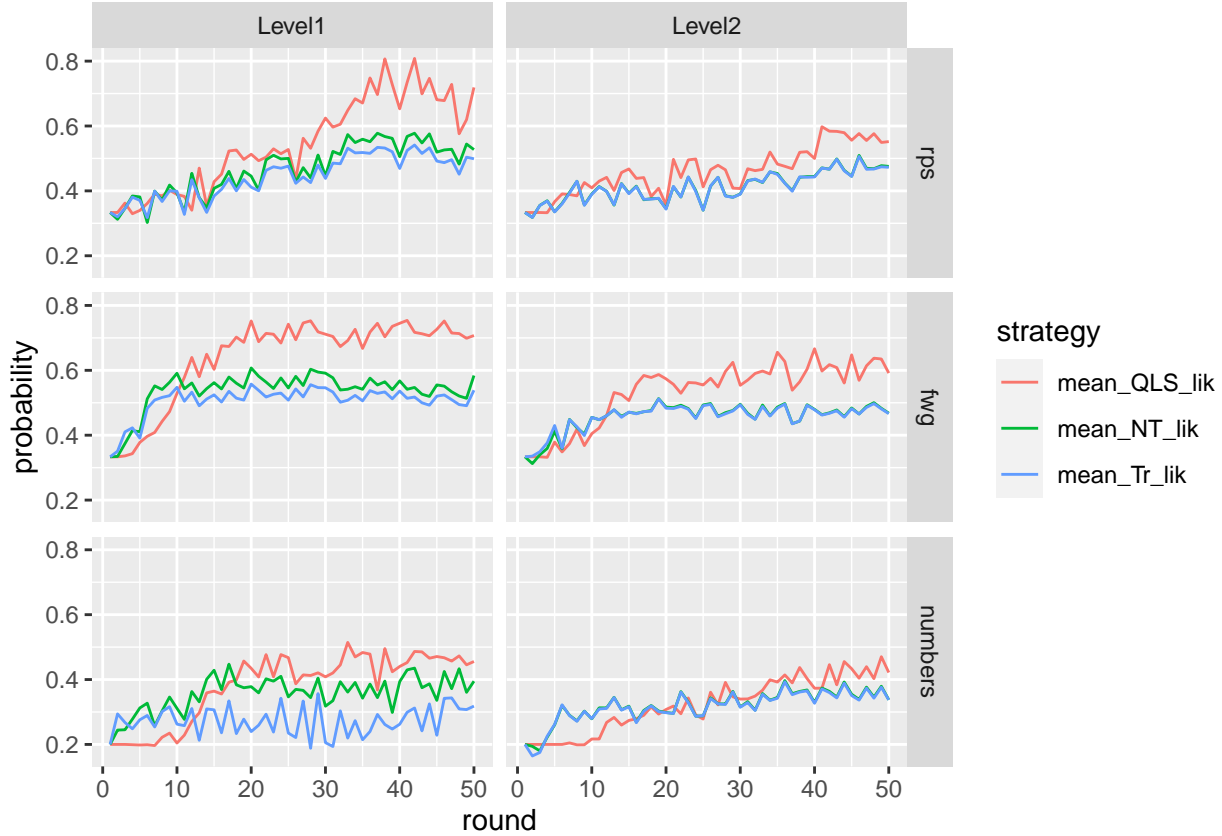489 Q-learning model becomes more important and exceeds that of ToM models.

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

Experiment 2

In experiment 2, we can see from Figure 11 that Q-learning with the aforementioned state-space was again more successful than the bayesian models in fitting player's action choices. In experiment 2 when participants faced both level-1 and level-2 agents sequentially, the bayesian models (with or without transfer) did not fit players observed data as well.

Plotting cumulative scores by best model for experiment 2, we see very similar results looking at Figure 12, in that participants whose behavior was best fit by a ToM model of learning the oppponent strategy had the highest cumulative performance. Out of these ToM models, the one in which there is within-game but no between-game transfer (NBT) had the best cumulative performance (although it only fit 2 participants best), followed by a model in which both within and between trasnfer of opponent models is allowed (BT). The next best
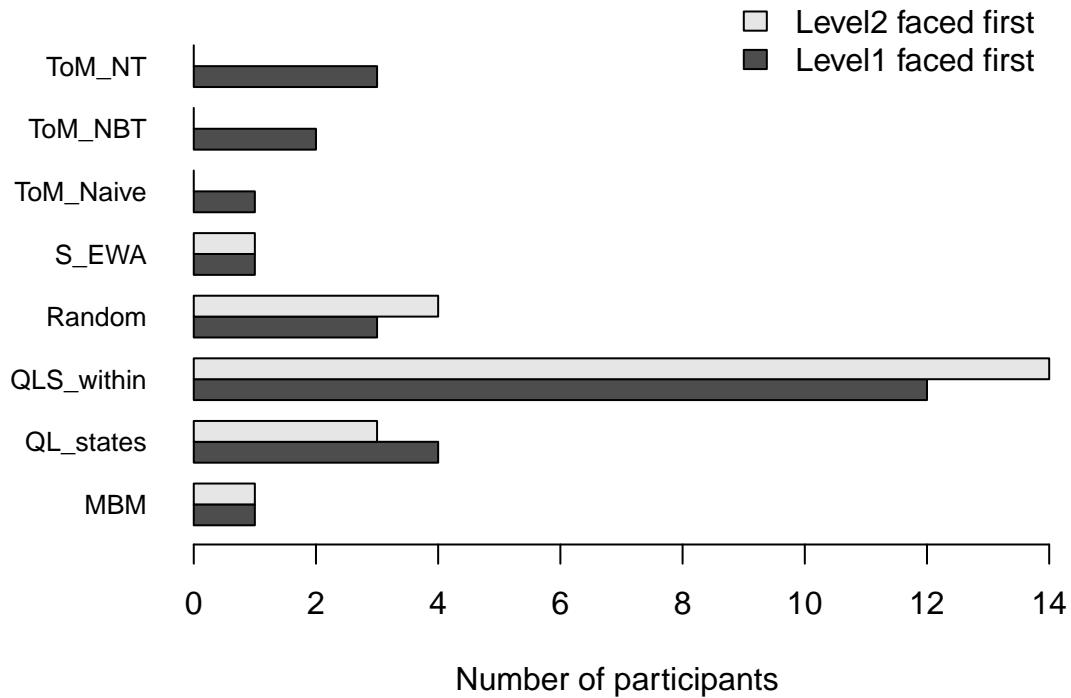
Figure 11. Experiment 2 Histogram of best fitting computational models by condition

₅₀₁ model from a performance perspective was a Q-learning model with states and within game

₅₀₂ transfer, followed by ToM models where either players started learning from scratch at each

₅₀₃ stage of each game (NT) and the model in which they didn't distinguish between opponenets

₅₀₄ (Naive). As expected, random play was at the bottom of cumulative performance.

₅₀₅   As in experiment 1, we want to understand the dynamic of strategy choice by plotting

₅₀₆ the likelihood by trial for each strategy, using the optimal parameters found when fitting the

₅₀₇ model. Figure 13 shows that, as in experiment 1, ToM models had higher likelihood in the

₅₀₈ early stages of the second (most similar) game, however the likelihood of Q-learning with

₅₀₉ states models increases steadily to be the highest in the later stages of all games. In the

₅₁₀ third and mosre dissimilar game, we get a result that is different from experiment 1. In this

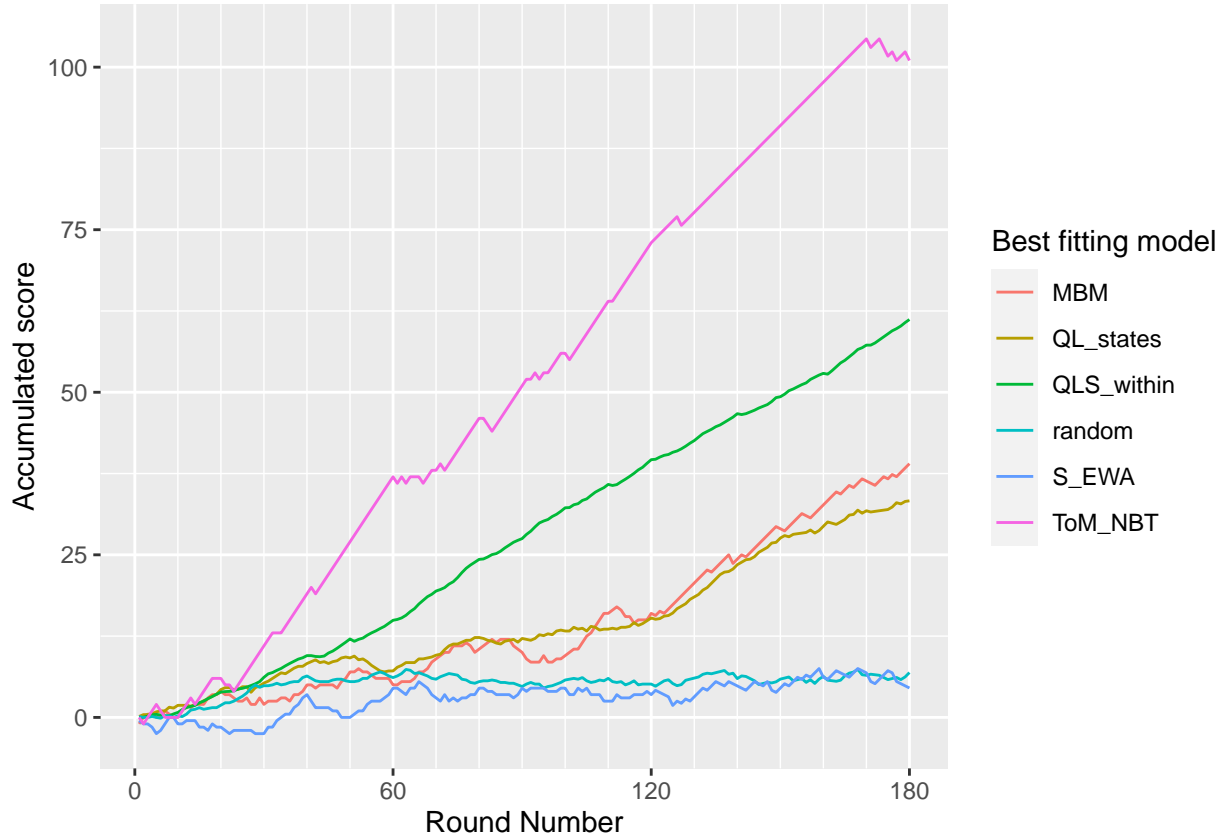₅₁₁ instance, the likelihoods of the ToM models stay constant and close to their initial values.

Figure 12. Experiment 2 Average cumulative scores of participants by best fitting model

Using Hidden Markov Model to explore strategy switching

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

In order to test for the existence of strategy switching in participants' play, we fit Hidden Markov Models in which the latent states are the 3 strategies used (Q learning with state space consisting of previous round play, ToM based model with and without opponent model transfer). Hidden Markov models are useful tools to explore structure in observed time series. They are named because of two properties: First, they make the assumption
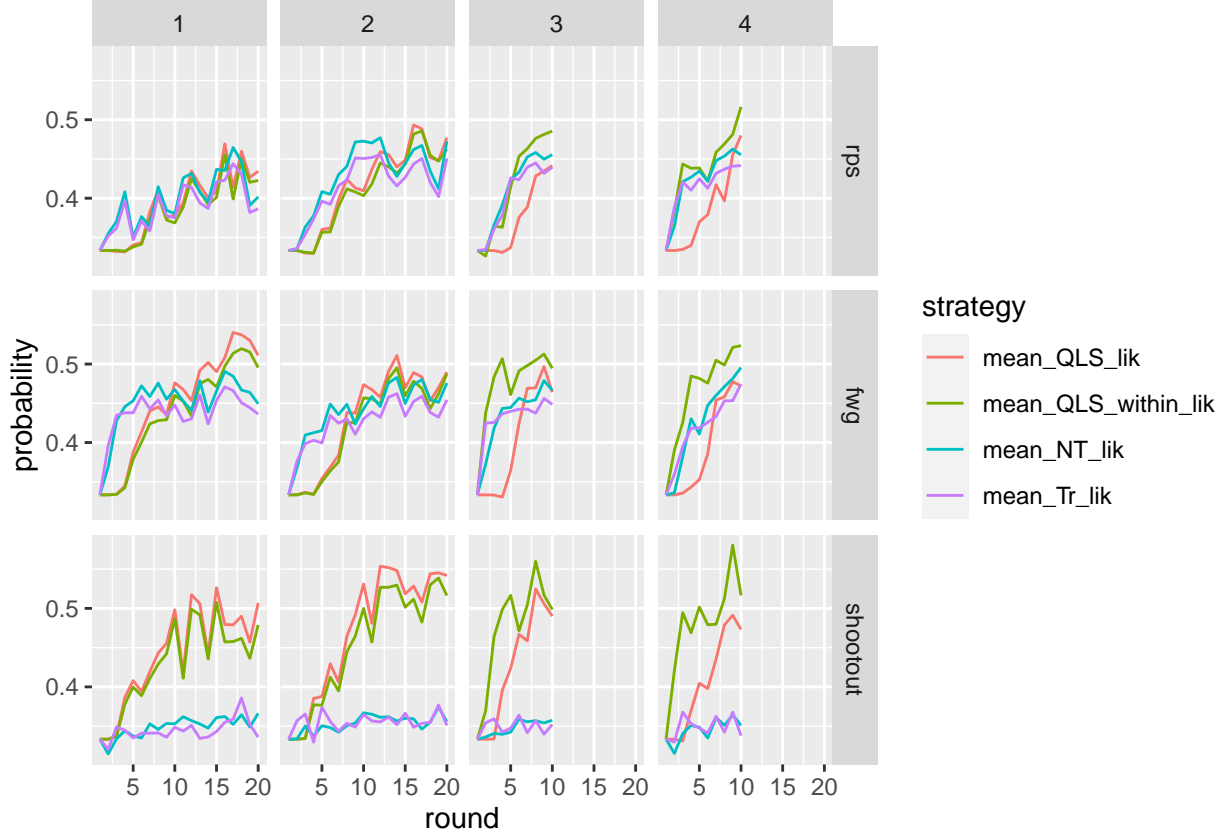
Figure 13. Experiment2 likelihood by trial by game and opponent faced

523 that any observable action at time t results from a process whose state at time t , named S_t

524 is "hidden" from the observer. Second, it also assumes that this hidden process has a Markov

525 property, meaning that given state S_{t-1}, the value of S_t is independent of all states occuring

526 before time $t-1$. We also assume that S_t has a discrete probability distribution in that it

527 take one of K discrete values. The model is therefore specified by initial probabilities of

528 being in each state $1, 2, ..., K$ and transition probabilities for moving from state i to state j.

529 These probabilities are fit using obsserved actiosn generated from these hidden states.

530 sing the R DepmixS4 package, we fit a total of two HMMs. In the first model, we allow

531 for a non-nil probability of players transitioning from one state (strategy) to another. In the

532 second model, we assume that such switching does not happen, and as such assume implicitly

533 that when players start with a particular strategy, they continue using it throughout the

534 game. We then compare the likleihoods of each HMM model using a likelihood ration test.
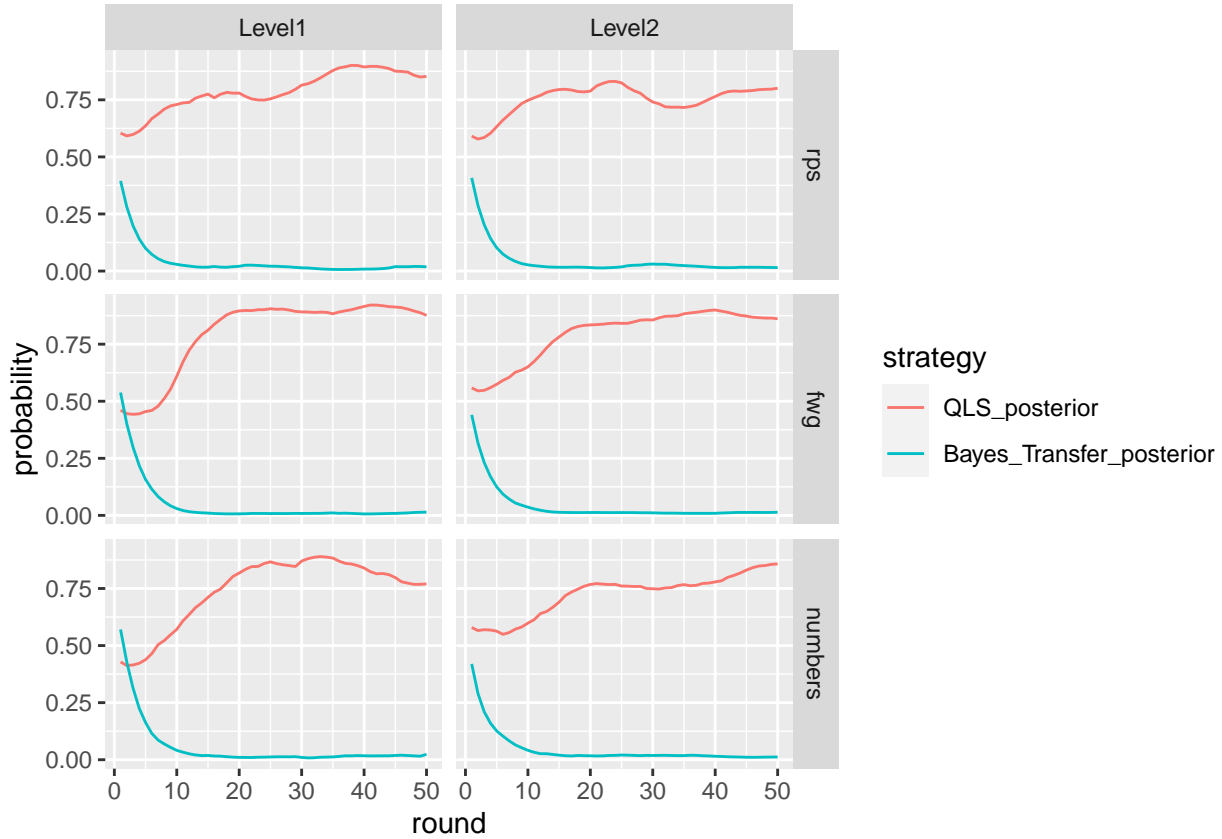

Figure 14. Experiment1 posterior probability of strategies by game and opponent faced

535   **Experiment 1:** In experiment 1, the HMM model with swithcing has a significantly

536 higher likelihood than the non-switching one (p-value = 0.00). This is further statistical

537 evidence in favour of the hypothesis that participants switch between strategies. In order to

538 understand at which stage of the games the switching might happen, and whether there are

539 any differences between games and type of opponents faced, we plot in Figure 14 the average

540 (across participants) posterior probabilities given our HMM trial by trial for each game and

541 opponent faced. Th eposterior probability is the probability that an observation comes from

542 a component distribution a posteriori, i.e. given the value of the observation. In the first

543 experiment, we can see from the plots of fwg and numbers games for level-1 opponent that

544 although the likelihoods are very close, the posterior probability of the Bayesian model with

545 transfer is slightly higher than that of the QLS model in the very early rounds, but decreases

₅₄₆ rapidly while the posterior probability of the QL-learning with states models keeps
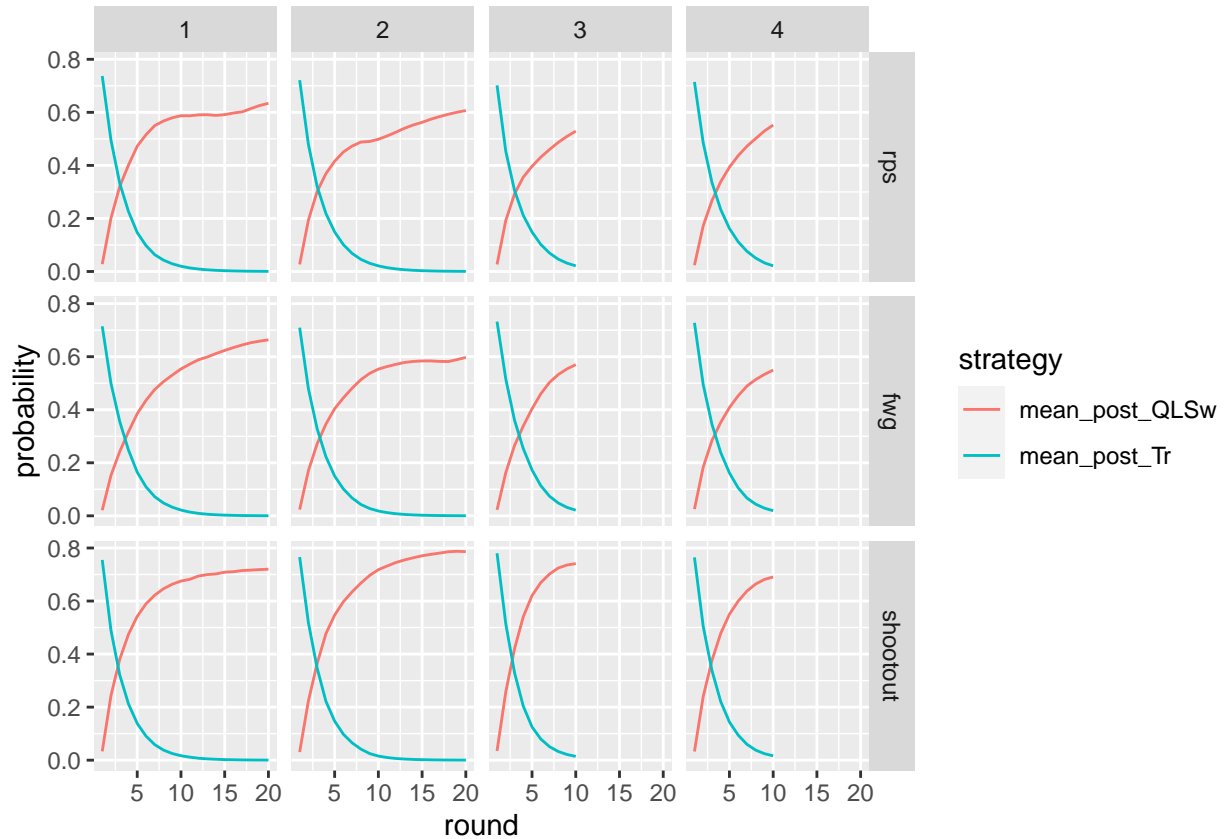
₅₄₇ increasing.

Figure 15. Experiment2 likelihood by trial by game and opponent faced

₅₄₈ **Experiment 2:** The switching model in experiment two has also significantly higher

₅₄₉ likelihood (p-value = 0.00). On top of indications from looking at the liklihood by trial

₅₅₀ graphs, we have therefore further evidence that participants did indeed switch their

₅₅₁ strategies as the games progressed. The posterior probability plot in Figure 15 shows

₅₅₂ switching much more clearly across games ans stages, indicating a robust process. The

₅₅₃ switchign also seems to happen very early one at the beginning of each game and stage, and

₅₅₄ is also consistently in the same direction: The probability of Bayesian models with transfer

₅₅₅ being initially high, then decreasing rapidly while the posterior probability of QL-Learning

₅₅₆ with states and within transfer learning increases rapidly.

₅₅₇   **Discussion**

₅₅₈   In this study, we investigated human learning transfer across games by making human

₅₅₉ participants play against computer agents with rule-based level-k strategies. We were

₅₆₀ interested in uncovering evidence for transfer and exploring whether it is modulated by the

₅₆₁ degree of similarity between games and the sophistication of the agent.

₅₆₂   The results of our first online experiment show that the majority of participants learn

₅₆₃ to adapt to the opponent strategy over multiple interactions and generalise this learning to

₅₆₄ the similar game. We found that using results on very early rounds is a better proxy for

₅₆₅ measuring transfer as it is not tainted by any within game learning. Using this approach, we

₅₆₆ showed that transfer to the more dissimilar game was modulated by the degree of

₅₆₇ sophistication of the agent, with evidence for transfer when players face the less

₅₆₈ sophisticated agent but not the more sophisticated one.

₅₆₉   In the second online experiment, the paradigm was better suited to explore our main

₅₇₀ research question, related to learning transfer rather than within-game learning. Indeed,

₅₇₁ when facing two opponents sequentially, there are many more opportunities to test transfer

₅₇₂ than before. Indeed, there are 2 opportunities to transfer knowledge within each game, and a

₅₇₃ total of three games, which means 6 opportunities to test transfer. When we made players

₅₇₄ face only one type of opponent each, we only had two possible learning transfers to test. In

₅₇₅ that regard, the results on learning transfer confirmed prior findings from the first

₅₇₆ experiment. While there was no evidence of learning transfer across interactions within the

₅₇₇ same game (likely due to the lower number of rounds per interaction and the higher

₅₇₈ cognitive load of facing two opponents rather than one), we found evidence for learning

₅₇₉ transfer across games as early round scores analysis confirmed. We also found that learning

₅₈₀ transfer is modulated by the type of opponent faced. When the players faced the level-1

₅₈₁ opponent, they were able to transfer learning. However, when they faced the level-2

₅₈₂ opponent, there was weaker evidence for transfer. The lack of transfer when facing the more

sophisticated opponent might be due to the difficulty of learning that opponent strategy to start with. A player cannot transfer what they have not learnt and as such, since it might be harder to learn the strategy of the level-2 opponent, this in turn might translate into weaker evidence for transfer.

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

We will proceed by considering multiple potential answers to this question. First, maybe players simply learn spatial heuristics that allow them to perform better than chance. An example is a spatial heuristic that consists of choosing "weapons" in a particular order (for instance left to right). This was one of the weaknesses in the design of experiment one, as it was indeed possible using very simple spatial sequences to beat the opponent on most rounds. We took this into account in designing experiment two by randomly shuffling the spatial order of action choices in each round. Still, the learning and conclusions were similar, so this could not explain both learning and transfer in experiments one and two.

A second possible hypothesis for learning the opponent's strategy is the use of simple rules based on last round play (for instance, I play scissors whenver opponent played rock in last round, or whenever the last round play was rock/scissors, I should play paper in this round, etc. . .). Our Q-learning with states as prior-round play model is a good proxy for this type of strategies. While this approach certainly seemed to be the best fit for some player's behavior, it is unsatisfactory in explaining some of the learning transfer evidence we showed. Indeed, learning the best action in a particular state cannot easily transfer to a new game since the state space is different and there is no simple mapping between the state spaces of

⁶⁰⁹ the initial and latter games. These rules would thereofre need to be learned anew in the

⁶¹⁰ latter game which is inconsistent with above chance performance in very early rounds.

⁶¹¹ Likewise, assuming that players learn a complete model of the environment (for

⁶¹² instance the transition probabilities from last round play to new play) might explain learning

⁶¹³ within games but is equally unable to account for early games transfer of learning as such

⁶¹⁴ models, besides being cognitively very expensive to learn, would require many rounds of

⁶¹⁵ practice. Another issue with these hypotheses is that they are not consistent with significant

⁶¹⁶ score differences between those facing level-2 and level-1 opponents. After all, if players were

⁶¹⁷ using some type of associative learning or spatial heuristics, then their scores should not

⁶¹⁸ depend on the degree of strategic sophistication of the opponent since their approaches would

⁶¹⁹ render this variable irrelevant. The fact that the degree of sophisitication of the opponent

⁶²⁰ matters points to the importance of opponent modelling to successful transfer of learning.

⁶²¹ We are left with two possible explanations: First, it is possible that the players have

⁶²² uncovered a heuristic that allows them to beat the opponent without explicitly modelling

⁶²³ their strategy, and is robust to transfer. Indeed, because of the cyclicality in action choices

⁶²⁴ (e.g : Rock beats Scissors beats Paper beats Rock), it is possible to beat level-2 opponents

⁶²⁵ most of the time by following a simple rule: Play in the next round whatever the opponent

⁶²⁶ played in the last round. This is a rule that wins and is also robust to transfer as it does not

⁶²⁷ depend on action labels and even works in the dissimilar game.

⁶²⁸ The second explanation of learning transfer is that it is driven by a group of

⁶²⁹ participants that are able to build a mental representation of what the strategy of the

⁶³⁰ opponent is. A successful mental representation would take the perspective of the opponent

⁶³¹ or endow it with intentionality in order to detect its strategy when the opponent is playing

⁶³² based on a level-k reasoning model. For instance, the player may think "My opponent is

⁶³³ always trying to be one step ahead of me, therefore, I will be one step ahead of where it

⁶³⁴ thinks I will be". This mental representation would facilitate the use of theory of mind

abilities and thus enable the players to learn opponent strategies when they are based on human-like reasoning models such as level-k or cognitive hierarchy. This type of learning would be deemed "explicit" in the psychology literature as a process through which knowledge consists of cognitive representations of concepts and rules, as well as the relationship between them. It involves the evaluation of explicit hypotheses and results in better problem-solving skills (Mandler, 2004). Since it is less context dependent, this type of learning is generalizable to new situations, akin to the more general framework of rule-based learning explored by Stahl (2000, 2003).

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

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

Our computational modelling allowed us to delve deeper into what might be driving the observed learning transfer. Initial modelling of observations using all available data seems to indicated that the most likely model was a Q-learning type model. However, as we argued above, that would be inconsistent with the evidence for learning transfer. Breaking down likelihoods by trial and fitting a hidden markov model to the data with states being

the various strategies that participants are assumes to be using, we showed evidence for within game switching of strategies. Participants start the early rounds of a new game acting in a way consistent with a Bayesian Theory of mind level, which would be accurate and generalisable but computationally expensive. However, as trials continue, participants seem to switch to a habitual type of learning (QL-models).

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

Conclusion

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

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

References

Arad, A., & Rubinstein, A. (2012). The 11-20 money request game: A level-k reasoning study. *American Economic Review, 102*(7), 3561–3573.

Baarslag, T., Hendrikx, M. J. C., Hindriks, K. V., & Jonker, C. M. (2016). Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques. *Autonomous Agents and Multi-Agent Systems, 30*(5), 849–898.

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

Cheung, Y.-W., & Friedman, D. (1994). *Learning in evolutionary games: some laboratory results.* University of California, Santa Cruz.

Cooper, D. J., & Kagel, J. H. (2003). Lessons learned: generalizing learning across games. *American Economic Review, 93*(2), 202–207.

Cooper, D. J., & Kagel, J. H. (2008). Learning and transfer in signaling games. *Economic Theory, 34*(3), 415–439. doi:10.1007/s00199-006-0192-5

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

Haruvy, E., & Stahl, D. O. (2012). Between-game rule learning in dissimilar symmetric normal-form games. *Games and Economic Behavior, 74*(1), 208–221. doi:10.1016/j.geb.2011.06.001

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

Ho, T.-H., Camerer, C., & Weigelt, K. (1998). Iterated dominance and iterated best

response in experimental" p-beauty contests". *The American Economic Review*, *88*(4), 947–969.

Juvina, I., Saleem, M., Martin, J. M., Gonzalez, C., & Lebiere, C. (2013). Reciprocal trust mediates deep transfer of learning between games of strategic interaction. *Organizational Behavior and Human Decision Processes*, *120*(2), 206–215. doi:10.1016/j.obhdp.2012.09.004

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*. doi: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.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and Organization*, *1*(1), 161–176.

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.

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

Weerd, H. de, Verbrugge, R., & Verheij, B. (2012). Higher-order social cognition in rock-paper-scissors: A simulation study. In *Proceedings of the 11th international conference on autonomous agents and multiagent systems-volume 3* (pp. 1195–1196).

Woodruff, G., Sciences, D. P. B., Brain, & 1978. (n.d.). Does the chimpanzee have a theory of mind.

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

Arad, A., & Rubinstein, A. (2012). The 11-20 money request game: A level-k reasoning study. *American Economic Review*, *102*(7), 3561–3573.

Baarslag, T., Hendrikx, M. J. C., Hindriks, K. V., & Jonker, C. M. (2016). Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques. *Autonomous Agents and Multi-Agent Systems*, *30*(5), 849–898.

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

Cheung, Y.-W., & Friedman, D. (1994). *Learning in evolutionary games: some laboratory results.* University of California, Santa Cruz.

Cooper, D. J., & Kagel, J. H. (2003). Lessons learned: generalizing learning across games. *American Economic Review*, *93*(2), 202–207.

Cooper, D. J., & Kagel, J. H. (2008). Learning and transfer in signaling games. *Economic*

765   *Theory, 34*(3), 415–439. doi:10.1007/s00199-006-0192-5

766 Goodie, A. S., Doshi, P., & Young, D. L. (2012). Levels of theory-of-mind reasoning in

767   competitive games. *Journal of Behavioral Decision Making, 25*(1), 95–108.

768 Haruvy, E., & Stahl, D. O. (2012). Between-game rule learning in dissimilar symmetric

769   normal-form games. *Games and Economic Behavior, 74*(1), 208–221.

770   doi:10.1016/j.geb.2011.06.001

771 Ho, T. H., Camerer, C. F., & Chong, J.-K. (2004). The economics of learning models: A

772   self-tuning theory of learning in games.

773 Ho, T.-H., Camerer, C., & Weigelt, K. (1998). Iterated dominance and iterated best

774   response in experimental" p-beauty contests". *The American Economic Review, 88*(4),

775   947–969.

776 Juvina, I., Saleem, M., Martin, J. M., Gonzalez, C., & Lebiere, C. (2013). Reciprocal trust

777   mediates deep transfer of learning between games of strategic interaction.

778   *Organizational Behavior and Human Decision Processes, 120*(2), 206–215.

779   doi:10.1016/j.obhdp.2012.09.004

780 Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building

781   machines that learn and think like people. *Behavioral and Brain Sciences, 40*.

782   doi:10.1017/S0140525X16001837

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

784   Elsevier.

785 Shachat, J., & Swarthout, J. T. (2004). Do we detect and exploit mixed strategy play by

786   opponents? *Mathematical Methods of Operations Research, 59*(3), 359–373.

787 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.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and Organization*, *1*(1), 161–176.

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.

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

Weerd, H. de, Verbrugge, R., & Verheij, B. (2012). Higher-order social cognition in rock-paper-scissors: A simulation study. In *Proceedings of the 11th international conference on autonomous agents and multiagent systems-volume 3* (pp. 1195–1196).

Woodruff, G., Sciences, D. P. B., Brain, & 1978. (n.d.). Does the chimpanzee have a theory of mind.

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