# Modeling Deep Strategic Reasoning by Humans in Competitive Games

# (Extended Abstract)

Xia Qu
Dept. of Computer Science
University of Georgia
Athens, GA 30602
quxia@uga.edu

Prashant Doshi
Dept. of Computer Science
University of Georgia
Athens, GA 30602
pdoshi@cs.uga.edu

Adam Goodie
Dept. of Psychology
University of Georgia
Athens, GA 30602
goodie@uga.edu

## ABSTRACT

The prior literature on strategic reasoning by humans of the sort, *what do you think that I think that you think*, is that humans generally do not reason beyond a single level. However, recent evidence suggests that if the games are made competitive and therefore representationally simpler, humans generally exhibited behavior that was more consistent with deeper levels of recursive reasoning. We seek to computationally model behavioral data that is consistent with deep recursive reasoning in competitive games. We use generative, process models built from agent frameworks that simulate the observed data well and also exhibit psychological intuition.

## Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent Systems

## General Terms

Experimentation, Performance

## Keywords

recursive reasoning, human decision making, modeling, games

## 1. INTRODUCTION

We model human judgment and behavioral data, reported by Goodie et al. [4], that is consistent with *three* levels of recursive reasoning in the context of fixed-sum games. In doing so, we investigate principled modeling of behavioral data consistent with levels rarely observed before. A previous model utilized underweighted belief learning, parameterized by $\gamma$, and a quantal response choice model [5] for the subject agent, parameterized by $\lambda$, within the framework of interactive partially observable Markov decision process (I-POMDP) [3]. We extend this model to make it applicable to games evaluating up to level 3 reasoning. Although it employs an empirically supported choice model for the subject agent, it does not ascribe plausible choice models to the opponent who in the experiments is also projected as being human. We hypothesize that an informed choice model for the opponent supports more nuanced explanations for observed opponent actions leading to improved performance. Hence, our second candidate model generalizes the previous by intuitively utilizing a quantal response choice model for selecting the opponent's actions at level 2. Finally, our third candidate model deviates from using I-POMDPs by utilizing weighted fictitious play [1], which predominantly relies on the past pattern of the opponent's observed actions to form a judgment about what the opponent will do next. This model differs from the previous two in that it does not seek to ascertain the mental models of the opponent but instead bases itself on the observed frequency of empirical play. The strictly competitive nature of the game discourages the influence of essentially cooperative social constructs such as positive reciprocity and altruism, otherwise observed in strategic games. While other processes such as inequality aversion may apply, an analysis of the data reveals that it did not play a role here.

**Appears in:** *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), 4-8 June 2012, Valencia, Spain.


## 2. COMPUTATIONAL MODELING

In order to computationally model the data, a multiagent decision making framework that integrates recursive reasoning in the decision process is needed. A finitely-nested I-POMDP$_{i,l}$ [3] for agent $i$ with a strategy level $l$ represents a choice which meets the requirements of explicit consideration of recursive beliefs and decision making based on such beliefs.

Because the opponent is thought to be human and guided by payoffs, we focus on intentional models only. Given that expectations about the opponent's action by the participants showed consistency with the opponent types used in the experimentation, intuitively, model set, $\Theta_j = \{\theta_{j,0}, \theta_{j,1}, \theta_{j,2}\}$, where $\theta_{j,0}$ is the level 0 (*myopic*) model of the opponent, $\theta_{j,1}$ is the level 1 (*predictive*) model and $\theta_{j,2}$ is the level 2 (*super-predictive*) model. Parameters of these models are analogous to the I-POMDP for agent $i$.

We observed that some of the participants learn about the opponent model as they continue to play. However, in general, the rate of learning is slow. This is indicative of the cognitive phenomenon that the participants could be underweighting the evidence that they observe. We may model this by augmenting normative Bayesian learning in the following way:

$$b'_{i,l}(s, \theta_{j,l-1}|o_i;\gamma) = \alpha\, b_{i,l}(s, \theta_{j,l-1}) \Big\{ \sum_{a_j} O_i(o_i|a_i, a_j, s') \times Pr(a_j|\theta_{j,l-1}) \Big\}^{\gamma} \quad (1)$$

where $\alpha$ is the normalization factor, state $s$ corresponds to A and $s'$ to B, action $a_i$ is to move, and if $\gamma < 1$, then the evidence $o_i \in \Omega_i$ is underweighted while updating the belief over $j$'s models. Furthermore, we observed significant rationality errors in the participants' decision making. We utilize the *quantal response* model [5] to simulate human non-normative choice. This model is based on

the finding that rather than always choosing the optimal action which maximizes the expected utility, individuals are known to select actions proportionally to their utilities. The quantal response model assigns a probability of choosing an action as a sigmoidal function of how close to optimal is the action. Previously, Doshi et al. [2] augmented I-POMDPs with both these models in order to simulate human recursive reasoning up to level two. As they continue to apply to our data, we extend the I-POMDP model to the longer games and label it as I-POMDP$^{\gamma,\lambda}_{i,3}$.

The methodology for the experiments reveals that the participants are deceived into thinking that the opponent is human. *Therefore, participants may justify unexpected actions of the opponent as errors in their decision making rather than due to their level of reasoning.* Hence, we generalize the previous model by attributing quantal response choice to opponent's action selection as well. Let $\lambda_1$ be the quantal response parameter for the participant and $\lambda_2$ be the parameter for the opponent's action. Then,

$$Q(a_i^*;\gamma,\lambda_1,\lambda_2) = \frac{e^{\lambda_1 \cdot U(b'_{i,3},a_i^*;\gamma,\lambda_2)}}{\sum_{a_i \in A_i} e^{\lambda_1 \cdot U(b'_{i,3},a_i;\gamma,\lambda_2)}} \quad (2)$$

parameters, $\lambda_1$, $\lambda_2 \in [-\infty,\infty]$; $a_i^*$ is the participant's action and $Q(a_i^*)$ is the probability assigned by the model. $U(b'_{i,3},a_i;\gamma,\lambda_2)$ is the utility for $i$ on performing action, $a_i$, given its updated belief, $b'_{i,3}$, with $\lambda_2$ parameterizing $j$'s action probabilities, $Pr(a_j|\theta_{j,l-1})$, present in Eq. 1 and in computation of the utility. We label this model as I-POMDP$^{\gamma,\lambda_1,\lambda_2}_{i,3}$.

A different reason for participant behavior that relies more heavily on past patterns of observed actions of the opponent, instead of ascertaining the mental models of the opponent as in the previous I-POMDP based models, is applicable. A well-known learning model in this regard is weighted (generalized) fictitious play [1]. Let $E_i(a_j)$ be the observed frequency of opponent's action, $a_j \in A_j$. We update this as:

$$E_i^t(a_j;\phi) = I(a_j,o_i) + \phi\, E_i^{t-1}(a_j) \quad t = 1,2,\ldots \quad (3)$$

where parameter, $\phi \in [0,1]$, is the weight put on the past observations; $I(a_j,o_i)$ is an indicator function that is 1 when $j$'s action in consideration is identical to the currently observed $j$'s action, $o_i$, and 0 otherwise. Due to the presence of rationality errors in the data, we combine the belief update of Eq. 3 with quantal response. We label this model as wFP$_i^{\phi,\lambda}$.

# 3. EVALUATION

To learn $\lambda_2$ in I-POMDP$^{\gamma,\lambda_1,\lambda_2}_{i,3}$, we use the expectations data of the "catch" games only. These are games in which no matter the type of the opponent, the rational action for the opponent is to move. Hence, expectations of opponent staying by the participants in the catch trials would signal non-normative action being attributed. This also permits learning a single $\lambda_2$ value across the three opponent types. However, this is not the case for the other parameters: for different opponent types, the learning rate is different. Also, we observed that the rationality errors differ considerably between the participant groups experiencing different opponent types. Therefore, we learn parameters, $\gamma$ and $\lambda_1$ given the value of $\lambda_2$ (and $\lambda$ in I-POMDP$^{\gamma,\lambda}_{i,3}$), separately from each group's diagnostic games. Analogously, we learn $\phi$ and $\lambda$ for wFP$_i^{\phi,\lambda}$ from the diagnostic games as well. We report the learned parameters in Table 1.

We utilize the learned values in Table 1 to parameterize the underweighting and quantal responses within the I-POMDP based models and fictitious play. We cross-validated the models on the test folds. Using a participant's actions in the first 5 trials, we initialized the prior belief distribution over the opponent types. We measure the goodness of the fit by computing the mean squared error (MSE) of the prediction by the models, and compare it to those of a random model (null hypothesis) for significance. We show the MSE in the achievement and prediction scores, as defined in Goodie et al. [4], based on the models in Table 2.

| model | param. | myopic | pred | super-pred |
|---|---|---|---|---|
| I-POMDP$^{\gamma,\lambda_1,\lambda_2}_{i,3}$ | $\lambda_2$ | 1.959 | | |
| | $\gamma$ | 0.164 | 0.049 | 0.221 |
| | $\lambda_1$ | 3.259 | 3.906 | 3.768 |
| I-POMDP$^{\gamma,\lambda}_{i,3}$ | $\gamma$ | 0.232 | 0.079 | 0.357 |
| | $\lambda$ | 2.985 | 3.826 | 3.667 |
| wFP$_i^{\phi,\lambda}$ | $\phi$ | 0.999 | 0.999 | 0.150 |
| | $\lambda$ | 2.127 | 3.107 | 3.165 |

**Table 1: Average group-level parameter values learned from the training folds of the experiment data for the three candidate models.**

| Opponent type | Mean Squared Error (MSE) | | | |
|---|---|---|---|---|
| | Achievement score | | Prediction score | |
| | myopic | super-pred | myopic | super-pred |
| **Random** | 0.0041 | 0.4502 | 0.0035 | 0.3807 |
| **I-POMDP$^{\gamma,\lambda_1,\lambda_2}_{i,3}$** | 0.0014 | 0.0009 | 0.0020 | 0.0010 |
| **I-POMDP$^{\gamma,\lambda}_{i,3}$** | 0.0025 | 0.0008 | 0.0016 | 0.0014 |
| **wFP$_i^{\phi,\lambda}$** | 0.0123 | 0.0082 | 0.0103 | 0.0120 |

**Table 2: MSE of the predictions by the different models.**

Notice from Table 2 that both I-POMDP based models have MSEs that are significantly lower than the random model. The difference in MSE of the achievement score for the myopic group between the two is significant (Student's paired t-test: $p = .015$). However, other MSE differences between the two models are insignificant and do not distinguish one model over the other across the scores and groups. Although attributing non-normative action selection to the opponent did not result in significantly more accurate expectations for any group, we think that it allowed the model to generate actions for agent $i$ that fit the data better by supporting an additional account of $j$'s (surprising) myopic behavior. Of course, this positive result should be placed in the context of increased expense of learning an additional parameter, $\lambda_2$. Large MSE of wFP$_i^{\phi,\lambda}$ reflects its weak simulation performance although it does improve on the par set by random for the super-predictive group.

**Acknowledgment** We acknowledge grant support from NSF, CAREER #IIS-0845036, and from USAF, #FA9550-08-1-0429.

# 4. REFERENCES

[1] Y.-W. Cheung and D. Friedman. Individual learning in normal-form games. *Games and Econ Behav*, 19:46–76, 1997.

[2] P. Doshi, X. Qu, A. Goodie, and D. Young. Modeling recursive reasoning in humans using empirically informed interactive POMDPs. In *AAMAS*, pages 1223–1230, 2010.

[3] P. Gmytrasiewicz and P. Doshi. A framework for sequential planning in multiagent settings. *JAIR*, 24:49–79, 2005.

[4] A. S. Goodie, P. Doshi, and D. L. Young. Levels of theory-of-mind reasoning in competitive games. *Behav Dec Making*, 24:95–108, 2012.

[5] R. McKelvey and T. Palfrey. Quantal response equilibria for normal form games. *Games and Econ Behav*, 10:6–38, 1995.