Modeling Recursive Reasoning by Humans Using Empirically Informed Interactive POMDPs

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

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

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

Diana Young
Dept. of Psychology
University of Georgia
Athens, GA 30602
dlyoung@uga.edu

ABSTRACT

Recursive reasoning of the form *what do I think that you think that I think* (and so on) arises often while acting rationally in multiagent settings. Several multiagent decision-making frameworks such as RMM, I-POMDP and the theory of mind model recursive reasoning as integral to an agent's rational choice. Real-world application settings for multiagent decision making are often *mixed* involving humans and human-controlled agents. In two large experiments, we studied the level of recursive reasoning generally displayed by humans while playing sequential general-sum and fi edsum, two-player games. Our results show that subjects experiencing a general-sum strategic game display firs or second level of recursive thinking with the firs level being more prominent. However, if the game is made simpler and more competitive with fi edsum payoffs, subjects predominantly attributed first-l vel recursive thinking to opponents thereby acting using second level of reasoning. Subsequently, we model the behavioral data obtained from the studies using the I-POMDP framework, appropriately augmented using well-known human judgment and decision models. Accuracy of the predictions by our models suggest that these could be viable ways for computationally modeling strategic behavioral data.

Categories and Subject Descriptors

I.2.11 [Distributed Artificial Intelligence]: Multiagent Systems

General Terms

Experimentation, Performance, Human Factors

Keywords

recursive reasoning, human decision making, models

1. INTRODUCTION

Strategic recursive reasoning of the form *what do I think that you think that I think* (and so on) arises naturally in multiagent settings. For example, an autonomous unmanned aerial vehicle (UAV)'s decision may differ if it believes that its reconnaissance target believes that it is not being spied upon in comparison to when the UAV believes that its target believes that it is under surveillance. Specifi cally, an agent's rational action in a two-agent game often depends on the action of the other agent, which, if the other is also rational, depends on the action of the subject agent.

Assumptions of *common knowledge* [8, 9] of elements of the game tend to preclude the modeling of recursive reasoning. However, not all elements are common knowledge. For example, an agent's belief is private especially in a non-cooperative setting. Multiple decision-making frameworks such as the recursive modeling method (RMM) [12, 13] and interactive partially observable Markov decision process (I-POMDP) [11] model recursive beliefs as an integral aspect of agents' decision making in multiagent settings.

Real-world applications of decision making often involve *mixed* settings that are populated by humans and human-controlled agents. Examples of such applications include UAV reconnaissance in an urban operating theater and online negotiations involving humans. The optimality of an agent's decisions as prescribed by frameworks such as RMM and I-POMDP in these settings depends on how accurately the agent models the strategic reasoning of others. A key aspect of this modeling is the depth of the recursive reasoning that is displayed by human agents.

Initial investigations into ascertaining the depth of strategic reasoning of humans by Stahl and Wilson [19] and more recently, by Hedden and Zhang [15] and Ficici and Pfeffer [10] show that humans generally operate at only firs or second level of recursive reasoning. Typically the firs level, which attributes no recursive reasoning to others, is more prominent. Evidence of these shallow levels of reasoning is not surprising, as humans are limited by bounded rationality.

In this paper, we report on two large studies that we conducted with human subjects to test levels of recursive reasoning. In the firs study, we constructed a task that resembled the two-player sequential, general-sum game as used by Hedden and Zhang [15]. Subjects played the game against a computer opponent, although they were led to believe that the opponent was human. Different groups of subjects were paired against an opponent that used no recursive reasoning (zero level) and opposite one that used first-l vel reasoning. Data collected on the decisions of the participants indicate that, (i) subjects generally attributed zero-level reasoning to the other *a'priori*; and (ii) subjects acted accurately significantl more times when the opponent displayed zero-level reasoning than when the opponent was at firs level. The participants also learned the

Cite as: Modeling Recursive Reasoning by Humans Using Empirically Informed Interactive POMDPs, Prashant Doshi, Xia Qu, Adam Goodie and Diana Young, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. 1223-1230

reasoning level of opponents slowly and incompletely as reported previously by Hedden and Zhang. In the second study, we made the game simultaneously simpler and more competitive by incorporating fi ed-sum outcomes. This had the surprising impact that subjects acted accurately more times when the opponent displayed first-l vel reasoning compared to zero level. Furthermore, the participants learned the reasoning level of opponents more quickly and completely than in the previous experiment.

Because I-POMDPs explicitly consider recursive beliefs of agents, they are a natural choice as a point of departure for computationally modeling the behavioral data collected in the experiments. We augment them with well-known human judgment and decision models that reflec the subrational behavior of humans, also observable in our data. We learn the parameters of these models by formulating the problem as a gradient descent through a subset of the data. Predictions by our models on the remaining data are significantl consistent with actual human decisions. In the absence of additional experimentation, this may not testify to the cognitive plausibility of our models. However, they represent a principled way to computationally model strategic behavioral data accurately.

2. RELATED WORK

Harsanyi [14] recognized that indefinit recursive thinking arises naturally among rational players, which leads to difficult in solving games. In order to, in part, avoid dealing with recursive reasoning, Harsanyi proposed the notion of types and common knowledge of the joint belief over the player types. However, common knowledge is itself modeled using an indefinit recursive system [8, 9].

Since Harsanyi's introduction of abstract agent types, researchers have sought to mathematically defin the type system. Beginning with Mertens and Zamer [18], who showed that a type could be define as a hierarchical belief system with strong assumptions on the underlying probability space, subsequent work [4, 16] has gradually relaxed the assumptions required on the state space while simultaneously preserving the desired properties of the hierarchical belief systems. Along a similar vein, Aumann define recursive beliefs using both a formal grammar [1] and probabilities [2] in an effort to formalize interactive epistemology.

Within the context of behavioral game theory [5], Stahl and Wilson [19] investigated the level of recursive thinking exhibited by humans. Stahl and Wilson found that only 2 out of 48 (4%) of their subjects attributed recursive reasoning to their opponents while playing 12 symmetric 3×3 matrix games. On the other hand, 34% of the subjects ascribed zero-level reasoning to others. Corroborating this evidence, Hedden and Zhang [15] in a study involving 70 subjects, found that subjects predominantly began with first-l vel reasoning. When pitted against first-l vel co-players, some began to gradually use second-level reasoning, although the percentage of such players remained generally low. Hedden and Zhang utilized a sequential, two-player, general-sum game – sometimes also called the Centipede game in the literature [3]. Ficici and Pfeffer [10] investigated whether human subjects displayed sophisticated strategic reasoning while playing 3-player, one-shot negotiation games. Although their subjects reasoned about others while negotiating, there was insufficien evidence to distinguish whether their level two models better fi the observed data than level one models.

Evidence of recursive reasoning in humans and investigations into the level of such reasoning is relevant to multiagent decision making in *mixed* settings. In particular, these results are directly applicable to computational frameworks such as RMM [12], I-POMDP [11] and cognitive ones such as theory of mind [7] that ascribe intentional models of behavior to other agents.

3. EXPERIMENTS: LEVELS OF RECURSIVE REASONING

In two large studies involving human subjects held simultaneously, we investigate the levels of recursive reasoning subjects would generally exhibit in particular interactions. We begin with a description of the problem setup followed by the participating population and our methodology for the firs experiment. We then provide similar information for our second experiment.

3.1 Study 1: General-Sum Game

In keeping with the tradition of experimental game research [5, 6] and the games used by Hedden and Zhang [15], we selected a two-player alternating-move game of complete and perfect information. In this sequential game, whose game tree is depicted in Fig. 1(a), player *I* (the leader) may elect to *move* or *stay*. If player *I* elects to move, player *II* (the follower) faces the choice of moving or staying, as well. An action of stay by any player terminates the game. Note that actions of all players are perfectly observable to each other. While the game may be extended to any number of moves, we terminate the game after two moves of player *I*.

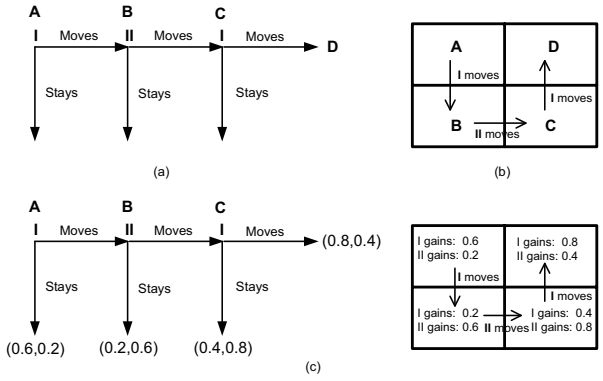


Figure 1: (a) A game tree representation (extensive form) of our two-player game. Because of its particular structure, such games are also sometimes called Centipede games. States of the game are indicated by the letters, A, B, C and D. (b) Arrows denote the progression of play in the game. An action of ***move*** by each player causes a transition of the state of the game. (c) An example general-sum game used in the study.

We set the outcomes as probabilities of gain for each player. Similar to magnitudes, rational choice involves selecting an action that maximizes the probability of success.

In order to decide whether to move or stay at state A, a rational player *I* must reason about whether player *II* will choose to move or stay at B. A rational player *II*'s choice in turn depends on whether player *I* will move or stay at C. Thus, the game lends itself naturally to recursive reasoning and the level of reasoning is governed by the height of the game tree.

For the example game in Fig. 1(c), a rational player *I* assuming that player *II* is rational and that *II* knows that *I* is rational, will choose to stay. This is because *I* thinks that if it chooses to move, player *II* will choose to stay to obtain a payoff of 0.6. A move by player *II* to C is not rational because player *I* will then choose to move as well with the payoff for *II* being only 0.4.

3.1.1 Participants

A total of 145 subjects participated in the study. The participants were undergraduate students enrolled in lower-level courses in our university. The students received performance-driven pay for par-

ticipating in the study.

All participants gave informed consent for their participation prior to admission into the study. They were appropriately debriefed at the conclusion of the study.

3.1.2 Methodology

Opponent models In order to test different levels of recursive reasoning, we designed the computer opponent (player *II*) to play a game in two ways: (i) If player *I* chooses to move, *II* decides on its action by simply choosing between the outcomes at states B and C in Fig. $1(b)$ rationally. Therefore, *II* is a zero-level player and we call it *myopic* (see Fig. $2(a)$). (ii) If player *I* chooses to move, the opponent decides on its action by reasoning what player *I* will do rationally at C. Based on the action of *I*, player *II* will select an action that maximizes its outcomes. Thus, player *II* is a first-l vel player, and we call it *predictive* (see Fig. $2(b)$).

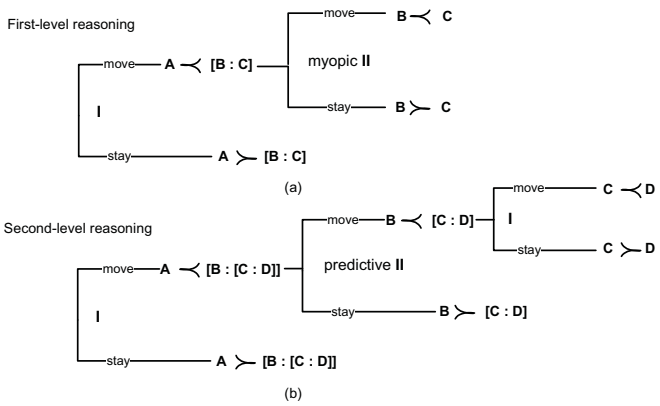


Figure 2: (a) A myopic player *II* decides on its action by comparing the payoff at state B with that at C. Here, B $\prec$ C denotes a preference of C over B for the player whose turn it is to play and B : C denotes the rational choice by the appropriate player between states B and C. Thus, player *I* exhibits first-level reasoning. (b) If player *II* is predictive, it reasons about *I*'s actions. Player *I* then exhibits second-level reasoning in deciding its action at state A.

To illustrate, in the game of Fig. $1(c)$ if player *I* decides to move, a myopic player *II* will move to obtain a payoff of 0.8, while a predictive *II* will choose to stay because it thinks that player *I* will choose to move from C to D, if it moved. By choosing to stay, *II* will obtain an outcome of 0.6 in comparison to 0.4 if it moves.

Payoff structure Notice that the rational choice of players in the game of Fig. 1 depends on the preferential ordering of states of the game rather than specifi probabilities. Let $a \prec b$ indicate that the player whose turn it is to play prefers state b over a. Games that exhibit a preference ordering of D $\prec$ C $\prec$ B $\prec$ A and A $\prec$ B $\prec$ C $\prec$ D for player *I* are trivial because player *I* will always opt to stay in the former case and move in the latter case, regardless of how *II* plays. Furthermore, consider the ordering C $\prec$ A $\prec$ B $\prec$ D for player *I* and an ordering D $\prec$ B $\prec$ A $\prec$ C for *II*. A myopic opponent will choose to move while a predictive opponent will stay. However, in both these cases player *I* will choose to move. Thus, games whose states display a preferential ordering of the type mentioned previously are not diagnostic – regardless of whether player *I* thinks that opponent is myopic or is predictive, *I* will select the same action precluding a diagnosis of *I*'s level of recursive reasoning. Of all the 576 distinct preferential orderings among states that are possible for both players, only 48 are diagnostic and not trivial – e.g., B $\prec$ C $\prec$ A $\prec$ D for player *I* and A $\prec$ D $\prec$ B $\prec$ C for *II*. For this ordering, player *I* will move if it thinks that the opponent is myopic, otherwise *I* will stay if the opponent is thought to be predictive. We note that the game in Fig. $1(c)$ follows this preference ordering.

Design of task Batches of participants played the game on computer terminals with each batch having an even number of players. Each batch was divided into two groups and members of the two groups were sent to different rooms. This was done to create the illusion that each subject was playing against another, although the opponent was in reality a computer program. This deception was revealed to the subjects during debriefing

Each subject experienced an initial *training phase* of at least 15 games that were trivial or those in which a myopic or predictive opponent behaved identically. These games served to acquaint the participants with the rules and goal of the task without unduly biasing them about the behavior of the opponent. Therefore, these games have no effect on the initial model of the opponent that participants may have. Participants who failed to choose the rational actions in any of the previous 5 games after the 15-game training phase continued with new training games until they met the criterion of no rationality errors in the 5 most recent games. Those who failed to meet this criterion after 40 total training games did not advance to the test phase, and were removed from the study.

In the *test phase*, each subject experienced 32 games instantiated with outcome probabilities that exhibited the diagnostic preferential orderings. The 32 critical games were divided into 4 blocks of 8 games each. In order to avoid subjects developing a mental set, we interspersed these games with 8 that exhibited orderings such as, D $\prec$ C $\prec$ A $\prec$ B and C $\prec$ B $\prec$ A $\prec$ D, resulting in a total of 40 games. These not only serve to distract the participants but also function as "catch" trials allowing us to identify participants who may not be attending to the games.

Approximately half the participants played against myopic opponents while the remaining played against predictive ones. In each group, all participants were presented with the Centipede and grid representation of the games with payoffs. All participants also experienced a screen asking them what they thought the opponent would play and their confidenc in the prediction, for each game.

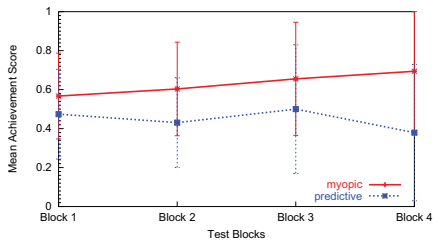
3.1.3 Results

Our study spanned a period of *three months* from February through April 2009. We report the results of this study below.

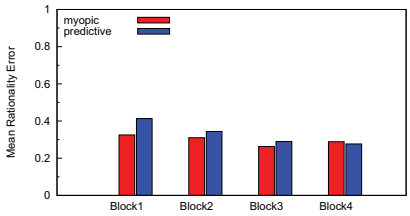
Training phase As mentioned before, each of the 145 human subjects initially played a series of 15 games in order to get acquainted with the general-sum and complete information structure, and objectives of the task at hand. After this initial phase, participants who continued to exhibit errors in any of the games up to 40 total games were eliminated. 31 participants did not progress further in the study. These participants either failed to understand how the game is played or exhibited excessive irrational behavior, which would have affected the validity of the results of this study.

Test phase Of the 114 participants (65 female) who completed the test phase, 58 experienced myopic opponents while the remaining 56 experienced predictive opponents.

Participants in each of the 2 groups were presented with 40 instances of the particular game type whose payoff structure is diagnostic. For the sake of analysis, we assembled 4 *test blocks* each comprising of 10 games – 8 test trials and 2 catch trials. For each participant, we measured the fraction of times that the subject played accurately in each test block. We defin the accurate choice as the action choice which is conditionally rational given the type of opponent. For example, in the game of Fig. $1(c)$, the accurate choice for player *I*, if the opponent is myopic, is to move. On the



(a)



(b)

Figure 3: (a) Mean achievement score of the participants for both groups across test blocks. Notice that subjects generally expected their opponents to play at zero level far more than at first level. (b) Rationality error rates among participants measured as actions which are not rational given predictions.

other hand, if the opponent is predictive, the accurate choice for *I* is to stay. We then compiled an *achievement score* which is the proportion of games in which the subject played the accurate choice given the opponent type.

In Fig. 3(a), we show the mean achievement scores across all participants in each of the 2 groups. We observe that the achievement score is significantl higher when the opponent is myopic as compared to when it is predictive. Statistical tests (F-test: F(1,112) = 34.31, p<0.001) confir that participants playing against myopic opponents have statistically significan higher achievement scores compared to predictive opponents across all test blocks. This implies that subjects predominantly displayed first-l vel reasoning when acting. They did not expect the opponent to reason about their subsequent play and acted accordingly. Data collected from the participants who experienced the screen asking about the opponents' possible action and their confidenc in the prediction confirme this expectation. Additionally, learning took place that was responsive to the opponent with achievement scores showing an increase across test blocks particularly for the myopic group.

Finally, we computed mean rationality errors, which reflec the proportion of times that participants' choices were irrational given their expectations about the opponents' decisions. As we show in Fig. 3(b), the error rates remained consistent across all test blocks and across both groups.

3.2 Study 2: Fixed-Sum Game

Outcomes from our study of recursive reasoning by humans in general-sum games are remarkably similar to previous results such as those of Hedden and Zhang [15]. They confir the predominance of low levels of reasoning by humans engaged in general strategic games. Our primary motivation behind the second study is to show demonstrations of higher levels of recursive thinking either by default or through learning. We sought to increase the level of reasoning observed in our participants by making the game more competitive, and subsequently simpler.

In order to be consistent with our previous study, we again selected the two-player alternating-move game of complete and perfect information. However, the game differed from the previous one in that the payoffs were probabilities of success for each player that summed to 1. In other words, if p is the probability of success for player *I* at a particular state, then $1 - p$ is the probability for player *II*. We show an example game in Fig. 4.

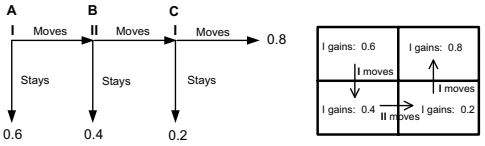


Figure 4: An example Centipede game used in the second study. The payoffs are probabilities of success for player *I*. The complement is the probability of success for player *II*.

In the game of Fig. 4, a rational player *I* assuming that player *II* is rational and that *II* knows that *I* is rational, will choose to stay in order to maximize his or her chances of success. This is because if *I* moves, then player *II* faced with a decision will choose to stay otherwise *I* will subsequently move at state C reducing *II*'s probability to 0.2.

3.2.1 *Participants and Methodology*

A different set of participants for Study 2 were drawn from the same pool of subjects as in the previous study. The students received performance-driven pay for their participation and provided informed consent for their participation prior to admission into the study. A total of 140 subjects participated in this study. They were debriefed at the end of the study.

Opponent models Models of the *opponent* were set to be identical to those in Study 1. A *myopic* player *II* decides its action by choosing between the outcomes at states B and C rationally. On the other hand, a *predictive* player *II* thinks about what *I* will do at C should he or she decide to move.

Payoff structure As with the general-sum game, the rational choice of players in the game of Fig. 4 depends on the preferential ordering of states of the game rather than actual probabilities. Of the 24 distinct preferential orderings of the states, only one ordering is diagnostic: C $\prec$ B $\prec$ A $\prec$ D. For this ordering, player *I* will move if it thinks that the opponent is myopic, otherwise *I* will stay if the opponent is thought to be predictive. Note that the game in Fig. 4 follows this preference ordering. To maintain the attention of subjects, we vary the actual probability values while following the diagnostic ordering, and include "catch" trials. Remaining orderings are either trivial for players *I* or *II*, or not diagnostic.

Design of task As in the previous study, batches of participants played the game on computers. Each batch was divided into two groups and members were sent to different rooms. We did this to create the illusion that each subject was playing against another. This deception was revealed to the subjects during debriefing

Each subject experienced an initial *training phase* of at least 15 games that were trivial or those in which a myopic or predictive opponent behaved identically. These games served to acquaint the participants with the rules and goal of the task without unduly biasing them about the behavior of the opponent. Participants who failed to choose the rational actions in any of the previous 5 games after the 15-game training phase continued with new training games until they met the criterion of no rationality errors in the 5 most recent games. Those who failed to meet this criterion after 25 total training games did not advance to the test phase.

In the *test phase*, each subject experienced 40 games instantiated with outcome probabilities that exhibited the diagnostic preferen-

tial ordering. The 40 critical games were divided into 4 blocks of 10 games each. In order to avoid subjects developing a mental set, we interspersed these games with 40 "catch" trials that exhibited the orderings, C $\prec$ A $\prec$ B $\prec$ D and D $\prec$ B $\prec$ A $\prec$ C. All of the participants were presented with the Centipede and grid representation of the games. All the participants experienced the screen asking them what they thought the opponent would play and their confidenc in the prediction, for the games.

3.2.2 Results

Study 2 was performed simultaneously with the previous study and also spanned a period of *three months* from February through April 2009. We report the results of this study below.

Training phase As we mentioned, each of the 140 human subjects initially played a series of 15 games in order to get acquainted with the fi ed-sum and complete information structure, and objectives of the task at hand. After this initial phase, participants who continued to exhibit errors in any of the games up to 25 total games were eliminated. 22 participants did not progress further in the study.

Test phase Of the 118 participants (60 female) who completed the test phase, 58 experienced a myopic opponent and 60 experienced a predictive opponent.

Participants in each of the 4 groups were presented with 40 instances of the game type whose payoff structure is diagnostic. For the sake of comparison, we again assembled 4 *test blocks* each comprising 10 games. For each participant, we measured the fraction of times that the subject played accurately in each test block given the opponent type, which we called the achievement score. For example, in the game of Fig. 4, the accurate choice for player *I*, if the opponent is myopic, is to move. On the other hand, if the opponent is predictive, the accurate choice for *I* is to stay.

Because opponents types are fi ed and participants experience 40 games, they have the opportunity to learn how their opponent might be playing the games. Consequently, participants may gradually make more accurate choices over time. Participants were deemed to have learnt the opponent's model at the game beyond which performance was always statistically significantl better than chance, as measured by a binomial test at the 0.05 level and one-tailed. This implies making no more than one inaccurate choice in any window of 10 games.

In Fig. 5, we show mean achievement scores across all participants for the 2 groups and the rationality error rates. Two group-level finding are evident from the results in Fig. 5(a): First, the mean achievement score is significantl higher when opponent is predictive as compared to when it is myopic. Student t-tests (t(116) = 9.22, $p < 0.001$) confir that participants playing against predictive opponents have statistically significan higher proportions of accuracy compared to myopic opponents across all test blocks.

The higher achievement score when the opponent is predictive in conjunction with the low score when the opponent is myopic implies that subjects predominantly displayed second-level reasoning when acting. An analysis of the subjects' predictions reveals that they generally expected the opponent to reason about their subsequent play (first-l vel reasoning). The fact that myopic opponents did not do this resulted in their choices being inaccurate.

Second, notice from Fig. 5(a) that the mean achievement score improves over successive test blocks in both groups. Multivariate t-tests of both the main effect of block position and the interaction between block position and opponent type reveal that the changes in both groups were significan ($p < 0.01$). Thus, learning took place that was responsive to the opponent, although the learning is slow and not all participants learnt the opponent type. However, 29 of the 58 subjects facing a myopic opponent never established

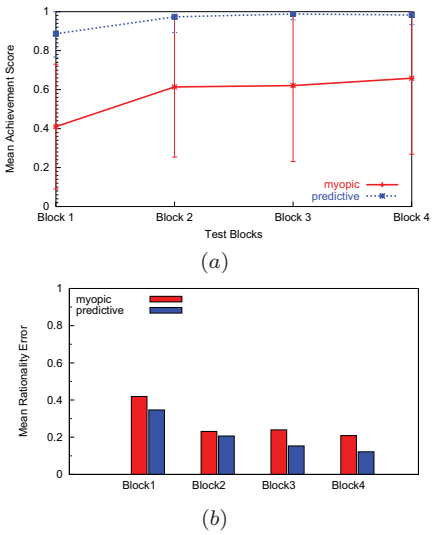


Figure 5: (a) Mean achievement score of the participants for both groups across test blocks. Notice that subjects generally expected their opponents to play at first level rather than at zero level. (b) Rationality error rates among participants.

statistically significan learning. On the other hand, only 2 among the predictive group did not establish learning, while 20 subjects achieved the fastest measurable learning possible. Consequently, participants learnt to play accurately significantl faster against a predictive opponent compared to a myopic one.

4. MODELING BEHAVIORAL DATA

Data produced by the studies described in Section 3 focus on human recursive thinking and subsequent action. In order to computationally model this data, we seek a multiagent decision-making framework capable of modeling recursive reasoning in the decision-making process. Finitely-nested interactive POMDPs (I-POMDPs) [11] are a natural choice because of their explicit consideration of recursive beliefs and decision making based on such beliefs.

However, I-POMDPs in their original form may not be applicable because human judgment and choice is not known to be normative. Thus, we will augment I-POMDPs with ways that are well known in the behavioral and cognitive literature to model relevant aspects of human decision making. Selection of these models and their parameters are informed by the data on hand.

4.1 Empirically Informed I-POMDP

Interactive POMDPs generalize POMDPs to multiagent settings by including other agents' models as part of the state space [11]. A finitel nested I-POMDP of agent i with a strategy level l interacting with another agent, j, is define as the tuple:

$$\text{I-POMDP}_{i,l} = \langle IS_{i,l}, A, \Omega_i, T_i, O_i, R_i \rangle$$

where: • $IS_{i,l}$ denotes a set of interactive states define as, $IS_{i,l} = S \times M_{j,l-1}$, where $M_{j,l-1} = \Theta_{j,l-1} \cup SM_j$, for $l \geq 1$, and $IS_{i,0} = S$, where S is the set of states of the physical environment. $\Theta_{j,l-1}$ is the set of computable *intentional models* of agent j: $\theta_{j,l-1} = \langle b_{j,l-1}, \hat{\theta}_j \rangle$ where $b_{j,l-1}$ is j's level $l-1$ belief and the *frame*, $\hat{\theta}_j = \langle A, \Omega_j, T_j, O_j, R_j, OC_j \rangle$. Here, j is Bayes rational and OC_j is j's optimality criterion. SM_j is the set of subintentional models of j. In this application, we focus on intentional models only.
• $A = A_i \times A_j$ is the set of joint actions of all agents. The remaining parameters have their usual meaning. For a description of

the belief update, additional details on I-POMDPs and how they compare with other multiagent frameworks, see [11].

Although the Centipede game involves multiple decision points, because these are alternating and occur at distinct states for different players, the decisions should be modeled recursively rather than sequentially. We model the Centipede games used in Studies 1 and 2 using the I-POMDP$_{i,2}$. We note that the physical state space, $S = \{A, B, C, D\}$, is perfectly observable; i's actions, $A_i = \{Stay, Move\}$ are deterministic and j has similar actions; i observes other's actions, $\Omega_i = \{Stay, Move\}$; and R_i captures the diagnostic preferential ordering of the states contingent on which of the two games is being considered.

Model set, $\Theta_j = \{\theta_{j,1}, \theta_{j,0}\}$, where $\theta_{j,1}$ is the level 1 predictive model of the opponent and $\theta_{j,0}$ is the level 0 myopic model. Parameters of these models are analogous to those for agent i, except for R_j which reflect the preferential ordering of the states for the opponent. Note that the predictive model, $\theta_{j,1}$, includes the level 0 model of i, $\theta_{i,0}$, in its interactive state space. Agent i's belief, $b_{i,2}$, assigns a varied distribution to j's models based on the game being modeled. This belief will reflec the general *de facto* thinking of the subjects about their opponent. It also assigns a marginal probability 1 to state A indicating that i decides at that state. Both $b_{j,1}$ and $b_{j,0}$ that are part of j's two models, respectively, assign a marginal probability 1 to B indicating that j acts at that state. Belief $b_{i,0}$, that is part of $\theta_{i,0}$, assigns probability of 1 to state C.

Notice from Figs. 3(a) and 5(a) and our discussion in Section 3.2.2 that some of the subjects learn about the opponent model as they continue to play. However, the rate of learning varies across subjects, and, in general, the learning is slow and partial. This is indicative of the well-known cognitive phenomenon that the subjects could be *underweighting* the evidence that they observe. We may model this by making the observations slightly noisy and augmenting normative Bayesian learning in the following way:

$$\frac{Pr(\theta_{j,1}|o_i)}{Pr(\theta_{j,0}|o_i)} = \frac{Pr(\theta_{j,1})}{Pr(\theta_{j,0})} \times \left\{ \frac{Pr(o_i|\theta_{j,1})}{Pr(o_i|\theta_{j,0})} \right\}^{\gamma} \quad (1)$$

where, if $\gamma < 1$, then the evidence $o_i \in \Omega_i$ is underweighted while updating the belief over j's models; if $\gamma = 1$, then the update is normative.

Figs. 3(b) and 5(b) reveal the significan presence of rationality errors in the participants' decision making. While several models exist that aim to simulate human non-normative choice, we utilize the *quantal response model* [17], which is well corroborated in behavioral game theory and applies to our problem. This model is based on the findin that rather than always choosing the decision that is of maximum expected utility, individuals are known to select actions proportionally to their utilities. The quantal response model accords a probability of choosing an action as a sigmoidal function of how close to optimal is the action. Mathematically,

$$q(a_i^* \in A_i) = \frac{e^{\lambda \cdot U(b_{i,2}, a_i^*)}}{\sum_{a_i \in A_i} e^{\lambda \cdot U(b_{i,2}, a_i)}} \quad (2)$$

where $q(a_i^* \in A_i)$ is the probability assigned to action, a_i^*, by the model, $U(b_{i,2}, a_i)$ is the utility for i performing the action, a_i, given its belief, $b_{i,2}$ as computed by the I-POMDP, and λ is the parameter that controls how responsive is the model to value differences. Within the I-POMDP, we may replace utility maximization with this model in a straightforward way.

In addition to the above models, we seek to ascertain the prior beliefs of agent i for the different games. This is informed, for e.g., by the fact that approximately 25% of the subjects in Study 2 thought that the opponent is predictive in the firs 5 games consistently.

4.2 Learning Parameters from Data

Augmenting I-POMDP based decision making with the models mentioned previously require us to fin values for the parameters γ (learning rate) and λ (non-normative choice). We may formulate the problem of findin these parameters as that of *gradient descent* over an error surface define by the inverse of the data likelihood. More formally, we seek the parameters that minimize the negative log likelihood of the data as predicted by the augmented I-POMDP:

$$\begin{aligned} X &= -\sum_{i=1}^{|Subj|} \sum_{g=1}^{N} log\ q(a_i^*|A_i) \\ &= -\sum_{i=1}^{|Subj|} \sum_{g=1}^{N} log\ \frac{e^{\lambda U(b_{i,2}^g, a_i^*)}}{\sum_{a_i \in A_i} e^{\lambda U(b_{i,2}^g, a_i)}} \end{aligned}$$

where a_i^* is the action from A_i selected by the subject i in the g^{th} game. Subject's belief, $b_{i,2}^g$, is updated according to Eq. 1 after each game instance, g. Notice that the ideal choice model, q, assigns a probability 1 to each of the actions played by each subject resulting in the minimum value of X (=0).

The gradient of the error function w.r.t model parameter, λ, is:

$$\begin{aligned} \frac{\partial X}{\partial \lambda} &= -\sum_{i=1}^{|Subj|} \sum_{g=1}^{N} \frac{\frac{\partial q}{\partial \lambda}}{q(a_i^*|A_i)} \\ &= -\sum_{i=1}^{|Subj|} \sum_{g=1}^{N} \left\{ \frac{U(b_{i,2}^g, a_i^*) \sum_{a_i \in A_i} e^{\lambda U(b_{i,2}^g, a_i)}}{\sum_{a_i \in A_i} e^{\lambda U(b_{i,2}^g, a_i)}} \right. \\ &\quad \left. - \frac{\sum_{a_i \in A_i} e^{\lambda U(b_{i,2}^g, a_i)} U(b_{i,2}^g, a_i)}{\sum_{a_i \in A_i} e^{\lambda U(b_{i,2}^g, a_i)}} \right\} \end{aligned}$$

Let $\alpha \in [0, 1]$ be the step size, then we update parameter λ as:

$$\lambda^{t+1} = \lambda^t - \alpha \cdot \frac{\partial X}{\partial \lambda}$$

In order to revise parameter γ, we note that the utility function becomes, $U(b_{i,2}^g, a_i) = \sum_{s,\theta_j} b_{i,2}^g(s, \theta_j)\ ER_i(s, \theta_j, a_i)$ where, $ER_i(s, \theta_j, a_i) = \sum_{a_j} R_i(s, a_i, a_j)\ Pr(a_j|\theta_j)$ is the expected reward. Here, $b_{i,2}^g$ is the updated belief using Eq. 1 if the current game is not the firs one. Then:

$$\begin{aligned} &\frac{\partial X}{\partial \gamma} = -\lambda \sum_{i=1}^{|Subj|} \sum_{g=1}^{N} \left\{ \sum_{s,\theta_j} b_{i,2}^{g-1}(s, \theta_j) Pr(o_i|\theta_j)^{\gamma} \right. \\ &\times log Pr(o_i|\theta_j) ER_i(s, \theta_j, a_i^*) \\ &\left. - \frac{\sum_{a_i \in A_i} \sum_{s,\theta_j} b_{i,2}^{g-1}(s,\theta_j) Pr(o_i|\theta_j)^{\gamma} log Pr(o_i|\theta_j) ER_i(s,\theta_j,a_i) e^{\lambda U(b_{i,2}^g, a_i)}}{\sum_{a_i \in A_i} e^{\lambda U(b_{i,2}^g, a_i)}} \right\} \end{aligned}$$

We perform the gradient descent until the values of λ and γ converge approximately. We utilize the converged values of the parameters within the augmented I-POMDP.

4.3 Model Performance

We integrated the models of human belief update and choice (Eqs. 1 and 2) within the finitel nested I-POMDP model of decision making. We report on the predictions of our model below.

4.3.1 Parameters

We randomly separated the behavioral data obtained from each study into training and test sets of approximately equal sizes. For this, we utilized only the subjects who successfully passed the training phase in the studies. We used the training sets to learn parameters, γ and λ, for the general-sum and fi ed-sum games.

General-sum game 57 of the 114 subjects who participated in Study 1 were randomly selected and their behavioral data was used to learn the parameters. Of these, 29 had experienced a myopic opponent while 28 had faced a predictive one. Because subjects fac-

ing myopic or predictive opponents displayed learning and acted significantl differently with different levels of rationality errors, we ran gradient descent separately on the two groups and report the parameter values in Table 1. Since each of the two parameters is obtained by holding the other fi ed, we ensured that the fina values are in equilibrium. The small γ for the myopic group results in a slightly updated belief about the myopic model after an observation. This reflect the increase in achievement scores observed across the test blocks in Fig. 3(a).

Parameters	General sum		Fixed sum	
	myopic	predictive	myopic	predictive
λ	0.77	0.69	0.43	3.02
γ	0.064	0.072	0.57	0.94

Table 1: Parameter values obtained by running gradient descent on the data from the two studies.

Fixed-sum game Of the 118 subjects who participated in the critical games, 59 were selected randomly and their data used to learn the parameters. We ran gradient descent separately on the two groups of subjects facing different opponents. In Fig. 6, we show plots of the negative log likelihood, X, with different values of the parameters for the two groups. Notice the presence of a distinct minima in each of the graphs. Gradient descent on these surfaces revealed the values for the parameters given in Table 1; column values are in equilibrium. We point out the relatively high value of λ for the predictive group, which reflect the very high achievement scores and low rationality errors for subjects in that group. Additionally, the low γ for the myopic group is indicative of the poor learning among subjects faced with a myopic opponent.

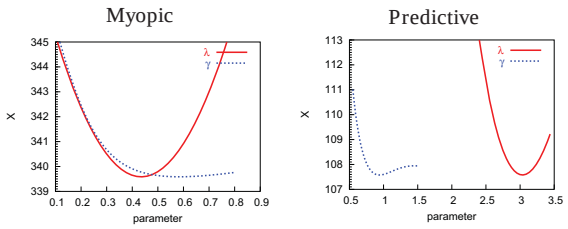


Figure 6: Negative log likelihoods for different values of the parameters for both the groups. We ran gradient descent to find parameters responsible for the minima.

4.3.2 Achievement Scores and Rationality Errors

We utilized the learnt values in Table 1 to parameterize the underweighting and quantal response choice models within the I-POMDP. We focused on the test set containing data of randomly picked subjects to evaluate the accuracy of our model. Using the subjects' expectations of the opponent type in the firs 5 games, we set the prior beliefs. For example, consistent expectations of the opponent type (regardless of being incorrect) resulted in assigning a highly informative prior. On the other hand, inconsistent expectations led to uninformative priors.

General-sum game We obtained model predictions that correspond to the 57 subjects in the test data, of which 29 faced a myopic opponent and 28 faced a predictive opponent. First, we visually compare the mean achievement scores of the model predictions with those of the study data. As we see in Fig. 7, model predictions closely align with the study data for the myopic group, with some difference between the two on the fina two test blocks for the predictive group, although the trend is consistent and the difference is not significant The difference is primarily due to the model predictions displaying less increase in the achievement score across test blocks 2 and 3. We quantify the closeness of the fi by computing mean squared error (MSE) of the predictions by our model (I-POMDP$_{i,2}$), and those of a random model (null hypothesis) for comparison. We show the MSE for both, the achievement score and the rationality error rate as predicted by the models, in Table 2. All differences in MSE between our model and the random model for both groups are significan ($p \leq 0.05$ on Student's t-test). In particular, level 1 predictions by our model when the opponent is myopic are highly consistent with the data; this was the dominant level of recursive reasoning in the general-sum game.

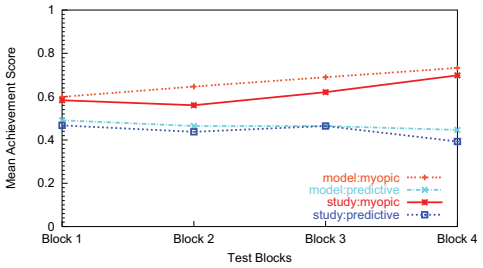


Figure 7: Comparison of model predictions with actual data for both groups in the general-sum game.

Fixed-sum game Predictions from our model were obtained that corresponded to the 40 game instances played by 29 subjects that faced a myopic opponent and 30 that faced a predictive opponent, for the fi ed-sum game. We compute the mean achievement score of our model predictions and visually compare it with that of the actual data, in Fig. 8. Notice the difference in the firs test block for the myopic group, which is due to the fact that subjects made more rationality errors initially compared to other test blocks. In Table 2, we show the MSE of the predictions by our model, and compare it with the MSE of the predictions by a random model (null hypothesis). Differences in the MSE between our model and the random model are statistically significan ($p < 0.05$ for t-test). Focusing on our model, we observe that the MSE of the predicted rationality errors is higher indicating that the quantal response model could be improved in how it models human choice and more appropriate models may be needed.

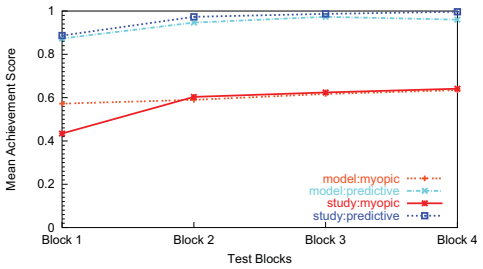


Figure 8: Comparison of model predictions with data for both groups in the fixed-sum game.

5. DISCUSSION

An alternate explanation of the high achievement scores observed in the fi ed-sum game for the predictive group could be that participants engaged in *backward induction (BI)* (or minimax) to solve the game rather than recursive thinking about opponent behavior. However, we provide three arguments against this explanation. First, in

Game type	Opponent type	Mean Squared Error (MSE)			
		Achievement score		Rationality error rate	
		Random	I-POMDP$_{i,2}$	Random	I-POMDP$_{i,2}$
General sum	myopic	0.01611	0.00341	0.03119	0.00087
	predictive	0.00443	0.00103	0.02264	0.02312
Fixed sum	myopic	0.01259	0.004828	0.08553	0.03527
	predictive	0.21490	0.0006278	0.09579	0.02853

Table 2: Goodness of the fit of our model with the study data. We include the random model for assessing significance.

anticipation of this argument participants were asked to reveal their line of thinking given an example game, during debriefing Two independent raters evaluated the participant responses for signs of rote use of BI. The raters agreed that 81.7% of the participants did not utilize BI. For the remaining participants, either the raters disagreed or were unable to clearly discern that BI was not utilized. Second, the relatively low achievement scores for the general-sum game – to which BI also applies – for both types of opponents indicate that the subject pool did not apply these methods. Finally, the presence of significan learning of opponent models as subjects played the games provides further evidence that participants were engaging in recursive reasoning.

A significan contributor to the error between the model predictions and study data was the initial performance of the participants in the firs few games. Although the participants underwent training, exposure to the diagnostic games resulted in rationality errors that were not consistent with the subjects' performance in remaining test blocks. Furthermore, subject's performance revealed a general trend of declining rationality errors, indicating that subjects became more attentive as the study progressed. Because we viewed the whole data and used a static λ, our models did not pick up this unexpected behavior. Investigating and modeling such phenomena is one aspect of our future work.

In comparison to many other two-player games, the Centipede game is particularly well-suited to rigorously measuring recursive thinking. This is because at each state, the corresponding player's rational action depends on how the other will act if given the chance, and not on other's previous action in that game. The game that we selected tested recursive reasoning up to two levels. Good performance by the subjects in the predictive group playing the fi ed-sum game suggests that they may think at higher levels; we are testing this hypothesis in an ongoing study.

Finally, it is tempting to include other types in the model set. However, this should be motivated by evidence of corresponding behavior in the data to avoid the set becoming intractably large.

Conclusion Recursive thinking of the form *what do I think that you think that I think* (and so on) arises naturally in interactive settings. While recognized frameworks for decision making in multiagent settings model recursive reasoning, the depth of the nesting is often arbitrary. We have shown that humans generally reason at low levels, but in simple settings exhibit higher levels of reasoning. These finding shed important insight into the type of models that should be ascribed to human agents in mixed settings. Our simplifie and augmented I-POMDP based model closely fitte the strategic behavioral data. Nevertheless, the good emulative behavior should not be interpreted as humans using POMDPs in their minds for decision making.

6. ACKNOWLEDGMENTS

This research is supported by grant FA9550-08-1-0429 from the Air Force Offic of Scientifi Research.

7. REFERENCES

[1] R. J. Aumann. Interactive epistemology i: Knowledge. *International Journal of Game Theory*, 28:263–300, 1999.

[2] R. J. Aumann. Interactive epistemology ii: Probability. *International Journal of Game Theory*, 28:301–314, 1999.

[3] A. Brandenburger. The power of paradox: Some recent developments in interactive epistemology. *International Journal of Game Theory*, 35:465–492, 2007.

[4] A. Brandenburger and E. Dekel. Hierarchies of beliefs and common knowledge. *Journal of Econ. Theory*, 59:189–198, 1993.

[5] C. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003.

[6] A. Colman. *Game Theory and Experimental Games*. Oxford: Pergamon Press, 1982.

[7] R. Dunbar. Theory of mind and evolution of language. In *Approaches to the Evolution of Language*. Cambridge University Press, 1998.

[8] R. Fagin, J. Geanakoplos, J. Halpern, and M. Vardi. The hierarchical approach to modeling knowledge and common knowledge. *International Journal of Game Theory*, 28, 1999.

[9] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge*. MIT Press, 1995.

[10] S. Ficici and A. Pfeffer. Modeling how humans reason about others with partial information. In *AAMAS*, pages 315–322, 2008.

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

[12] P. J. Gmytrasiewicz and E. H. Durfee. A rigorous, operational formalization of recursive modeling. In *ICMAS*, pages 125–132, 1995.

[13] P. J. Gmytrasiewicz and E. H. Durfee. Rational interaction in multiagent environments: Coordination. *Journal of AAMAS*, 3:319–350, 2000.

[14] J. C. Harsanyi. Games with incomplete information played by bayesian players. *Management Sc.*, 14(3):159–182, 1967.

[15] T. Hedden and J. Zhang. What do you think i think you think?: Strategic reasoning in matrix games. *Cognition*, 85:1–36, 2002.

[16] A. Heifetz and D. Samet. Topology-free typology of beliefs. *Journal of Econ. Theory*, 82:324–341, 1998.

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

[18] J. Mertens and S. Zamir. Formulation of bayesian analysis for games with incomplete information. *International Journal of Game Theory*, 14:1–29, 1985.

[19] D. Stahl and P. Wilson. On player's models of other players: Theory and experimental evidence. *Games and Econ. Behavior*, 10:218–254, 1995.