Bayesian Inference of Other Minds Explains Human Choices in Group Decision Making

Koosha Khalvati[1], Seongmin A. Park[2,3], Saghar Mirbagheri[4], Remi Philippe[3], Mariateresa Sestito[3], Jean-Claude Dreher[3,*], and Rajesh P. N. Rao[1,5,*]

[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] *Department of Psychology, New York University, New York, USA*
[5] *Center for Neurotechnology, University of Washington, Seattle, USA*
[*] *Joint senior authors*

Abstract

To make decisions in a social context, humans have to predict the behavior of others, an ability that is thought to rely on having a model of other minds known as theory of mind. Such a model becomes especially complex when the number of people one simultaneously interacts is large and the actions are anonymous. Here, we show that in order to make decisions within a large group, humans employ Bayesian inference to model the "mind of the group," making predictions of others' decisions while also considering the effects of their own actions on the group as a whole. We present results from a group decision making task known as the Volunteers Dilemma and demonstrate that a Bayesian model based on partially observable Markov decision processes outperforms existing models in quantitatively explaining human behavior. Our results suggest that in group decision making, rather than acting based solely on the rewards received thus far, humans maintain a model of the group and simulate the group's dynamics into the future in order to choose an action as a member of the group.

1 Introduction

The importance of social decision making in human behavior has spawned a large body of research in social neuroscience and decision making (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 and Daw, 2008; 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). However, model-based decision making requires more cognitive resources, for example, for simulation of future events. Thus, there is an inherent trade-off between the two approaches, and determining which approach humans adopt for different situations is an important open area of research.

Several studies have presented evidence in favor of the model-based approach by quantifying the similarity between probabilistic model-based methods and human behavior when the subject interacts with or reasons about another human (Ray et al., 2009; Yoshida et al., 2010; Xiang et al., 2012; Hula et al., 2015; Baker et al., 2017). Additional evidence in favor of a model-based approach comes from studies suggesting that ToM, autobiographical memory, and thinking about the future are all supported by a network of interconnected nodes in the brain known as the Default Mode Network (DMN) (Spreng et al., 2009; Spreng and Grady, 2010).

Compared to reasoning about a single person, decision making in a group with a large number of members can get complicated. On the one hand, having more group members disproportionately increases the cognitive cost of tracking minds (as in model-based approaches) compared to the cost of tracking the reward history of each action (as in model-free approaches). On the other hand, consistent with the importance that human society places on group decisions, a model-based approach might be the optimal strategy.

How does one extend a model-based approach for reasoning about a single person to the case of decision making within a large group? Group decision making becomes even more challenging when the actions of others in the group are anonymous (e.g., voting as part of a jury) (Park et al., 2017). In such situations, reasoning about the state of mind of individual group members is not possible but the dynamics of group decisions do depend on each individual's actions.

To investigate these complexities that arise in group decision making, we focused on the Volunteer's Dilemma task, wherein a few individuals endure some costs to benefit the whole group (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). In order to mimic the Volunteer's Dilemma in a laboratory setting, we use the binary version of a multiround Public Goods Game (PGG) where the actions of each individual are hidden from others (Olson, 1971; Palfrey and Rosenthal, 1984; Diekmann, 1985; Archetti, 2009b,a; Archetti and Scheuring, 2011).

Using an optimal Bayesian framework based on Partially Observable Markov Decision Processes (POMDPs) (Kaelbling et al., 1998), we propose that in group decision making, humans simulate the "mind of the group" by modeling an *average group member* when making their current choices. Our model incorporates prior knowledge, current observations, and a simulation of the future for modeling human decisions within a group. We compared our model to a model-free approach based on the reward history of each action as well as a state-of-the-art descriptive method for fitting human behaviour in the Public Goods Game. Our model explains human behaviour significantly better than the model-free and descriptive approaches. Furthermore, by leveraging

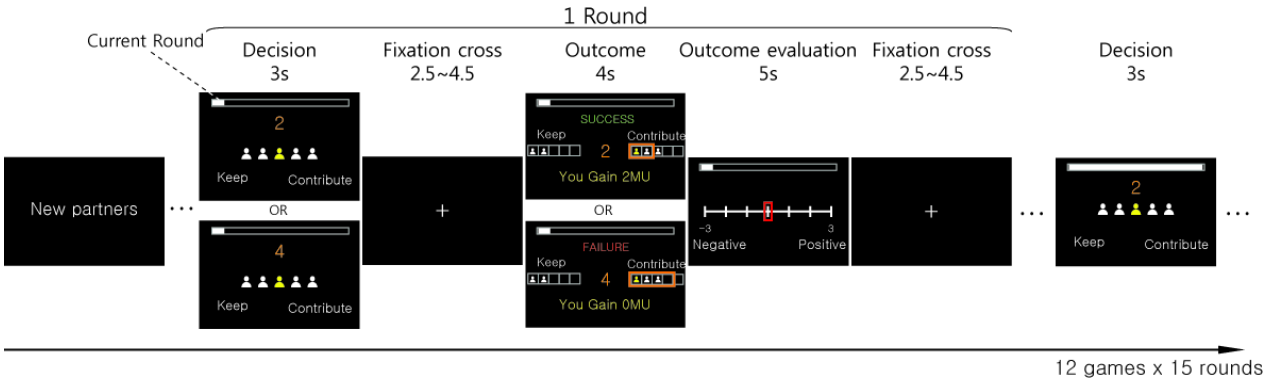


Figure 1: Multi-Round Public Goods Game (PGG). The figure depicts the sequence of computer screens a subject sees in one round of the PGG. The subject is assigned 4 other players as partners and each round requires the subject to make a decision: Keep 1 monetary unit (i.e., free-ride) or contribute. 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 total number of contributions and overall outcome of the round (SUCCESS or FAILURE) are revealed.

the interpretable nature of our model, we are able to show a potential underlying computational mechanism for the group decision making process.

2 Results

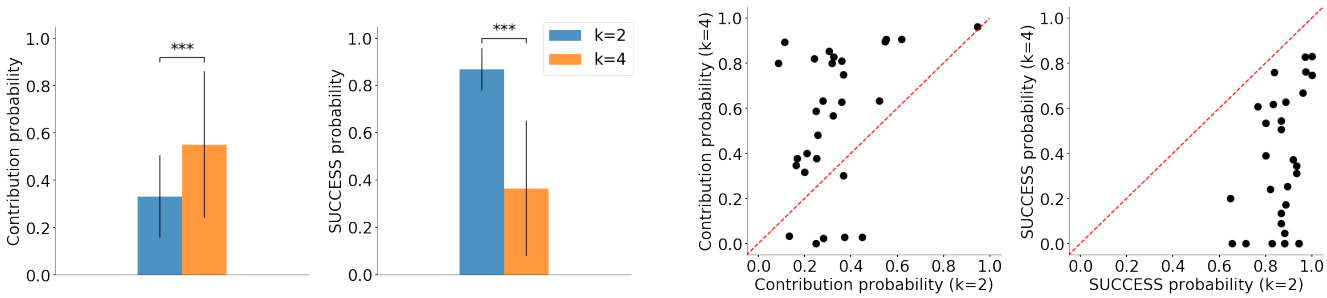
2.1 Human Behavior in Binary Public Goods Game

Participants were 29 adults (mean age 22.97 years old $\pm$ 0.37, 14 women). 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 ($N = 5$).

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. In contrast to the classical PGG where the group reward is a linear function of total contributions (Fehr and Gachter, 2000; Krajbich et al., 2009; Hauert et al., 2002), in our PGG, 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 had E = 1 MU when no public good was produced (the round was a FAILURE).

Figure 1 depicts one round of the PGG task. After the subject acted, the total number of contributions, free rides, and the overall outcome of the round was revealed (success or failure in securing the 2 MU group reward) but each individual players actions remained unknown. 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). In each session, the subject played with a different group of players.

As shown in Figure 2a, subjects contributed significantly more when the number of required



(a) **(b)** **(c)** **(d)**

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 $k = 4$.

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 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, $t = -0.69$, $df = 26$, $p = 0.25$).

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 the number of contributions they anticipate 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 the others, compare it 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 the 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:

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

Starting with a prior probability $Beta(\alpha_1, \beta_1)$ for θ, the player updates their belief about θ after observing the number of contributions from the 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) (Murphy, 2012). In our case, with a prior of $Beta(\alpha_t, \beta_t)$ after observing c contributions (including one's own when made) in round t, the posterior probability of θ becomes $Beta(\alpha_{t+1}, \beta_{t+1})$ where $\alpha_{t+1} = \alpha_t + c$ and $\beta_{t+1} = \beta_t + N - c$. 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 the others.

Intuitively, α represents the number of contributions made thus far and β the number of free rides. α_1 and β_1 (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_1 is larger than β_1, the player starts the task with the belief that people will contribute more than free ride. Large values of α_1 and β_1 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 the 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 the 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 $Beta(\gamma\alpha_t, \gamma\beta_t)$, 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 $Beta(\alpha_t, \beta_t)$ to $Beta(\alpha_{t+1}, \beta_{t+1})$ where $\alpha_{t+1} = \gamma\alpha_t + c$ and $\beta_{t+1} = \gamma\beta_t + N - c$.

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

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

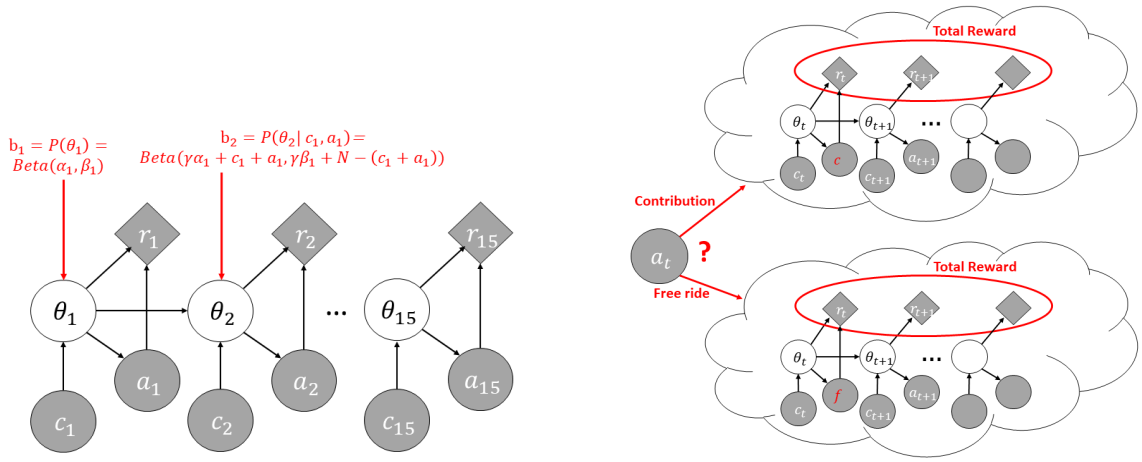
One can calculate the expected reward for the *contribute* versus *free-ride* actions in the current round based on the above equation. Maximizing this reward however is not the best strategy. As alluded to earlier, the actions of each player can change the behavior of other group members in future rounds. 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) (Kaelbling et al., 1998). Further details can be found in Methods 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 3a shows a schematic of the PGG experiment modeled using a POMDP, while Figure 3b illustrates the mechanism of action selection in this model.

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*.

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.

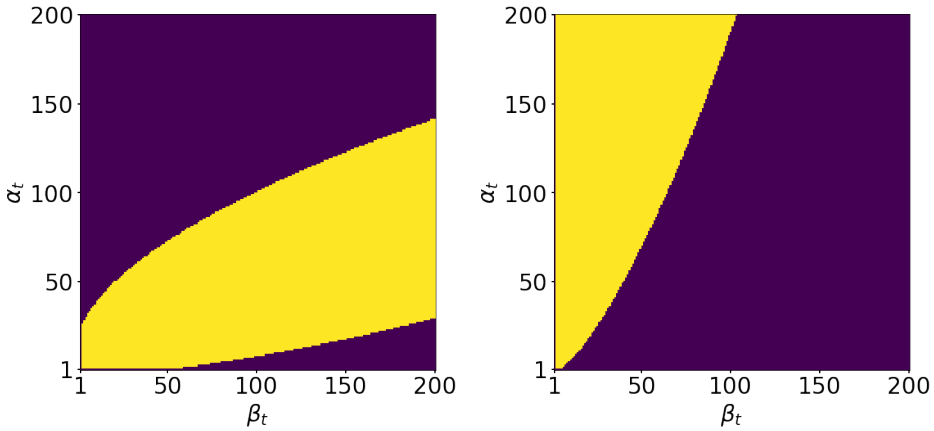
2.4 POMDP Model Explains Human Behaviour in Volunteer's Dilemma Task

The POMDP model has three parameters, α_1, β_1, 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



(a) **(b)**

Figure 3: POMDP Model of the Multi-Round Public Goods Game. (a) 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. (b) 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 the group behavior in future rounds.

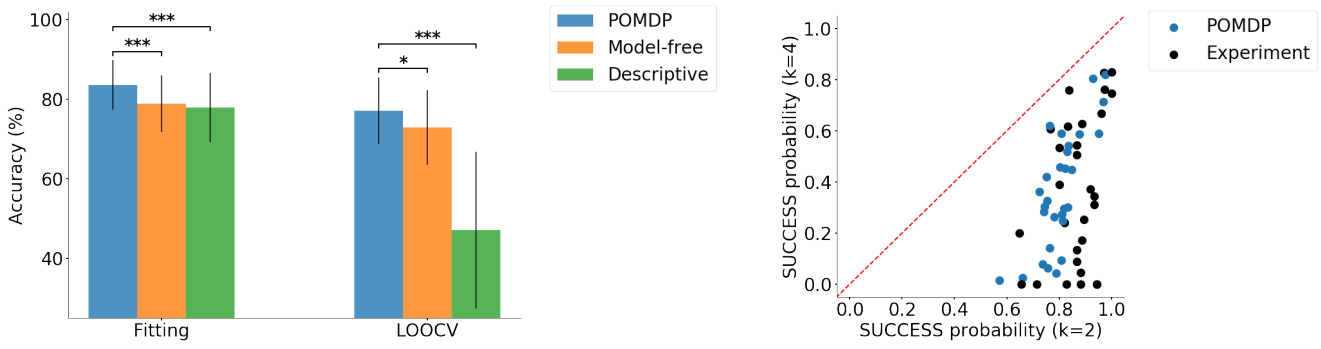


(a) **(b)**

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. For the two plots, the decay rate was 1 and t was 9.

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 (Murphy, 2012).

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



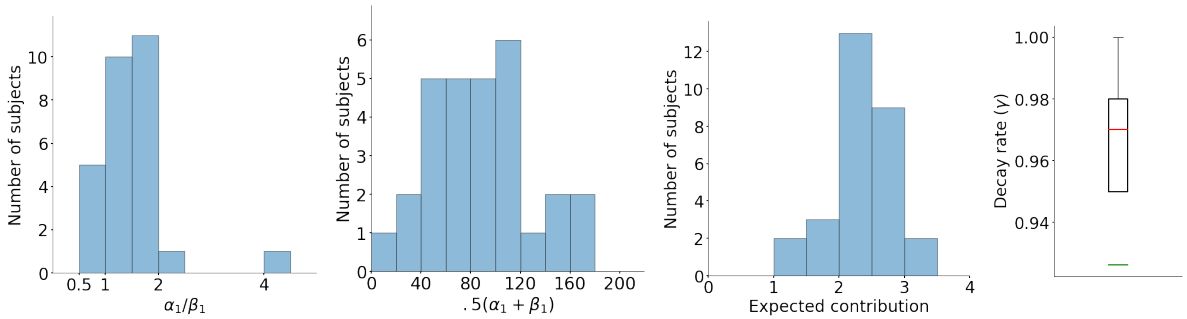
(a) **(b)**

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).

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). 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 $= [-0.08, -0.01]$). 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, $t = 2.20$, $df = 29$, $p = 0.0187$, 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 the others in the previous round (this model has three free parameters per subject – see Methods). 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 $= [-0.38, -0.22]$). The main reason for this result, especially the lower LOOCV accuracy, is that group SUCCESS also depends on the required number of volunteers (k). This value is automatically incorporated in the POMDP's calculation of expected reward. Also, reinforcement learning works directly with rewards and therefore does not need explicit knowledge of k (however, a separate parameter for each k is needed in the initial value function for Q-learning – see Methods).

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 action chosen by the POMDP model, one can predict the subject's belief about the probability of SUCCESS in the current round. This prediction cannot be directly validated but it can be compared to actual SUCCESS. 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 (Figure 5b). The other models presented above are not capable of making such a prediction.



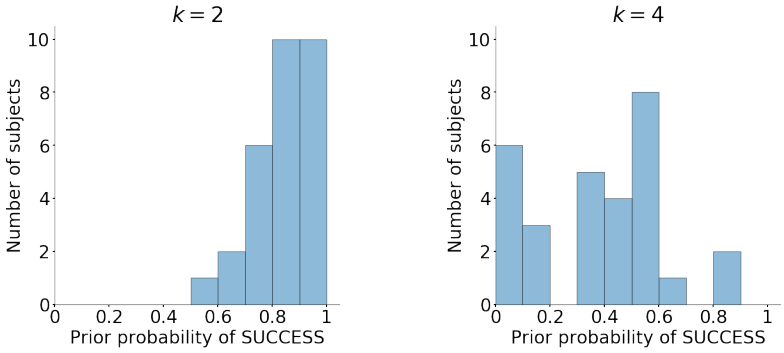
(a) **(b)** **(c)** **(d)**

Figure 6: Distribution of POMDP Parameters across Subjects. (a) Histogram of ratio α_1/β_1 shows a value between .5 and 2 for almost all subjects. (b) Histogram of average of α_1 and β_1. For the majority of subjects, this value is between 40 to 120. (c) Histogram of prior belief $Beta(\alpha_1, \beta_1)$ translated into expected contribution by the 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).

2.5 Distribution of POMDP Parameters

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_1 and β_1 represent the subject's prior expectations of contributions and free-rides respectively. Therefore, the ratio α_1/β_1 characterizes the subject's expectation of contributions by group members while the average of these parameters, $(\alpha_1 + \beta_1)/2$, 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_1/β_1 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). The value of $(\alpha_1 + \beta_1)/2$ 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 the others in the first round, which is between 0 and $N - 1 = 4$ based on the values of α_1 and β_1 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 in the first round) based on α_1, β_1, 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_1 and β_1. 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 $k = 4$. While these beliefs cannot be directly



(a) **(b)**

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.

validated, the results point to the importance of the required number of volunteers in shaping the subjects' behavior.

3 Discussion

We introduced a normative model based on POMDPs for explaining human behavior in a group decision making task. Our model combines probabilistic reasoning about the group with long-term reward maximization by simulating the effect of each action on the future behaviour of the group. The greater accuracy of our model in explaining the subjects' behaviour compared to the other models suggests that humans make decisions in group settings by reasoning about the group as a whole. This mechanism is analogous to maintaining a theory of mind about another person, except theory of mind pertains to a group member on average.

This is the first time, to our knowledge, that a normative model has been proposed for a group decision making task. All existing models to explain human behaviour in the Public Goods game, for example, are descriptive and do not provide insights into the computational mechanisms underlying the decisions (Wanga et al., 2012; Wunder et al., 2013; Fischbacher et al., 2001). Moreover, these methods are based on linear regression of previous events to predict the subject's action in the current round. These regression-based methods assume independent data points but in a multi-round game, the rounds are not independent. Nonetheless, our model outperforms this family of methods as well.

In addition to providing a better fit to the subject's behavior, our model, when fit to the subject's actions can predict SUCCESS rate in each round without being explicitly trained for such predictions, in contrast to the other methods. Also, as alluded to in figures 6a, 6b and 6d, when fit to the subjects' actions, the parameters were all within a reasonable range, showing the importance of prior knowledge and multiple observations in decision making. These observations imply that our high accuracy in explaining subjects' behaviour is not an artifact of the complexity of the model or of the Beta distribution.

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 model is able to explain the data by using a longer horizon than the number of rounds in the game. While this may make the policy deviate from optimality, it provides a significant computational benefit by making the policies for different rounds similar to each other, avoiding re-calculation of a policy for each single round. Recent studies in human decision making have demonstrated that humans may use such minimal modifications of model-based policies for efficiency (Momennejad et al., 2017; Russek et al., 2017). An additional reason behind favoring contribution in the PGG may be an incomplete understanding of the task by some subjects (Burton-Chellew and West, 2013).

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, 2010; Hula et al., 2015). Moreover, we expect less levels of theory of mind in group decision making in general compared to two-person interactions due to the complexity of the group decision making process and the anonymity of individuals.

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. This gives the model-based approach a huge advantage over model-free methods in terms of reward gathering, thus making it more beneficial for the brain to endure the extra cognitive cost. It is possible that in simpler tasks where the accurate prediction of minds is less crucial, the brain adopts a model-free approach.

Our model was based on the binomial and beta distributions for binary values due to nature of the task, but it 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 can be modeled a joint probability distribution over all parameters of the state. This however incurs a significant computational cost. An interesting area for future research is investigating whether under some

circumstances, humans model group members with similar behaviour as one subgroup in order to reduce the number of minds one should reason about.

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 sitations (Koster-Hale and Saxe, 2013; Tamir and Thornton, 2018). In these cases, we would need to modify the model's definition of the 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).

The interpretability of the POMDP framework offers an opportunity to study the neurocognitive mechanisms of group decision making in healthy and diseased brains. POMDPs and similar Bayesian models have previously proved useful in understanding neural responses in sensory decision making (Rao, 2010; Huang et al., 2012; Huang and Rao, 2013; Khalvati and Rao, 2015) and in tasks involving interactions with a single individual (Xiang et al., 2012; Yoshida et al., 2010; Hula et al., 2015; Baker et al., 2017). We believe the POMDP model we have proposed can likewise prove useful in interpreting neural responses and data from neuroimaging studies of group decision making tasks. Additionally, the model can be used for Bayesian theory-driven investigations in the field of computational psychiatry (Huys et al., 2016). For example, ToM deficits are a key feature of autism spectrum disorder (Baron-Cohen et al., 2001) but it is unclear what computational components are impaired and how they are affected. The POMDP model may provide a new avenue for computational studies of such neuropsychiatric disorders.

4 Methods

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. This study was approved by the Institutional Review Board of the local ethics committee from Parma University (IRB no. A13-37030), which was carried out according to the ethical standards of the 2013 Declaration of Helsinki. All participants gave their informed written consent. 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.

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 $S \times S \times A \rightarrow [0, 1]$ 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 $S \times A \rightarrow \mathbb{R}$ representing the reward associated with each state and action (Bellman, 1957). In an MDP with horizon H (total number of performed actions), given the initial state s_1, the goal is to choose a sequence of actions that maximizes the total expected reward:

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

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

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

Starting from the initial state s_1 at time 1, 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 $Z \times S \rightarrow [0, 1]$

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 $b_t(s) = P(s|z_1, a_1, z_2, ..., a_{t-1})$. 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_1, the optimal policy for the POMDP is given by:

$$\pi^* = \underset{a_1, a_2, ..., a_H}{\arg\max} \sum_{t=1}^{H} E_{s_t}[R(b_t, a_t)]. \tag{9}$$

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 Thrun et al. (2005). 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 Murphy (2012). 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 $Beta(\alpha_1, \beta_1)$ 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_1, β_1, 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_1 and β_1 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.

A preliminary version of the above model but without the γ parameter was presented in (Khalvati et al., 2016).

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:

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

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:

$$1 \leq t \leq 15 : \begin{cases} \hat{a}^t = \arg\max_{a \in \{c,f\}} Q^t(a) \\ Q^{t+1}(a^t) \leftarrow (1-\eta^t) Q^t(a^t) + \eta^t r^t \end{cases} \tag{11}$$

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 (implementation by Scikit-learn (Pedregosa et al., 2011)) that predicts the subject's action in the current round based on their own previous action and the total number of contributions by the 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:

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

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. S.M. interpreted the computational results in the context of social neuroscience. K.K. developed the reinforcement learning model after discussions with R.P. K.K., S.M., and R.P.N.R. wrote the manuscript in collaboration with S.A.P., R.P., and J.C.D.

References

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.

Bellman, R. (1957). A markovian decision process. *Journal of Mathematics and Mechanics*, 6(5):679–684.

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

Burton-Chellew, M. N. and West, S. A. (2013). Prosocial preferences do not explain human cooperation in public-goods games. *Proceedings of the National Academy of Sciences*, 110(1):216–221.

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.

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.

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.

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

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.

Huang, Y., Friesen, A. L., Hanks, T. D., Shadlen, M. N., and Rao, R. P. N. (2012). How prior probability influences decision making: A unifying probabilistic model. In *Proceedings of The Twenty-sixth Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1277–1285.

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.

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.

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.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N., and Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9):680.

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.

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

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.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

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.

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

Ray, D., King-Casas, B., Montague, P. R., and Dayan, P. (2009). Bayesian model of behaviour in economic games. In *Advances in neural information processing systems*, pages 1345–1352.

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.

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., and Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS computational biology*, 13(9):e1005768.

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

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.

Spreng, R. N. and Grady, C. L. (2010). Patterns of brain activity supporting autobiographical memory, prospection, and theory of mind, and their relationship to the default mode network. *Journal of cognitive neuroscience*, 22(6):1112–1123.

Spreng, R. N., Mar, R. A., and Kim, A. S. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *Journal of cognitive neuroscience*, 21(3):489–510.

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

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.

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.

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.

Xiang, T., Ray, D., Lohrenz, T., Dayan, P., and Montague, P. R. (2012). Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS computational biology*, 8(12):e1002841.

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.