Adaptive Behavior in Variable Games Requires Theory of Mind

Wenhao Qi (wqi@ucsd.edu) and **Edward Vul (evul@ucsd.edu)**
Department of Psychology, University of California, San Diego
La Jolla, CA 92093 USA

Abstract

People seem to infer each others' beliefs and desires when navigating social interactions, perhaps because such a "theory of mind" can guide cooperation and coordination. However, such strategic, altruistic interactions fall naturally out of evolutionary game theory without invoking any theory of mind; so why is theory of mind useful? Here we show that the interactions studied in game theory have been too impoverished to require theory of mind, but when interacting in *variable games*, agents with theory of mind have a clear advantage. We use simulated tournaments to demonstrate that traditional action-level strategies such as tit-for-tat fare miserably in variable games, that goal-based agents can adapt to new games instantly, and that having a theory of mind is increasingly helpful for coping with a variety of opponents as the variability in games increases. Our work suggests that variable games merit further investigation in game theory and social sciences.

Keywords: repeated games; theory of mind; social value orientation

Introduction

Theory of mind refers to one's ability to impute mental states—desires and beliefs in particular—to oneself and others, in order to explain and predict behavior (Premack & Woodruff, 1978). It is considered universal in human adults, plays an important role in social development (Wellman, 1992), and its deficiency is associated with profound social impairments (Baron-Cohen, Leslie, & Frith, 1985; Brüne, 2005).

Despite its prominence in psychology, theory of mind has not been adequately studied in game theory, a formal framework for studying social interactions. To our knowledge, two lines of research in game theory relate to theory of mind. The first approach deals with games with incomplete information, i.e., in which the payoff functions of other players are unknown. It follows the central theme in classic game theory, which is to reduce the infinitely recursive reasoning about the players' beliefs ("I think that you think that I think. . . ") to a solvable equilibrium. A classic solution to this problem is the Bayesian game, in which "theory of mind" refers to the inference of others' payoff functions and a Bayesian Nash equilibrium can be derived (Harsanyi, 1968; Aumann, 1987; Robalino & Robson, 2012).

The second approach begins with the observation that people do not play at the theoretical equilibrium in many games and tend to operate on a very shallow level of recursion (Hedden & Zhang, 2002; Goodie, Doshi, & Young, 2012). This pattern of behavior is characterized by different recursion depths across individuals (Stahl, 1993; Camerer, Ho, & Chong, 2004; Yoshida, Dolan, & Friston, 2008; Mohlin, 2012). On this account, theory of mind pertains to a general belief about the recursion depth of opponents a given player may face. A common limitation of these two approaches is that they are not concerned with learning a model of other players or adapting to a particular opponent, which humans often do, and which the psychological notion of theory of mind seems particularly well-suited for. We will examine this limitation more closely for Bayesian games in the Discussion.

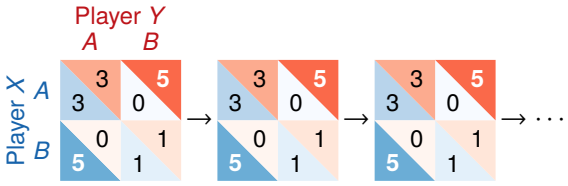

Figure 1: A *fixed* repeated game (2×2 normal form). More specifically, this is an iterated Prisoner's Dilemma. In each round of game, player X chooses either row A or row B, and player Y chooses either column A or column B. They make simultaneous choices, and the payoff for either player is determined by the values in the resulting cell. They repeatedly play games with the same payoff structure.

What remains unanswered is why evolution has endowed humans with theory of mind. A natural place to seek an evolutionary explanation in game theory is *repeated games*, in which different agents with different strategies play games with each other repeatedly. One of the most studied repeated games, iterated Prisoner's Dilemma (IPD, Figure 1), has been used as an explanation for the emergence of cooperation among selfish agents (Axelrod, 1984; Nowak, 2006). In Prisoner's Dilemma, the two possible actions for each player are usually designated as *cooperation* and *defection*. Tit-for-tat (TfT)—a strategy that starts with cooperation and then repeats the opponent's action from the previous turn—is a strong and robust strategy in evolutionary IPD (Axelrod, 1984).

The majority of repeated games studied thus far are *fixed* games, where the game is repeated with exactly the same payoff structure. Fixed repeated games do not incentivize theory of mind because simple action-level strategies like TfT are already unbeatable in such games. Moreover, real-world interactions, although often repeated with the same players, are

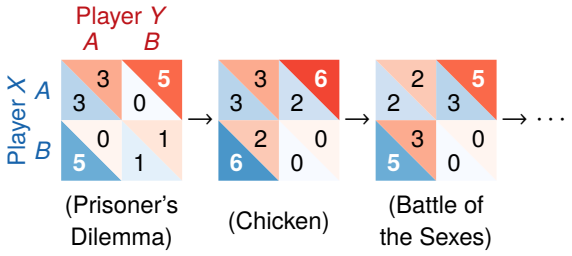

Figure 2: A *variable* 2×2 game. The payoff structure is changing across different rounds of games, and the colloquial names of the three sample rounds are given below. Although the examples are all symmetric games and the payoffs for AA are all greater than BB, such constraints do not hold generally in variable games.

far from fixed in their payoff structures. For instance, two PhD students collaborating on a research project may become direct competitors on the job market. We call this class of repeated games with variable payoff structure *variable games* (Figure 2; Kleiman-Weiner, 2018, Chapter 2). In this work, we use variable games (2×2 normal form) to provide a natural account of the advantage of having a theory of mind. Our work shares the spirit in Axelrod's pioneering work and the tradition of evolutionary game theory: identifying what conditions favor a particular behavioral capacity. People seem to cooperate even in *one-shot* Prisoner's Dilemmas that penalize cooperation, and Axelrod showed that cooperation is advantageous when the game is played *repeatedly*. People seem to have theory of mind, but *fixed* repeated games do not require or reward theory of mind, and we show that having a theory of mind is advantageous when the games *vary*.

The paper is organized as follows. First, we define goal-based agents with or without theory of mind that can play variable games adaptively. Second, we add variability to IPD in three steps and use a simulated tournament to demonstrate that having a theory of mind is beneficial as the variability in games increases. Finally, we discuss alternative explanations for theory of mind and implications of our work.

Goal-based Agents in Variable Games

From a cognitive science perspective, the *strategies* in game theory can be seen as defined on Marr's algorithmic level in terms of the raw actions (A or B in 2×2 games; Marr, 1982). In variable games, it is tricky, if not impossible, to define strategies on this level, since the payoffs corresponding to the actions, and thus the meaning of those actions, vary from game to game. This problem is obviated if we define strategies on the computational level. Concretely, each agent makes its decision by maximizing its expected utility in each round of the game. We call this type of agents *goal-based agents*.

An agent's utility function can take various forms. Here we draw inspiration from the literature on social value orientation (Van Lange, De Bruin, Otten, & Joireman, 1997), social discounting (Jones & Rachlin, 2006) and welfare tradeoff ratios (Delton & Robertson, 2016), in which one's utility function is a weighting of payoffs between self and others. We use a single parameter λ to account for this weighting, i.e., different degrees of altruism or spite directed toward the other player. In each 2×2 game, player X's utility function is

$$u_X(x, y) = v_X(x, y) + \lambda v_Y(x, y),$$

in which v_X and v_Y are the two players' actual payoffs in the game determined by their choices x and y, and λ remains the same through the sequence of games. For example, in the second game of Figure 2, $v_X(A, B) = 2$ while $v_Y(A, B) = 6$. The definition of u_Y is similar. This utility function captures both (i) every creature's desire to maximize its own payoff, and (ii) the consideration of other's payoffs as an altruist ($\lambda > 0$), an egoist ($\lambda = 0$), or a nemesis ($\lambda < 0$).

Now that we have a utility function, we can define theory of mind (ToM), which essentially means having a model of the opponent to help predict its choices. A level-0 agent is equivalent to having no theory of how the opponent will behave, so it assumes its opponent will choose A or B with equal probability. A level-1 ToM agent treats its opponent as a level-0 agent and does iterated Bayesian inference on its λ from its choices. In this way, we can define higher-level ToM, but we do not do that in the current work for three reasons. First, our main goal here is to show that having ToM is beneficial, and level-1 ToM is sufficient in this respect. Second, higher-level ToM introduces the problem of infinite recursion as in traditional game theory (Nash, 1951), and does not match people's tendency to use lower-level ToM (Hedden & Zhang, 2002). If we assume people's behavior is well adapted to the environment and the computational constraints (i.e., rational in the sense of Anderson (1990)), we can expect lower-level ToM to perform nearly as well as higher-level ToM, but with much less computational burden. Third, the parameters in higher-level ToM models are much harder to infer from the sparse information in 2×2 games (at most one bit of information in each round). In reality, people play much more complex games that provide rich information, and higher-level ToM would be more plausible in those situations.

To simplify the terms, we call a level-0 agent a "λ agent", and a level-1 ToM agent a "ToM agent". To describe the behavior of these agents in 2×2 normal-form games, we parameterize such games as in Figure 3a. A λ agent's expected utility for making either choice as player X is then

$$\begin{aligned} u_\lambda(A) &= \frac{1}{2} v_X(A, A) + \frac{1}{2} v_X(A, B) \\ &= \frac{1}{2}\left[(w_1 + w_3) + \lambda(w_2 + w_4)\right], \end{aligned}$$

and similarly for $u_\lambda(B)$. In the deterministic case, the agent would choose A whenever $u_\lambda(A) > u_\lambda(B)$ and vice versa. But to account for a noisy, and potentially imperfect maximizer, a λ agent uses a softmax function to determine its choice. Given

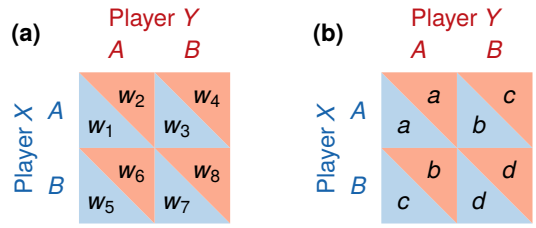

Figure 3: Parameterization of 2×2 normal-form games. **(a)** General form. **(b)** *Symmetric* games.

λ, its probability of choosing A is

$$p_A(\lambda) = \frac{\exp\left[\alpha u_\lambda(A)\right]}{\exp\left[\alpha u_\lambda(A)\right] + \exp\left[\alpha u_\lambda(B)\right]}, \qquad (1)$$

in which $\alpha > 0$ is the softmax parameter. Likewise, $p_B(\lambda) = 1 - p_A(\lambda)$. Note that this is the same decision rule as in the quantal response equilibrium in game theory (McKelvey & Palfrey, 1995).

A ToM agent treats its opponent as a λ agent and does Bayesian inference on λ. Its goal is to maximize its own utility function parameterized by λ_s ("s" for "self"). In each round of the game, the ToM agent samples a $\hat{\lambda}_o$ ("o" for "opponent") from its posterior distribution of λ_o and uses Equation 1 to compute the likelihood $p_A(\hat{\lambda}_o)$ and $p_B(\hat{\lambda}_o)$ for its opponent to choose either action[1]. Then it computes its own expected utility (assuming it is player Y)

$$\begin{aligned} u_{\text{ToM}}(A) &= p_A(\hat{\lambda}_o)u_Y(A,A) + p_B(\hat{\lambda}_o)u_Y(B,A) \\ &= p_A(\hat{\lambda}_o)(w_2 + \lambda_s w_1) + p_B(\hat{\lambda}_o)(w_6 + \lambda_s w_5), \end{aligned}$$

(similar for $u_{\text{ToM}}(B)$) and uses softmax to sample a choice. After its opponent makes a choice x and the game is played, it updates the posterior using x and the likelihood function (Equation 1).

Simulation

To compare the two goal-based agents with the action-level strategies, we pit them against an agent that plays the tit-for-tat strategy, the epitome of action-level strategies, named "TfT agent". To give TfT some advantage at the beginning, we start from IPD and add variability to it in three steps to form three game spaces: Variable Prisoner's Dilemma, Symmetric Games, and All Games[2]. We use a computer-simulated tournament to compare the performance of the three agents.

Variable Prisoner's Dilemma

The first step is to make the payoff values variable across rounds of games while ensuring that each game is a Prisoner's Dilemma. A symmetric 2×2 game is parameterized by a, b, c, d (Figure 3b). Prisoner's Dilemma requires $c > a > d > b$ and $2a > b + c$ [3]. In order to generate each game in the sequence, we sample 4 numbers independently from the uniform distribution between 0 and 1 ($U(0,1)$), sort them in descending order, and assign them to c, a, d, b. If $2a > b + c$ is not satisfied, we repeat the sampling until it is satisfied. We then apply a linear transformation to a, b, c, d to satisfy the normalizing constraints

$$a + b + c + d = 0, \quad a^2 + b^2 + c^2 + d^2 = 1.$$

We call such a distribution of games Variable Prisoner's Dilemma.

Simulation Details We include a TfT agent, λ agents with $\lambda = 0, 1, -1$, and ToM agents with $\lambda_s = 0, 1, -1$ in the tournament, let each pair of agents (including pairs of the same type) play 100 rounds of games sampled independently from Variable Prisoner's Dilemma, and repeat it 20 times. We set the softmax parameter $\alpha = 10$ everywhere. We set the ToM agents' prior distribution of λ_o to be uniform between -2 and 2, a range wide enough to include the nicest and the nastiest agents that are plausible. We discretize the posterior distribution with a grid step of 0.02.

Results Figure 4a shows the average per-game payoff each agent gains when playing against each agent. The scores shown are relative to the baseline score earned by a random-choosing agent when playing against each agent (because, e.g., playing against a $\lambda = 1$ opponent will yield a higher score for all agents as compared to playing against a $\lambda = -1$ agent). We are only comparing the three agents with $\lambda = 0$, for whom the average payoffs they gain reflect their ability to reach their goal. TfT can be seen as having an effective $\lambda = 0$, because it is an evolutionarily selected strategy, where agents maximizing their own payoffs survive in the long run.

From the "Mean" panel in Figure 4a, we can get a general idea of how well the three types of agents perform in an evolutionary sense. For Variable Prisoner's Dilemma, the three types of agents have similar performance. Note that this result is specific to the uniform distribution of the 7 types of agents in the environment. If, say, the distribution of agents is dominated by TfT, TfT would perform much better than the other two types of agents, who behave like defectors in Prisoner's Dilemma. On the other hand, if the distribution is dominated by $\lambda(1)$ or ToM(1), TfT would perform much worse, because it cannot exploit a generous agent. In other words, in Variable

[1]Calculating p_A and p_B via the sampled $\hat{\lambda}_o$ yields an unbiased estimate of the expectation of these quantities using the full posterior distribution over λ_o. Moreover, as the posterior becomes more concentrated through iterated Bayesian inference, there is less Monte Carlo variability in the samples.

[2]Other action-level strategies, such as "win-stay, lose-shift", outperform TfT in some variants of IPD (Nowak & Sigmund, 1993), and yet other strategies are tailored for other specific games. Each of these specialized strategies would maximize returns in a specific type of game, but these same strategies would fail to generalize across game types. For our present purposes it is sufficient to show this failure of generalization for TfT, but similar simulations can be done for other specialized games and strategies.

[3]This latter constraint is not required in the broad sense of Prisoner's Dilemma, but previous work on IPD usually imposes it to make mutual cooperation the only outcome with maximum total payoff for both players (Axelrod, 1984).

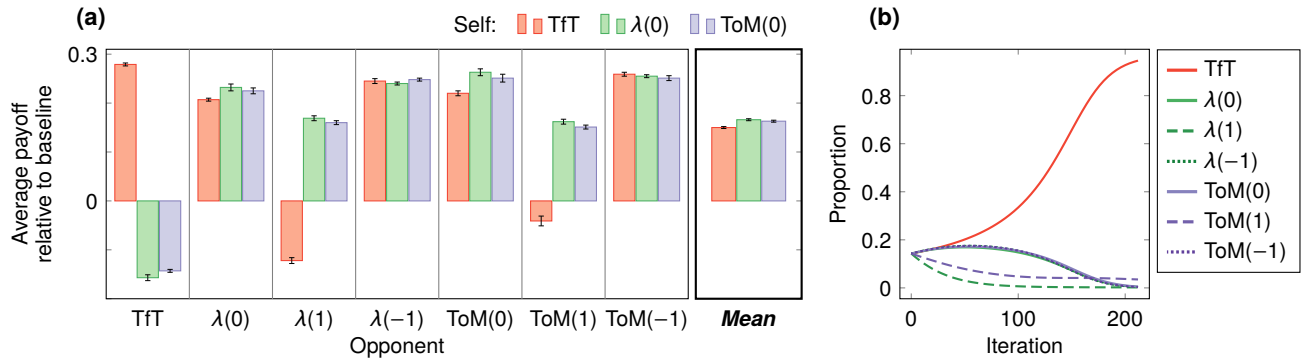

Figure 4: Simulation results for Variable Prisoner's Dilemma. $\lambda(0)$ refers to the λ agent with $\lambda = 0$; ToM(1) refers to the ToM agent with $\lambda_s = 1$; etc. **(a)** Comparison of performance of TfT, $\lambda(0)$ and ToM(0) when playing against the same type of opponent. The plotted values are the average per-game payoff "Self" earns after 100 rounds of the game with "Opponent", relative to the baseline score earned by a random-choosing agent. In each panel, the payoffs earned by the three types of agents playing against the same type of opponent are compared directly. Each value in the "Mean" panel is an average of the payoffs in the other 7 panels. The error bars indicate the standard error of the means (SEM) resulting from 20 repetitions. **(b)** Evolutionary process. This is a deterministic version of the Moran process that starts with equal proportions of the agents in the environment.

Prisoner's Dilemma, the performance of TfT relative to the other two types is contingent on the distribution of agents, while $\lambda(0)$ and ToM(0) always have similar performance.

Based on the pairwise average payoffs for the 7 agents, we also simulate an evolutionary process to compare their performance (Figure 4b). Let v_{ij} be the average payoff that agent i gets when playing against agent j. The environment starts with equal proportions of the 7 types of agents, and evolves according to a deterministic version of the Moran process (Moran, 1958). Let $p_i(t)$ be the proportion of agent i in the environment at iteration t. The fitness of agent i, $f_i(t)$, is defined as

$$f_i(t) = \sum_{j=1}^{7} v_{ij} p_j(t) \,,$$

i.e., the mean payoffs weighted by the proportion of the opponent. Then the proportions are updated as

$$p_i(t+1) = \frac{p_i(t) \cdot \exp f_i(t)}{\sum_{j=1}^{7} p_j(t) \cdot \exp f_j(t)} \,,$$

which means each agent's proportion is scaled by the softmax function of its fitness, and normalized to sum to one. Note that a thorough evolutionary analysis would require simulation based on different initial proportions, but equal initial proportions can serve as a good first-level analysis. As shown in Figure 4b, for Variable Prisoner's Dilemma, TfT dominates the population, and $\lambda(1)$ and ToM(1), whose dominance would put $\lambda(0)$ and ToM(0) at an advantage, die out very quickly. This replicates previous findings that TfT is strong and robust in IPD (Axelrod, 1984).

Symmetric Games

The second step is to consider all symmetric 2×2 games as parameterized in Figure 3b. We introduce an inequality $a > d$ to make TfT definable, since both players choosing A is better than both players choosing B. This also eliminates one of the redundant, symmetric halves of the game space. To generate each game, we sample 4 numbers independently from $U(0, 1)$ and assign them to a, b, c, d. If $a < d$, we reverse the order of the 4 numbers. Then we apply the linear transformation to a, b, c, d as in Variable Prisoner's Dilemma. We call this distribution Symmetric Games.

Results Apart from the game distribution, the simulation is identical to Variable Prisoner's Dilemma, and the results are shown in Figure 5. $\lambda(0)$ and ToM(0) significantly outperform TfT in terms of the mean payoff across opponents. This is the case because, unlike Variable Prisoner's Dilemma, the off-diagonal outcomes in Symmetric Games can change the expected value ordering of the two actions. Consequently, TfT's alternation between actions with no regard for the present game's payoffs yields systematically worse outcomes than agents that consider the payoffs. ToM(0) does better than $\lambda(0)$ when playing against λ agents, but worse against TfT and, even more, ToM(0) and ToM(1). The primary reason is that ToM agents do not have a correct model of TfT or ToM, which results in an error in prediction.

Another factor contributes to the particularly large difference when playing against ToM(0) and ToM(1), which, when computing the "Mean", cancels out ToM(0)'s advantage over $\lambda(0)$. Consider the comparison when the opponent is ToM(0). It will be shown later in Figure 7 that ToM(0) infers a λ close to 0 of ToM(0). It can be proved that it is a property of symmetric games that when playing against another ToM(0) agent, ToM(0) can never gain a higher score in any symmetric game than $\lambda(0)$, provided that in the ToM–ToM rivalry either ToM infers $\lambda = 0$ of its opponent. When we lift the constraint of symmetry in the next game space, ToM(0) overtakes $\lambda(0)$ even without an accurate model of the ToM opponent.

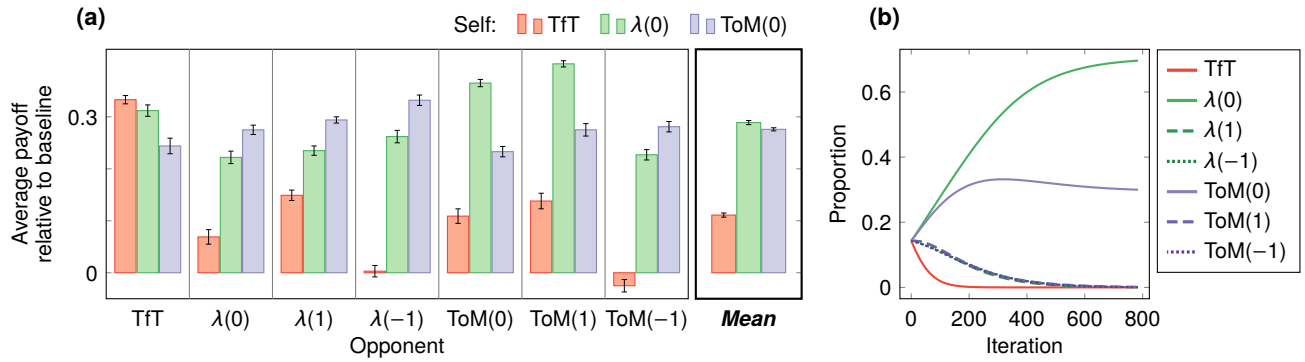

Figure 5: Simulation results for Symmetric Games. **(a)** Comparison of performance of TfT, $\lambda(0)$ and ToM(0). **(b)** Evolutionary process.

In the evolutionary simulation of Symmetric Games (Figure 5b), TfT quickly loses population share, even though we tried to help it by setting $a > d$; even nice agents like $\lambda(1)$ and ToM(1) fare better than TfT. The evolutionary process converges on a mixture of about 70% of $\lambda(0)$ and 30% of ToM(0). $\lambda(0)$'s advantage over ToM(0) is due to the ToM disadvantage specific to symmetric games discussed above.

All Games

The third step in extending IPD is to lift the constraint of symmetry. The game is then parameterized by the 8 payoff values $w_1, w_2, \ldots, w_8$ (Figure 3a). For each game, we draw 8 independent samples from $U(0,1)$, assign them to the 8 payoffs, and impose the normalizing constraints

$$\sum_{i=1}^{8} w_i = 0\,, \quad \sum_{i=1}^{8} w_i^2 = 2\,.$$

We call this distribution All Games.

Results Again, apart from the game distribution, the simulation is identical to Variable Prisoner's Dilemma, and the results are shown in Figure 6. In this game distribution, TfT is no different from the random-choosing baseline, and ToM(0) generally outperforms $\lambda(0)$. In the evolutionary simulation, TfT still dies out first, and ToM(0) dominates in the end.

λ inferred by ToM(0)

To verify that ToM agents' superior performance stems from having an accurate model of the opponent, we inspected ToM(0)'s posterior of the opponent's λ at the 100th round for each type of opponent in each game space, as sketched in Figure 7.

In Variable Prisoner's Dilemma, the ToM agent can infer $\lambda(0)$'s and $\lambda(1)$'s λs accurately and consistently, but not for $\lambda(-1)$. This is because in a Prisoner's Dilemma, a λ agent will always defect as long as its $\lambda < 0$; therefore, consistent defecting provides no information about the exact value of λ. When the opponent is TfT or ToM, a ToM agent does not have an accurate model of it, but we can still interpret the inferred λ as indicating how the opponent generally behaves in a particular distribution of games. When playing against ToM(0) in Variable Prisoner's Dilemma, TfT behaves like a selfish agent, ToM(0) and ToM(−1) behave just like $\lambda(0)$ and $\lambda(-1)$, and ToM(1) behaves like a somewhat nice agent.

In Symmetric Games, the patterns of λ inferred by ToM(0) for different opponents are similar to Variable Prisoner's Dilemma. TfT behaves slightly more nicely. The inferences for $\lambda(-1)$ and ToM(−1) are much more accurate than in Vari-

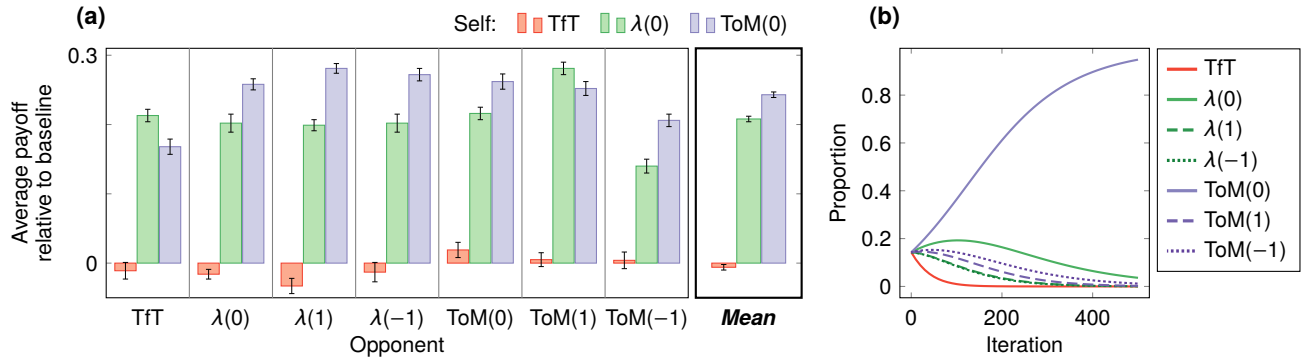

Figure 6: Simulation results for All Games. **(a)** Comparison of performance of TfT, $\lambda(0)$ and ToM(0). **(b)** Evolutionary process.

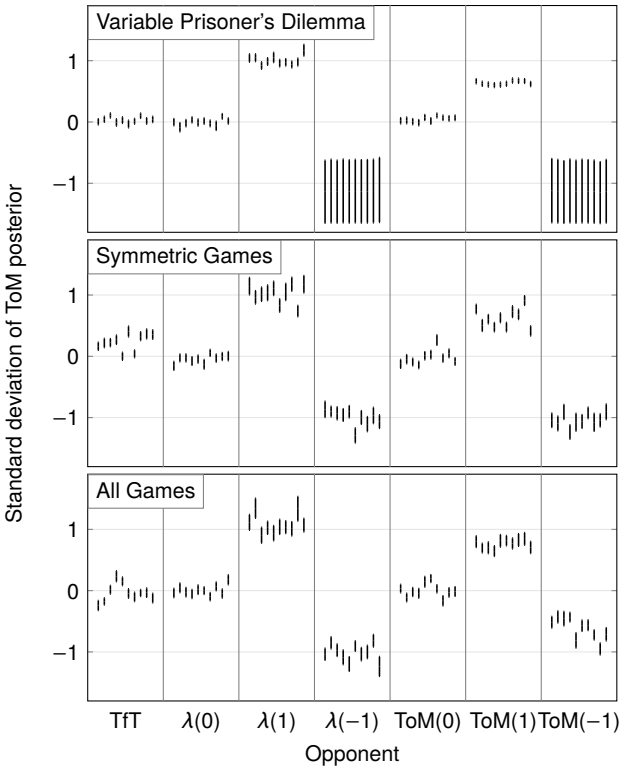

Figure 7: ToM(0)'s posterior distribution of λ_o for each type of opponent in each game space. Each vertical line segment reflects the overall shape of the posterior at the last (100th) round of the game. The segment is centered on the mean of the posterior, and its length is twice the standard deviation. In each panel, the 10 segments represent 10 repetitions.

able Prisoner's Dilemma as the more varied payoff structures differentiate degrees of spite.

Finally, in All Games, the λs inferred by ToM(0) is similar to before, except being more symmetric as the game space is completely symmetric (in a different sense than symmetric games).

Discussion

We have presented a preliminary demonstration that action-level strategies like tit-for-tat cannot adapt to the variability in the games as opposed to goal-based agents, and that having a theory of mind is beneficial as the variability increases, which provides an explanation for why people have theory of mind. This work also suggests that variable games are an important setting to study in both theoretical and empirical game theory.

Bayesian games seem to explain people's theory of mind by suggesting that people use theory of mind to deal with games with incomplete information (Robalino & Robson, 2012). However, Bayesian games are not concerned with theory of mind in the sense of learning a model of other players. In one-shot Bayesian games, there is no learning whatsoever (Harsanyi, 1968). The likelihood function in Bayesian games is the distribution of the other player's type given one's own type, instead of the distribution of the other player's choices given its type as discussed in this paper. In repeated Bayesian games, each player learns others' payoff functions through repeated play (Kalai & Lehrer, 1993; Jordan, 1991). But since the repeated games are *fixed*, the payoff function is learned on the action level, which is not a very useful model for generalization.

Recent work by Robalino and Robson (2016) aims to provide a similar evolutionary explanation of theory of mind as ours. In their work, however, learning still occurs on the action level and does not generalize to truly novel situations. Concretely, the "sequentially rational theory of preference" player (SR-ToP, comparable to the ToM agent in the current work) memorizes the other player's preference over outcomes, so when the subgame with the same two possible outcomes occurs again, SR-ToP can predict its opponent's choice. The environment is "variable" in that new outcomes are introduced gradually, giving SR-ToP an evolutionary advantage over a naïve player that does not memorize its opponent's preferences at all. In this framework, each new outcome has to be learned anew, therefore still no parsimonious model of other players is learned and the SR-ToP cannot adapt to brand-new games instantly, which humans must do in a dynamic social environment.

Acknowledgments

We thank three anonymous reviewers for insightful comments. This work was supported by UCSD Academic Senate Grant RG095178.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Aumann, R. J. (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica*, *55*(1), 1–18.

Axelrod, R. (1984). *The evolution of cooperation*. New York, NY: Basic Books.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Brüne, M. (2005). "Theory of mind" in schizophrenia: A review of the literature. *Schizophrenia Bulletin*, *31*(1), 21–42.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, *119*(3), 861–898.

Delton, A. W., & Robertson, T. E. (2016). How the mind makes welfare tradeoffs: Evolution, computation, and emotion. *Current Opinion in Psychology*, *7*, 12–16.

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.

Harsanyi, J. C. (1968). Games with incomplete informa-

tion played by "Bayesian" players, Part I–III. *Management Science*, *14*(3,5,7), 159–182,320–334,486–502.

Hedden, T., & Zhang, J. (2002). What do you think I think you think?: Strategic reasoning in matrix games. *Cognition*, *85*(1), 1–36.

Jones, B., & Rachlin, H. (2006). Social discounting. *Psychological Science*, *17*(4), 283–286.

Jordan, J. S. (1991). Bayesian learning in normal form games. *Games and Economic Behavior*, *3*(1), 60–81.

Kalai, E., & Lehrer, E. (1993). Rational learning leads to Nash equilibrium. *Econometrica*, *61*(5), 1019–1045.

Kleiman-Weiner, M. (2018). *Computational foundations of human social intelligence*. Thesis, Massachusetts Institute of Technology.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W. H. Freeman and Company.

McKelvey, R. D., & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, *10*(1), 6–38.

Mohlin, E. (2012). Evolution of theories of mind. *Games and Economic Behavior*, *75*(1), 299–318.

Moran, P. A. P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, *54*(1), 60–71.

Nash, J. (1951). Non-cooperative games. *Annals of Mathematics*, *54*(2), 286–295.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560–1563.

Nowak, M. A., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, *364*(6432), 56–58.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

Robalino, N., & Robson, A. (2012). The economic approach to "theory of mind". *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2224–2233.

Robalino, N., & Robson, A. (2016). The evolution of strategic sophistication. *American Economic Review*, *106*(4), 1046–1072.

Stahl, D. O. (1993). Evolution of smart$_n$ players. *Games and Economic Behavior*, *5*(4), 604–617.

Van Lange, P. A. M., De Bruin, E. M. N., Otten, W., & Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of Personality and Social Psychology*, *73*(4), 733–746.

Wellman, H. M. (1992). *The child's theory of mind*. Cambridge, MA: The MIT Press.

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLOS Computational Biology*, *4*(12), 1–14.