COGNITIVE NEUROSCIENCE

Modeling other minds: Bayesian inference 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]*, Rajesh P. N. Rao[1,5]*†

Copyright © 2019 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

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 with is large and actions are anonymous. Here, we present results from a group decision-making task known as the volunteer's dilemma and demonstrate that a Bayesian model based on partially observable Markov decision processes outperforms existing models in quantitatively predicting human behavior and outcomes of group interactions. Our results suggest that in decision-making tasks involving large groups with anonymous members, humans use Bayesian inference to model the "mind of the group," making predictions of others' decisions while also simulating the effects of their own actions on the group's dynamics in the future.

Downloaded from http://advances.sciencemag.org/ on February 8, 2021

INTRODUCTION

The importance of social decision-making in human behavior has spawned a large body of research in social neuroscience and decision-making (*1*, *2*). 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 drastically as the behavior of others is much less predictable than the physics of the environment.

One approach to handling uncertainty in social settings is to act based on a belief about others. This approach includes inferring the consequences of one's own behavior under uncertainty as opposed to "belief-free" models (*3*) that simply select the action that has been rewarding in the past, given current observations (*4*, *5*). The difference between "belief-based" and belief-free models in social decision-making is closely related to "model-based" and "model-free" approaches (*6*, *7*) in nonsocial decision-making but with a greater emphasis on uncertainty due to the greater unpredictability of human behavior in social tasks.

In belief-based decision-making, the subject learns a model of the environment, updates the model based on observations and rewards, and chooses actions based on a probabilistic "belief" about the current state of the world (*5*, *8*, *9*). As a result, the relationship of the current action with rewards received and current observations is indirect. Besides the history of rewards received and the current observation, the learned model can also include other factors such as potential future rewards and more general rules about the environment. Therefore, the belief-based (model-based) approach is more flexible than belief-free (model-free) decision-making (*10*, *11*). However, belief-based decision-making requires more cognitive resources, for example, for simulation of future events. Thus, there is an inherent trade-off between the two types of approaches, and determining which approach humans adopt for different situations is an important open area of research (*12*).

Several studies have presented evidence in favor of the belief-based approach by quantifying the similarity between probabilistic model-based methods and human behavior when the subject interacts with or reasons about another human (*5*, *13*–*18*). 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 compared to the cost of only tracking the reward history of each action given the current observations. On the other hand, consistent with the importance that human society places on group decisions, a belief-based approach might be the optimal strategy.

How does one extend a belief-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) (*19*, *20*). 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 (*21*). Examples of the task include guarding duty, blood donation, and stepping forward to stop an act of violence in a public place (*22*). To mimic the volunteer's dilemma in a laboratory setting, we used the thresholded binary version of a multiround public goods game (PGG) where the actions of each individual are hidden from others (*21*, *23*).

Using an optimal Bayesian framework based on partially observable Markov decision processes (POMDPs) (*24*), we propose that in group decision-making, humans simulate the "mind of the group" by modeling an average group member's mind when making their current choices. Our model incorporates prior knowledge, current observations, and a simulation of the future based on the current actions for modeling human decisions within a group. We compared our model to a model-free reinforcement learning approach based on the reward history of each action as well as a previous descriptive method for fitting human behavior in the PGG. Our model predicts

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. [2]Center for Mind and Brain, University of California, Davis, Davis, CA, USA. [3]Neuroeconomics Laboratory, Institut des Sciences Cognitives Marc Jeannerod, Lyon, France. [4]Department of Psychology, New York University, New York, NY, USA. [5]Center for Neurotechnology, University of Washington, Seattle, WA, USA.
*Joint senior authors.
†Corresponding author. Email: rao@cs.washington.edu

human behavior significantly better than the model-free reinforcement learning and descriptive approaches. Furthermore, by leveraging the interpretable nature of our model, we are able to show a potential underlying computational mechanism for the group decision-making process.

RESULTS

Human behavior in a binary PGG

The participants were 29 adults (mean age, 22.97 years old ± 0.37; 14 women). We analyzed the behavioral data of 12 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 (*25*), in our PGG, public goods were produced as a group reward (G = 2 MU to each player) if and only if at least k players each contributed 1 MU. k was set to two or four randomly for each session and conveyed to group members before the start of the session. The resultant amount after one round is therefore E − C + G = 2 MU for the contributor and E + G = 3 MU for the free-rider when public goods were produced (the round was a SUCCESS). On the other hand, the contributor has E − C = 0 MU and the free-rider has E = 1 MU when no public goods were produced (the round was a FAILURE).

Figure 1 depicts one round of the PGG task. After the subject acts, the total number of contributions, free-rides, and the overall outcome of the round is revealed (success or failure in securing the 2 MU group reward), but each individual player's actions remained unknown. In addition, as shown in the figure, the value of k for the current session was always presented on the screen to ensure that the subjects had it in mind when making decisions. 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 using an algorithm based on human data (see Methods). In each session, the subject played with a different group of players.

As shown in Fig. 2A, subjects contributed significantly more when the number of required volunteers was higher with an average contribution rate of 55% (SD = 0.31) for $k = 4$ in comparison to 33% (SD = 0.18) for $k = 2$ {two-tailed paired sample t test, $t(28) = 3.94$, $P = 5.0 \times 10^{-4}$, 95% confidence interval (CI) difference = [0.11,0.33]}. In addition, Fig. 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 = 0.29) when $k = 4$ {two-tailed paired sample t test, $t(28) = 10.08$, $P = 8.0 \times 10^{-11}$, 95% CI difference = [0.40,0.60]}. All but six of the subjects contributed more when $k = 4$ (Fig. 2C). Of these six players, five chose to free-ride more than 95% of the time. In addition, success rate was higher when $k = 2$ for all players (Fig. 2D).

The contribution rate of the subjects dropped during the course of the trial on average, especially for $k = 4$, but remained above zero. Figure 2E shows the average contribution rate across all subjects as a function of round number (1 to 15). We also compared the average contribution for the first five rounds with that for the last five rounds. For $k = 4$, the average contribution probability across all subjects for the first five rounds was 0.6 (SD = 0.20) and significantly higher than that for the last five rounds (average across subjects = 0.49, SD = 0.19) {two-tailed paired sample t test, $t(28) = 3.65$, $P = 0.001$, 95% CI difference = [0.05,0.17]}. For $k = 2$, the difference between the first five rounds (average = 0.53, SD = 0.32) and the last five rounds (average = 0.50, SD = 0.33) was insignificant {two-tailed paired sample t test, $t(28) = 1.51$, $P = 0.14$, 95% CI difference = [− 0.01,0.06]}.

The average contribution probability did not change significantly as subjects played more games (Fig. 2F). In each condition, most of the players played at least five games (27 players for $k = 2$ and 26 for $k = 4$). For $k = 2$, in their first game, the average contribution rate of players was 0.37 (SD = 0.25), while in their fifth game, it was 0.30 (SD = 0.24) {two-tailed paired sample t test, $t(26) = 1.34$, $P = 0.19$, 95% CI difference = [− 0.03,0.17]}. When $k = 4$, the average contribution rate was 0.57 (SD = 0.30) in the first game and 0.61 (SD = 0.35) in the fifth game {two-tailed paired sample t test, $t(25) = -0.69$, $P = 0.50$, 95% CI difference = [− 0.16,0.08]}.

Probabilistic model of theory of mind for the group in the PGG

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

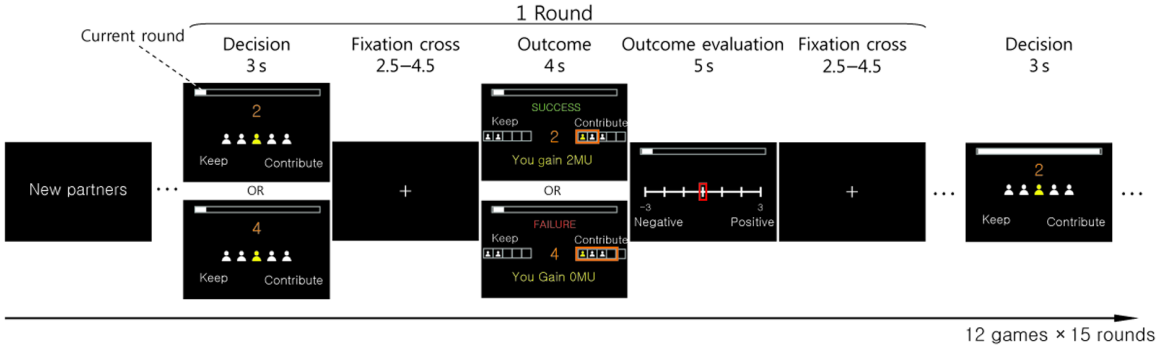

Fig. 1. Multiround PGG. The figure depicts the sequence of computer screens a subject sees in one round of the PGG. The subject is assigned four other players as partners, and each round requires the subject to make a decision: Keep 1 MU (i.e., free-ride) or contribute 1 MU. The subject knows whether the threshold to generate public goods (reward of 2 MU for each player) is two or four contributions (from the five players). After the subject acts, the total number of contributions and overall outcome of the round (success or failure) are revealed.

Downloaded from http://advances.sciencemag.org/ on February 8, 2021

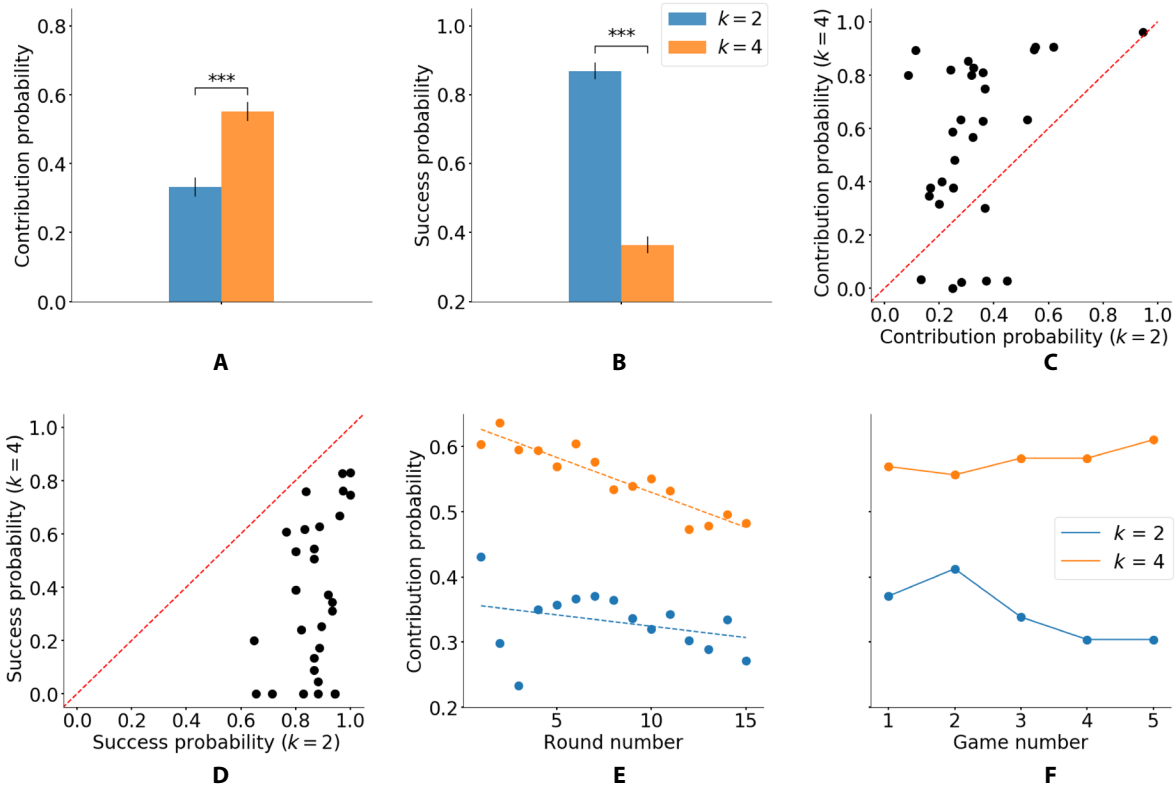

Fig. 2. Human behavior in the PGG Task. (**A**) Average contribution probability across subjects is significantly higher when the task requires more volunteers (*k*) to generate the group reward. (**B**) Average probability of success across all subjects in generating the group reward is significantly higher when *k* is lower. Error bars indicate within-subject SE (*52*). (**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 $k = 4$. **(E)** Average probability of contribution across subjects decreases throughout a game, especially for $k = 4$. Dotted lines are linear functions showing this trend for each *k*. (**F**) Average contribution probability across subjects as a function of number of games played. The contribution probability does not change significantly as subjects play more games.

contributions they anticipate the others to make in that round. Because 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

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

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 that 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. Second, other group members also have a theory of mind. Therefore, they can be expected to change their strategy based on the actions of others. Because of this ability in other group members, each player needs to simulate the effect of their action on the group's behavior in the future.

To model the uncertainty in θ, we assume that a probability distribution over θ is maintained 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 prior life experience of the player, or what they believe others would do through fictitious play (*26*). 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 Eq. 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

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

Starting with a prior probability $\mathrm{Beta}(\alpha_1, \beta_1)$ 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.

Downloaded from http://advances.sciencemag.org/ on February 8, 2021

In economics, the ability to infer the belief of others is sometimes called sophistication (*27*, *28*). Here, we consider a simple form of sophistication: We assume that each player thinks other group members have the same model as themselves (α and β). This is justifiable due to computational efficiency and more importantly anonymity of players. As a result, with a prior of Beta(α_t, β_t) after observing c contributions (including one's own when made) in round t, the posterior probability of θ for the subject becomes Beta(α_{t+1}, β_{t+1}), where $\alpha_{t+1} = \alpha_t + c$ and $\beta_{t+1} = \beta_t + N - c$. Technically, this follows because the beta distribution is conjugate to the binomial distribution (*29*). 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 represents 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 subject thinks that the average contribution probability 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 $0 \leq \gamma \leq 1$, which serves as a decay rate: 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(α_t, β_t) to Beta(α_{t+1}, β_{t+1}) where $\alpha_{t+1} = \gamma\alpha_t + c$ and $\beta_{t+1} = \gamma\beta_t + N - c$.

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 Eq. 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(α, β) is given by

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

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. Specifically, our model assumes that its own contribution in the current round increases the average contribution rate of the group in the future rounds. Equation 10 in Methods shows the exact assumptions of our model (with updates of $\alpha_{t+1} = \gamma\alpha_t + c$ and $\beta_{t+1} = \gamma\beta_t + N - c$ for its belief) about the dynamics of the actual (hidden) state of the environment. 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. Thus, an optimal agent would contribute for two reasons. First, contributing could enable the group to reach at least k volunteers in the current round. Second, contributing encourages other members to contribute in future rounds. Specifically, a contribution by the subject increases the average contribution rate for the next round by increasing α in the next round (see the transition function in Methods).

Long-term reward maximization (as discussed above) based on probabilistic inference of hidden state in an environment (here, θ, the probability of contribution of group members) can be modeled using the framework of POMDPs (*24*). Further details can be found in Methods, but briefly, to maximize the total expected reward, our model starts from the last round, the reward is calculated for each action and state, and then the model steps 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, and Fig. 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, Fig. 4 (A and B) shows 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 beliefs to actions is called a policy.

Our simulations using the POMDP model showed that considering a much longer horizon (e.g., 50 rounds) instead of just 15 rounds gave a better fit to the subjects' behavior, suggesting that human subjects may be inclined to use long horizons for group decision-making tasks (see Discussion). Such a long horizon for determining the optimal policy makes the model similar to an infinite horizon POMDP model (*30*). As a result, the optimal policy for all rounds in our model is very similar to the policy for round 9 shown in Fig. 4 (A and B).

In summary, the POMDP model performs two computations simultaneously. The first computation is probabilistic estimation of the (hidden) average contribution rate through belief updates. The average contribution rate changes during the course of the game as players interact with each other. The second computation involves selecting actions to influence this average contribution rate and to maximize total expected reward. This is the action selection component, which is performed by backward reasoning from the last round.

POMDP model predicts human behavior 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 fit these parameters to the subject's actions by minimizing the error, i.e., the difference between the POMDP model's predicted action and the

Downloaded from http://advances.sciencemag.org/ on February 8, 2021

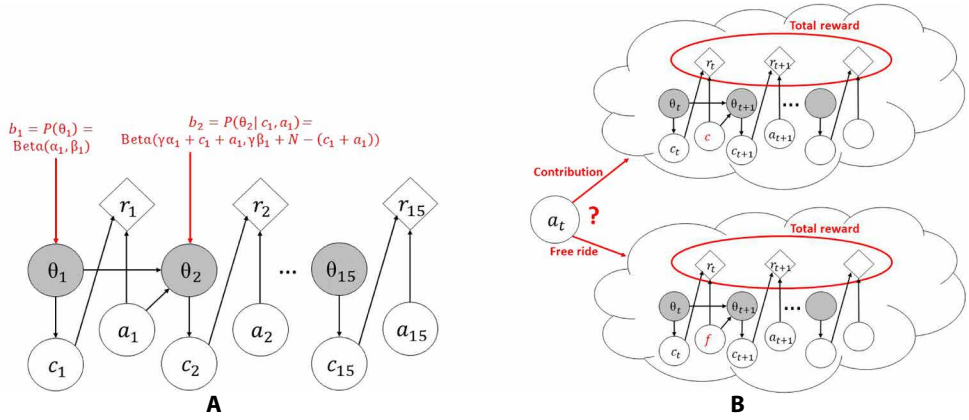

Fig. 3. POMDP model of the multiround PGG. (**A**) Model: The subject does not know the average probability of contribution of the group. The POMDP model assumes that the subject maintains a probability distribution ("belief," denoted by b_t) about the group's average probability of contribution (denoted by θ_t) and updates this belief after observing the outcome c_t (contribution by others) in each round. (**B**) Action selection: The POMDP model chooses an action (a_t) that maximizes the expected total reward ($\cdot r_i$) across all rounds based on the current belief and the consequence of the action (contribution "c" or free-ride "f") on group behavior in future rounds.

subject's action in each round. The average percentage error across all rounds is then the percentage of rounds that the model incorrectly predicts (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 (*29*), where each "left out" data point is one whole game and the parameters were fit to the other 11 games of the subject. Note that our LOOCV accuracy is a prediction of the subject's behavior in a game without any parameter tuning based on this game. In addition, while different rounds of each game are highly correlated, the games of each subject are independent from each other (given the parameters of that subject) as the other group members change in each game.

We found that the POMDP model had an average fitting accuracy across subjects of 84% (SD = 0.06), while the average LOOCV accuracy was 77% (SD = 0.08). Figure 5A compares the average fitting and LOOCV accuracies of the POMDP model with two other models. The first is a model-free reinforcement learning model known as Q-learning: Actions are chosen on the basis of their rewards in previous rounds (*31*), with the utility of group reward, initial values, and learning rate as free parameters (five parameters per subject; see Methods).

The average fitting accuracy of the Q-learning model was 79% (SD = 0.07), which is significantly worse than the POMDP model's fitting accuracy given above {two-tailed paired t test, $t(28) = -6.75$, $P = 2.52 \times 10^{-7}$, 95% CI difference = [−0.06, −0.03]}. In addition, the average LOOCV accuracy of the POMDP model was significantly higher than the average LOOCV accuracy of Q-learning, which was 73% (SD = 0.09) {two-tailed paired t test, $t(28) = 2.20$, $P = 0.037$, 95% CI difference =[0.004,0.08]}.

We additionally tested a previously explored descriptive model in the PGG literature known as the linear two-factor model (*32*), 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 fitting accuracy of the two-factor model was 78% (SD = 0.09), which is significantly lower than the POMDP model's fitting accuracy {two-tailed paired t test, $t(28) = -4.86$, $P = 4.1 \times 10^{-5}$, 95% CI difference = [−0.08, −0.03]}. Moreover, the LOOCV accuracy of the two-factor model was 47% (SD = 20), significantly lower than the POMDP model {two-tailed paired t test, $t(28) = -7.61$, $P = 2.7 \times 10^{-8}$, 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). Given that the number of free parameters for the descriptive and model-free approaches is greater than or equal to the number of free parameters in the POMDP model, the higher accuracy of POMDP is notable in terms of model comparison.

We tested the POMDP model's predictions of contribution probability for each subject for the two k values with experimental data (same data as in Fig. 2C; see Methods). As shown in Fig. 5 (B and C), the POMDP model's predictions match the pattern of distribution of actual data from the experiments.

The POMDP model, when fit to a subject's actions, can also explain other events during the PGG task in contrast to the other models described above. For example, based on Eq. 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 the POMDP model's prediction of success probability across subjects was 71% (SD = 0.07). Moreover, the predictions matched the pattern of success rate data from the experiment (Fig. 5, D and E). The other models presented above are not capable of making such a prediction.

The POMDP model's predictions also match experimental data when the data points are binned on the basis of round of the game. The model correctly predicts a decrease in contribution for $k = 4$ and lack of significant change in contribution rate on average for $k = 2$ (Fig. 5F). Moreover, the model's prediction of a subject's belief about group success matches the actual data round by round (Fig. 5G). Further comparisons to other models, such as the interactive-POMDP model (*33*), are provided in the Supplementary Materials.

Downloaded from http://advances.sciencemag.org/ on February 8, 2021

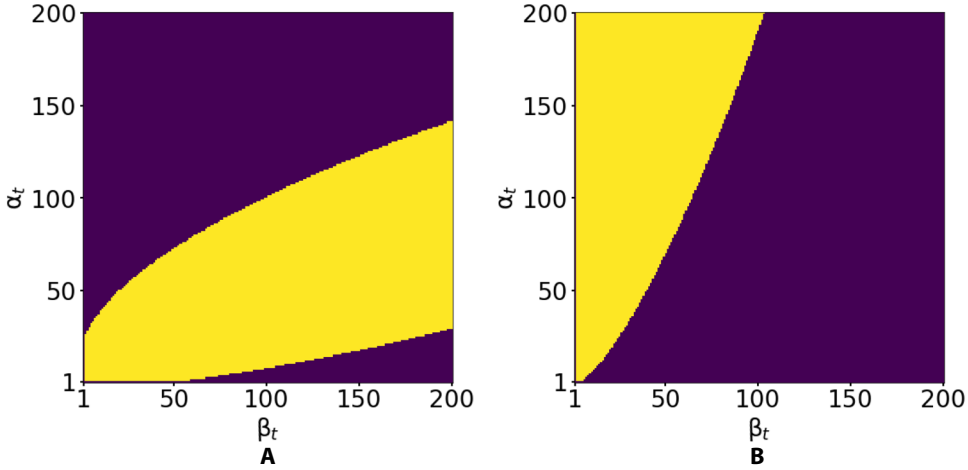

Fig. 4. Optimal actions prescribed by the POMDP policy as a function of belief state. Plot (**A**) shows the policy for $k = 2$ and plot (**B**) for $k = 4$. The purple regions represent those belief states (defined by α_t and β_t) for which free-riding is the optimal action; the yellow regions represent 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.

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 decay rate γ determines the weight given to past observations compared to new ones: The smaller the decay rate, 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 (Fig. 6, A and B). The ratio α_1/β_1 was in the reasonable range of 0.5 to 2 for almost all subjects (Fig. 6C; in our algorithm, 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 (Fig. 6D), suggesting that prior belief about groups did have a significant role in players' strategy, but it was not the only factor because observations over multiple rounds can still alter this initial belief. To confirm the effect of actions during the game, we performed a comparison with a POMDP model that does not update α and β over time and only uses its prior. The accuracy of this modified POMDP model was 66% (SD = 0.17), significantly lower than our original model {two-tailed paired t test, $t(28) = -5.47$, $P = 7.64 \times 10^{-6}$, 95% CI difference =[− 0.23, − 0.11]}. The average α_t and β_t for each of the 15 rounds, as well as distributions of their difference with the prior values α_1 and β_1 are presented in the Supplementary Materials.

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 two and three (Fig. 6E).

In addition, we 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. Figure 6 (F and G) shows 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 most of the subjects expected less than 60% chance for success when $k = 4$. While these beliefs cannot be directly validated, the results point to the importance of the required number of volunteers in shaping the subjects' behavior.

Additionally, the decay rate γ, which determines the weight accorded to the prior and previous observations compared to the most recent observation, was almost always above 0.95, with a mean of 0.93 and a median of 0.97 (Fig. 6H). Only three subjects had a decay rate less than 0.95 (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.

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 behavior of the group. The greater accuracy of our model in explaining and predicting the subjects' behavior 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 that the 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. Existing models to explain human behavior in the PGG, for example, are descriptive and do not provide insights into the computational mechanisms underlying the decisions (*32*). While the regression-based descriptive method we compared our POMDP model to can potentially be seen as a "learned" model-free approach to mapping observations to choice in the next round, our model was also able to outperform this method.

Downloaded from http://advances.sciencemag.org/ on February 8, 2021

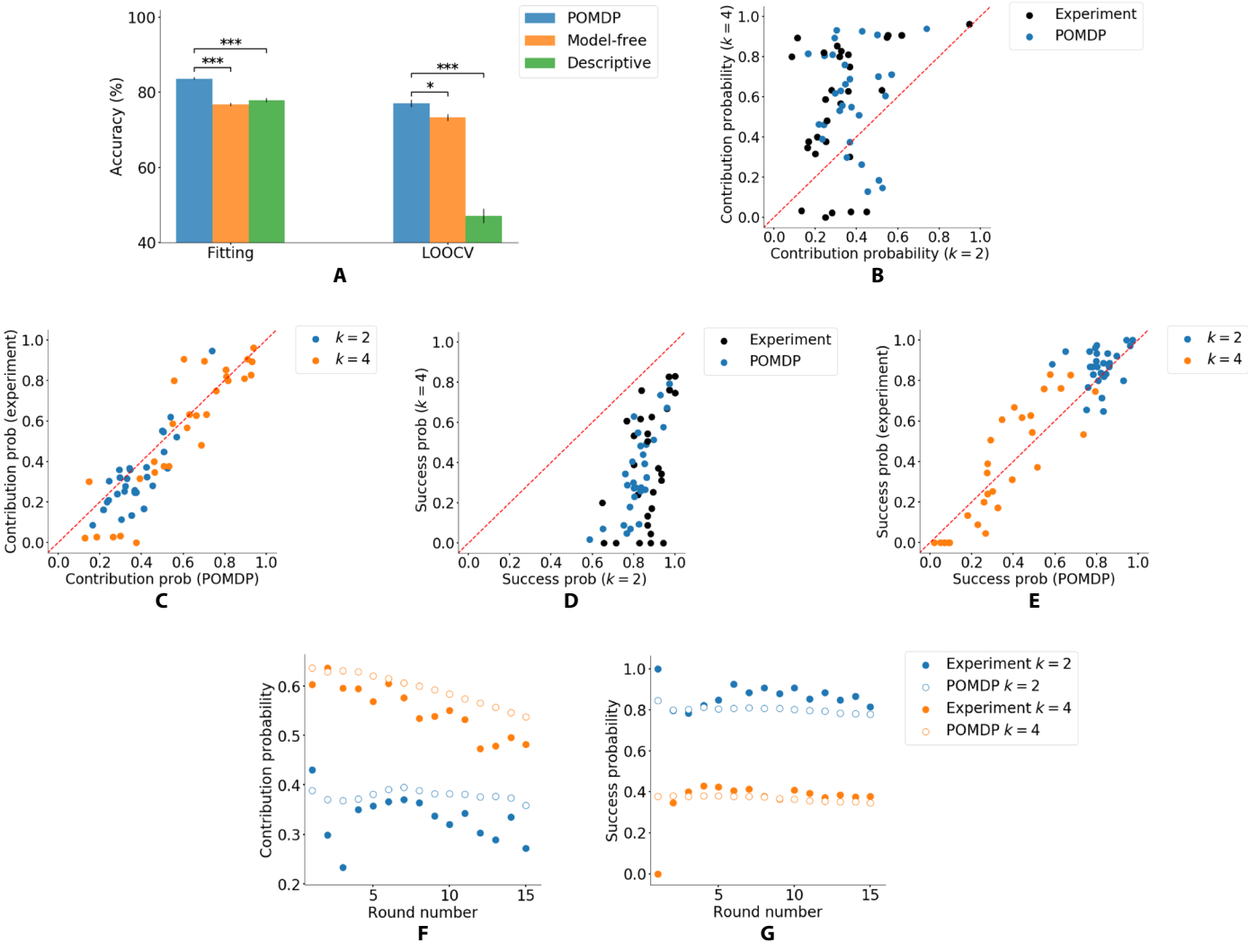

Fig. 5. POMDP model's performance and predictions. (**A**) Average fitting and LOOCV accuracy across all models. The POMDP model has significantly higher accuracy compared to the other models (*$P < 0.05$ and ***$P < 0.001$). Error bars indicate within-subject SE (*52*). (**B**) POMDP model's prediction of a subject's probability of contribution compared to experimental data for the two k values [black circles: same data as in Fig. 2C). (**C**) Same data as (B) but the POMDP model's prediction and the experimental data are shown for each k separately (blue for $k = 2$ and orange for $k = 4$]. (**D**) POMDP model's prediction (blue circles) of a subject's belief about group success in each round (on average) compared to actual data (black circles, same data as in Fig. 2D). (**E**) Same data as (D), but the POMDP model's prediction and actual data are shown for each k separately (blue for $k = 2$ and orange for $k = 4$). (**F**) Same data as (B) and (C) but with the data points binned on the basis of round of the game. (**G**) Same data as (D) and (E) but with the data points binned based on round of the game.

In addition to providing a better fit and prediction of 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. In addition, as alluded to in Fig. 6 (C, D, and H), 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. The POMDP model is normative and strictly constrained by probability theory and optimal control theory. The beta distribution is used because it is the conjugate prior of the binomial distribution (*29*) and not due to better fits compared to other distributions.

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 Fig. 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 five 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 Fig. 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 Fig. 4A). The POMDP model predicts that free-riding is not a viable action in the $k = 4$ case (Fig. 4B) because not only does this action require all the other four players

Downloaded from http://advances.sciencemag.org/ on February 8, 2021

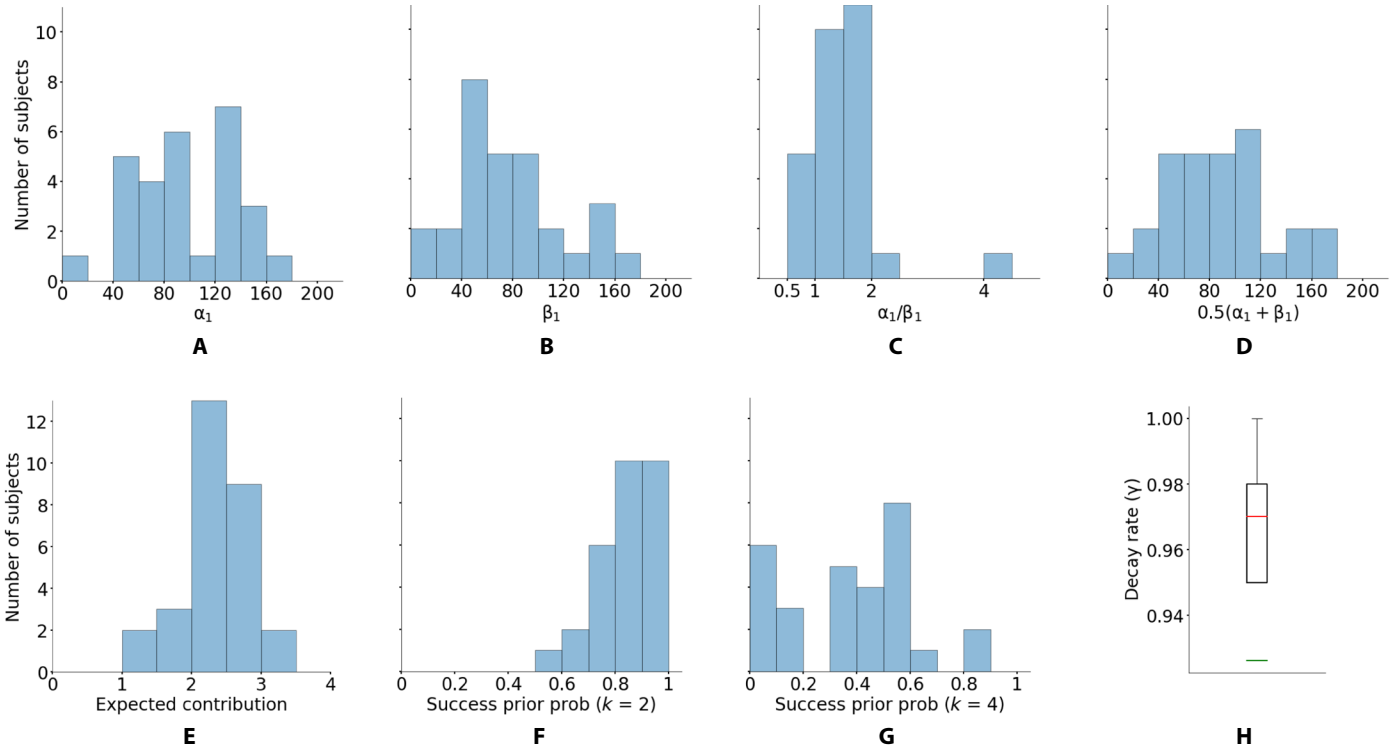

Fig. 6. Distribution of POMDP parameters across subjects. (**A**) Histogram of α_1 across all subjects. (**B**) Histogram of β_1 across all subjects. (**C**) Histogram of the ratio α_1/β_1 shows a value between 0.5 and 2 for almost all subjects. (**D**) Histogram of $(\alpha_1 + \beta_1)/2$. For most subjects, this value is between 40 and 120. (**E**) Histogram of prior belief Beta(α_1, β_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. (**F**) When $k = 2$, all subjects expected a high probability of group success in the first round (before making any observations about the group). (**G**) When $k = 4$, almost all subjects assigned a chance of less than 60% to group success in the first round. (**H**) Box plot of decay rate γ across subjects shows that this value is almost always above 0.95. The median is 0.97 (orange line) and the mean is 0.93 (green line).

to contribute to generate the group reward in the current round but also such an action increases the chances that the group contribution will be lower in the next round, resulting in lesser expected reward in future rounds. The opposite situation can also occur especially when $k = 2$. A player may contribute not to gain the group reward in the current round but to encourage others to contribute in the next rounds. When an optimal player chooses free-riding due to low cooperativeness of the group, the estimated average contribution is so low that the group is not likely to get the group reward in the next rounds even with an increase in the average contribution due to the player's contribution. On the other hand, when an optimal player chooses to free-ride due to high cooperativeness of the group, the estimated average contribution rate is so high that the chance of success remains high in future rounds even with a decrease in average contribution rate due the player free-riding in the current round.

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 experimental data did not support this conclusion. Our model is able to explain these data using the hypothesis that subjects may use a longer horizon than the exact number of rounds in a game. Such a strategy provides a significant computational benefit by making the policies for different rounds similar to each other, avoiding recalculation 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 (*34*, *35*). More broadly, group decision-making occurs among groups of humans (and animals) that live together. Thus, any group decision-making involves practically an infinite horizon, i.e., there is always a future interaction even after the current task has ended, justifying the use of long horizons.

In the volunteer's dilemma, not only is the common goal not reached when there are not enough volunteers but also 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 the 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 (*36*). In addition, 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 as 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 behavior as one subgroup to reduce the number of minds one should reason about.

Downloaded from http://advances.sciencemag.org/ on February 8, 2021

Our POMDP framework assumes that each subject starts with the same prior about average group member contribution probability at the beginning of each game. However, subjects might try to estimate this prior for a new group in the first few rounds, i.e., "explore" their new environment before seeking to maximize their reward ("exploit") based on this prior (*5*). Such an "active inference" approach has been studied in two-person interactions (*15*, *16*) and is an interesting direction of research in group decision-making.

Mimicking human behavior does not guarantee that a POMDP model (or any model) is being implemented in the brain. However, the POMDP model's generalizability and the interpretability of its components, such as existence of a prior or simulation of the future, make it a useful tool for understanding the decision-making process.

The POMDP framework can model social tasks beyond economic decision-making, such as prediction of others' intentions and actions in everyday situations (*37*). 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 (*38*).

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 (*39*–*41*) and in tasks involving interactions with a single individual (*13*, *17*, *18*). We believe that the POMDP model we have proposed can likewise prove useful in interpreting neural responses and data from neuroimaging studies of group decision-making tasks. In addition, the model can be used for Bayesian theory-driven investigations in the field of computational psychiatry (*42*). For example, theory of mind deficits are a key feature of autism spectrum disorder (*43*), 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 (*44*).

METHODS

Experiment

Thirty 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. On the basis of 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 PGG (i.e., the volunteer's dilemma), each containing 15 rounds. In the first two 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 Fig. 1), the participant had to make a decision within 3 s by pressing a key; otherwise, the round was repeated. After the action selection (2.5 to 4 s), the outcome of the round was shown to the subject for 4 s. 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 four 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 $a_i^t \quad (1 \le i \le N)$, 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

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

where $K = k/N$, in which k is the required number of contributors.

This model has three free parameters: e_0, e_1, and e_2. These were obtained by fitting the above function to the actual actions of subjects in another PGG study (*45*), making this function a simulation of human behavior in the PGG task. Specifically, to generate the actions of others, we fixed e_2 to 1 for all games. e_0 was drawn randomly from the range of [0.15,0.35] for each game, and e_1 was set to $1 - e_0$. This combination and the random sampling of e_0 in each game simulated different response strategies for the others in each game, simulating new sets of group members. Higher values of e_0 make the algorithm more likely to choose its next action based on the result of the group interaction in the previous round (especially the action of the subject). On the other hand, lower values of e_0 make the algorithm more likely to stick to its previous action. For the first round of each game, we used the mean contribution rate of each subject as their fellow members' decision.

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., $T(s', s, a) = P(s' \mid s, a)$, and R is the reward function $S \times A \rightarrow R$ representing the reward associated with each state and action (*30*). 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 (*30*). For an MDP with time horizon H, the Q value, value function V, and action function U at the last time step $t = H$ are defined as

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

Downloaded from http://advances.sciencemag.org/ on February 8, 2021

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

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

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, and 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., $O(z, s) = P(z \mid s)$. To find the optimal policy, the POMDP model uses the posterior probability of states, known as the belief state, where $b_t(s) = P(s \mid z_1, a_1, z_2, \ldots, a_{t-1})$. 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

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

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, unless the belief state can be represented by a few parameters as in our case (*30*). For solving larger POMDP problems, various approximation and learning algorithms have been proposed. We refer the reader to the growing literature on this topic (*46–48*).

POMDP for binary PGG

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. Because 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 (*29*). 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(α_1, β_1) and with an additional free parameter γ, the next belief states are 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 PGG. Therefore, the elements of our new MDP derived from the PGG POMDP are as follows

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

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

The POMDP model above assumes that the hidden state, i.e. θ, is a random variable following a Bernoulli distribution, which changes with the actions of all players in each round. These actions serve as samples from this distribution, with α_1 and β_1 being the initial samples. Also, the decay rate γ controls the weights of previous samples. Using maximum likelihood estimation, for any t, θ_t equals $\alpha_t/(\alpha_t + \beta_t)$. One can also estimate θ in a recursive fashion

$$\theta_{t+1} \leftarrow \frac{1}{\gamma\alpha_t + \gamma\beta_t + N}\left((\gamma\alpha_t + \gamma\beta_t)\,\theta_t + \sum_{i=1}^{N} a_i^t\right) \tag{10}$$

where a_i^t is the action of player i in round t ($a_i^t = 1$ for contribution and 0 for free-ride).

According to the experiment, the time horizon should be 15 time steps. However, we found that a longer horizon ($H = 50$) for all players provides a better fit to the subjects' data, potentially reflecting an intrinsic bias in humans for using longer horizons for social decision-making. 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 and 1 with a precision of 0.01 (0.01,0.02, …,0.99,1.0) 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 of predicting it). Because the utility value for public good for a subject can be higher than the monetary reward due to social or cultural reasons (*49*), 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. A preliminary version of the above model but without the γ parameter was presented in (*50*).

As specified above, the best action for each state in round t is $U^t(s)$. The probability of contribution (choice probability) can be calculated using a logit function: $1/(1 + exp\,(z(Q^t(s,f) - Q^t(s,c))$ (*19*). For each k, we used one free parameter z across all subjects to maximize the likelihood of contribution probability given the experimental data [implementation by scikit-learn (*51*)]. Note that the parameter z does not affect the accuracy of fits and predictions because it does not affect the action with the maximum expected total reward.

In round t, if the POMDP model selects the action "contribution," the probability of success can be calculated as $\sum_{m=k-1}^{N-1} P(m \mid \alpha_t, \beta_t)$ (see Eq. 3). Otherwise, the probability of success is $\sum_{m=k}^{N-1} P(m \mid \alpha_t, \beta_t)$. 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.

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

Downloaded from http://advances.sciencemag.org/ on February 8, 2021

"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

$$\begin{cases} Q^1(c) \leftarrow p(E - C + G) + (1 - p)(E - C) \\ Q^1(f) \leftarrow p(E + G) + (1 - p)E \end{cases} \tag{11}$$

We customized the utility function for each subject by making the group reward G a free parameter to account for possible prosocial intent (*49*). 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 on the basis of 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

$$1 \le t \le 15: \begin{cases} \hat{a}^t = \arg\max_{a \in \{c,f\}} Q^t(a) \\ Q^{t+1}(a^t) \leftarrow (1 - \eta^t) Q^t(a^t) + \eta^t r^t \end{cases} \tag{12}$$

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.

Descriptive model

Our descriptive model was based on a logistic regression [implementation by scikit-learn (*51*)] 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 three 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

$$\hat{a}_1^{t+1} = \begin{cases} c & \kappa_0 + \kappa_1 a_1^t + \kappa_2 \left(\sum_{i=2}^N a_i^t \right) > 0 \\ f & \text{otherwise} \end{cases} \tag{13}$$

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.

Leave-one-out cross-validation

For all three approaches, LOOCV was computed on the basis of the games played by each subject. For each subject, we set aside one game, fitted the parameters to the other 11 games, and computed the error of the model with fitted parameters on the game that was set aside. We repeated this for all games and reported the average of the 12 errors as LOOCV error for the subject.

Static probability distribution and greedy strategy

If a player does not consider the future and solely maximizes the expected reward in the current round (greedy strategy) or ignores the effect of an action on others, 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 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 one 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 0.5. 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

$$P(k-1 \mid \theta) = \binom{N-1}{k-1} \theta^{k-1} (1-\theta)^{N-k} \le \binom{N-1}{k-1} \left(\frac{k-1}{N-1} \right)^{k-1} \left(\frac{N-k}{N-1} \right)^{N-k} = 0.75^3 < 0.5 \tag{14}$$

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

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/11/eaax8783/DC1

Supplementary Text

Fig. S1. Distribution and change in belief parameters over multiple rounds.

Fig. S2. Data generated by the POMDP model compared to experimental data.

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. A. G. Sanfey, Social decision-making: Insights from game theory and neuroscience. *Science* **318**, 598–602 (2007).
2. J. Joiner, M. Piva, C. Turrin, S. W. C. Chang, Social learning through prediction error in the brain. *npj Sci. Learn.* **2**, 8 (2017).
3. D. Mookherjee, B. Sopher, Learning and decision costs in experimental constant sum games. *Games Econ. Behav.* **19**, 97–132 (1997).
4. C. F. Camerer, T. H. Ho, Experience-weighted attraction learning in normal form games. *Econometrica* **67**, 827–874 (1999).
5. K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, J. O'Doherty, G. Pezzulo, Active inference and learning. *Neurosci. Biobehav. Rev.* **68**, 862–879 (2016).
6. P. Dayan, N. D. Daw, Decision theory, reinforcement learning, and the brain. *Cogn. Affect. Behav. Neurosci.* **8**, 429–453 (2008).
7. N. D. Daw, P. Dayan, The algorithmic anatomy of model-based evaluation. *Philos. Trans. R. Soc. B* **369**, 20130478 (2014).
8. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
9. A. J. Culbreth, A. Westbrook, N. D. Daw, M. Botvinick, D. M. Barch, Reduced model-based decision-making in schizophrenia. *J. Abnorm. Psychol.* **125**, 777–787 (2016).
10. B. B. Doll, D. A. Simon, N. D. Daw, The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol.* **22**, 1075–1081 (2012).
11. R. J. Dolan, P. Dayan, Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
12. C. J. Charpentier, J. P. ODoherty, The application of computational models to social neuroscience: Promises and pitfalls. *Soc. Neurosci.* **13**, 637–647 (2018).
13. W. Yoshida, B. Seymour, K. J. Friston, R. J. Dolan, Neural mechanisms of belief inference during cooperative games. *J. Neurosci.* **30**, 10744–10751 (2010).
14. T. Xiang, D. Ray, T. Lohrenz, P. Dayan, P. R. Montague, Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLOS Comput. Biol.* **8**, e1002841 (2012).
15. M. Moutoussis, P. Fearon, W. El-Deredy, R. J. Dolan, K. J. Friston, Bayesian inferences about the self (and others): A review. *Conscious. Cogn.* **25**, 67–76 (2014).
16. M. Moutoussis, N. J. Trujillo-Barreto, W. El-Deredy, R. J. Dolan, K. J. Friston, A formal model of interpersonal inference. *Front. Hum. Neurosci.* **8**, 160 (2014).
17. A. Hula, P. R. Montague, P. Dayan, Monte carlo planning method estimates planning horizons during interactive social exchange. *PLoS Comput. Biol.* **11**, e1004254 (2015).
18. C. L. Baker, J. Jara-Ettinger, R. Saxe, J. B. Tenenbaum, Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* **1**, 0064 (2017).

Downloaded from http://advances.sciencemag.org/ on February 8, 2021

19. S. Suzuki, R. Adachi, S. Dunne, P. Bossaerts, J. P. O'Doherty, Neural mechanisms underlying human consensus decision-making. *Neuron* **86**, 591–602 (2015).
20. S. A. Park, S. Goïame, D. A. O'Connor, J.-C. Dreher, Integration of individual and social information for decision-making in groups of different sizes. *PLOS Biol.* **15**, e2001958 (2017).
21. D. Diekmann, Volunteer's dilemma. *J. Confl. Resolut.* **29**, 605–610 (1985).
22. J. M. Darley, B. Latane, Bystander intervention in emergencies: Diffusion of responsibility. *J. Pers. Soc. Psychol.* **8**, 377–383 (1968).
23. M. Archetti, I. Scheuring, Coexistence of cooperation and defection in public goods games. *Evolution* **65**, 1140–1148 (2011).
24. L. P. Kaelbling, M. L. Littman, A. R. Cassandra, Planning and acting in partially observable stochastic domains. *Artif. Intell.* **101**, 99–134 (1998).
25. E. Fehr, S. Gachter, Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994 (2000).
26. G. W. Brown, Iterative solution of games by fictitious play, in *Activity Analysis of Production and Allocation*, T. C. Koopmans, Ed. (Wiley, 1951), pp. 374–376.
27. M. Costa-Gomes, V. P. Crawford, B. Broseta, Cognition and behavior in normal-form games: An experimental study. *Econometrica* **69**, 1193–1235 (2001).
28. M. Devaine, G. Hollard, J. Daunizeau, Theory of mind: Did evolution fool us? *PLOS ONE* **9**, e87619 (2014).
29. K. P. Murphy, Machine learning: A probabilistic perspective, in *Adaptive Computation and Machine Learning* (MIT Press, 2012).
30. S. Thrun, W. Burgard, D. Fox, *Probabilistic Robotics* (MIT Press, 2005).
31. J. N. Tsitsiklis, Asynchronous stochastic approximation and Q-learning. *Mach. Learn.* **16**, 185–202 (1994).
32. M. Wunder, S. Suri, D. J. Watts. Empirical agent based models of cooperation in public goods games, in *Proceedings of the Fourteenth ACM Conference on Electronic Commerce (EC)* (ACM, 2013), pp. 891–908.
33. P. J. Gmytrasiewicz, P. Doshi, Interactive POMDPs: Properties and preliminary results, in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 3* (IEEE Computer Society, 2004), pp. 1374–1375.
34. I. Momennejad, E. M. Russek, J. H. Cheong, M. M. Botvinick, N. D. Daw, S. J. Gershman, The successor representation in human reinforcement learning. *Nat. Hum. Behav.* **1**, 680–692 (2017).
35. E. M. Russek, I. Momennejad, M. M. Botvinick, S. J. Gershman, N. D. Daw, Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLOS Comput. Biol.* **13**, e1005768 (2017).
36. H. Attias, Planning by probabilistic inference, in *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, Key West, FL, 3 to 6 January 2003.
37. D. I. Tamir, M. A. Thornton, Modeling the predictive social mind. *Trends Cogn. Sci.* **22**, 201–212 (2018).
38. D. I. Tamir, M. A. Thornton, J. M. Contreras, J. P. Mitchell, Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 194–199 (2016).
39. R. P. N. Rao, Decision making under uncertainty: A neural model based on partially observable Markov decision processes. *Front. Comput. Neurosci.* **4**, 146 (2010).
40. Y. Huang, R. P. N. Rao, Reward optimization in the primate brain: A probabilistic model of decision making under uncertainty. *PLOS ONE* **8**, e53344 (2013).
41. K. Khalvati, R. P. Rao, A Bayesian framework for modeling confidence in perceptual decision making, in *Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, 7 to 12 December 2015, pp. 2413–2421.
42. Q. J. M. Huys, T. V. Maia, M. J. Frank, Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* **19**, 404–413 (2016).
43. S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, I. Plumb, The "reading the mind in the eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J. Child Psychol. Psychiatry* **42**, 241–251 (2001).
44. P. Schwartenbeck, K. Friston, Computational phenotyping in psychiatry: A worked example. *eNeuro* **3**, ENEURO.0049-16.2016 (2016).
45. S. A. Park, S. Jeong, J. Jeong, TV programs that denounce unfair advantage impact women's sensitivity to defection in the public goods game. *Soc. Neurosci.* **8**, 568–582 (2013).
46. K. Khalvati, A. K. Mackworth, A fast pairwise heuristic for planning under uncertainty, in *Proceedings of The Twenty-Seventh AAAI Conference on Artificial Intelligence* (Association for the Advancement of Artificial Intelligence, 2013), pp. 187–193.
47. G. Shani, J. Pineau, R. Kaplow, A survey of point-based POMDP solvers. *Auton. Agent. Multi-Agent Syst.* **27**, 1–51 (2013).
48. Y. Luo, H. Bai, D. Hsu, W. S. Lee, Importance sampling for online planning under uncertainty. *Int. J. Robot. Res.* **38**, 162–181 (2018).
49. E. Fehr, U. Fischbacher, S. Gächter, Strong reciprocity, human cooperation, and the enforcement of social norms. *Hum. Nat.* **13**, 1–25 (2002).
50. K. Khalvati, S. A. Park, J.-C. Dreher, R. P. Rao, A probabilistic model of social decision making based on reward maximization, in *Advances in Neural Information Processing Systems*, Barcelona, Spain, 5 to 10 December 2016, pp. 2901–2909.
51. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
52. D. Cousineau, Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutor. Quant. Methods Psychol.* **1**, 42–45 (2005).

Acknowledgments

Funding: 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). **Author contributions:** R.P.N.R. and J.-C.D. developed the general research concept. S.A.P. designed and programmed the task under the supervision of J.-C.D., and M.S. ran the experiment under the supervision of J.-C.D. K.K. developed the model under the supervision of R.P.N.R., 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. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Both data and code are available upon request from the corresponding author. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 3 May 2019
Accepted 19 September 2019
Published 27 November 2019
10.1126/sciadv.aax8783

Citation: K. Khalvati, S. A. Park, S. Mirbagheri, R. Philippe, M. Sestito, J.-C. Dreher, R. P. N. Rao, Modeling other minds: Bayesian inference explains human choices in group decision-making. *Sci. Adv.* **5**, eaax8783 (2019).

Downloaded from http://advances.sciencemag.org/ on February 8, 2021

Science Advances

Modeling other minds: Bayesian inference explains human choices in group decision-making

Koosha Khalvati, Seongmin A. Park, Saghar Mirbagheri, Remi Philippe, Mariateresa Sestito, Jean-Claude Dreher and Rajesh P. N. Rao

Sci Adv **5** (11), eaax8783.
DOI: 10.1126/sciadv.aax8783

ARTICLE TOOLS http://advances.sciencemag.org/content/5/11/eaax8783

SUPPLEMENTARY MATERIALS http://advances.sciencemag.org/content/suppl/2019/11/21/5.11.eaax8783.DC1

REFERENCES This article cites 43 articles, 3 of which you can access for free
http://advances.sciencemag.org/content/5/11/eaax8783#BIBL

PERMISSIONS http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.

Copyright © 2019 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

Downloaded from http://advances.sciencemag.org/ on February 8, 2021