Bayesian Inference of Other Minds Explains Human Decisions in a Group Decision Making Task

Koosha Khalvati[1], Seongmin A. Park[2,3], Remi Philippe[3], Mariateresa Sestito[3], Jean-Claude Dreher[3,*], and Rajesh P. N. Rao[1,4,*]

[1] *Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, USA*
[2] *Center for Mind and Brain,University of California, Davis, USA*
[3] *Neuroeconomics Lab, Institut des Sciences Cognitives Marc Jeannerod, Lyon, France*
[4] *Center for Neurotechnology, University of Washington, Seattle, USA*
[*] *Joint senior authors*

Abstract

To make decisions in a social context, humans try to predict the behavior of others, an ability that is thought to rely on having a model of other minds. Having such a model, known as a theory of mind, allows one to infer the intentions of others, simulate their beliefs, and predict their possible actions, taking into account the fact that others too have a similar theory of mind. Here we show that Bayesian inference of other minds explains human decisions in a group decision making task known as the Volunteer's Dilemma. Our Bayesian model incorporates the effect of ones own actions on future rewards that could accrue to the social group one belongs to. Quantitative results from our normative model of human social decision making suggest that humans maintain a model of other minds and use this model to infer the future actions of others when deciding on their current action. We show that our model explains human data significantly better than model-free reinforcement learning and other previous models.

1 Introduction

The importance of social decision making in human behavior has spawned a large body of research in social and decision neuroscience (Sanfey, 2007; Camerer, 2011; Henrich et al., 2005; Behrens et al., 2009; Rilling and Sanfey, 2011; Dunne and ODoherty, 2013; Ruff and Fehr, 2014; Joiner et al., 2017). Human behavior relies heavily on predicting future states of the environment under uncertainty and choosing appropriate actions to achieve a goal. In a social context, the degree of uncertainty about the possible outcomes increases dramatically because the behavior of other human beings can be much more difficult to predict than the physics of the environment. There are two major approaches for action selection in decision making: "model-free" and "model-based" (Dayan, 2012; Sutton and Barto, 1998; Daw and Dayan, 2014). In the model-free approach, the decision maker performs actions based on the past history of actions and rewards. At each step, the chosen action is the one with the maximum average (or total) reward obtained from previous trials (McAllister, 1991; Mookherjee and Sopher, 1997; Dickinson and Balleine, 2002). In this approach, there is a direct link between choices and past rewards. On the other hand, in model-based decision making, the subject learns a model of the environment, updates the model based on observations and rewards, and chooses actions based on the current state of the model (Daw et al., 2011; Culbreth et al., 2016). As a result, the relationship between rewards received and current choice is indirect. Besides the history of rewards received, the learned model can include other factors such as potential future rewards and more general rules about the environment. Therefore, the model-based approach is more flexible than model-free decision making (Doll et al., 2012; Dolan and Dayan, 2013). On the other hand, model-based decision making requires more cognitive resources, for example, for simulation of future events. The basic assumption underlying both approaches, and decision making in general, is that the decision maker seeks to maximize a utility function (Glscher et al., 2010; Seo and Lee, 2017).

The model-free approach is similar in social versus non-social decisions, involving in both cases a mapping from states to choices based on reward history. In the model-based approach, the main difference is that in social decisions, the learned model includes models for other humans. Having a theory of mind (ToM, i.e, the ability to model others' minds) allows one to infer intentions of others, simulate their beliefs and predict their possible actions, taking into account the fact that they too have a similar theory of mind (Wimmer and Perner, 1983; Amodio and Frith, 2006). Having a theory of mind requires the ability to handle uncertainty about others' intentions and behavior, thereby introducing a need for probabilistic reasoning. Such reasoning involves combining probabilistically any prior knowledge about others (based on past experiences with them) with their more recent actions to update one's belief about their intentions and impending actions. Additionally, assuming others have similar reasoning abilities as our own, one must also take into account others' beliefs about our own intentions and actions. Thus, in multiple-round social decision making tasks, an individual must simulate the effect of one's own decision on the future decisions of others (Coricelli and Nagel, 2009; Yoshida et al., 2010; Hill et al., 2017; Nagel et al., 2018).

The majority of theoretical frameworks used to model feedback-dependent changes in decision making strategies, such as reinforcement learning (RL) and related Markov Decision Process (MDP) models, assume that optimal decisions can be determined from the state of the decision maker's environment and that this state is fully observable to the decision maker (Costa and Averbeck, 2013; Ligneul et al., 2016; O'doherty et al., 2017). Clearly, these assumptions do not capture the reality and complexity of human social decision making. One powerful approach to studying social decision making has been to use tasks developed in behavioral economics using the theoretical framework of game theory (Hampton et al., 2008; Coricelli and Nagel, 2009; Yoshida et al., 2010). This approach has proven successful for studying social decision-making in laboratory settings and to probe the underlying neural systems using fMRI or electrophysiology.

Here, we adopt a different, more general theoretical approach based on the hypothesis that the brain performs Bayesian inference based on observations using probabilistic representations of the world and utilizes the results of Bayesian inference to choose optimal actions. Our approach is grounded in the rigorous theoretical framework of Partially Observable Markov Decision Processes (POMDPs) developed in the field of Artificial Intelligence for solving tasks involving decision making under uncertainty (Kaelbling et al., 1998). In POMDP models, it is assumed that the state dynamics are Markov (next state only depends on the current state) and the agent cannot directly observe the state. Instead, it must use its sensors to obtain observations and maintain a probability distribution (”belief”) over the set of possible states, based on the observations, observation probabilities, and the underlying dynamics. An exact solution to a POMDP yields the optimal action for each possible belief over possible states. The optimal action maximizhttps://v2.overleaf.com/project/5b96b54298d6c075a131c1e4es (or minimizes) the expected reward (or cost) of the agent over a possibly infinite horizon. The mapping of each possible belief to an optimal action is known as an optimal policy for the agent for interacting with its environment.

In this article, we investigate whether POMDPs can provide a formal probabilistic framework for understanding human behavior in the context of group decision making tasks involving reasoning about others' intentions. Our work builds on previous research on optimal inference about a hidden variable in the environment in perceptual decision making and simple social decision making tasks involving two players (Dayan and Daw, 2008; Rao, 2010; Huang and Rao, 2013; Khalvati and Rao, 2015; Hula et al., 2015; Baker et al., 2017). We extend this line of research to tasks involving reasoning about the behavior of groups of several people. A preliminary version of this approach was presented in Khalvati et al. (2016).

Here we show that when success in a group decision making task requires accurate prediction of others' intentions and long-term outcomes, human behavior is very similar to a POMDP strategy based on optimal inference about others' beliefs. We compare the POMDP model to a model-free method and other state-of-the-art methods for fitting human behavior in our social decision making task. Unlike the POMDP model, these methods are not normative and are descriptive models that derive the action of a player in the current round from their actions in previous rounds (Fischbacher et al., 2001; Wanga et al., 2012; Wunder et al., 2013). We focus specifically on the Volunteer's Dilemma task, wherein a few individuals endure some costs to benefit the whole group (Olson, 1971; Diekmann, 1985; Archetti and Scheuring, 2011). Examples of the task include guarding duty, blood donation, and stepping forward to stop an act of violence in a public place (Darley and Latane, 1968). Volunteering leads to a dilemma wherein the decision that maximizes an individual's utility differs from the strategy which maximizes benefits to the group to which the individual belongs. Moreover, in the Volunteer's Dilemma, not only is the common goal not reached when there are not enough volunteers, but having more than the required number of volunteers leads to a waste of resources. As a result, an accurate prediction of others' intentions based on one's beliefs is crucial to make accurate decisions. In order to mimic the Volunteer's Dilemma in a laboratory setting, we use the binary version of a multiround Public Goods Game (PGG) (Fehr and Gachter, 2000; Krajbich et al., 2009; Hauert et al., 2002). In the classical PGG task, the group reward is a linear function of total contribution from group members. In each round of the binary PGG task we use, at least k players (out of N total players, $N > k$) need to contribute in order to generate a bigger reward for the entire group (Palfrey and Rosenthal, 1984; Diekmann, 1985; Archetti, 2009b,a; Archetti and Scheuring, 2011).

We show that the traditional POMDP model can be extended to social decision making scenarios by incorporating models of other agents, allowing us to make testable predictions about how the brain makes inferences about others' behaviors. Our POMDP model not only explains actions of subjects, but also the underlying computational mechanism behind those actions. Also, in contrast to other methods, when fit to a subject's behavior, it can predict other events during the experiment such as success or failure of producing the group reward.

2 Results

2.1 Human Behavior in Public Goods Game

Participants were 29 adults (mean age 22.97 years old ± 0.37, 14 women). Based on self-reported questionnaires, none of them reported a history of neurological or psychiatric disorders. This

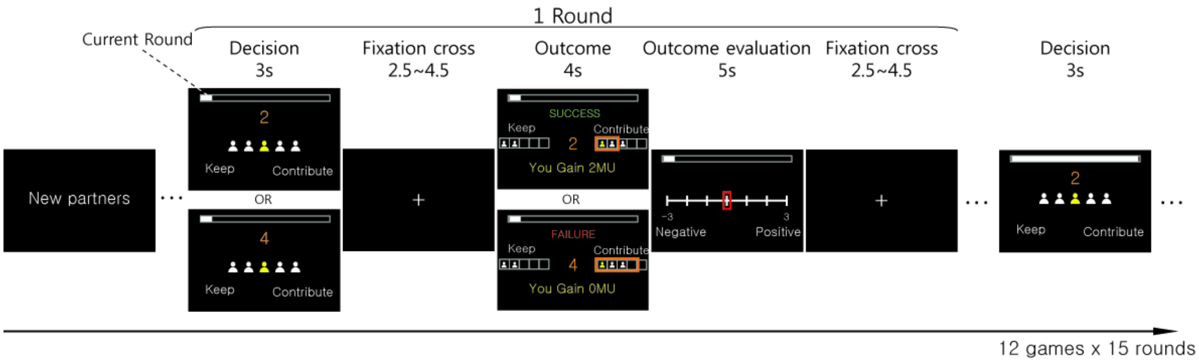

Figure 1: Multi-Round Public Goods Game (PGG). The figure depicts the sequence of computer screens a subject sees in one round of the PGG. The subject is assigned 4 other players as partners and each round requires the subject to make a decision: Keep 1 monetary unit (i.e., free-ride) or contribute. The subject knows whether the threshold to generate public goods (reward of 2 MU for each player) is 2 or 4 contributions (from the 5 players). After the subject acts, the overall outcome of the round is revealed (success or failure in securing the 2 MU group reward) but each individual player's actions remain unknown.

study was approved by the Institutional Review Board of the local ethics committee from Parma University (IRB no. A13-37030), and all participants gave their informed written consent. We analyzed the behavioral data of 12 Public Goods Games (PGGs) in which participants played 15 rounds of the game within the same group of N players ($N = 5$); feedback was given at the end of every trial. Figure 1 depicts one round of the PGG task. In each session, the subject played with a different group of players. No communication among players was allowed before or during the experiment. At the beginning of each round, 1 monetary unit (MU) was endowed (E) to each player. In each round, a player could choose between two options: *contribute or* free-ride.

Contribution had a cost of C = 1 MU, implying that the player could choose between keeping their initial endowment or giving it up. Public goods were produced as the group reward (G = 2 MU to each player) if and only if at least k players each contributed 1 MU. k was set to 2 or 4 randomly for each session and conveyed to group members before the start of the session. The resultant reward after one round was E - C + G = 2 MU for the contributor and E + G = 3 MU for the free-rider when public good was produced (the round was a SUCCESS). On the other hand, the contributor had E - C = 0 MU and the free-rider has E = 1 MU when no public good was produced (the round was a FAILURE). Although subjects were told that they were playing with other humans, in reality, they were playing with a computer that generated the actions of all the other $N - 1 = 4$ players based on an algorithm (see Methods and Materials).

As shown in Figure 2a, subjects contributed significantly more when the number of required volunteers was higher with an average contribution rate of 55% ($SD = .31$) for $k = 4$ in comparison to 33% ($SD = .18$) for $k = 2$ (one-tailed paired sample t-test, $t = 3.94$, $df = 29$, $p = 0.00025$, 95% CI difference = $[0.08, 0.36]$). In addition, Figure 2b shows that the probability of generating public good was significantly higher when $k = 2$ with a SUCCESS rate of 87% ($SD = 0.09$) compared to 36% ($SD = .29$) when $k = 4$ (one-tailed paired sample t-test, $t = 10.08$, $df = 29$, $p = 4.0 \times 10^{11}$, 95% CI difference =$[0.39, 0.62]$) (Figure 2b). All but 6 of the subjects contributed more when $k = 4$ (Figure 2c). Of these 6 players, 5 chose to free-ride more than 95% of the time. Also, SUCCESS rate was higher when $k = 2$ for all players (Figure 2d).

Average contribution did not change significantly as subjects played more games. In each

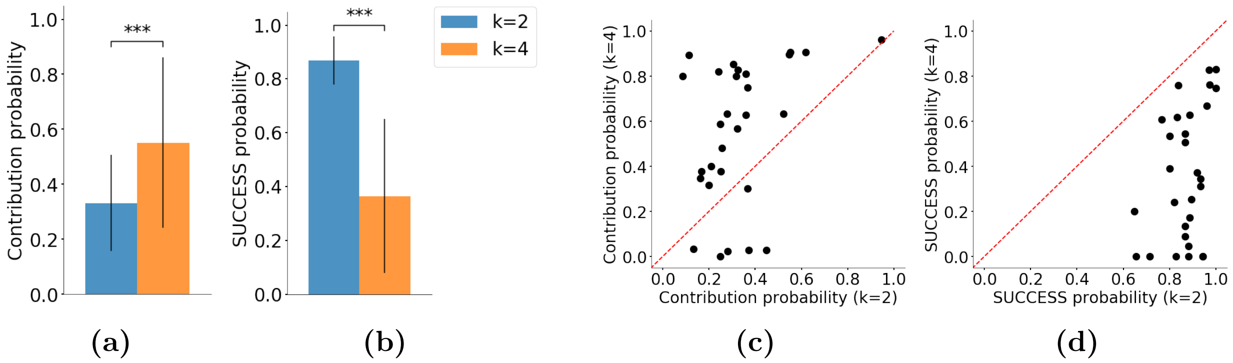

Figure 2: Human Behavior in the PGG Task (a) The average contribution probability across subjects is significantly higher when the task requires more volunteers (k) to generate the group reward. (b) Average probability of SUCCESS across all subjects in generating the group reward is significantly higher when k is lower. (c) Average probability of contribution for each subject for $k = 2$ versus $k = 4$. Each point represents a subject. Subjects tend to contribute more often when the task requires more volunteers. (d) Average SUCCESS rate for each subject was higher for $k = 2$ versus $k = 4$.

condition, most of the players played at least 5 games (27 for $k = 2$ and 26 for $k = 4$). For $k = 2$, in their first game, the average contribution rate of players was .37 ($SD = .25$) while in their fifth game, it was .30 ($SD = .24$) (one-tailed paired sample t-test, $t = 1.34$, $df = 27$, $p = 0.095$). When $k = 4$, the average contribution rate was .57 ($SD = .30$) in the first game and .61 ($SD = .35$) in the fifth game (one-tailed paired sample t-test, $t = .69$, $df = 26$, $p = 0.25$).

2.2 Probabilistic Model of Theory of Mind for the Public Goods Game

Consider one round of the PGG task. A player can be expected to choose an action (*contribute* or *free ride*) based on their prediction of the number of contributions they expect the others to make in that round. Since the actions of individual players remain unknown through the game, the only observable parameter is the total number of contributions. One can therefore model this situation using a single random variable θ, denoting the average probability of contribution by any group member. With this definition, the total number of contributions by all the other members of the group can be expressed as a binomial distribution. Specifically, if θ is the probability of contribution by each group member, the probability of observing m contributions from the $N-1$ others in a group of N people is:

$$P(m|\theta) = \binom{N-1}{m} \theta^m (1-\theta)^{N-1-m}. \tag{1}$$

Using this probability, a player can calculate the expected number of contributions from others, compare with k and decide whether to contribute or free-ride accordingly. For example, if θ is very low, there is not a high probability of observing $k-1$ contributions by others, implying free riding is the best strategy.

There are two important facts that make this decision making more complex. First, the player does not know θ. θ must be estimated from the behavior of the group members. We assume that there is a probability distribution over θ in the player's mind, representing their belief about the cooperativeness of the group. Each player starts with an initial probability distribution, called the prior belief about θ and updates this belief over successive rounds based on the actions of the others. The prior belief may be based on the previous life experience of the player, or what they believe others would do through fictitious play (Brown, 1951). For example, the player may start with a prior belief that the group will be a cooperative one but change this belief after observing low numbers of contributions by the others. Such belief updates can be performed using Bayes' rule to invert the probabilistic relationship between θ and the number of contributions given by Equation 1.

A suitable prior probability distribution for estimating the parameter θ of a Binomial distribution is the Beta distribution, which is itself determined by two (hyper) parameters α and β:

$$\begin{aligned} \theta &\sim Beta(\alpha, \beta). \\ Beta(\alpha, \beta) &: P(x|\alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1} \end{aligned} \tag{2}$$

Starting with a prior probability $Beta(\alpha_0, \beta_0)$ for θ, the player updates their belief about θ after observing the number of contributions from others in each round through Bayes' rule. This updated belief is called the posterior probability of θ. The posterior probability of θ in each round serves as the prior for the next round. With the Beta distribution as the prior and the Binomial distribution as the distribution of observations (here, the number of contributions), the posterior probability distribution remains a Beta distribution (details in Methods and Materials) (Murphy, 2012). In our case, with a prior of $Beta(\alpha_{t-1}, \beta_{t-1})$ after observing c contributions (including one's own when made) in round t, the posterior probability of θ becomes $Beta(\alpha_t, \beta_t)$ where $\alpha_t = \alpha_{t-1} + c$ and $\beta_t = \beta_{t-1} + N - c$. Note that we include one's own action in the update of the belief because one's own action can change the future contribution level of others.

Intuitively, α represents the number of contributions made thus far and β the number of free rides. α_0 and β_0 (that define prior belief) represent the player's a priori expectation about the relative number of contributions versus free-rides respectively before the session begins. For example, when α_0 is larger than β_0, the player starts the task with the belief that people will contribute more than free ride. Large values of α_0 and β_0 imply that the player comes to the game with a lot of prior observations about other players and therefore, the player's belief will not change significantly after one round of the game when updated with the relatively small number c as above.

Decision making in the PGG task is also made complex by the fact that the actual cooperativeness of the group itself (not just the player's belief about it) may change from one round to the next: players observe the contributions of others and may change their own strategy for the next round. For example, players may start the game making contributions but change their strategy to free riding if they observe a large number of contributions by others. We model this phenomenon using a parameter $\gamma \leq 1$, which serves as a discount factor for the prior: the prior probability for round t is modeled as $Beta(\gamma\alpha_{t-1}, \gamma\beta_{t-1})$, which allows recent observations about the contributions of other players to be given more importance than observations from the more distant past. Thus, in a round with c total contributions (including the subject's own contribution when made), the subject's belief about the cooperativeness of the group as a whole changes from $Beta(\alpha_{t-1}, \beta_{t-1})$ to $Beta(\alpha_t, \beta_t)$ where $\alpha_t = \gamma\alpha_{t-1} + c$ and $\beta_t = \gamma\beta_{t-1} + N - c$.

2.3 Action Selection

How should a player decide whether to contribute or free ride in each round? One possible strategy is to maximize the reward for the current round by calculating the expected number of

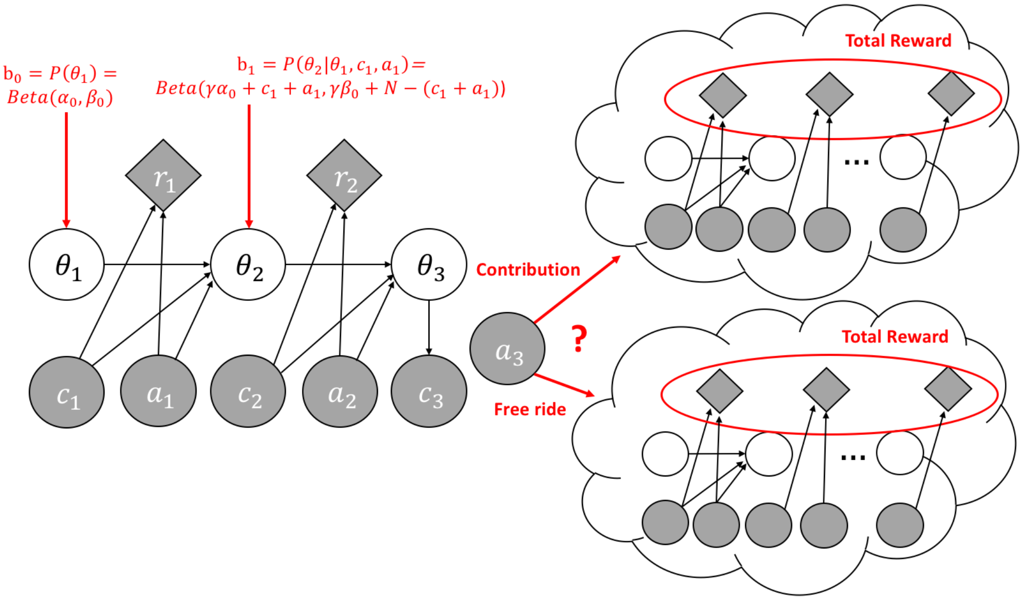
$b_0 = P(\theta_1) = Beta(\alpha_0, \beta_0)$
$b_1 = P(\theta_2|\theta_1, c_1, a_1) = Beta(\gamma\alpha_0 + c_1 + a_1, \gamma\beta_0 + N - (c_1 + a_1))$
r_1
r_2
θ_1
θ_2
θ_3
c_1
a_1
c_2
a_2
c_3
a_3
Contribution
?
Free ride
Total Reward
Total Reward
...
...

Figure 3: POMDP Model of the Multi-Round Public Goods Game. The subject does not know the average probability of contribution of the group. The POMDP model assumes the subject maintains a probability distribution (”belief”) about the group’s average probability of contribution (denoted here by θ) and updates this belief about θ after observing the outcome of each round. The POMDP model chooses an action that maximizes total expected reward across all rounds based on the current belief and the consequence of the action on group behavior in future rounds.

contributions by others based on the current belief. Using equation 1 and the prior probability distribution over θ, the probability of seeing m contributions by others when the belief about the cooperativeness of the group is $Beta(\alpha, \beta)$ is given by:

$$\begin{aligned} P(m|\alpha,\beta) &= \int_0^1 P(m|\theta)P(\theta|\alpha,\beta)d\theta \propto \int_0^1 \binom{N-1}{m}\theta^m(1-\theta)^{N-1-m}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\ &\propto \binom{N-1}{m}\int_0^1 \theta^{\alpha+m-1}(1-\theta)^{\beta+N-m-2}d\theta \end{aligned} \tag{3}$$

One can calculate the expected reward for the *contribute* versus *free-ride* actions in the current round based on the above equation. Maximizing this reward however is not the best strategy. As alluded to earlier, the actions of each player can change the behavior of other group members in future rounds. The optimal strategy therefore is to calculate the cooperativeness of the group through the end of the session and consider the reward over all future rounds in the session before selecting the current action. Such long-term reward maximization based on probabilistic inference of hidden state in an environment (here, θ, the probability of contribution of group members) can be modelled using the framework of Partially Observable Markov Decision Processes (POMDPs). Further details can be found in Methods and Materials but briefly, to maximize the total expected reward, Markov decision processes start from the last round, calculate the reward for each action and state, and then step back one time step to find the optimal action for each state in that round. This process is repeated in a recursive fashion. Figure 3 shows a schematic of the PGG experiment modeled using a POMDP.

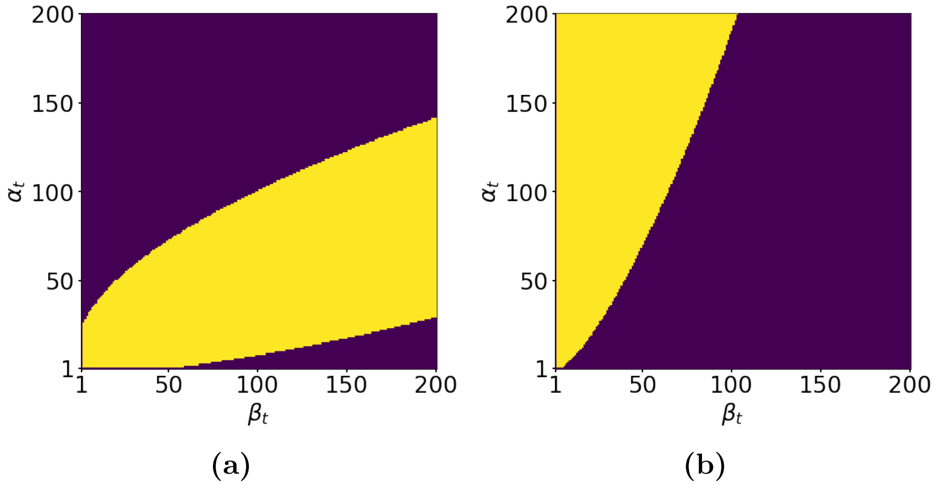

Figure 4: Optimal Actions prescribed by the POMDP Policy as a Function of Belief State. Plot (a) shows the policy for $k = 2$ and plot (b) for $k = 4$. The purple regions represent those belief states (defined by α_t and β_t) for which free-riding is the optimal action; the yellow regions represent belief states for which the optimal action is contributing. These plots confirm that the optimal policy depends highly on k, the number of required volunteers. The policies make sense intuitively: for example, when the contribution rate from others is so low that getting the group reward is unlikely, it makes sense to free-ride (regime of large β_t, small α_t in (a)) and likewise when the contribution rate is so high that a contribution from the subject is unnecessary (regime of large α_t, small β_t in (a)). (For the two plots, the decay rate was 1 and t was 9).

As an example of the POMDP model's ability to select actions for the PGG task, Figures 4a and 4b show the best actions for a given round (here, round 9) as prescribed by the POMDP model for $k = 2$ and $k = 4$ respectively (the number of minimum volunteers needed); the best actions are shown as a function of different belief states the subject may have, expressed in terms of the different values possible for belief parameters α_t and β_t. This mapping from belief to actions is called a *policy*.

We find that the POMDP policy aligns with our intuition about action selection in the Volunteer's Dilemma task. A player chooses to free ride for two reasons: (i) when the cooperativeness of the group is low and therefore there is no benefit in contributing, and (ii) when the player knows there are already enough volunteers and contributing leads to a waste of resources. The two purple areas of Figure 4a represent these two conditions for $k = 2$. The upper left part represents large α_t and small β_t, implying a high contribution rate, while the bottom right part represents small α_t and large β_t implying a low contribution rate. When $k = 4$, all but one of the 5 players must contribute for group success - this causes a significant difference in the optimal POMDP policy compared to the $k = 2$ condition. As seen in Figure 4b, there is only a single region of belief space for which free-riding is the best strategy, namely, when the player does not expect contributions by enough players (relatively large β_t). On the other hand, as expected, this region is much larger compared to the same region for $k = 2$ (see Figure 4a). The POMDP model predicts that free-riding is not a viable action in the $k = 4$ case (Figure 4b) because not only does this action require all the other 4 players to contribute in order to generate the group reward in the current round, but such an action also increases the chances that the group contribution will be lower in the next round, resulting in lesser expected reward in future rounds.

In a game with a predetermined and known number of rounds, even if the player considers the future, one might expect the most rewarding action in the last rounds to be free riding as there is little or no future to consider. However, our data did not support this conclusion. Our simulations using the POMDP model showed that considering a much longer horizon (50 rounds) instead of just 15 rounds gave a much better fit to the subjects' behavior. Such a long horizon for determining the optimal policy makes the model similar to an infinite horizon POMDP model (Thrun et al., 2005). As a result, the optimal policy for all rounds in our model is very similar to the policy for round 9 shown in Figures 4a and 4b.

2.4 POMDP Model Explains Human Behaviour in Volunteer's Dilemma Task

The POMDP model has three parameters, α_0, β_0, and γ which determine the subject's actions and belief in each round. We fitted these parameters to the subject's actions by minimizing the error, i.e. the difference between the POMDP model's predicted action and the subject's action in each round. The average percentage error across all rounds is then the percentage of rounds that the model predicts incorrectly (*contribute* instead of *free-ride* or vice versa). We defined accuracy as the percentage of the rounds that the model predicts correctly. We also calculated the leave-one-out cross validated (LOOCV) accuracy of our fits.

We found that the POMDP model had an average accuracy across subjects of 84% ($SD = 0.06$) while the average LOOCV accuracy was 77% ($SD = 0.08$). Figure 5a compares the average accuracy and LOOCV accuracy of the POMDP model with two other models. The first is a "model-free" reinforcement learning model known as Q-learning: actions are chosen based on their rewards in previous rounds (Tsitsiklis, 1994) with the utility of group reward, initial values, and learning rate as free parameters (5 parameters per subject – see Methods and Materials). The average accuracy of the Q-learning model was 79% ($SD = .07$) which is significantly worse than the POMDP model's accuracy given above (one-tailed paired t-test, $t = 6.75$, $df = 29$, $p = 1.26 \times 10^{-7}$, 95% CI difference = $[0.01, 0.08]$). Also, the average LOOCV accuracy of the POMDP model was significantly higher than the average LOOCV accuracy of Q-learning, which was 73% ($SD = .09$) (one-tailed paired t-test, $t = 2.20$, $df = 29$, $p = 0.02$, 95% CI difference =$[0.00, 0.09]$).

We additionally tested a state-of-the-art descriptive model known as the linear two-factor model (Wunder et al., 2013), which predicts the current action of each player based on the player's own action and contributions by others in the previous round (this model has three free parameters per subject – see Methods and Materials). The average accuracy of the two-factor model was 78% ($SD = 0.09$) which is significantly lower than the POMDP model's accuracy (one-tailed paired t-test, $t = -4.86$, $df = 29$, $p = 2.1 \times 10^{-5}$, .95% CI difference = $[-0.10, 0.02]$. Moreover, the LOOCV accuracy of the two-factor model was 47% ($SD = 20$), significantly lower than the POMDP model (one-tailed paired t-test, $t = -7.61$, $df = 29$, $p = 1.4 \times 10^{-8}$, 95% CI difference = $[-.38, -.22]$). The main reason behind the failure of the linear two-factor model, especially in LOOCV accuracy, is that group SUCCESS depends on the required number of volunteers (k) as well. This value is automatically incorporated into POMDP in calculation of expected reward. Also, reinforcement learning works directly with rewards and therefore does not need explicit knowledge of k (however, we need separate parameter for each k in the initial value function of Q-learning, see Methods and Materials).

The POMDP model, when fit to a subject's actions, can also explain other events during the PGG in contrast to the other models described above. For example, based on equation 3 and the

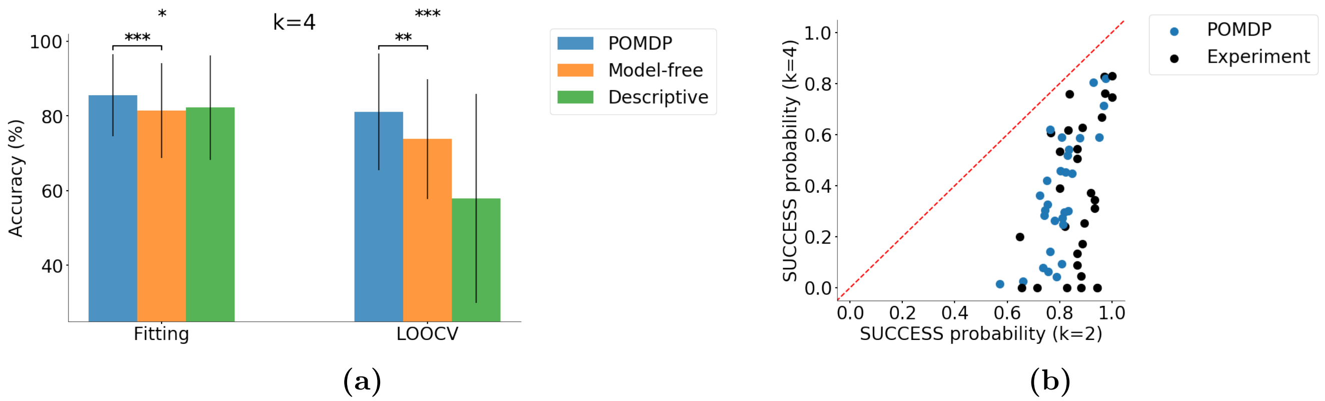

Figure 5: POMDP Model's Performance and Predictions (a) Average and LOOCV accuracy across all models. The POMDP model has significantly higher accuracy compared to the other models. (b) POMDP prediction of a subject's belief about group success in each round (on average) for different k's (blue circles) compared to actual data (black circles, same data as in figure 2d).

action chosen by the POMDP model, one can predict the subject's belief about the probability of group SUCCESS in the current round. This prediction cannot be directly validated but it can be compared to actual SUCCESS in generating the group reward. If we consider actual SUCCESS as the ground truth, the average accuracy of POMDP model's prediction across subjects was 71% ($SD = .07$). Moreover, the prediction matched the pattern of SUCCESS rate data from the experiments for $k = 2$ versus $k = 4$ (Figure 5b). The other models presented above are not capable of making such a prediction.

Finally, we can gain insights into the subject's behavior by interpreting the parameters of our POMDP model in the context of the task. As alluded to above, the prior parameters α_0 and β_0 represent the subject's prior expectations of contributions and free-rides respectively. Therefore, the ratio α_0/β_0 characterizes the subject's expectation of contributions by group members while the average of these parameters, $(\alpha_0 + \beta_0)/2$, indicates the weight the subject gives to prior experience with similar groups before the start of the game. The discount factor (or decay rate) γ determines the weight given to past observations compared to new ones: the smaller the discount factor, the more weight the subject gives to new observations. We examined the distribution of these parameter values for our subjects after fitting the POMDP model to their behavior. The ratio α_0/β_0 was in the reasonable range of .5 to 2 for almost all subjects (Figure 6a; in theory the ratio can be as high as 200 or as low as 1/200 (see Methods and Materials). The value of $(\alpha_0 + \beta_0)/2$ across subjects was mostly between 40 to 120 (Figure 6b) suggesting that prior belief about groups did have a significant role in players' strategy, but it was not the only factor since observations over multiple rounds can still alter this initial belief. We also calculated the expected value of contribution by others in the first round, which is between 0 and $N - 1 = 4$ based on the values of α_0 and β_0 for the subjects. For almost all subjects, this expected value was between 2 and 3 (Figure 6c). Moreover, the discount factor γ was almost always above .95, with a mean of .93 and a median of .97 (Figure 6d). Only three subjects had a discount factor less than .95 (not shown in the figure), suggesting that almost all subjects relied on observations made across multiple rounds when computing their beliefs rather than reasoning based solely on the current or most recent observations.

We also calculated each subject's prior belief about group SUCCESS (probability of SUCCESS

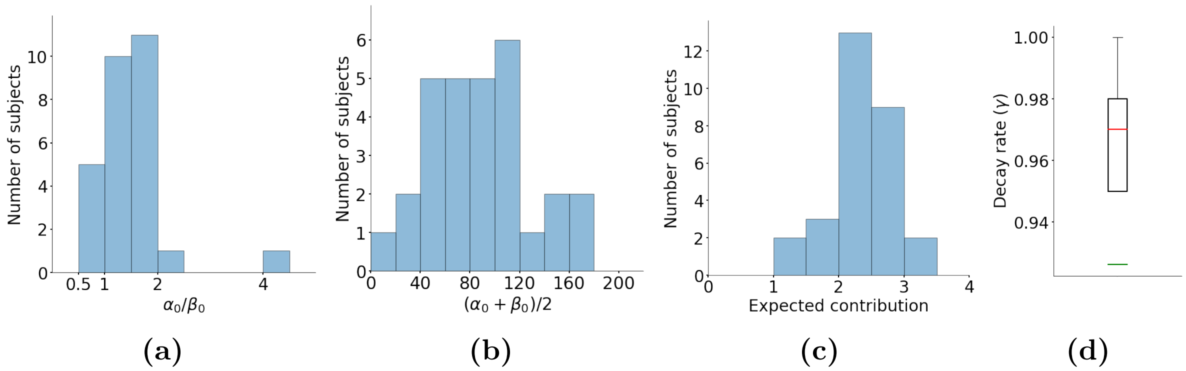

Figure 6: Distribution of POMDP Parameters across Subjects. (a) Histogram of ratio α_0/β_0 shows a value between .5 and 2 for almost all subjects, suggesting that the POMDP model assigns a reasonable prior probability for contributions versus free rides for these subjects. (b) Histogram of average of α_0 and β_0. For the majority of subjects, this value is between 40 to 120, suggesting that the prior belief that the subject brings to the group decision making task plays an important role in the subject's decisions, but the values are small enough for observations to influence decisions over multiple rounds. (c) Histogram of prior belief $Beta(\alpha_0, \beta_0)$ translated into expected contribution by others in the first round. Note that the values, after fitting to the subjects' behavior, are mostly between 2 and 3. (d) The box plot of discount factor or decay rate γ across subjects shows that this value is almost always above .95. The median is .97 (orange line) and the mean is .93 (green line).

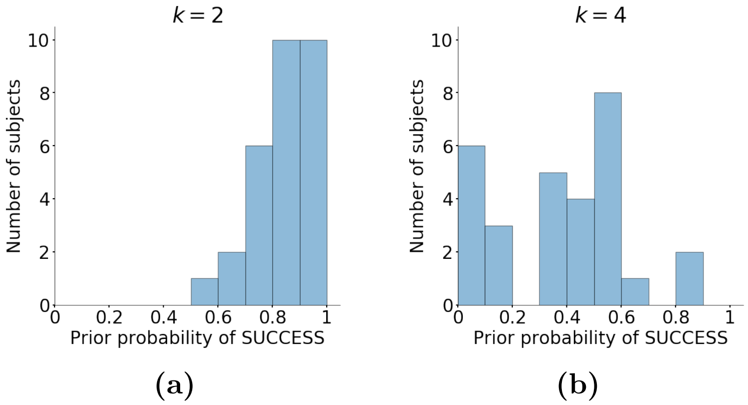

Figure 7: Prior Belief about the Group SUCCESS (a) When $k = 2$, all subjects expected a high probability of group SUCCESS in the first round (before making any observations about the group). (b) When $k = 4$, almost all subjects assigned a chance of less than 60% to group SUCCESS in the first round.

in the first round) based on α_0, β_0, and the subject's POMDP policy in the first round. As group SUCCESS depends on the required number of volunteers (k), probability of SUCCESS is different for $k = 2$ and $k = 4$ even with the same α_0 and β_0. Figures 7a and 7b show the distribution of this prior probability of SUCCESS across all subjects for $k = 2$ and $k = 4$. For $k = 2$, all subjects expected a high probability of SUCCESS in the first round, whereas a majority of the subjects expected less than 60% chance for SUCCESS when $k = 4$.

3 Discussion

We introduced a normative model based on POMDPs for explaining human behavior in a group decision making task. The POMDP model combines probabilistic reasoning about the environment with long-term reward maximization. In the model, the environment consists of multiple states hidden to the decision maker. Based on observations and known transitions between the states, the POMDP model maintains a belief (posterior probability distribution) over the hidden states. Using this probability distribution and simulation of future states, the POMDP model finds for each time step an action that maximizes the expected total reward through the end of the task. In the case of social decision making, we assumed subjects maintain a model of others' intentions (a theory of mind) to predict their actions and possible changes in their intentions as social interactions progress. For the Volunteer's Dilemma, we used a POMDP model in which the hidden state is the intention of a group member to contribute, captured in our model as the average probability that a group member will contribute. Our results show that the POMDP model significantly outperforms model-free reinforcement learning and a state-of-the-art descriptive model in fitting human behavior. In addition, the POMDP model when fit to a subject's actions accurately predicts SUCCESS rate in each round without being explicitly trained for such predictions. A potential drawback of the POMDP model is that maintaining a mental model has a cognitive and computational cost. However, having such a model facilitates the learning process significantly, especially in a highly dynamic environment such as group decision making (Hu and Wellman, 1998; Frank and Goodman, 2012; Lowe et al., 2017).

A strength of the POMDP model is that it can model social tasks beyond economic decision making, such as prediction of others' intentions and actions in everyday interactions (Koster-Hale and Saxe, 2013; Tamir and Thornton, 2018). In these cases, we would need to modify the model's definition of state of other minds to include dimensions such as valence, competence, and social impact instead of propensity to contribute monetary units as in the PGG task (Posner et al., 2005; Gray et al., 2007; Cuddy et al., 2008; Tamir et al., 2016).

Our model assumes that the subject estimates the cooperativeness of others in each round before choosing the next action. It is possible to extend this reasoning by considering the effect of the chosen action on other minds in terms of their cooperativeness, and so on in a recursive fashion. If such a multi-level theory of mind is extended to infinite depth, the game converges to the Nash equilibrium (Nash, 1950). In reality, however, such an infinite-depth theory of mind appears not to occur in actual social interactions among humans (Kagel and Roth, 2016; Camerer, 2011; Henrich et al., 2005), with multi-level theory of mind limited to very few levels as observed in some experiments (Camerer et al., 2004; Yoshida et al., 2008; Hula et al., 2015).

The idea of combining belief with rewards in social decision making has been previously proposed (Camerer and Hua Ho, 1999) but this relationship was based on heuristics rather than a formal mathematical framework as in the case of the POMDP model. Our specific POMDP model was based on the binomial and beta distributions for binary values due to nature of the task, but the model can be easily extended to the more general case of a discrete set of actions using multinomial and Dirichlet distributions. Additionally, the model can be extended to multivariate states, e.g., when the players are no longer anonymous. In such cases, the belief is a joint probability distribution over all parameters of the state.

An important open question is how the various components of the POMDP model could be neurally implemented in the brain. Some suggestions on neural implementation were made in a previous POMDP model for sensory decision making (Rao, 2010; Huang and Rao, 2013).

Specifically, belief computation was ascribed to cortical circuits (Park et al., 2017) while policy learning based on cortical beliefs was assigned to the basal ganglia network. In the context of the present model, belief update corresponds to updating α_t as $\gamma\alpha_{t-1} + c$ and β_t as $\gamma\beta_{t-1} + N - c$. Such an update can be implemented neurally using two leaky accumulators (Usher and McClelland, 2001), which in turn can be mapped into a probability (Kiani and Shadlen, 2009; Kiani et al., 2014). Such accumulators have been widely used in modelling other simpler decision making tasks (Ratcliff and McKoon, 2008; Brunton et al., 2013; Purcell and Kiani, 2016; Tajima et al., 2016) and there is neurophysiological evidence supporting such an implementation in cortical circuits (Gold and Shadlen, 2007; Purcell et al., 2010; Churchland and Ditterich, 2012; Domenech et al., 2018).

Deficits in theory of mind (ToM) are a key feature of autism spectrum disorder. Yet, tasks such the false belief task usually employed to measure such ToM deficits do not provide quantitative measurements of the underlying neurocognitive operations (Baron-Cohen et al., 2001). In contrast, the strength of the POMDP approach to modeling ToM is that it allows a precise deconstruction of the computational signals involved. The present findings thus suggest that POMDP models may be used in the field of computational psychiatry to characterize putative abnormalities in specific computational mechanisms, thereby opening the door to a new classification of mental disorders in terms of dysfunctional computational mechanisms (Huys et al., 2016).

4 Methods and Materials

4.1 Experiment

30 right-handed students at the University of Parma were recruited for this study. One of them aborted the experiment due to anxiety. Data from the other 29 participants were collected, analyzed, and reported. Based on self-reported questionnaires, none of the participants had a history of neurological or psychiatric disorders. The local Ethics Committee approved the study, which was carried out according to the ethical standards of the 2013 Declaration of Helsinki. As mentioned in Results, each subject played 14 sessions of the Public Goods Game (PGG) (i.e., the Volunteer's Dilemma), each containing 15 rounds. In the first 2 sessions, subjects received no feedback about the result of each round. However, in the following 12 sessions, social and monetary feedback were provided to the subject. The feedback included the number of contributors and free riders, and the subject's reward in that round. Each individual player's action, however, remained unknown to the others. Therefore, individual players could not be tracked. We present analyses from the games with feedback.

In each round (see Figure 1), the participant had to make a decision within three seconds by pressing a key; otherwise the round was repeated. 2.5 to 4 seconds after the action selection, the outcome of the round was shown to the subject for 4 seconds. Then, players evaluated the outcome of the round before the next round started. Subjects were told that they were playing with 19 other participants located in other rooms. Overall, 20 players were playing the PGG in 4 different groups simultaneously. These groups were randomly chosen by a computer at the beginning of each session. In reality, subjects were playing with a computer. In other words, a computer algorithm was generating all the actions of others for each subject. Each subject got a final monetary reward equal to the result of one PGG randomly selected by the computer at the end of the study.

In a PGG with $N = 5$ players, we denote the action of player i in round t with the binary value of a_i^t $(1 \leq i \leq N)$ with $a_i^t = 1$ representing contribution and $a_i^t = 0$ representing free-riding. The human subject is assumed to be player 1. We define the average contribution rate of others $\bar{a}_{2:N}^t = \frac{\sum_{i=2}^N a_i^t}{N-1}$ and generate each of the $N - 1$ actions of others in round t using the following probabilistic function:

$$logit(\bar{a}_{2:N}^t) = e_0 a_1^{t-1} + e_1((\frac{1 - K^{T-t+1}}{1 - K})^{e_2} \bar{a}_{2:N}^{t-1} - K) \tag{4}$$

where $K = k/N$ where k is the required number of contributors ($k = 2$ or $k = 4$ in our experiments).

This model has 3 free parameters: e_0, e_1, e_2. These were obtained by fitting the above function to the actual actions of subjects in another PGG study (Park et al., 2013), making this function a simulation of human behavior in the PGG task. For the first round, we used the mean contribution rate of each subject as their fellow members' decision.

4.2 Markov Decision Processes

A Markov Decision Process (MDP) is a tuple (S, A, T, R) where S represents the set of states of the environment, A is the set of actions, T is the transition function $S \times S \times A \to [0, 1]$ that determines the probability of the next state given the current state and action, i.e. $T(s', s, a) = P(s'|s, a)$, and R is the reward function $S \times A \to \mathbb{R}$ representing the reward associated with each state and action. In an MDP with horizon H (total number of performed actions), given the initial state s_0, the goal is to choose a sequence of actions that maximizes the total expected reward:

$$\pi^* = \underset{a_0, a_1, a_2, \ldots, a_{H-1}}{\arg\max} \sum_{t=0}^{H-1} E_{s_t}[R(s_t, a_t)]. \tag{5}$$

This sequence, called the optimal policy, can be found using the technique of dynamic programming (Thrun et al., 2005). For an MDP with time horizon H, define the value function V and action function U at the last time step $t = H$ as:

$$\forall s \in S : \begin{cases} V^H(s) \leftarrow \max_a R(s, a) \\ U^H(s) \leftarrow \arg\max_a R(s, a) \end{cases} \tag{6}$$

For any t from 0 to $H - 1$, the value function V^t and action function U^t is defined recursively as:

$$\forall s \in S, 0 \leq t < H : \begin{cases} V^t(s) \leftarrow \max_a \{R(s, a) + \sum_{s' \in S} T(s', s, a) V^{t+1}(s')\} \\ U^H(s) \leftarrow \arg\max_a \{R(s, a) + \sum_{s' \in S} T(s', s, a) V^{t+1}(s')\} \end{cases} \tag{7}$$

Starting from the initial state s_0 at time 0, the action chosen by the optimal policy π^* at time step t is $U^t(s_t)$.

When the state of the environment is hidden, the MDP turns into a Partially Observable MDP (POMDP) where the state is estimated probabilistically from observations or measurements from sensors. Formally, a POMDP is defined as (S, A, Z, T, O, R) where S, A, T, R are defined as in the case of MDPs, Z is the set of possible observations, and O is the observation function $Z \times S \to [0, 1]$ that determines the probability of any observation z given a state s, i.e., $O(z, s) = P(z|s)$. In order to find the optimal policy, the POMDP model uses the posterior probability of states, known as the belief state, where $b_t(s) = P(s|z_0, a_0, z_1, ..., a_{t1})$. Belief states can be computed recursively as follows:

$$\forall s \in S : B_{t+1}(s) \propto O(z_t, s) \sum_{s' \in S} T(s, s', a_t) B_t(s') \tag{8}$$

If we define $R(b_t, a_t)$ as the expected reward of a_t, i.e., $E_{s_t}[R(s_t, a_t)]$, starting from initial belief state, b_0, the optimal policy for the POMDP is given by:

$$\pi^* = \underset{a_0, a_2, \ldots, a_{H-1}}{\arg\max} \sum_{t=0}^{H-1} E_{s_t}[R(b_t, a_t)]. \tag{9}$$

A POMDP can be considered an MDP whose states are belief states. This belief state space however is exponentially larger than the underlying state space. Therefore, solving a POMDP optimally is computationally expensive (Smallwood and Sondik, 1973), unless the belief state can be represented by a few parameters as in our case. For solving larger POMDP problems, various approximation and learning algorithms have been proposed - we refer the reader to the growing literature on this topic (Ross et al., 2008; Silver and Veness, 2010; Khalvati and Mackworth, 2013; Shani et al., 2013; Luo et al., 2018).

4.3 POMDP for Binary Public Goods Game

The state of the environment is represented by the average cooperativeness of the group, or equivalently, the average probability θ of contribution by a group member. Since θ is not observable, the task is a POMDP and one must maintain a probability distribution (belief) over θ. The Beta distribution, represented by two free parameters (α and β), is the conjugate prior for binomial distribution. Therefore, when performing Bayesian inference to obtain the belief state over θ, combining the Beta distribution as the prior belief and the binomial distribution as the likelihood results in another Beta distribution as the posterior belief. Using the Beta distribution for the belief state, our POMDP turns into an MDP with a two-dimensional state space represented by α and β. Starting from an initial belief state $Beta(\alpha_0, \beta_0)$ and with an additional free parameter γ, the next belief states is determined by the actions of all players at each round as described in Results. For the reward function, we used the monetary reward function of the Public Goods Game (PGG). Therefore, the elements of our new MDP derived from the PGG POMDP are as following:

- $S = (\alpha, \beta)$
- $A = \{c, f\}$
- $T(s', s, a) : \begin{cases} P((\gamma\alpha + k' + 1, \gamma\beta + N - 1 - k')|(\alpha, \beta), c) = \binom{N-1}{k'} \frac{B(\gamma\alpha + k', \gamma\beta + N - 1 - k')}{B(\gamma\alpha, \gamma\beta)} \\ P((\gamma\alpha + k', \gamma\beta + N - k')|(\alpha, \beta), f) = \binom{N-1}{k'} \frac{B(\gamma\alpha + k', \gamma\beta + N - 1 - k')}{B(\gamma\alpha, \gamma\beta)} \end{cases}$
- $R(s, a) : \begin{cases} R((\alpha, \beta), c) = E - C + \sum_{k'=k-1}^{N} \binom{N-1}{k'} \frac{B(\alpha + k', \beta + N - 1 - k')}{B(\alpha, \beta)} G \\ R((\alpha, \beta), f) = E + \sum_{k'=k}^{N} \binom{N-1}{k'} \frac{B(\alpha + k', \beta + N - 1 - k')}{B(\alpha, \beta)} G \end{cases}$

$B(\alpha, \beta)$ is the normalizing constant: $B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$.

According to the experiment, the time horizon should be 15 time steps. However, we found that a very long horizon ($H = 50$) for all players provides a better fit to the subjects' data. For each subject, we found α_0, β_0, and γ that made our POMDP's optimal policy fit the subject's actions as much as possible. For simplicity, we only considered integer values for states (integer α and β). The fitting process involved searching over integer values from 1 to 200 for α_0 and β_0 and values between 0 to 1 with a precision of .01 $(.01, .02, \ldots, .99, 1.0)$ for γ. The fitting criterion was round-by-round accuracy. For consistency with the descriptive model, the first round was not included (despite the POMDP model's capability for predicting it). Since the utility value for public good for a subject can be higher than the monetary reward due to social or cultural reasons (Fehr et al., 2002; Rilling et al., 2002), we investigated the effect of higher values for the group reward G in the reward function of the POMDP. This however did not improve the fit.

In round t, if the POMDP model selects the action "contribution", the probability of SUCCESS can be calculated as $\sum_{m=k-1}^{N-1} P(m|\alpha_t, \beta_t)$ (see Equation 3). Otherwise, the probability of SUCCESS is $\sum_{m=k}^{N-1} P(m|\alpha_t, \beta_t)$. This probability value was compared to the actual SUCCESS and FAILURE of each round to compute the accuracy of SUCCESS prediction by the POMDP model.

4.4 Model-Free Method: Q-Learning

We used Q-learning as our model-free approach. There are two Q values in the PGG task, one for each action, i.e., Q(c) and Q(f) for "contribute" and "free-ride" respectively. At the beginning of each PGG, Q(c) and Q(f) are initialized to the expected reward for a subject for that action based on a free parameter p which represents the prior probability of group SUCCESS. As a result, we have:

$$\begin{cases} Q^0(c) \leftarrow p(E - C + G) + (1-p)(E - C) \\ Q^0(f) \leftarrow p(E + G) + (1-p)E \end{cases} \tag{10}$$

We customized the utility function for each subject by making the group reward G a free parameter. Moreover, as the probability of SUCCESS is different for $k = 2$ and $k = 4$, we used two separate parameters p_2 and p_4 instead of p, depending on the value of k in the PGG.

In each round of the game, the action with the maximum Q value was chosen. The Q value for that action was then updated based on the subject's action and group SUCCESS/FAILURE, with a learning rate η^t. This learning rate was a function of the round number, i.e. $\eta^t = \frac{1}{\lambda_0 + \lambda_1 t}$ where λ_0 and λ_1 are free parameters and t is the number of the current round. Let the subject's action in round t be a^t, the Q-learning model's chosen action be $\hat{a}^t$, and the reward obtained be r^t. We have:

$$\begin{cases} \hat{a}^t = \arg\max_{a \in \{c,f\}} Q^t(a) \\ Q^{t+1}(a^t) \leftarrow (1 - \eta^t) Q^t(a^t) + \eta^t r^t \end{cases} \tag{11}$$

For each subject, we searched for the values of λ_0, λ_1, the group reward G, and the probability of group SUCCESS p_2 or p_4 that maximize the round-by-round accuracy of the Q-learning model. Similar to the other models, the first round was not included in this fitting process.

4.5 Descriptive Model

Our descriptive model was based on a logistic regression that predicts the subject's action in the current round based on their own previous action and the total number of contributions by others in the previous round. As a result, this model has 3 free parameters (two features and a bias parameter). Let a_1^t be the subject's action in round t and $a_{2:N}^t$ be the actions of others in the same round. The subject's predicted action in the next round $t+1$ is then given by:

$$\hat{a}_1^{t+1} = \begin{cases} c & \kappa_0 + \kappa_1 a_1^t + \kappa_2 (\sum_{i=2}^N a_i^t) > 0 \\ f & \text{otherwise} \end{cases} \tag{12}$$

We used one separate regression model for each subject. As the model's predicted action is based on the previous round's actions, the subject's action in the first round cannot be predicted by this model.

4.6 Greedy Strategy

If a player wants to solely maximize the expected reward in the current round and ignores the future, the optimal action is always free-riding independent of the average probability of contribution by a group member. This is because free-riding always results in one unit more monetary reward (3 MU for SUCCESS or 1 MU for FAILURE) compared to contribution (2 MU or 0 MU), except in the case where the total number of contributions by others is *exactly* $k-1$. In the latter case, choosing contribution yields 1 unit more reward (2 MU) compared to free-riding (1 MU). This means that the expected reward for free-riding is always more than that for contribution unless the probability of observing exactly $k-1$ contributions by others is greater than .5. We show that this is impossible for any value of θ. First, note that the probability of exactly $k-1$ contributions from $N-1$ players is maximized when $\theta = (k-1)/(N-1)$. Next, for any θ, the probability of $k-1$ contributions from $N-1$ players is:

$$\binom{N-1}{k-1}\theta^{k-1}(1-\theta)^{N-k} \leq \binom{N-1}{k-1}(\frac{k-1}{N-1})^{k-1}(\frac{N-k}{N-1})^{N-k} = .75^3 < .5 \tag{13}$$

for $N = 5$ and for either $k = 2$ or $k = 4$.

5 Acknowledgments

This work was funded by the NSF-ANR Collaborative Research in Computational Neuroscience 'CRCNS SOCIAL POMDP' $n°$16-NEUC to J.C.D. and CRCNS NIMH grant no. 5R01MH112166-03, NSF grant no. EEC-1028725, and a Templeton World Charity Foundation grant to R.P.N.R. The experiments were performed within the framework of the Laboratory of Excellence 'LABEX ANR-11-LABEX-0042' of Universite de Lyon, attributed to J.C.D., within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). J.C.D. was also funded by the IDEX University Lyon 1 (project INDEPTH).

6 Contributions

R.P.N.R. and J.C.D. developed the general research concept. S.A.P. designed and programmed the task under J.C.D.'s supervision and M.S. ran the experiment under J.C.D.'s supervision. K.K. developed the model under R.P.N.R.'s supervision, implemented the algorithms and analyzed the data in collaboration with R.P.N.R. K.K. developed the reinforcement learning model after discussions with R.P. K.K. and R.P.N.R. wrote the manuscript in collaboration with S.A.P., R.P., and J.C.D.

References

Amodio, D. M. and Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, 7(4):268.

Archetti, M. (2009a). Cooperation as a volunteers dilemma and the strategy of conflict in public goods games. *Journal of Evolutionary Biology*, 22(11):2192–2200.

Archetti, M. (2009b). The volunteer's dilemma and the optimal size of a social group. *Journal of Theoretical Biology*, 261(3):475–480.

Archetti, M. and Scheuring, I. (2011). Coexistence of cooperation and defection in public goods games. *Evolution*, 65(4):1140–1148.

Baker, C. L., Jara-Ettinger, J., Saxe, R., and Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):0064.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. (2001). The "reading the mind in the eyes" test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 42(2):241–251.

Behrens, T. E. J., Hunt, L. T., and Rushworth, M. F. S. (2009). The computation of social behavior. *Science*, 324(5931):1160–1164.

Brown, G. W. (1951). Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation*, 13(1):374–376.

Brunton, B. W., Botvinick, M. M., and Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science*, 340(6128):95–98.

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction.* Princeton University Press.

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

Camerer, C. F. and Hua Ho, T. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874.

Churchland, A. K. and Ditterich, J. (2012). New advances in understanding decisions among multiple alternatives. *Current Opinion in Neurobiology*, 22(6):920–926.

Coricelli, G. and Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, 106(23):9163–9168.

Costa, V. D. and Averbeck, B. B. (2013). Frontal–parietal and limbic-striatal activity underlies information sampling in the best choice problem. *Cerebral Cortex*, 25(4):972–982.

Cuddy, A. J., Fiske, S. T., and Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in Experimental Social Psychology*, 40:61–149.

Culbreth, A. J., Westbrook, A., Daw, N. D., Botvinick, M., and Barch, D. M. (2016). Reduced model-based decision-making in schizophrenia. *Journal of Abnormal Psychology*, 125(6):777.

Darley, J. M. and Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 8(4p1):377.

Daw, N. D. and Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Phil. Trans. R. Soc. B*, 369(1655):20130478.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215.

Dayan, P. (2012). Twenty-five lessons from computational neuromodulation. *Neuron*, 76(1):240–256.

Dayan, P. and Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453.

Dickinson, A. and Balleine, B. (2002). The role of learning in the operation of motivational systems. *Stevens' Handbook of Experimental Psychology*.

Diekmann, D. (1985). Volunteer's dilemma. *The Journal of Conflict Resolution*, 29(4):605–610.

Dolan, R. J. and Dayan, P. (2013). Goals and habits in the brain. *Neuron*, 80(2):312–325.

Doll, B. B., Simon, D. A., and Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, 22(6):1075–1081.

Domenech, P., Redout, J., Koechlin, E., and Dreher, J.-C. (2018). The neuro-computational architecture of value-based selection in the human brain. *Cerebral Cortex*, 28(2):585–601.

Dunne, S. and ODoherty, J. P. (2013). Insights from the application of computational neuroimaging to social neuroscience. *Current Opinion in Neurobiology*, 23(3):387–392.

Fehr, E., Fischbacher, U., and Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, 13(1):1–25.

Fehr, E. and Gachter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.

Fischbacher, U., Gatcher, S., and Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71(3):397 – 404.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Glscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–595.

Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30.

Gray, H. M., Gray, K., and Wegner, D. M. (2007). Dimensions of mind perception. *science*, 315(5812):619–619.

Hampton, A. N., Bossaerts, P., and O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences*.

Hauert, C., De Monte, S., Hofbauer, J., and Sigmund, K. (2002). Volunteering as red queen mechanism for cooperation in public goods games. *Science*, 296(5570):1129–1132.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., et al. (2005). economic man in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(6):795–815.

Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J. P., and Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience*, 20(8):1142–1149.

Hu, J. and Wellman, M. P. (1998). Online learning about other agents in a dynamic multiagent system. In *Proceedings of the Second International Conference on Autonomous Agents*, pages 239–246. ACM.

Huang, Y. and Rao, R. P. N. (2013). Reward optimization in the primate brain: A probabilistic model of decision making under uncertainty. *PLoS ONE*, 8(1):e53344.

Hula, A., Montague, P. R., and Dayan, P. (2015). Monte carlo planning method estimates planning horizons during interactive social exchange. *PLoS Computational Biology*, 11(6):e1004254.

Huys, Q. J. M., Maia, T. V., and Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19(3):404–413.

Joiner, J., Piva, M., Turrin, C., and Chang, S. W. (2017). Social learning through prediction error in the brain. *npj Science of Learning*, 2(1):8.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99 – 134.

Kagel, J. H. and Roth, A. E. (2016). *The Handbook of Experimental Economics, Volume 2: The Handbook of Experimental Economics*. Princeton university press.

Khalvati, K. and Mackworth, A. K. (2013). A fast pairwise heuristic for planning under uncertainty. In *Proceedings of The Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 187–193.

Khalvati, K., Park, S. A., Dreher, J.-C., and Rao, R. P. (2016). A probabilistic model of social decision making based on reward maximization. In *Advances in Neural Information Processing Systems*, pages 2901–2909.

Khalvati, K. and Rao, R. P. (2015). A bayesian framework for modeling confidence in perceptual decision making. In *Advances in Neural Information Processing Systems*, pages 2413–2421.

Kiani, R., Corthell, L., and Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, 84(6):1329–1342.

Kiani, R. and Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, 324(5928):759–764.

Koster-Hale, J. and Saxe, R. (2013). Theory of mind: a neural prediction problem. *Neuron*, 79(5):836–848.

Krajbich, I., Camerer, C., Ledyard, J., and Rangel, A. (2009). Using neural measures of economic value to solve the public goods free-rider problem. *Science*, 326(5952):596–599.

Ligneul, R., Obeso, I., Ruff, C. C., and Dreher, J.-C. (2016). Dynamical representation of dominance relationships in the human rostromedial prefrontal cortex. *Current Biology*, 26(23):3107–3115.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, O. P., and Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pages 6382–6393.

Luo, Y., Bai, H., Hsu, D., and Lee, W. S. (2018). Importance sampling for online planning under uncertainty. *The International Journal of Robotics Research*.

McAllister, P. H. (1991). Adaptive approaches to stochastic programming. *Annals of Operations Research*, 30(1):45–62.

Mookherjee, D. and Sopher, B. (1997). Learning and decision costs in experimental constant sum games. *Games and Economic Behavior*, 19(1):97–132.

Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive computation and machine learning. MIT Press.

Nagel, R., Brovelli, A., Heinemann, F., and Coricelli, G. (2018). Neural mechanisms mediating degrees of strategic uncertainty. *Social Cognitive and Affective Neuroscience*, 13(1):52–62.

Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49.

O'doherty, J. P., Cockburn, J., and Pauli, W. M. (2017). Learning, reward, and decision making. *Annual Review of Psychology*, 68:73–100.

Olson, M. (1971). *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press.

Palfrey, T. R. and Rosenthal, H. (1984). Participation and the provision of discrete public goods: a strategic analysis. *Journal of Public Economics*, 24(2):171–193.

Park, S. A., Goame, S., O'Connor, D. A., and Dreher, J.-C. (2017). Integration of individual and social information for decision-making in groups of different sizes. *PLOS Biology*, 15(6):e2001958.

Park, S. A., Jeong, S., and Jeong, J. (2013). TV programs that denounce unfair advantage impact women's sensitivity to defection in the public goods game. *Social Neuroscience*, 8.

Posner, J., Russell, J. A., and Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734.

Purcell, B. A., Heitz, R. P., Cohen, J. Y., Schall, J. D., Logan, G. D., and Palmeri, T. J. (2010). Neurally constrained modeling of perceptual decision making. *Psychological Review*, 117(4):1113.

Purcell, B. A. and Kiani, R. (2016). Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proceedings of the National Academy of Sciences*, 113(31):E4531–E4540.

Rao, R. P. N. (2010). Decision making under uncertainty: a neural model based on partially observable Markov decision processes. *Frontiers in Computational Neuroscience*, 4.

Ratcliff, R. and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, 20(4):873–922.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., and Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron*, 35(2):395–405.

Rilling, J. K. and Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology*, 62:23–48.

Ross, S., Pineau, J., Paquet, S., and Chaib-draa, B. (2008). Online planning algorithms for POMDPs. *Journal of Artificial Intelligence Research*, 32(1).

Ruff, C. C. and Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, 15(8):549.

Sanfey, A. G. (2007). Social decision-making: insights from game theory and neuroscience. *Science*, 318(5850):598–602.

Seo, H. and Lee, D. (2017). Chapter 18 - reinforcement learning and strategic reasoning during social decision-making. In Dreher, J.-C. and Tremblay, L., editors, *Decision Neuroscience*, pages 225–231. Academic Press.

Shani, G., Pineau, J., and Kaplow, R. (2013). A survey of point-based pomdp solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1):1–51.

Silver, D. and Veness, J. (2010). Monte-carlo planning in large pomdps. In *Advances in Neural Information Processing Systems*, pages 2164–2172.

Smallwood, R. D. and Sondik, E. J. (1973). The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Tajima, S., Drugowitsch, J., and Pouget, A. (2016). Optimal policy for value-based decision-making. *Nature Communications*, 7:12400.

Tamir, D. I. and Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences*.

Tamir, D. I., Thornton, M. A., Contreras, J. M., and Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences*, 113(1):194–199.

Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics*. MIT Press, Cambridge, MA,.

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and q-learning. *Machine learning*, 16(3):185–202.

Usher, M. and McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, 108(3):550.

Wanga, J., Surib, S., and Wattsb, D. J. (2012). Cooperation and assortativity with dynamic partner updating. *Proceedings of the National Academy of Sciences*, 109(36):14363–14368.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Wunder, M., Suri, S., and Watts, D. J. (2013). Empirical agent based models of cooperation in public goods games. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce (EC)*, pages 891–908.

Yoshida, W., Dolan, R. J., and Friston, K. J. (2008). Game theory of mind. *PLoS Computational Biology*, 4(12):e1000254.

Yoshida, W., Seymour, B., Friston, K. J., and Dolan, R. J. (2010). Neural mechanisms of belief inference during cooperative games. *Journal of Neuroscience*, 30(32):10744–10751.