Transfer of Learned Opponent Models in Zero Sum Games

Ismail Guennouni (i.guennouni.17@ucl.ac.uk)
Maarten Speekenbrink (m.speekenbrink@ucl.ac.uk)
Department of Experimental Psychology, University College London

Abstract

Human learning transfer takes 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 role in learning transfer by letting human participants play different games against the same computer agent, who possesses human-like theory of mind abilities with a limited degree of iterated reasoning. We find that participants deviate from Nash equilibrium play and learn to adapt to the 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 level of sophistication of the opponent. Computational modelling shows that it is likely that players start each game using a model-based learning strategy that facilitates generalisation and opponent model transfer, but then switch to behaviour that is consistent with a model-free learning strategy in the later stages of the interaction.

Keywords: Opponent modelling; Zero-sum games, Learning transfer; Hidden Markov models

Introduction

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

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

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

We measure transfer of learning about the opponent's strategy between games with varying degrees of similarity. The first two games we use are structurally identical, except for action labels. The third game is strategically similar to the first two, but descriptively different. Participants face the same opponent throughout the three games, and the opponents are randomised to be either level-1 or level-2 players.

Method

Participants and design

A total of 52 (28 female, 24 male) participants were recruited on the Prolific Academic platform. The mean age of participants was 31.2 years. Participants were paid a fixed fee of £2.5 plus a bonus dependent on their performance which averaged £1.06. The experiment used a 2 (computer opponent: level 1 or level 2) by 3 (games: rock-paper-scissors,

fire-water-grass, numbers) design, with repeated measures on the second factor. Participants were randomly assigned to one of the two levels of the first factor.

Task

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 whether learning is transferred in a fundamentally similar game where the only difference is in the description of the possible actions. This should make it relatively easy to generalize knowledge of the opponent's strategy, provided this knowledge is on a sufficiently abstract level, such as knowing the opponent is a level 1 or 2 player. Crucially, learning simple contingencies such as "If I played Rock on the previous round, playing Scissors next will likely result in a win", as might be learned by a simple reinforcement learning algorithm, will not be able to generalize to such a game, as these contingencies are tied to the labels of the actions. The numbers game is a generalization of RPS. In the variant we use, 2 participants concurrently pick a number between 1 and 5. To win in this game, a participant needs to pick a number exactly 1 higher than the number chosen by their opponent. For example, if a participant thinks their opponent will pick 3, they ought to choose 4 to win the round. To make the strategies cyclical as in RPS, the game stipulates that the lowest number (1) beats the highest number (5), so if the participant thinks the opponent will play 5, then the winning choice is to pick 1. This game has a structure similar to RPS in which every action is dominated by exactly one other action. All other possible combinations of choices are considered ties. Similar to RPS and FWG, the mixed-strategy Nash equilibrium is to play each action with equal probability in a random way.

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

Procedure

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

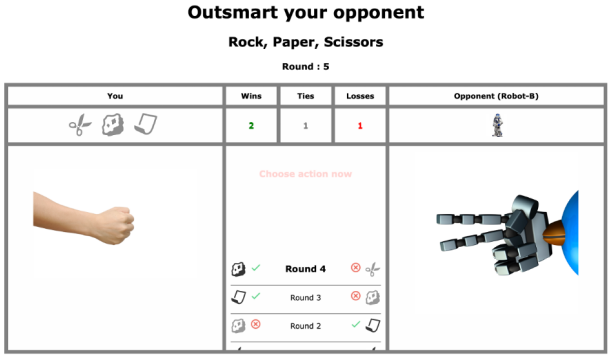

Figure 1: Screenshot of the Rock-Paper-Scissors game

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

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

Results

On average, participants obtained the lowest score in the RPS game ($M = 0.289$, $SD = 0.348$), followed by NUMBERS ($M = 0.31$, $SD = 0.347$). Participants' performance was highest in the FWG game ($M = 0.454$, $SD = 0.354$). Scores in each game were significantly different from 0, the expected score of random play (RPS: $t(51) = 7.26$, $p < .001$; FWG: $t(51) = 10.04$, $p < .001$; NUMBERS: $t(51) = 7.17$, $p < .001$). To assess learning within and between games, we used a 2 (condition: level-1, level-2) by 3 (game: RPS, FWG, NUMBERS) by 2 (block: first half, second half) repeated-measures ANOVA, with the first factor varying between participants. This showed a main effect of Game ($F(2,100) = 8.54$, $\eta^2 = 0.05$, $p < .001$), indicating that average scores varied significantly over the games. Post-hoc pairwise comparisons showed that performance in the FWG game was significantly higher than in the RPS game ($t(100) = 3.78$, $p < .001$) and the NUMBERS game ($t(100) = 3.32$, $p = .002$). The score in RPS was not significantly different from the score in NUMBERS ($t(100) = 0.45$, $p = .65$). The main effect of Block ($F(1,50) = 22.51$, $\eta^2 = 0.03$, $p < .001$) shows that the score in the first half of each game ($M = 0.29$) was significantly lower than in the second half ($M = 0.40$), which indicates within-game learning. The main effect of Condition ($F(1,50) = 5.44$, $\eta^2 = 0.05$, $p = .024$) indicates that scores were higher against the level-1 player ($M = 0.43$) than against the level-2 player ($M = 0.27$). Thus, it appears that is was harder for participants to exploit the strategy of the more sophisticated level-2 opponent than the comparatively less sophisticated level-1 opponent.

Learning Transfer

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

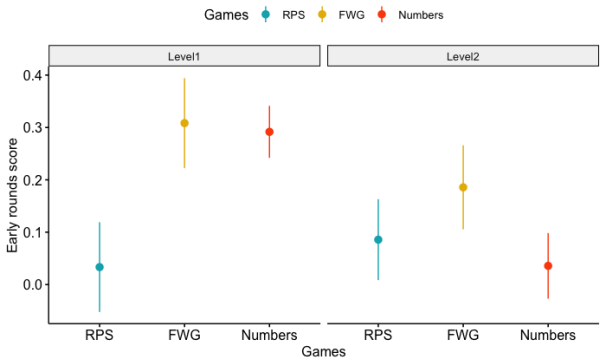

Figure 2: Average scores in rounds 2-6 by game and type of opponent. Error bars reflect 95% confidence intervals for the mean.

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

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

Computational Modelling

To gain more insight into participants' strategies against their computer opponents, we constructed and tested several computational models of strategy learning. The baseline model assumes play is random, and each potential action is chosen with equal probability. Note that this corresponds to the Nash equilibrium strategy. The other models adapted their play to

the opponent, either by reinforcing successful actions in each game (reinforcement learning), or by determining the type of opponent through Bayesian learning (Bayesian Cognitive Hierarchy models). We also include the Expected Weighted Attraction (EWA), which is a popular model in behavioral economics.

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

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

Reinforcement learning (RL) model

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

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

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

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

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

The RL model has two free parameters: the learning rate (α) and the inverse temperature parameter of the softmax decision rule (λ).

Experience-weighted attraction (EWA) model

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

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

where $I(x)$ is an indicator function which takes the value 1 is its argument is true, and 0 otherwise, and $R(a, o_t)$ is the reward that would be obtained from playing action a against opponent action o_t, which equals the actual obtained reward when $a = a_t$, and otherwise is a counterfactual reward that would have been obtained if a different action were taken. Unlike reinforcement learning, this uses knowledge of the rules of the game to allow reinforcing actions that were not taken. We can see that setting $\delta = 0$ leads to reinforcement of past actions, while positive and high delta parameters make the update rule take into account foregone pay-offs, which is similar to weighted fictitious play (Cheung & Friedman, 1994). While the assumption in expanding the update rule above is that ϕ and δ are free parameters (Camerer, Ho, & Others, 1997), the self-tuning aspect of the model comes from the fact that these are now self-tuned using the formulas expanded in Ho et al. (2004). $N(t)$ represents an experience weights and can be interpreted as the number of "observation-equivalents" of past experience. We initialise it to 1 so initial attractions and reinforcement from payoffs are weighted equally.

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

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

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

Bayesian Cognitive Hierarchy (BCH) model

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

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

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

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

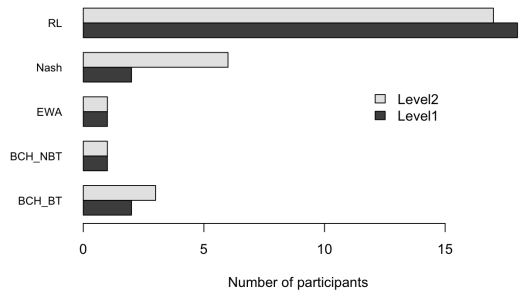

Figure 3: Histogram of best fitting computational models by condition

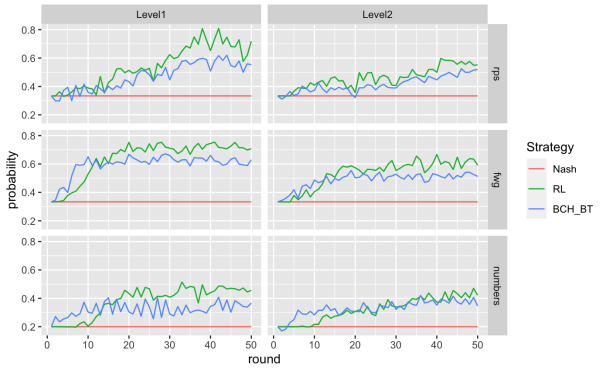

Figure 4: Likelihood by trial by game and opponent faced

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

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

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

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

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

Estimation and model comparison

We fitted all models to the data of each individual participant, across all three games, estimating the model parameters by maximum likelihood. We used the Bayesian Information Criterion (BIC) to determine the best fitting model for each participant. Figure 3 shows the number of participants best fit by each model. We can see that the RL model clearly described most participants' behaviour best, followed by the random (Nash) model. Only a few participants were best described by one of the BCH models, or the EWA model.

Using Hidden Markov Model to explore strategy switching

The computational modelling indicates that most players are best fit by an RL type model which reinforces successful actions within each game. As this model does not allow for between-game transfer, this finding is at odds with the behavioral results of learning transfer. To gain more insight into this discrepancy, Figure 4 plots the average likelihood by trial and game, according to the Nash, RL, and BCH_BT model. In the FWG and NUMBERS game, we see that in the initial rounds of these games, the likelihood of actions is highest according to the BCH_BT model (which incorporates between-game transfer), but that over time, the RL model exceeds the predictive quality of the BCH model. The fact that the likelihoods of these strategies cross over is consistent with participants switching between strategies as each game progress.

In order to more formally assess evidence for such strategy switching by participants, we fitted hidden Markov models in which the latent states are the 3 strategies (Nash, RL, and BCH_BT). Hidden Markov models are a useful framework to model switches between latent strategies. We used the three models fitted to the individual participant data above to define the likelihood of actions according to each latent state (strategy). The model then contains free parameters for the initial probability of each state $1,2,3$ at the start of each game, and the transition probabilities for switching from state i to state j during the games. These parameters were estimated by maximum likelihood with the depmixS4 R package (Visser & Speekenbrink, 2010).

As a test of strategy switching, we also fitted a restricted version of the hidden Markov model, which only allows

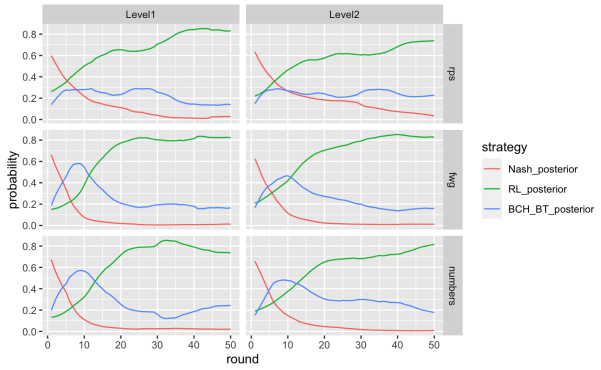

Figure 5: Posterior probability of strategies by game and opponent faced

self-transitions between the states (i.e., no switches between strategies during each game). An approximate likelihood-ratio test shows that the full HMM, which allows for strategy switches during the games, fitted significantly better than the restricted model ($\chi^2(6) = 167$, $p < .001$). This is further statistical evidence in favour of the hypothesis that participants switch between strategies during the games. In order to understand at which stage of the games the switching occurs, and whether there are any differences between games and type of opponents faced, Figure 5 shows the posterior probability of each strategy (state) at each trial of a game, averaged over participants. This figure shows a similar pattern for all games and opponents, in that in the initial rounds, the probability of the BCH model is highest, while in later rounds, the RL model takes over.

The HMM model thus shows clear evidence in favour of strategy switching by participants, from a Bayesian Cognitive Hierarchy strategy at the initial stages of each game, to a model-free RL model.

Discussion

The results of our experiment show that the majority of participants learn to adapt their play to their opponent's strategy, and generalise knowledge of their opponent's strategy to other games. Transfer to the more dissimilar NUMBERS game was moderated by the degree of sophistication of the agent, with evidence for transfer when players face the less sophisticated Level-1 agent but not the more sophisticated Level-2 player.

Initial computational modelling of observations using all available data seems to indicate that the most likely model was a reinforcement learning model, which learns rewarding actions based on the previous round's play. However, as this model is unable to generalize to new games, this finding is inconsistent with the behavioural evidence for learning transfer. Using a hidden Markov model, we showed that this discrepancy appears to be due to participants switching between strategies during the games. They start the early rounds of a new game acting in a way consistent with a Bayesian Cognitive Hierarchy learning strategy, which determines the opponent's level of iterative reasoning, and best responds to the action predicted from this opponent model. While accurate and generalisable, working through the required steps of iterative reasoning ("I think that you think that I think...") may be cognitively expensive. Model-free RL, to which participants switch in later rounds, is consistent with a more habitual type of learning and may be cognitively less taxing. Switching between these strategies then shows flexibility in the use of learning strategies. As the games progress, more and more information is acquired about rewarding actions in the current game, which provides efficient training data to a model-free RL learning algorithm, allowing participants to successfully rely on a this cognitively less taxing strategy. The preference for less computationally demanding strategies is well established (Kool, McGuire, Rosen, & Botvinick, 2010), and the ability to flexibly switch between different learning strategies is consistent with evidence of switching between model-based and model-free RL strategies according to environmental demands (Simon & Daw, 2011).

It can be argued that participants are learning simple behavioral rules (Brockbank & Vul, 2020), rather than a model of the opponent's strategy or the value of particular actions as in the RL framework. The best response to a level-1 opponent would be to choose the action that beats the opponent's previous action, and for a level-2 strategy, a winning rule would be to choose the action that would be beaten by the agent's own previous action. Such rules can mimic the predictions of our iterative reasoning and Bayesian Cognitive Hierarchy account, without seemingly requiring the cognitive effort needed to reason through what another player knows about what you know about them. On the other hand, the number of possible behavioural rules is much larger than the set of contingent actions predicted by a cognitive hierarchy model. If the generation of a constrained set of plausible behavioural rules rests upon a form of iterative reasoning as we have proposed here, then those rules could be reinforced afterwards, much like successful actions can be reinforced in a model-free RL strategy. Such an account would then be mostly equivalent to ours.

In conclusion, the results of our experiment are consistent with work in behavioural game theory showing that human players can deviate from Nash equilibrium play when their opponent does so also. In these cases, it may be possible to adapt to the opponent's strategy and exploit their deviations from equilibrium play. There is however a high degree of heterogeneity amongst players. As such, some players may have a higher ability to detect patterns in their opponent's play and learn how to exploit them. We plan to run a future experiment to address this individual heterogeneity by designing tasks where the participants face different types of opponents sequentially. This will allow for more opportunities to measure and model learning transfer as well as explore its determinants in a within-subject design.

References

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

Brockbank, E., & Vul, E. (2020). Recursive adversarial reasoning in the rock, paper, scissors game.

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

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

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

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

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

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

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of experimental psychology: general*, *139*(4), 665.

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.

Spiliopoulos, L. (2013). Strategic adaptation of humans playing computer algorithms in a repeated constant-sum game. *Autonomous agents and multi-agent systems*, *27*(1), 131–160.

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

Visser, I., & Speekenbrink, M. (2010). depmixS4: an R package for hidden Markov models. *Journal of statistical Software*, *36*(7), 1–21.

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

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