Article

Neuron

# Neural Mechanisms Underlying Human Consensus Decision-Making

## Highlights

- A task is used to study how the brain implements consensus decision-making
- Consensus decision-making depends on three distinct computational processes
- These different signals are encoded in distinct brain regions
- Integration of these signals occurs in the dorsal anterior cingulate cortex

## Authors

Shinsuke Suzuki, Ryo Adachi, ..., Peter Bossaerts, John P. O'Doherty

## Correspondence

shinsuke.szk@gmail.com

## In Brief

Suzuki et al. provide insight into the neural computations underlying human consensus formation. They implicate three different computational variables in this capacity. Each variable is encoded in distinct brain regions yet integrated within the dorsal anterior cingulate cortex.

Suzuki et al., 2015, Neuron *86*, 1–12
April 22, 2015 
http://dx.doi.org/10.1016/j.neuron.2015.03.019

CellPress

Please cite this article in press as: Suzuki et al., Neural Mechanisms Underlying Human Consensus Decision-Making, Neuron (2015), http://dx.doi.org/10.1016/j.neuron.2015.03.019

Neuron
Article

CellPress

# Neural Mechanisms Underlying Human Consensus Decision-Making

Shinsuke Suzuki,[1,2,*] Ryo Adachi,[1] Simon Dunne,[3] Peter Bossaerts,[4,5,6] and John P. O'Doherty[1,3]
[1]Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA
[2]JSPS Postdoctoral Fellow, Graduate School of Letters, Hokkaido University, Sapporo, Hokkaido 060-0810, Japan
[3]Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125, USA
[4]David Eccles School of Business, University of Utah, Salt Lake City, UT 84112, USA
[5]Faculty of Business and Economics, The University of Melbourne, Carlton, VIC 3010, Australia
[6]Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Parkville, VIC 3052, Australia
*Correspondence: shinsuke.szk@gmail.com
http://dx.doi.org/10.1016/j.neuron.2015.03.019

## SUMMARY

**Consensus building in a group is a hallmark of animal societies, yet little is known about its underlying computational and neural mechanisms. Here, we applied a computational framework to behavioral and fMRI data from human participants performing a consensus decision-making task with up to five other participants. We found that participants reached consensus decisions through integrating their own preferences with information about the majority group members' prior choices, as well as inferences about how much each option was stuck to by the other people. These distinct decision variables were separately encoded in distinct brain areas—the ventromedial prefrontal cortex, posterior superior temporal sulcus/temporoparietal junction, and intraparietal sulcus—and were integrated in the dorsal anterior cingulate cortex. Our findings provide support for a theoretical account in which collective decisions are made through integrating multiple types of inference about oneself, others, and environments, processed in distinct brain modules.**

## INTRODUCTION

In our daily life, we build consensus with other people in order to make collective decisions (Kerr and Tindale, 2004; Krause and Ruxton, 2002; Sumpter, 2010). This type of consensus decision-making has been widely observed in social animals from insects to primates (Conradt and Roper, 2005). Examples include nest-site selection in swarms of honey bees (Seeley and Visscher, 2004), coherent movement of individuals in schools of fish (Ward et al., 2011), collective choices of travel route in flocks of migrating birds (Black, 1988), and jury systems in humans (Devine et al., 2001). As the French philosopher the Marquis de Condorcet suggested (McLean, 1994), consensus decision-making can offer various advantages, such as a reduction of risk from predators and an enhancement of decision accuracy (Bahrami et al., 2010; Ioannou et al., 2012; Krause et al., 2010; Ward et al., 2011). Consensus formation is hence fundamental in human and animal social behavior.

Given its importance, consensus decision-making has been of considerable interest to many fields. Traditionally, social psychologists have studied mock juries (Davis et al., 1976), and economists have pursued theoretical aspects (Arrow, 1963). In biology, while researchers have primarily focused on cases of eusocial insects, recent studies have begun to investigate vertebrate animals (Conradt and Roper, 2005). However, it still remains unclear how consensus arises from interactions between human group members. Critically, no study to date in either animals or humans has examined the underlying neural mechanisms (Raafat et al., 2009).

Here, by combining behavior and fMRI with a computational model, we provide an account of consensus decision-making and its neural implementation. Our model stands on the following three hypotheses about key factors necessary for guiding decisions. The first hypothesis is that an individual's decision-making is guided by that individual's own preferences. The second is that there is a tendency to follow the majority's choice during consensus formation. These hypotheses are motivated by the results of classical human behavioral studies using mock juries (Davis et al., 1976), as well as recent findings in vertebrate animal studies (Sumpter and Pratt, 2009; Ward et al., 2011). The third hypothesis is based on recent findings in decision neuroscience that our brain is capable of inferences about hidden structures of the environment, including mental states of other people (Dayan and Daw, 2008; Yoshida et al., 2010). In the context of consensus formation with other people, we hypothesize that it is possible to infer others' preferences for each of the available options, or in other words, the "stickiness" of the options (i.e., how much each option was stuck to by the other people).

In the remainder of this paper, following the short description of our experimental task, we test the above hypotheses by simple model-free analyses; then we show that our model can well capture human behavior in a process of consensus formation; and finally we reveal the underlying neural mechanisms.

## RESULTS

### Experimental Design

To validate the computational model and examine its neural underpinnings, we developed an experimental paradigm in which

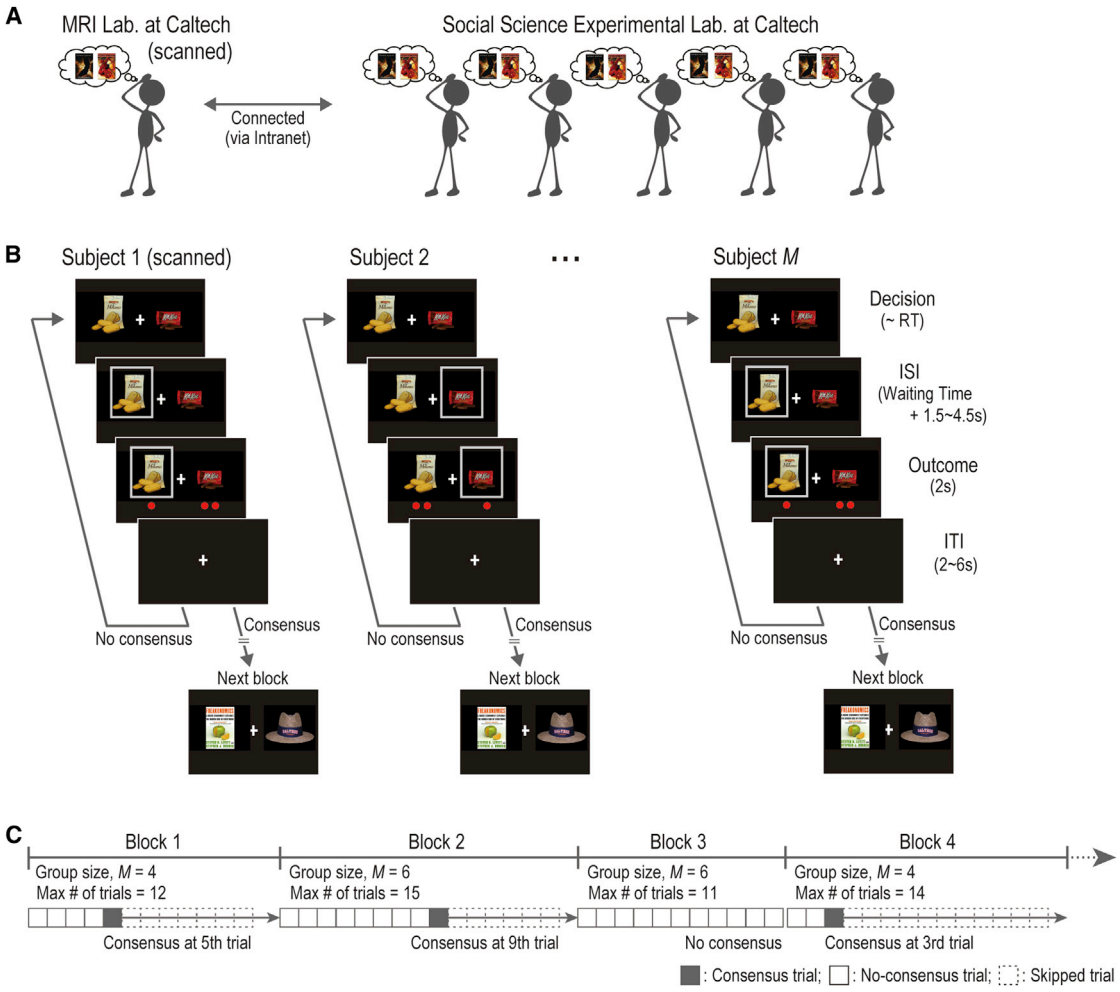


**Figure 1. Experimental Task**

(A) Illustration of the experimental setting. One participant inside the MRI scanner interacts with other participants. They try to build consensus on a choice between two items.

(B) Timeline of one trial. On each trial, participants choose between two items (Decision), and the item chosen is then highlighted by a gray frame. After a waiting time for the others' choices and a jittered delay (ISI), the other participants' choices are indicated by red dots (Outcome). Notably, participants are not able to identify each of the others; they were informed only about the distribution of the red dots (i.e., the number of others choosing each of the two items). If they reach a consensus, they move to the next block; otherwise, they again made a choice between the same items on the next trial in the same block. RT, reaction time; ISI, interstimulus interval; ITI, intertrial interval.

(C) Overall timeline of the experiment. The experiment consists of 40 blocks: 20 six-person blocks involving six participants (M = 6) and 20 four-person blocks involving four participants (M = 4). In each block, the maximum number of trials is determined randomly. Once participants reach a consensus, the remaining trials in the blocks are skipped and they move to the next block (e.g., block 1). If they do not build a consensus before the end of the block (e.g., block 3), they move to the next block and have no possibility to obtain any items for that block.

one participant was scanned with fMRI (20 participants were scanned in total), while interacting with five other participants outside the scanner (Figure 1A; see Experimental Procedures for details). In this experiment, the group was asked to come to a unanimous consensus on a choice between two everyday items (Figure 1B). The experiment consisted of 40 blocks of trials (Figure 1C). Each block was associated with a unique pair of items, and on each trial in a block every participant in the group made a choice between those two items (Figure 1B). If the group reached a unanimous consensus on a trial, they obtained the item and moved to the next block; otherwise they moved to the next trial in the same block and made another choice between the same pair of items (Figures 1B and 1C). If they did not reach consensus before the end of the block, they did not get anything and moved to the next block (e.g., block 3 in Figure 1C). As the maximum number of trials in each block was determined randomly, participants could not exploit the information about the number of trials left in the block. Notably, we measured participants' preference for each item beforehand by using a Becker-DeGroot-Marschack (BDM) auction (Becker et al., 1964; Chib et al., 2009) (see Figures S1A and S1B and Experimental Procedures for details).

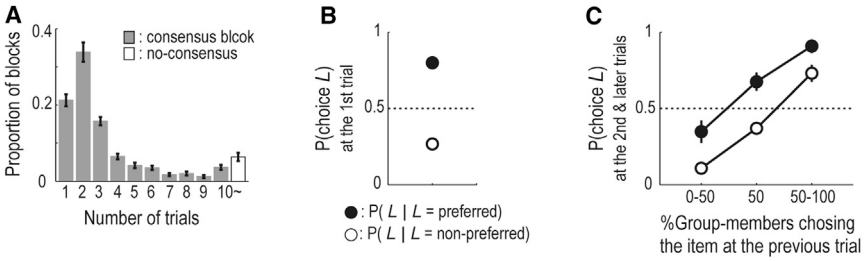


**Figure 2. Behavioral Results**

(A) Histogram of the number of trials in each block (mean ± SEM across participants; n = 20). Consensus block, participants reached a consensus; no-consensus block, participants failed to reach a consensus.

(B) Participants' choices on the first trial in each block. Probabilities of choosing the item presented at the left side of the screen are shown (mean ± SEM across participants). A filled circle denotes the probability when the left item was preferred by the participant, $P(\text{choice} = L \mid L = \text{preferred})$, and an open circle represents the probability when the item was not preferred, $P(\text{choice} = L \mid L = \text{non-preferred})$. The circles overlap the error bars.

(C) Participants' choices on the second and later trials in each block. Probabilities of choosing the left item are plotted as a function of percentages of the group members who chose the item on the previous trial (mean ± SEM across participants). As in (B), filled and open circles denote the probabilities when the left item was preferred and when the item was not preferred, respectively.

In addition to the main experiment, we conducted a control experiment in which participants were asked to build a consensus with a computer algorithm (see Experimental Procedures), to examine whether or not the behavior and neural activity observed in the main experiment were dedicated to social inferences about other people (Carter et al., 2012; Gallagher and Frith, 2003; Mitchell, 2008; Saxe and Kanwisher, 2003). The computer algorithm was designed to mimic human participants' actual tendency to follow the majority and to simulate the reaction times in the main experiment (Figure S1C). Here, it is worth noting that different sets of participants took part in the control and the main experiment. We employed this between-participants design, instead of the within-participants designs used in previous studies on this issue (Delgado et al., 2005; Gallagher et al., 2002; Sanfey et al., 2003), so as to prevent cross-contamination of task set between the two experiments (e.g., attributing agency to the computer algorithms).

Moreover, to explore possible effects of group size, in half of the 40 blocks all six of the participants (one scanned and five not scanned, six-person block) were engaged; while four participants (one scanned and three not scanned selected randomly from the five, four-person block) were involved in the other half of the blocks (Figure 1C). However, as no significant effect of group size was found in the main analyses, we pool together results from the two group sizes, unless specifically mentioned otherwise.

In the main experiment, the behaviors of the scanned (n = 20) and the not-scanned participants (n = 100) were highly consistent with each other, and we therefore report only the scanned participants' data in the main text (see Figure S2 for the not-scanned participants' data).

## Group-Level Overall Behavior in the Main Experiment

Participants quickly reached a consensus in most of the blocks. The success rate of the consensus was 0.94 ± 0.01 (mean ± SEM), and the consensus formation required fewer than five trials in 77.37% ± 2.21% of the 40 blocks (Figure 2A). On the other hand, in some blocks, they had a hard time building a consensus. The number of trials in a block was equal to or greater than ten in 10.63% ± 1.45% of the blocks (Figure 2A), and they failed to reach a consensus in 6.38% ± 1.06% of the blocks (Figure 2A, open bar).

## Individual Behavior in the Main Experiment: Effect of Participants' Own Preference and Group Members' Prior Choice

We hypothesized that participants' choices would be guided by their own preference for each item and the group members' prior choices. Consistent with this, on the first trial in each block, participants were more likely to choose their preferred item. The probability of choosing the item presented at the left side of the screen was greater when the item was preferred by the participant, compared with when the left item was nonpreferred ($p < 0.01$, two-tailed t test; Figure 2B). Furthermore, within each participant, the probability of choosing the preferred item was significantly greater than 0.5 ($p < 0.05$, two-tailed binomial test) in 19 out of the 20 participants.

On the second and later trials, participants took into account the group members' prior choice, as well as their own preference (Figure 2C). The probability of choosing the left item was modulated by the participant's preference and the percentage of group members' who had chosen that item on the previous trial (ANOVA, $p < 0.01$ for the preference effect, $p < 0.01$ for the group members' prior choice effect, $p = 0.26$ for their interaction; Figure 2C). The positive effect of the group members' prior choice indicates that participants tended to follow the majority as well as to choose their preferred item.

## Individual Behavior in the Main Experiment: Effect of Hidden Stickiness of the Items

We further hypothesized that participants tracked a hidden variable, the "stickiness" of the two items, which potentially reflects other participants' preference. In other words, participants' own choices would be modulated by how much the other participants tended to stick to their choice of one or other of the items as the round progressed.

We assume participants inferred the stickiness by a simple Bayesian learning algorithm (see Figure 3A for the graphical description of the inference, and Supplemental Experimental Procedures for details). In the formulation of the inference, the stickiness reflects the other group members' relative preference for the item (i.e., positive values denote that they prefer that item to the other item, and negative values indicate the opposite). Estimates of the stickiness are updated based on the belief that the others' choices on the current trial, $Y$, were generated by the group members' choices on the previous trial, $G$, and the hidden

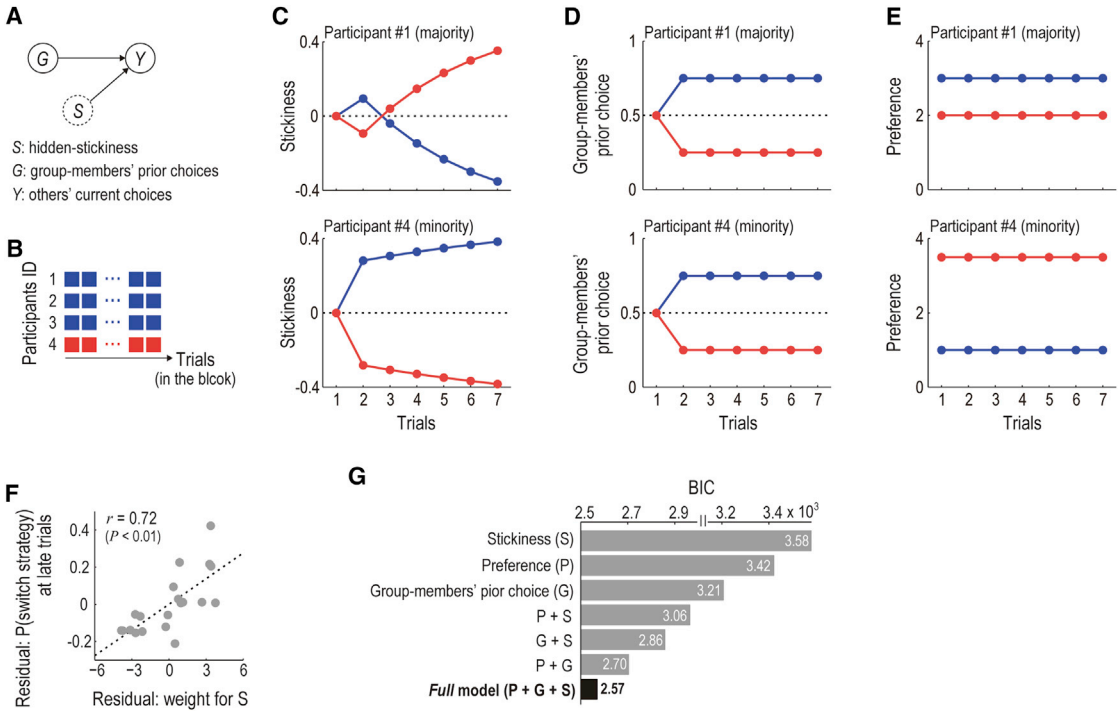


**Figure 3. Computational Model**

(A) Graphical description of the inference about the hidden stickiness of the items. The Bayesian learner infers the hidden stickiness, *S*, based on the belief that the others' current choices, *Y*, are generated by the stickiness, *S*, and the group members' prior choice, *G*. Dashed circle, a hidden variable; solid circles, observable variables.

(B) Example block. Three participants continue to choose the blue item, while the other one chooses the red.

(C) Estimated stickiness of each item in the example block (B). Top, from a viewpoint of participant #1; bottom, from a viewpoint of participant #4; same for (D) and (E).

(D) Group members' prior choice in the example block. Proportions of group members who chose each item on the previous trial are plotted.

(E) Participants' own preference for each item in the example block. Participant #1 prefers the blue item to the red, while participant #4 has the opposite preference.

(F) Across-participants correlation between the decision weight for the stickiness and the tendency to change their default behavior in later trials ($t \geq 4$). The decision weight was estimated using the best-fitting model in (G). The partial correlation coefficient controlling for the decision weights of the other variables was significantly positive ($r = 0.72$, $p < 0.01$).

(G) Computational models fit to participants' choices. Each bar denotes BIC of each model. BIC, Bayesian information criterion (smaller values indicate better fit).

stickiness of each item, *S*. This assumption about the participants' belief is reasonable, given that the stickiness reflected the others' preferences, and as shown in the previous section (Figure 2C) participants' choices were actually guided by their own preference and the group members' prior choice. That is, they updated their estimate of the hidden variable *S* from the observable variables *Y* and *G* when they got a new piece of information about *Y* at the outcome phase (Figure 1B).

For example, suppose that in a block, one minority participant continues to choose one item, say, red, and that the majority participants choose the other item, say, blue (Figure 3B). In this case, the majority participants estimate stickiness of the red item as high (Figure 3C, top), because the minority's choice of the red cannot be attributed to conforming to a majority. On the other hand, from a viewpoint of the minority participant, all of the others choose the blue item, and none of them choose the red. The stickiness of the blue is therefore judged as high (Figure 3C, bottom). Notably, at the outcome phase of the first trial, the estimated stickiness is updated always in favor of the item chosen by the majority (i.e., the blue item; see Figure 3C) whether the participant is in the majority side or in the minority side. This is because there is no information about the group members' prior choice on the first trial; and so only the others' current choice governs an update of the stickiness.

This learning algorithm has two important properties. First, a trial-by-trial signal of the estimated stickiness was not highly correlated with the other two decision variables: the participant's own preference (mean correlation coefficient $r = -0.11 \pm 0.04$) and the percentage of group members who chose each item on the previous trial (mean $r = 0.35 \pm 0.08$), making it possible to identify neural activity related to each of the three variables (see Figure S3A and the section of neural results for further analyses and discussion). We plot the time course of the three variables in the example block (Figures 3C–3E).

Second, the estimation of the stickiness guided participants' decisions on whether to change their behavioral strategy when they had a hard time reaching a consensus (i.e., later trials in a block). Again, consider the example in Figure 3B. If participants'

choices are positively modulated by the estimated stickiness, the majority participants would change their behavior from blue to red in later trials (Figure 3C, top), and the minority participant would also change his/her behavior from red to blue (Figure 3C, bottom). Conversely, in the presence of a negative modulation of stickiness, they would not change their behavior (Figure 3C). Thus, the behavioral effect of the estimated stickiness captures the participants' tendency to change their default behavior in later trials. Indeed, we found a significant correlation between the tendency to change behavior and the decision weight of our stickiness variable across participants (Figure 3F, partial correlation after controlling for the effect of the other two decision variables, $r' = 0.72$, $p < 0.01$, two-tailed; also see Figures S3B and S3C for distributions of each decision weight and the cross-correlations across participants). The relation remained significant when we assessed it based on a conventional correlation coefficient not controlling for the effect of the other variables ($r = 0.68$. $p < 0.01$, two-tailed).

## Individual Behavior in the Main Experiment: Computational Model Fits

To ascertain contributions of the estimated stickiness to participants' decision-making, we fit various computational models to the participants' actual choice data and compared their goodness of fits (see Supplemental Experimental Procedures for details). We first constructed a full model in which the decision value of each item was computed as a weighted sum of the three computational variables: the participants' own preference for the item, percentage of the group members who had chosen the item on the previous trial, and estimated stickiness of the item. We then considered alternative partial models that include only one or two of the three decision variables. In the behavioral model fitting procedure, a hierarchical modeling approach was employed to reduce the estimation noise in the parameter estimates (Daw, 2011) (Figure S4A; Supplemental Experimental Procedures). Furthermore, each model's goodness of fit was assessed by Bayesian information criterion (BIC), which penalizes additional free parameters.

The model comparison revealed that our full model provided the best fit to the participants' actual choices than the other alternative models (Figure 3G; comparison against the second-best model, Bayes factor > 150; $p < 0.01$, likelihood ratio test), suggesting that their decision-making was guided by their own preference, the group members' prior choice, and the estimated stickiness of the items. This conclusion did not change if we applied the same model-fitting analysis to the data for early and late blocks separately (Figure S4C). Furthermore, we analyzed the data in the four-person and the six-person blocks separately, and confirmed that the full model best fit both block types (Figure S4D). This result implies that the three decision variables guide the participants' behavior independent of group size.

We also tested several variations of these models. The variants include a model suggested by a theoretical study (Couzin et al., 2005) in which the behavioral weight of the group members' prior choice was modulated by trial-by-trial feedback. None of these alternative models outperformed the original full model (Figure S4C). Finally, for further confirmation, we fit the models to each participant's choice data individually (Figure S4B; c.f., hierarchical modeling approach) and compared the goodness of fits by using Bayesian model selection (Stephan et al., 2009). The result obtained was consistent with those based on the hierarchical modeling approach (Figure S4E). These complementary analyses together support the notion that participants' decision-making was modulated by their own preference, the group members' prior choice, and the estimated stickiness.

We also examined the possibilities that participants employed more complicated strategies and that their preferences for each item were altered during the experiment. The results of these additional analyses confirmed that these factors do not appear to be playing a major role in explaining participants' choice behavior (see additional behavioral analyses in Supplemental Experimental Procedures and Figures S1D–S1G).

## Individual Behavior in the Control Experiment

In the control experiment where each participant interacted with a computer algorithm, we found the same qualitative result. That is, the full model provided the best fit to the participants' choice data (Figure S4). This suggests that even in the nonsocial experiment, as well as in the main social experiment, participants' choices were guided by the three computational variables: their own preference, group members' (computer algorithms') prior choice, and the estimated stickiness of the items.

## Neural Signals Encoding Participants' Own Preference

We next analyzed the fMRI data to test for brain regions tracking the key computational variables identified in the behavioral analyses, by regressing these variables against the BOLD signal across the whole brain (see Experimental Procedures and Figure S3A). The regression analysis was performed by SPM8 without serial orthogonalization of parametric modulators.

Based on previous findings, we predicted that participants' preference for each item would correlate with activity in the ventromedial prefrontal cortex (vmPFC) independently of social or nonsocial contexts (Chib et al., 2009; Smith et al., 2010; Strait et al., 2014). We analyzed the data in the main and the control experiment together, and consistent with our hypothesis, we found that the BOLD signal in the vmPFC at the time of decision was significantly correlated with the participants' preference for the chosen item (Figure 4A, $p < 0.05$ small-volume corrected).

A closer examination of an independently identified ROI in the vmPFC (see Supplemental Experimental Procedures) revealed that the effect of preference on the neural activity was significant on the first trial in each block, but not in the second and later trials (Figure 4B), while behaviorally the preference guided the participants' choices also in the later trials (Figure 2C). One account for this result could be "repetition suppression," in that the neural response is attenuated by repetition of the same computation or the presentation of the same item within a block, which is often accompanied by performance improvements such as a decrease in reaction time (Grill-Spector et al., 2006). An alternative explanation is that participants had a lapse in concentration or were bored by making a decision between the same items repeatedly, which might result in increased reaction time. To test these two alternatives, we compared reaction times on the

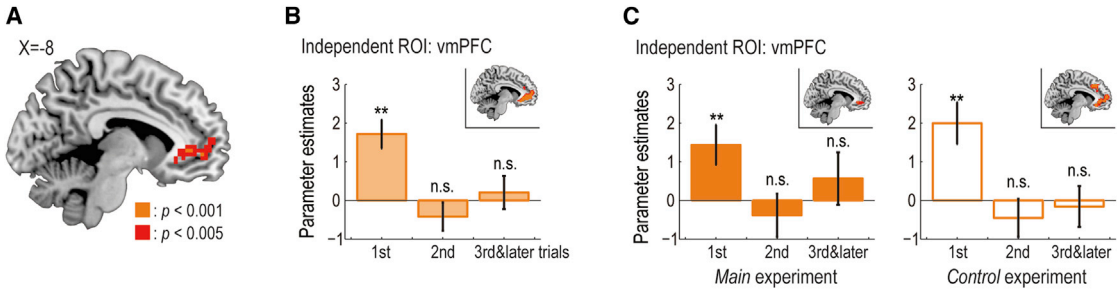


**Figure 4. Neural Correlates of Participants' Own Preference**

(A) Activity in the vmPFC significantly correlated with preference for the chosen item at the time of decision. The vmPFC activation map is thresholded at $p < 0.005$ uncorrected for display purpose.

(B) Effect sizes of the preference in the independently identified vmPFC ROI. The effect sizes are plotted separately for the first, the second, and the later trials in each block (mean ± SEM across participants; n = 40). **$p < 0.01$, and n.s., nonsignificant as $p > 0.05$. Inset, activated voxels in response to the preference on the first trial ($p < 0.005$ uncorrected). vmPFC, ventromedial prefrontal cortex.

(C) Effect sizes of the preference in the vmPFC ROI for the main and the control experiment. Left, the main experiment (n = 20); right, the control experiment (n = 20). The format is the same as in (B).

first trial with those on the second and later trials in each block. The comparison showed a significant decrease in log reaction time ($p < 0.01$, two-tailed t test), consistent with the repetition suppression account.

vmPFC activity exhibited the same pattern when we analyzed the data for the main and the control experiment separately (Figure 4C). Indeed, a two-way ANOVA on the vmPFC activity revealed no significant effect of experimental type (main versus control, $p = 0.85$), group size (four versus six-person, $p = 0.89$) or their interaction ($p = 0.40$). Moreover, no significant difference was found in activity in this region between the two experiments in a direct statistical comparison (even at $p > 0.005$ uncorrected).

## Neural Signals Encoding Group Members' Prior Choice

In the main social experiment, the second key computational variable, group members' prior choice, was correlated with activity in the right posterior superior temporal sulcus (pSTS) and the adjacent area, temporoparietal junction (TPJ). We found at the time of decision a significant correlation between the BOLD signal in the right pSTS/TPJ and the percentage of group members who had previously selected the item that was chosen by the participant on the current trial (Figure 5A, $p < 0.05$ whole-brain corrected at cluster level; see Table S1 for other activated areas, including the central sulcus). Furthermore, as a robustness check, we confirmed that the right pSTS/TPJ activity remained significant ($p < 0.05$ corrected) when the relevant regressor variable was orthogonalized against the other two key computational variables (i.e., the participant's own preference and the estimated stickiness), so that those other variables subsumed all of the common variance. The right pSTS/TPJ activity also remained significant ($p < 0.05$ corrected) even when we included the following decision-irrelevant variables into our regression analysis as regressors of no interest: overall motivation (sum of the preference values for the two items), cognitive load (log reaction time), and motor response (1 for choosing the left item, 0 for the right).

On the other hand, in the control nonsocial experiment, we did not find the right pSTS/TPJ activity to be significantly correlated with the group members' prior choice at our whole-brain corrected significance threshold (Figure 5B; see Table S1 for a list of activated areas, including the left central sulcus). Furthermore, a direct comparison of the whole-brain activation maps between the two experiments revealed a significantly greater effect of the group members' prior choice on the pSTS/TPJ activity in the main experiment (Figure 5C, $p < 0.05$, small-volume corrected). This differential effect was also shown in an independent ROI analysis (Figure 5D): the effect was significantly positive only in the main experiment ($p < 0.01$, one-tailed), and the effect was significantly greater in the main experiment compared with the control experiment ($p < 0.05$, two-tailed). Consistent with this, using a two-way ANOVA on the pSTS/TPJ activity, we found a significant main effect of experimental type (main versus control, $p = 0.03$); but no effect of group size (four- versus six-person, $p = 0.38$) or their interaction ($p = 0.40$). These results together demonstrate that pSTS/TPJ encoded group members' prior choice selectively only in the main social experiment.

Importantly, such differential activity cannot be attributed to a difference in the behavioral effect of the group members' prior choice or to the characteristics of the participants. There was no significant difference between the two experiments in the behavioral weight of the group members' prior choice estimated by the model fitting (Figure S3D, $p > 0.4$, two-tailed). Also, participants in the control experiment matched those who were scanned in the main experiment in many aspects, such as age, sex, education level, income level, IQ, hunger-rating score, and self-reported sociality scales (see Supplemental Experimental Procedures).

It is also worth noting that activity in the left-central sulcus was found to be significant in both the main and the control experiments (Table S1). This activity, however, vanished when we included decision-irrelevant regressors described above in the regression analyses. Combining this finding with prior evidence about the central sulcus implicated in primary motor/sensor processing, we speculate the activation reflected a basic sensorimotor process, not directly related to decision-making, such as pressing a key in the keypad or perceiving information about red dots (Figure 1B; Experimental Procedures).

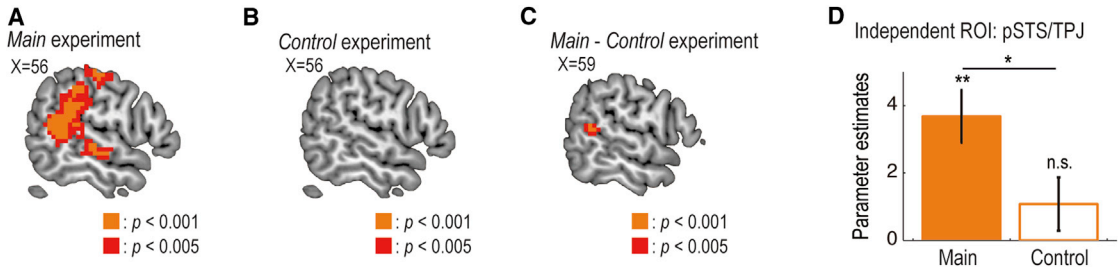


**Figure 5. Neural Correlates of the Group Members' Prior Choice**

(A) Main experiment: activity in the right pSTS/TPJ at the time of decision significantly correlated with the percentage of group members who had previously selected the item that was chosen by the participant on the current trial. The map is thresholded at $p < 0.005$ uncorrected for display purpose.

(B) Control experiment: no activity in the right pSTS/TPJ significantly correlated.

(C) Main versus control experiments: activity in the right pSTS/TPJ significantly better correlated in the main experiment.

(D) Effect sizes of the group members' prior choice in the independently identified right pSTS/TPJ ROI for the main and the control experiment (mean ± SEM across participants; n = 20). *$p < 0.05$, **$p < 0.01$, and n.s., nonsignificant as $p > 0.05$. pSTS, posterior superior temporal sulcus; TPJ, temporoparietal junction.

### Neural Signals Encoding Estimated Stickiness

The third variable, estimated stickiness of the chosen item, was significantly correlated with the BOLD signal in the bilateral intraparietal sulcus (IPS) at the time of decision in the main experiment (Figure 6A, $p < 0.05$ whole-brain corrected at cluster level; see Table S2 for other activated areas). The bilateral IPS activations survived ($p < 0.05$ corrected), even when the regressor value of the stickiness was orthogonalized to the other two variables, and even when decision-irrelevant potential confounds (see above) were included in the regression analysis as regressors of no interest.

In the control experiment, the whole-brain analysis revealed the BOLD signal in the right IPS to be significantly correlated with the estimated stickiness (Figure 6B, $p < 0.05$ whole-brain corrected at cluster level). Although we did not detect a significant effect in the left IPS under our statistical threshold for the whole-brain analysis, an independent ROI analysis showed a significant effect also in the left IPS (Figure 6C, left; $p < 0.05$, one-tailed). Furthermore, the ROI analysis demonstrated no significant difference between the two experiments in the effect size of the estimated stickiness in either the right (Figure 6C, right; $p > 0.3$, two-tailed) or the left IPS (Figure 6C, left; $p > 0.4$, two-tailed). We also confirmed, by a two-way ANOVA, that there was no significant effect of experimental type (main versus control, $p = 0.46$), group size (four- versus six-person, $p = 0.08$) or their interaction ($p = 0.53$) on the bilateral IPS activity. Consistent with this, in the whole-brain direct comparison between the two experiments, we did not find any significantly differential activities in the right or left IPS ($p > 0.005$, uncorrected). Taken together, neural activity in bilateral IPS was modulated by the estimated stickiness of the chosen item in both the main and the control experiment, suggesting that the IPS tracked the computational variable irrespective of social or nonsocial contexts.

### Neural Integration of the Decision Variables

Computationally, the three key variables need to be integrated in order to enable an overall decision about whether or not to choose a given item. We tested for brain regions implicated in the integration process during the main social experiment. To this end, we reasoned that if a region is engaged in the integration, the region must (1) encode the integrated choice probability assigned by the computational model to the participant's chosen item and (2) have functional connectivity with regions tracking each of the individual key decision variables (i.e., vmPFC, right pSTS/TPJ and bilateral IPS) at the time of decision.

When including the modeled choice probability, orthogonalized to the other three variables, into the fMRI regression analysis (see Supplemental Experimental Procedures), we found that a region of rostral anterior cingulate cortex (rACC), as well as a region of dorsal anterior cingulate cortex (dACC) extending into the adjacent presupplementary motor areas, satisfies the first criterion. That is, BOLD signal in the rACC and the dACC significantly correlated with the modeled choice probability (Figure 7A, $p < 0.05$, whole-brain corrected at cluster level). Next, to test for the second criterion, we conducted a connectivity analysis, psychophysiological interaction (PPI). The PPI analysis examined whether each of the three seed regions, the vmPFC, the right pSTS/TPJ, and the bilateral IPS signaling the three variables, respectively, had increased connectivity at the time of decision with the two regions encoding the choice probability (see Supplemental Experimental Procedures). Results of the analysis showed a significant increase in the functional connectivity between the three seed regions and the dACC at the time of decision (Figure 7B). On the other hand, we did not find significant modulation in the connectivity between rACC and the right TPJ or the bilateral IPS (Figure 7C).

These results together indicate that only the dACC satisfies both of the two criteria, supporting the notion that the three key computational variables involved in consensus decision-making are integrated in dACC.

## DISCUSSION

This study provides insight into the computational and neural mechanisms underlying group consensus formation. The present findings go beyond results from other tasks in social neuroscience that have hitherto focused on dyadic interactions, and have hence not been designed to address the neural or computational mechanisms underlying decision-making in groups (Behrens et al., 2009; Fehr and Camerer, 2007; Lee, 2008).

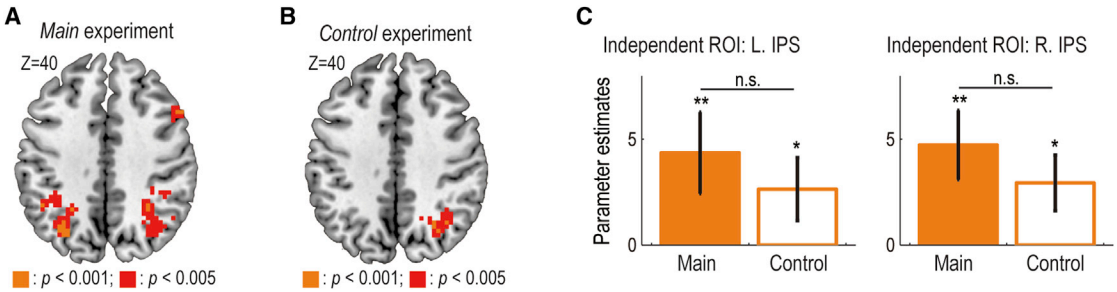


**Figure 6. Neural Correlates of the Estimated Stickiness**

(A) Main experiment: activity in the bilateral IPS significantly correlated with the estimated stickiness of the chosen item at the time of decision. The map is thresholded at $p < 0.005$ uncorrected for display purpose.

(B) Control experiment: activity in the right IPS significantly correlated with the estimated stickiness of the chosen item.

(C) Effect sizes of the stickiness in the independently identified IPS ROIs for the main and the control experiment. R. IPS, right intraparietal sulcus; L. IPS, left intraparietal sulcus. The format is the same as in Figure 5D.

Using model-based fMRI, we elucidated a role for several computational variables in human consensus decision-making, as well as determining how those variables are encoded at the neural level. Participants' choices were guided by their own preferences, the group members' prior choices, and the estimated stickiness of the items. These variables were each encoded in distinct brain structures, with vmPFC representing the participant's own preference, pSTS/TPJ tracking the group members' prior choice, and IPS tracking the stickiness. Furthermore, functional connectivity analysis combined with additional model-based fMRI analysis revealed that these computational signals were integrated in dACC, demonstrating not only what computations were implemented in individual brain regions but also how those computations were combined to drive consensus decision-making.

## Stimulus Valuation Signals in the vmPFC

As expected, participants were more likely to choose their preferred item in our task. Further, an individual's preference for each item was represented in the vmPFC irrespective of social or nonsocial contex, consistent with prior evidence implicating vmPFC in the valuation of many types of goods at the time of decision-making (Chib et al., 2009; Levy and Glimcher, 2011; Tom et al., 2007). Here we show that valuation signals in the vmPFC are present even during complex group decision-making.

We further found that value signals in the vmPFC were attenuated by repeated choices between the same items, in a manner consistent with repetition suppression (Grill-Spector et al., 2006). Given ambiguity in the precise physiological mechanism underlying repetition suppression, we cannot completely exclude other accounts for this effect, such as the possibility that activity decreases in vmPFC relate to a transition to a more habitual form of behavioral control (Daw et al., 2005).

## Computations Pertaining to an Inference about Group Behavior in the pSTS/TPJ

Participants took into account the choice tendencies of the group participants when making their own choices (Figure 2C): they were likely to choose a particular item when the majority of the group members had chosen the item on the previous trial. A key computational variable underpinning this behavior is a representation of the percentage of the group members who had previously selected the item. This variable was found to be encoded in the right pSTS/TPJ (Figure 5A), areas previously implicated in mentalizing (Frith and Frith, 2003; Gallagher and Frith, 2003; Saxe, 2010). Recent studies using formal mathematical models (Behrens et al., 2009; Dunne and O'Doherty, 2013) have demonstrated that pSTS/TPJ encodes learning signals for the prediction of other people's behavior, such as prediction error about the influence of one's own action on the opponent's next move (Hampton et al., 2008), others' intentions (Behrens et al., 2008; Suzuki et al., 2012), and others' expertise (Boorman et al., 2013). The present finding that pSTS/TPJ tracked group members' prior choice at the time of decision suggests this region plays a pivotal computational role not only in learning and updating but also in encoding information necessary for guiding choices in a social context.

Is the pSTS/TPJ specifically recruited for social cognition? The issue of domain specificity of this region has spurred heated debates in social neuroscience (Mitchell, 2008; Saxe, 2010), with some studies reporting evidence for social specificity (Coricelli and Nagel, 2009; Rilling et al., 2004; Carter et al., 2012; Saxe and Kanwisher, 2003; Saxe, 2010), while others have reported evidence for domain generality (Mitchell, 2008). In the current study, consistent with the social-specificity hypothesis, we found that the pSTS/TPJ selectively represented group members' prior choice in the main social experiment, but not in the control nonsocial experiment, even though this control experiment was matched in every other way to the main experiment except for the social component. Different sets of participants took part in the two experiments, and therefore crosscontamination of task set was unlikely to occur. It is thus likely that differential activity in the pSTS/TPJ between the tasks emerged naturally because of how the tasks were framed. We can thus conclude that the right pSTS/TPJ does indeed have socially specific contributions, at least with regard to the computations required for consensus decision-making.

There is a large body of cross-species work from insects to primates showing that an individual's probability of choosing a

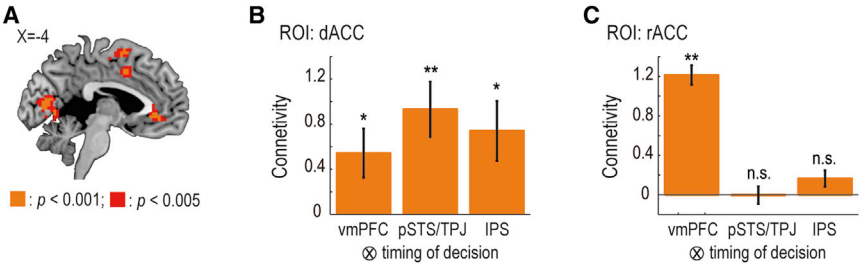


**Figure 7. Neural Correlates of the Integrated Signal in the Main Experiment**

(A) Activity in the dACC and the rACC at the time of decision significantly correlated with the choice probability assigned by the computational model to the participant's choice. The map is thresholded at $p < 0.005$ uncorrected for display purpose. dACC, dorsal anterior cingulate cortex; rACC, rostral ACC.

(B) Functional connectivity between the dACC and the other regions at the time of decision. Effect sizes of the PPI regressors in the dACC ROI are plotted (mean ± SEM across participants; n = 20). $^{**}p < 0.01$, and $^{*}p < 0.05$. vmPFC, ventromedial prefrontal cortex; pSTS/TPJ, right posterior superior temporal sulcus and temporoparietal junction; IPS, bilateral intraparietal sulcus. PPI, psychophysiological interaction.

(C) Functional connectivity between the rACC and the other regions at the time of decision. The format is the same as in (B). n.s., nonsignificant as $p > 0.05$.

particular option increases as a function of the number of conspecifics already choosing the option (Halloy et al., 2007; Sumpter, 2010;Ward et al., 2011; Sueur et al., 2010). Despite behavioral concordance across species, to our knowledge little is known about the neural mechanisms underlying group consensus formation in animals. An interesting avenue for future research would be to examine the degree of homology of neural encoding for group members' prior choice in the brains of humans and other social animals such as nonhuman primates.

## Signals Tracking a Hidden Structure of the Environment in the IPS

Participants tracked the stickiness of the presented items during consensus formation, suggesting that they did not simply respond to their own preference or group members' prior choice, but also tracked and utilized the hidden structure of the environment. The stickiness indicates how much each item was stuck to by the other group members, which potentially reflects the others' preference for the item. Computationally, the stickiness was estimated by a Bayesian learning algorithm that took into account the degree to which others conform to the majority's choice (Figure 3A). The estimated stickiness guided participants' decisions on whether to change their behavioral strategy when they had a hard time reaching a consensus (Figures 3B, 3C, and 3F). One interesting question for future theoretical studies is if and how the inference about the hidden structure of the environment facilitates or suppresses the consensus formation.

The estimated stickiness for the chosen item was encoded in the bilateral IPS and the adjacent inferior parietal lobule (IPL). Because the IPS/IPL activation was present both in the main social and the control nonsocial experiment (in contrast to the social-specific pSTS/TPJ activity), our findings suggest that neural computations in the IPS/IPL are domain general. Note that it is unlikely that the domain-general activation pattern results from participants in the control experiment assuming that they were playing against a human-like agent. Different sets of participants took part in the main and the control experiments (so as to prevent crosscontamination of task set between the two experiments), and in the instruction for the control experiment we did not use any suggestion of human-likeness in the computer algorithms; e.g., a part of the instruction was "You will get the item if all the red dots are located below the image of the item you choose" (see Supplemental Experimental Procedures). Nevertheless, the bilateral IPS/IPL was recruited in both the main and the control experiments.

The IPS/IPL has previously been implicated in evidence accumulation of both sensory (Gold and Shadlen, 2007) and value information (Sugrue et al., 2005) in monkeys and humans (Shadlen and Newsome, 2001; Platt and Glimcher, 1999; Sugrue et al., 2004; Hare et al., 2011; Heekeren et al., 2004). Recent studies have also implicated this region in learning about the abstract structure of the environment, in a manner not necessarily related to sensory or value information directly, such as in updating state transitions (Gläscher et al., 2010) or when encoding the probability of events (d'Acremont et al., 2013). Taken together, these findings and ours suggest that the bilateral IPS/IPL could be involved in facilitating inference about environmental structure in a domain-general manner.

## Integration of the Three Computational Signals

Finally, we demonstrated that the three key computational variables we identified are integrated in dACC to compute the choice probability of each item. While regions of dACC and rACC both tracked the choice probability, the dACC, but not the rACC, had connectivity with other regions encoding each of the three key computational variables.

The present results are broadly consistent with studies on simple decision-making, suggesting that the valuation of goals and stimuli in vmPFC provides input for the computation of action value in dorsomedial prefrontal cortex including dACC before finally being transformed to a motor command in motor cortex (Hare et al., 2011; Rangel and Clithero, 2013). This view is consistent with the strong anatomical connections between dmPFC and motor-related areas (Beckmann et al., 2009). Other studies on foraging or decision-making requiring cost-benefit consideration have reported results consistent with value integration at the action-value level in dACC (Kolling et al., 2012; Wallis and Rushworth, 2013).

In a social context, several studies have suggested that vmPFC plays a pivotal role in value integration by employing simple experimental tasks, which do not involve actual interactions with other people, such as learning from social information, decision-making on behalf of others, valuation of social stimuli, or charitable giving (Behrens et al., 2008; Hare et al., 2010; Janowski et al., 2013; Smith et al., 2014). However, no study to date has addressed how value integration occurs for decision-making in

real social strategic interactions (c.f. van den Bos et al. [2013] for integration of multiplex learning signals). Our finding provides evidence for value integration during social interactions, and supports the notion that multiple types of information are integrated at the level of action values in dACC, thereby providing mechanistic insights into the neural computations underlying social decision-making.

To conclude, in this study we provide a theoretical account of human consensus decision-making by identifying a key role for three distinct computational processes. This framework is further validated empirically by the finding that these variables are separately encoded in three distinct brain systems. More broadly, our findings provide direct evidence that multiple types of inference about oneself, others, and the environments are processed in parallel and integrated in our brain to guide decision-making in a social context. Moving beyond the dyadic interactions that have already been extensively studied in social neuroscience (Behrens et al., 2009; Fehr and Camerer, 2007; Lee, 2008), the present study suggests the importance of examining decision-making in larger group contexts in order to gain broader insight into the nature of human social intelligence (Krause et al., 2010).

## EXPERIMENTAL PROCEDURES

We provide a comprehensive description of the methods in the Supplemental Experimental Procedures.

### Participants

In our main experiment, 120 healthy, normal volunteers participated. Twenty out of the 120 participants were scanned with fMRI while they performed an experimental task. The remaining 100 participants were engaged in the same task outside the MRI scanner. A control experiment involved scanning 20 additional volunteers with fMRI who did not participate in the main experiment. The study was approved by the Institutional Review Board of the California Institute of Technology.

### Experimental Tasks

Participants performed three tasks: prescanning BDM auction task, consensus decision-making task, and postscanning BDM auction task.

#### *Pre- and Postscanning BDM Auction Task*

We measured participants' preference for each of the 40 items by using a BDM auction (Becker et al., 1964).

#### *Consensus Decision-Making Task*

In the main experiment, each participant tried to build a consensus with other participants on a choice between two items (Figure 1A). The task consisted of 40 blocks of trials (Figures 1B and 1C): 20 six-person and 20 four-person blocks.

In each block of trials, participants simultaneously chose between two items repeatedly until they reached a consensus, i.e., choosing the same item (Figures 1B and 1C). If they reached a consensus on a trial, they got the item and moved to the next block; otherwise, they moved to the next trial in the same block and made another choice between the same pair of the items. If they did not reach consensus before the end of the block, they did not get anything and moved to the next block (e.g., block 3 in Figure 1C). In the next block, participants made choices between a different pair of the items repeatedly, again, until they reached a consensus. Pairs of items were pseudorandomly assigned so that the same pair was never presented again. Importantly, the maximum number of trials in each block was not instructed to participants, and in actuality was determined stochastically.

At the beginning of each trial, each participant was asked to make a choice between the pair of items by pressing a button with their right hand with no time constraint (decision phase; Figure 1B). The chosen item was immediately highlighted by a gray frame, initiating the interstimulus interval (ISI) phase. After a waiting time for the other group members' decisions and a jittered interval (1.5–4.5 s), the others' choices were revealed to the participant via placement of red dots under the chosen items (outcome phase, 2 s). Notably, participants were not able to identify each of the other group members; they were informed only about the distribution of the red dots (i.e., the number of participants choosing each of the two items). If all the dots were located below the image of the item the participant had chosen, i.e., consensus, the participant was informed that she/he obtained the item (instruction phase, 3 s) and moved to the next block after the jittered ITI (2–6 s). Otherwise, decision phase on the next trial in the same block was initiated following the ITI.

The control experiment was almost the same as the main experiment, except that each participant tried to build a consensus with a computer algorithm instead of with other human participants. The computer algorithm to determine the location of each red dot was designed to mimic the not-scanned participants' actual choice behavior in the main experiment, in terms both of the tendency to follow the majority's choice and reaction times (Figure S1C).

### Computational Models

To determine the key computational variables involved in consensus decision-making, we constructed a family of computational models and fit those models to the participants' actual choice behaviors.

### fMRI Data Analysis

We used SPM8 for image processing and statistical analysis. A separate general linear model (GLM) was defined for each participant. The GLM contained parametric regressors representing the three key computational variables at the trial onset (Figure 1B): the participant's preference for the chosen item, the percentage of group members who had previously selected the item that was chosen by the participant on the current trial, and the estimated stickiness of the chosen item.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, two tables, and Supplemental Experimental Procedures and can be found with this article at http://dx.doi.org/10.1016/j.neuron.2015.03.019.

## ACKNOWLEDGMENTS

This work was supported by the Grant-in-Aid for JSPS Fellows 232648 (S.S.), the JSPS Postdoctoral Fellowship for Research Abroad (S.S.), the Suntory Foundation Grant-in-Aid for Young Scientists (S.S.), the Nakajima Foundation (R.A.), and the NIMH Caltech Conte Center for the Neurobiology of Social Decision Making (J.P.O.). We thank Tim Armstrong and Lynn K. Paul for support with the participant recruitment, and Ralph E. Lee and Chris Crabbe for assistance with the experiments.

Received: December 3, 2014
Revised: February 11, 2015
Accepted: March 4, 2015
Published: April 9, 2015

## REFERENCES

Arrow, K.J. (1963). Social Choice and Individual Values. (New York: John Wiley & Sons, Inc.).

Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G., and Frith, C.D. (2010). Optimally interacting minds. Science *329*, 1081–1085.

Becker, G.M., DeGroot, M.H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. Behav. Sci. *9*, 226–232.

Beckmann, M., Johansen-Berg, H., and Rushworth, M.F.S. (2009). Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. J. Neurosci. *29*, 1175–1190.

Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., and Rushworth, M.F.S. (2008). Associative learning of social value. Nature *456*, 245–249.

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

Black, J.M. (1988). Preflight signalling in swans: a mechanism for group cohesion and flock formation. Ethology *79*, 143–157.

Boorman, E.D., O'Doherty, J.P., Adolphs, R., and Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. Neuron *80*, 1558–1571.

Carter, R.M., Bowling, D.L., Reeck, C., and Huettel, S.A. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. Science *337*, 109–111.

Chib, V.S., Rangel, A., Shimojo, S., and O'Doherty, J.P. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. J. Neurosci. *29*, 12315–12320.

Conradt, L., and Roper, T.J. (2005). Consensus decision making in animals. Trends Ecol. Evol. *20*, 449–456.

Coricelli, G., and Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. Proc. Natl. Acad. Sci. USA *106*, 9163–9168.

Couzin, I.D., Krause, J., Franks, N.R., and Levin, S.A. (2005). Effective leadership and decision-making in animal groups on the move. Nature *433*, 513–516.

d'Acremont, M., Fornari, E., and Bossaerts, P. (2013). Activity in inferior parietal and medial prefrontal cortex signals the accumulation of evidence in a probability learning task. PLoS Comput. Biol. *9*, e1002895.

Davis, J.H., Stasser, G., and Spitzer, C.E. (1976). Changes in group members' decision preferences during discussion: An illustration with mock juries. J. Pers. Soc. Psychol. *34*, 1177–1187.

Daw, N.D. (2011). Trial-by-trial data analysis using computational models. In Decision Making, Affect, and Learning Attention and Performance XXIII, M.R. Delgado, E.A. Phelps, and T.W. Robbins, eds. (Oxford: Oxford University Press), pp. 3–38.

Daw, N.D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. *8*, 1704–1711.

Dayan, P., and Daw, N.D. (2008). Decision theory, reinforcement learning, and the brain. Cogn. Affect. Behav. Neurosci. *8*, 429–453.

Delgado, M.R., Frank, R.H., and Phelps, E.A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. Nat. Neurosci. *8*, 1611–1618.

Devine, D.J., Clayton, L.D., Dunford, B.B., Seying, R., and Pryce, J. (2001). Jury decision making: 45 years of empirical research on deliberating groups. Psychol. Public Policy Law *7*, 622–727.

Dunne, S., and O'Doherty, J.P. (2013). Insights from the application of computational neuroimaging to social neuroscience. Curr. Opin. Neurobiol. *23*, 387–392.

Fehr, E., and Camerer, C.F. (2007). Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn. Sci. *11*, 419–427.

Frith, U., and Frith, C.D. (2003). Development and neurophysiology of mentalizing. Philos. Trans. R. Soc. Lond. B Biol. Sci. *358*, 459–473.

Gallagher, H.L., and Frith, C.D. (2003). Functional imaging of 'theory of mind'. Trends Cogn. Sci. *7*, 77–83.

Gallagher, H.L., Jack, A.I., Roepstorff, A., and Frith, C.D. (2002). Imaging the intentional stance in a competitive game. Neuroimage *16*, 814–821.

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

Gold, J.I., and Shadlen, M.N. (2007). The neural basis of decision making. Annu. Rev. Neurosci. *30*, 535–574.

Grill-Spector, K., Henson, R., and Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn. Sci. *10*, 14–23.

Halloy, J., Sempo, G., Caprari, G., Rivault, C., Asadpour, M., Tâche, F., Saïd, I., Durier, V., Canonge, S., Amé, J.M., et al. (2007). Social integration of robots into groups of cockroaches to control self-organized choices. Science *318*, 1155–1158.

Hampton, A.N., Bossaerts, P., and O'Doherty, J.P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc. Natl. Acad. Sci. USA *105*, 6741–6746.

Hare, T.A., Camerer, C.F., Knoepfle, D.T., and Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. J. Neurosci. *30*, 583–590.

Hare, T.A., Schultz, W., Camerer, C.F., O'Doherty, J.P., and Rangel, A. (2011). Transformation of stimulus value signals into motor commands during simple choice. Proc. Natl. Acad. Sci. USA *108*, 18120–18125.

Heekeren, H.R., Marrett, S., Bandettini, P.A., and Ungerleider, L.G. (2004). A general mechanism for perceptual decision-making in the human brain. Nature *431*, 859–862.

Ioannou, C.C., Guttal, V., and Couzin, I.D. (2012). Predatory fish select for coordinated collective motion in virtual prey. Science *337*, 1212–1215.

Janowski, V., Camerer, C., and Rangel, A. (2013). Empathic choice involves vmPFC value signals that are modulated by social processing implemented in IPL. Soc. Cogn. Affect. Neurosci. *8*, 201–208.

Kerr, N.L., and Tindale, R.S. (2004). Group performance and decision making. Annu. Rev. Psychol. *55*, 623–655.

Kolling, N., Behrens, T.E.J., Mars, R.B., and Rushworth, M.F.S. (2012). Neural mechanisms of foraging. Science *336*, 95–98.

Krause, J., and Ruxton, G.D. (2002). Living in Groups. (Oxford: Oxford University Press).

Krause, J., Ruxton, G.D., and Krause, S. (2010). Swarm intelligence in animals and humans. Trends Ecol. Evol. *25*, 28–34.

Lee, D. (2008). Game theory and neural basis of social decision making. Nat. Neurosci. *11*, 404–409.

Levy, D.J., and Glimcher, P.W. (2011). Comparing apples and oranges: using reward-specific and reward-general subjective value representation in the brain. J. Neurosci. *31*, 14693–14707.

McLean, I. (1994). Condorcet: Foundations of Social Choice and Political Theory (Northampton, MA: Edward Elgar Pub).

Mitchell, J.P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. Cereb. Cortex *18*, 262–271.

Platt, M.L., and Glimcher, P.W. (1999). Neural correlates of decision variables in parietal cortex. Nature *400*, 233–238.

Raafat, R.M., Chater, N., and Frith, C. (2009). Herding in humans. Trends Cogn. Sci. *13*, 420–428.

Rangel, A., and Clithero, J.A. (2013). The computation of stimulus values in simple choice. In Neuroeconomics, Second Edition, E. Fehr and P.W. Glimcher, eds., pp. 125–148.

Rilling, J.K., Sanfey, A.G., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. (2004). The neural correlates of theory of mind within interpersonal interactions. Neuroimage *22*, 1694–1703.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. (2003). The neural basis of economic decision-making in the Ultimatum Game. Science *300*, 1755–1758.

Saxe, R. (2010). The right temporo-parietal junction: a specific brain region for thinking about thoughts. In Handbook of Theory of Mind, A. Leslie and T. German, eds. (Oxford: Psychology Press).

Saxe, R., and Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". Neuroimage *19*, 1835–1842.

Seeley, T.D., and Visscher, P.K. (2004). Group decision making in nest-site selection by honey bees. Apidologie (Celle) *35*, 101–116.

Shadlen, M.N., and Newsome, W.T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J. Neurophysiol. *86*, 1916–1936.

Smith, D.V., Hayden, B.Y., Truong, T.-K., Song, A.W., Platt, M.L., and Huettel, S.A. (2010). Distinct value signals in anterior and posterior ventromedial prefrontal cortex. J. Neurosci. *30*, 2490–2495.

Smith, D.V., Clithero, J.A., Boltuck, S.E., and Huettel, S.A. (2014). Functional connectivity with ventromedial prefrontal cortex reflects subjective value for social rewards. Soc. Cogn. Affect. Neurosci. *9*, 2017–2025.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., and Friston, K.J. (2009). Bayesian model selection for group studies. Neuroimage *46*, 1004–1017.

Strait, C.E., Blanchard, T.C., and Hayden, B.Y. (2014). Reward value comparison via mutual inhibition in ventromedial prefrontal cortex. Neuron *82*, 1357–1366.

Sueur, C., Deneubourg, J.-L., and Petit, O. (2010). Sequence of quorums during collective decision making in macaques. Behav. Ecol. Sociobiol. *64*, 1875–1885.

Sugrue, L.P., Corrado, G.S., and Newsome, W.T. (2004). Matching behavior and the representation of value in the parietal cortex. Science *304*, 1782–1787.

Sugrue, L.P., Corrado, G.S., and Newsome, W.T. (2005). Choosing the greater of two goods: neural currencies for valuation and decision making. Nat. Rev. Neurosci. *6*, 363–375.

Sumpter, D.J.T. (2010). Collective Animal Behavior. (Princeton, NJ: Princeton University Press).

Sumpter, D.J.T., and Pratt, S.C. (2009). Quorum responses and consensus decision making. Philos. Trans. R. Soc. Lond. B Biol. Sci. *364*, 743–753.

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J.L., Ichinohe, N., Haruno, M., Cheng, K., and Nakahara, H. (2012). Learning to simulate others' decisions. Neuron *74*, 1125–1137.

Tom, S.M., Fox, C.R., Trepel, C., and Poldrack, R.A. (2007). The neural basis of loss aversion in decision-making under risk. Science *315*, 515–518.

van den Bos, W., Talwar, A., and McClure, S.M. (2013). Neural correlates of reinforcement learning and social preferences in competitive bidding. Journal of Neuroscience *33*, 2137–2146.

Wallis, J.D., and Rushworth, M.F. (2013). Integrating benefits and costs in decision making. In Neuroeconomics, Second Edition, E. Fehr and P.W. Glimcher, eds. (Waltham, MA: Academic Press).

Ward, A.J.W., Herbert-Read, J.E., Sumpter, D.J.T., and Krause, J. (2011). Fast and accurate decisions through collective vigilance in fish shoals. Proc. Natl. Acad. Sci. USA *108*, 2312–2315.

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

Neuron, Volume *86*

## Supplemental Information

# Neural Mechanisms Underlying Human Consensus Decision-Making

Shinsuke Suzuki, Ryo Adachi, Simon Dunne, Peter Bossaerts, and John P. O'Doherty

## Supplemental Figures

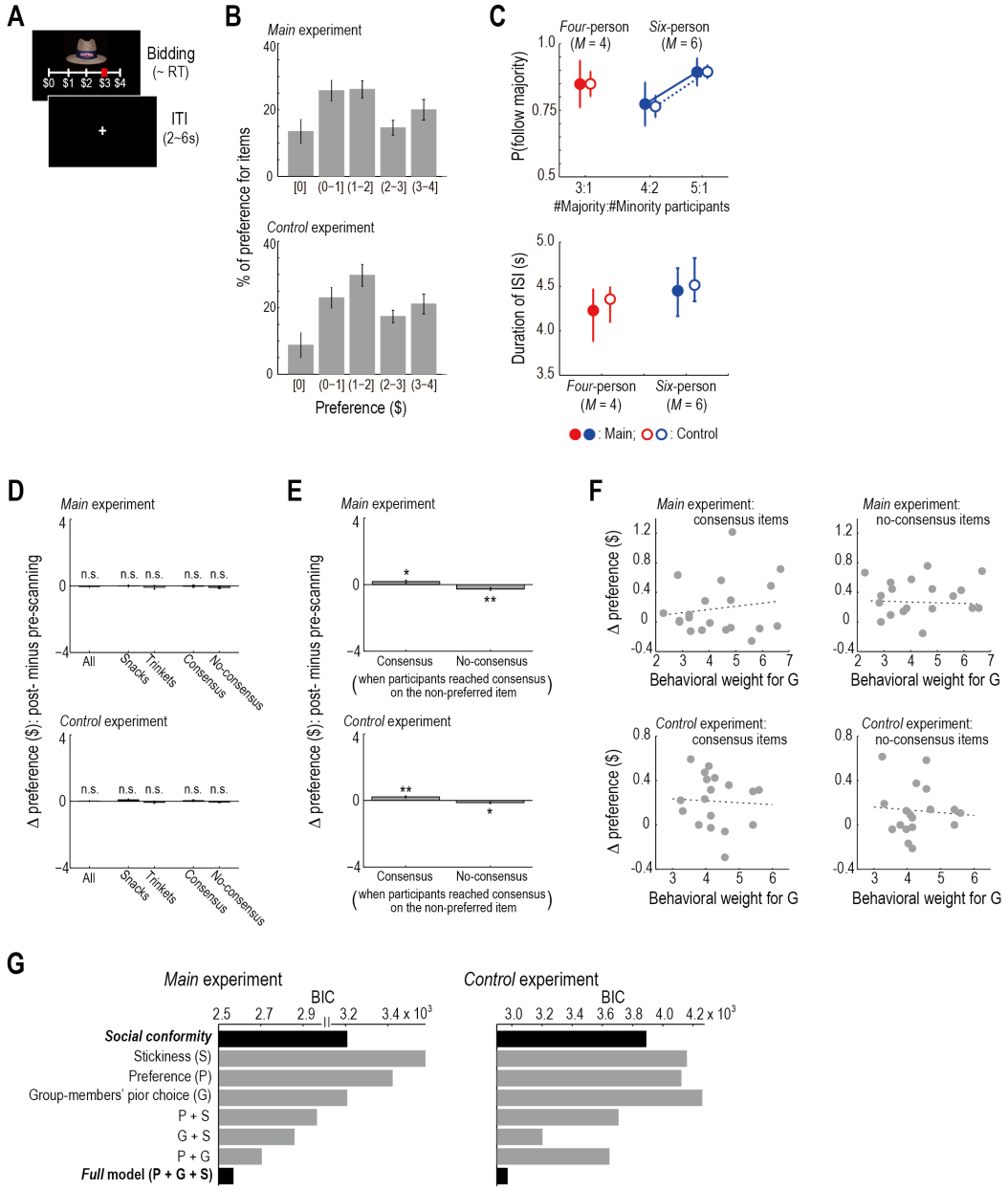


**Figure S1 – related to Figure 1: BDM auction task, *control* experiment, and stability of the preference and social conformity**

(A) Timeline of one trial in the BDM auction task. Participants were asked to make self-paced bids for 40 items. The 40 items were presented in random order. On each trial, they made a bid for the item by moving a red cursor. The initial position of the cursor was randomly chosen in each trial. RT, reaction time; ITI, inter-trial-interval.

(B) Distributions of the preference for each of the 40 items (mean ± SEM across participants;

*n* = 20; *Top*, the main experiment; *Bottom*, the control experiment).

(C) Choices of the human not-scanned participants in the main experiment and those of the computer algorithms in the control experiment. *Red*, *four*-person blocks; *Blue*, *six*-person blocks. *Filled* circles, human participants' choice behaviors in the main experiment; *Open* circles, computer algorithms' choice behaviors in the control experiment. *Top*, probabilities of following the majority's prior choice (mean ± SD across participants) as a function of the number of participants/algorithms in the majority side. "4:2", for example, means that the majority contains *four* while the minority has *two* participants. *Bottom*, durations of the ISI phase, i.e., the other participants/algorithms' reaction time and a jittered interval (mean ± 25%-75% quantiles). ISI, inter-stimulus-interval.

(D) Changes of the preference during the experiment. Differences in the preference between pre- and post-scanning BDM auction task are plotted for each category of items (mean ± SEM across participants; *Top*, the main experiment; Bottom, the control experiment). Consensus, items on which participants reached a consensus; No-consensus, items on which participants did not reach a consensus. n.s., non-significant as $p > 0.05$.

(E) Changes of the preference during the experiment when participants reached a consensus on the non-preferred item by compromise. Consensus, consensus/non-preferred items. No-consensus, no-consensus/preferred items paired with the consensus items. The format is the same for panel *D*.

(F) Across-participants correlation between the weight for the group members' prior choice and the social conformity effect (i.e., degree of the preference change shown in panel *E*). The correlation coefficients were not significant (consensus items in the main experiment: $r = 0.16$, $p = 0.51$, two-tailed; no-consensus items in the main experiment: $r = -0.04$, $p = 0.88$, two-tailed; consensus items in the control experiment: $r = -0.08$, $p = 0.75$, two-tailed; no-consensus items in the control experiment: $r = -0.18$, $p = 0.46$, two-tailed).

(G) Behavioral fits of computational models including the social conformity model. *Left*, the main experiment; *Right*, the control experiment. BIC, Bayesian information criterion (the smaller the value, the better the model fit). The format is the same for Figure 3G.

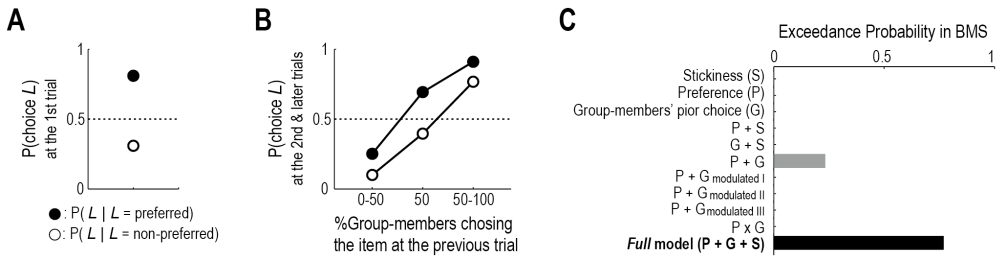


**Figure S2 – related to Figure 2: Behavioral results of the not-scanned participants in the *main* experiment**

(A) Participants' choices on the first trials in each block. The format is the same for Figure 2B.

(B) Participants' choices on the second and later trials in each block. The format is the same for Figure 2C.

(C) Computational models' fit to participants' choices. Each model was separately fitted to each individual participant's choice behavior (i.e., individual modeling approach in Figure S4B), and then each model's goodness-of-fit was compared by using Bayesian Model Selection (BMS) (Stephan et al., 2009). Each bar denotes exceedance probability in BMS (the probability that a model is more likely than the other models; larger values closer to *one* indicate better fit). The format is the same for Figure S4E.

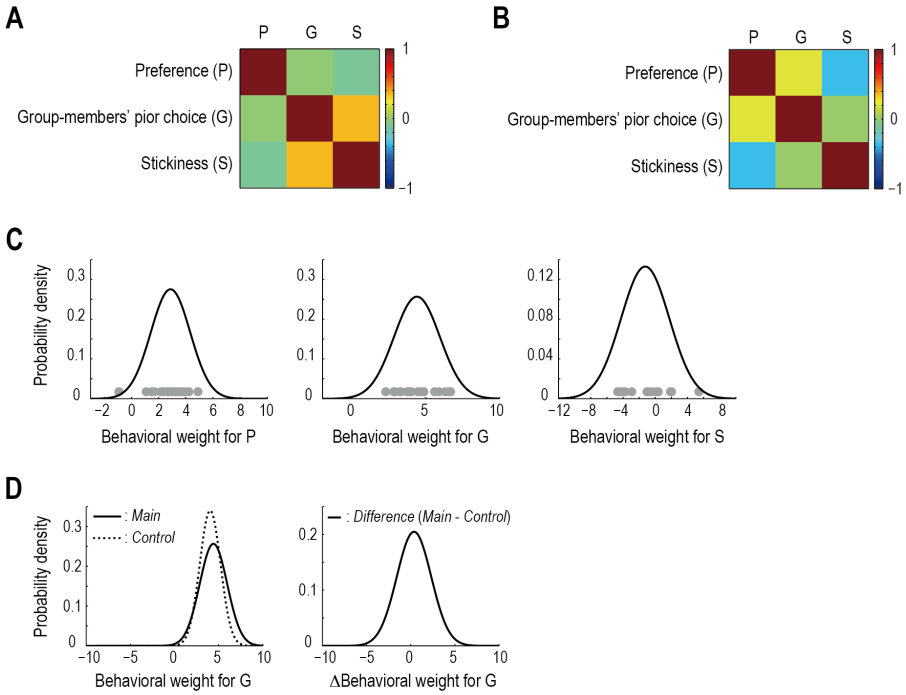


**Figure S3 – related to Figure 4: Correlation between key computational variables and the behavioral weights for the variables**

(A) Trial-by-trial cross correlation of the three computational variables (i.e., regressors of interest in our fMRI regression analysis). Preference (P), preference value of the chosen item; Group members' prior choice (G), percentage of group-members who had previously selected the item that was chosen by the participant on the current trial; Stickiness (S), estimated stickiness of the chosen item.

(B) Across-participants cross correlation of the behavioral weights for the individual key computational variables.

(C) Probability distributions of the individual behavioral weights estimated by the hierarchical modeling approach in Figure S4A. Each gray dot denotes each participant's behavioral weight for the corresponding variable.

(D) Main vs. control experiments: behavioral weight for the group-members' prior choice. *Left*, probability distributions of the behavioral weight (*solid* line, the main experiment; *dashed* line, the control experiment). *Right*, distribution of the difference in the weight between the main and the control experiment.

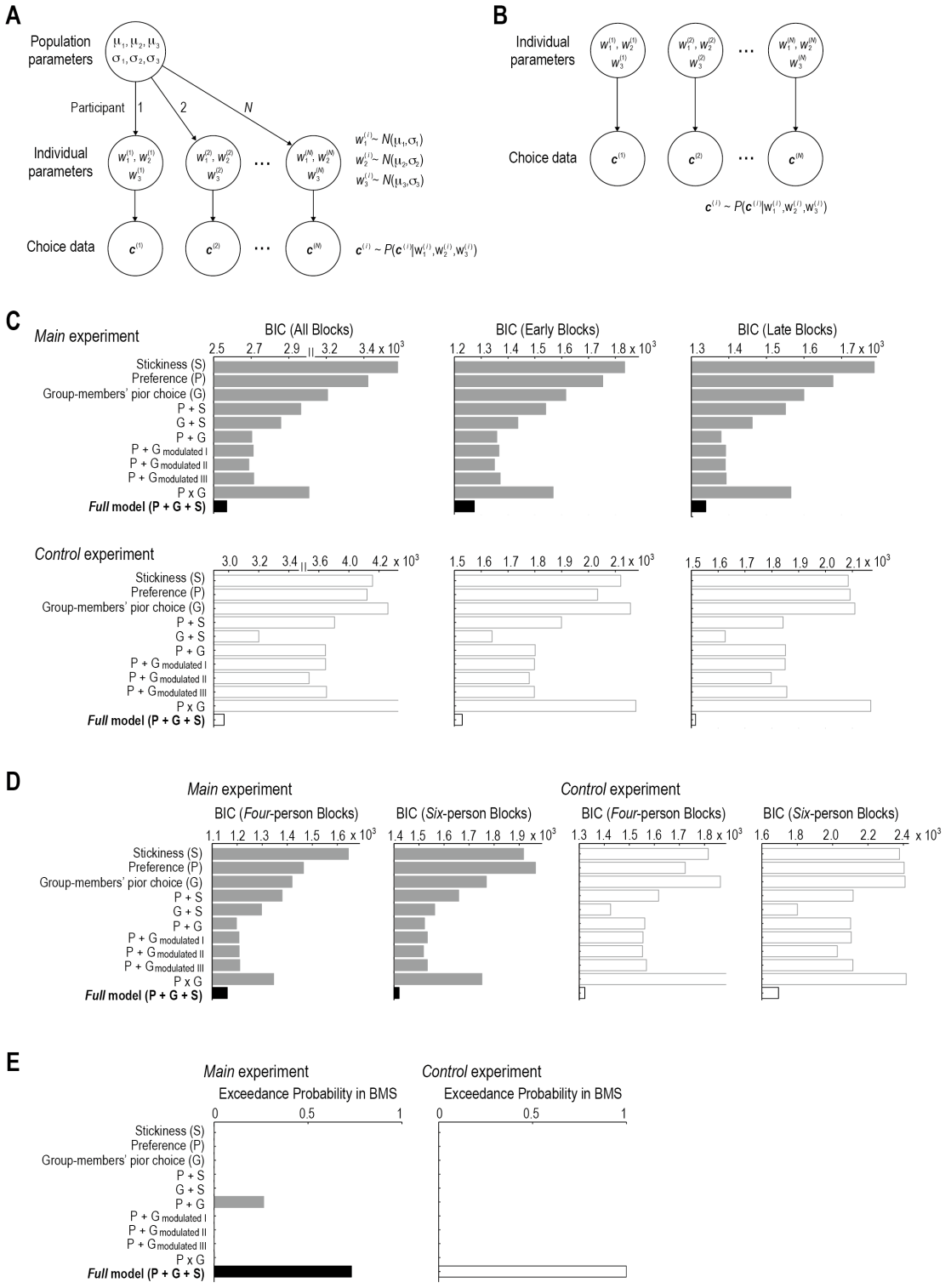


**Figure S4 – related to Figure 3: Computational models' fit**

(A) Graphical illustration of the *hierarchical modeling* approach to model fitting (Daw, 2011). Each participant's individual-level parameters, $w_1$, $w_2$ and $w_3$, are drawn from common population-level *normal* distributions.

(B) Illustration of the *individual modeling* approach to model fitting. Individual-level parameters are separately estimated for each participant.

(C) Computational models' fit based on the hierarchical modeling approach separately for *early* and *late* blocks. *Top*, the main experiment; *Bottom*, the control experiment. *Left*, all blocks; *Middle*, early blocks; *Right*, late blocks. BIC, Bayesian information criterion (smaller values indicate better fit). The format is the same for Figure 3G.

(D) Computational models' fit based on the hierarchical modeling approach separately for *four*-person and *six*-person blocks. *Left*, *four*-person blocks; *Right*, *six*-person blocks. The format is the same for Figure 3G.

(E) Computational models' fit based on the individual modeling approach. *Left*, the main experiment; *Right*, the control experiment. Each bar denotes the exceedance probability (the probability that a model is more likely than the other models; larger values closer to *one* indicate better fit) in Bayesian Model Selection (Stephan et al., 2009).

## Supplemental Tables

**Table S1. Areas exhibiting significant changes in BOLD associated with the group-members' prior choices**

| | Region | Hemi | BA | x | y | z | *t*-statistic | *p*-value | Voxels |
|---|---|---|---|---|---|---|---|---|---|
| *Main* experiment | **pSTS/TPJ** | R | 22/39/40 | 63 | -46 | 10 | 5.88 | 0.000 | 874 |
| | Central sulcus | R | 1/2/3/4 | 22 | -43 | 67 | 5.64 | 0.000 | 204 |
| | | L | 1/2/3/4 | -42 | -25 | 43 | 3.95 | 0.000 | 120 |
| | Premotor cortex | L | 6 | -45 | 2 | 7 | 4.34 | 0.000 | 193 |
| | | R | 6 | 54 | 8 | 1 | 3.72 | 0.001 | 84 |
| *Control* experiment | Lateral sulcus | L | 42 | -51 | -31 | 25 | 3.89 | 0.000 | 112 |
| | Central sulcus | L | 1/2/3/4 | -30 | -28 | 49 | 3.71 | 0.001 | 80 |

**Table S1 – related to Figure 5:** Activated clusters observed in the whole-brain analysis ($p < 0.05$, corrected at cluster level) for the main and the control experiment. The stereotaxic coordinates are in accordance with MNI space. *t*-statistics, uncorrected *p*-values at the peak of each locus are shown. In the far right column, the number of voxels in each cluster is shown. The region of interest discussed in the main text is shown in bold. pSTS, posterior superior temporal sulcus; TPJ, temporoparietal junction; Hemi, hemisphere; BA, Brodmann area.

**Table S2. Areas exhibiting significant changes in BOLD associated with the estimated-stickiness**

| | Region | Hemi | BA | x | y | z | *t*-statistic | *p*-value | Voxels |
|---|---|---|---|---|---|---|---|---|---|
| *Main* experiment | **IPS** | R | 40 | 30 | -52 | 34 | 4.68 | 0.000 | 183 |
| | | L | 40 | -30 | -67 | 40 | 4.16 | 0.000 | 157 |
| | Cerebellum | L | - | -42 | -70 | -29 | 4.91 | 0.000 | 82 |
| | dlPFC | R | 8 | 51 | 17 | 43 | 4.83 | 0.000 | 107 |
| | Fusiform gyrus | L | 19/37 | -21 | -61 | -8 | 4.28 | 0.000 | 151 |
| *Control* experiment | **IPS** | R | 40 | 27 | -64 | 40 | 3.98 | 0.000 | 107 |

**Table S2 – related to Figure 6:** Activated clusters observed in the whole-brain analysis ($p < 0.05$, corrected at cluster level) for the main and the control experiment. The format is the same for Table S1. IPS, left intraparietal sulcus; dlPFC, dorsolateral prefrontal cortex; Hemi, hemisphere; BA, Brodmann area.

## Supplemental Experimental Procedures

### Participants

In our *main* experiment, 120 healthy, normal volunteers (48 females; age range 18-38 years; mean ± SD, 22.11 ± 4.61) participated. Twenty out of the 120 participants (10 females; age range 19-38 years; mean ± SD, 28.40 ± 5.03) were scanned with functional magnetic resonance imaging (fMRI), while they performed an experimental task (see below). The remaining 100 participants were engaged in the same task outside the MRI scanner.

A *control* experiment involved 20 normal volunteers (9 females; age range, 20-39 years; mean, 27.75 ± 5.67 years) who did not participate in the main experiment. Importantly, the participants in the control experiment matched those who were scanned in the main experiment in terms of mean age ($p > 0.7$, two-tailed *t*-test), gender ratio ($p > 0.9$, two-tailed, Fisher's exact test), handedness ratio ($p > 0.9$, two-tailed, Fisher's exact test), hunger-rating score ($p > 0.1$, two-tailed *t*-test), education level ($p > 0.5$, two-tailed, Wilcoxon rank sum test), income level ($p > 0.1$, two-tailed, Wilcoxon rank sum test), Baron-Cohen Autism Quotient scale (Wheelwright et al., 2006) ($p > 0.9$, two-tailed *t*-test), degree of risk-aversion ($p > 0.4$, two-tailed, Wilcoxon rank sum test) and loss-aversion ($p > 0.1$, two-tailed, Wilcoxon rank sum test), Full-Scale IQ ($p > 0.6$, two-tailed *t*-test), and self-reported sociality scales (Russell and Karol, 2002) such as size of the social network, social-boldness, social-sensitivity, social-adjustment, social-control and social-expressivity ($p > 0.2$ for all the comparisons, two-tailed *t*-test).

All participants were pre-assessed to exclude those with a special diet, allergies to the type of foods used in the experiment or any previous history of neurological/psychiatric

illness, and they gave their informed written consent. The study was approved by the Institutional Review Board of the California Institute of Technology.

**Stimuli**

In our experiments, participants evaluated and made choices with 20 snack food items (e.g. chips and chocolate bars) and 20 non-food items, termed "trinkets" (e.g. Caltech memorabilia, DVDs and books). These items were highly familiar and available at campus and local stores, and mostly overlapped with those used in the previous study (Chib et al., 2009). All items were presented to participants as high-resolution color images (72 dpi).

**Experimental Tasks**

Participants performed three tasks: *pre-scanning BDM auction task*, *consensus decision-making task*, and *post-scanning BDM auction task*.

In the main experiment, six participants came to the Caltech Social Science Experimental Laboratory, SSEL (Pasadena, CA). In SSEL each participant was separated by a partition and was not allowed to make any verbal or gestural communication with other participants. The participants were taken through the experimental instructions together and then engaged in the pre-scanning BDM auction task. After that, one participant moved to the Caltech Brain Imaging Center, CBIC (Pasadena, CA), and performed the consensus decision-making task inside the MRI scanner, while the remaining five participants performed the same task at SSEL (CBIC and SSEL were interconnected via intranet). Finally, they were engaged in the post-scanning BDM auction task at SSEL. On the other hand, in the control

experiment, each participant never met any other participants and performed the three tasks individually.

Notably, to enhance participants' motivation to food items, we asked them to refrain from eating or drinking any liquids, besides water, for three hours before the experiment. Furthermore, they were asked to stay at SSEL for 30 min after the experiment, during which time the only thing they were able to eat was the snack obtained in the experiment.

***Pre- and Post-scanning BDM auction task.*** We measured participants' preference for each of the 40 items by using a Becker-DeGroot-Marschack (BDM) auction (Becker et al., 1964). We followed the experimental procedure used in Chib et al. (2009). In the BDM auction task, participants were asked to make self-paced bids, ranging from $0 to $4, for 40 items (see Figure S1AB and Chib et al. (2009) for details). The auction mechanism has been mathematically proven to be incentive compatible in a sense that the optimal strategy for the participants is to always bid the number closest to their true willingness to pay for obtaining that item (Becker et al., 1964). The optimal strategy was explicitly instructed to participants, and by using a questionnaire we confirmed that they correctly understood the experimental mechanism. In the present study, we refer to the amount of bid (i.e., willingness to pay) as "preference" for the item.

***Consensus decision-making task.*** In the main experiment, each participant tried to build a consensus with other participants on a choice between two items (Figure 1A). The task consisted of 40 blocks of trials (Figure 1BC): 20 *six*-person and 20 *four*-person blocks. The *six*-person blocks involved all *six* of the participants (*one* scanned and *five* not-scanned), while

the *four*-person blocks involved *four* participants (*one* scanned and *three* not-scanned selected randomly from the five). In conducting the experiment, we ran *four* fMRI scanning sessions, each of which had five *four*-person and five *six*-person blocks in a random order.

In each block of trials, participants simultaneously chose between two items repeatedly until they reached a consensus, i.e., choosing the same item (Figure 1BC). If they reached a consensus on a trial, they got the item and moved to the next block; otherwise they moved to the next trial in the same block and made another choice between the same pair of the items. If they did not reach consensus before the end of the block, they did not get anything and moved to the next block (e.g. Block 3 in Figure 1C). In the next block, participants made choices between a different pair of the items repeatedly, again, until they reached a consensus. Pairs of items were pseudo-randomly assigned so that the same pair was never presented again; that each item was selected only twice; and that a snack (trinket) item was always paired with another snack (trinket). Importantly, the maximum number of trials in each block was not instructed to participants, and in actuality was determined stochastically, as follows: if after 10 trials consensus was not yet reached, subsequent trials were triggered with a probability of 0.75, or the block was terminated with a probability of 0.25. The probability that the maximum number of trials is *k* is therefore, $P(k) = \begin{cases} 0 & \text{if } k < 10 \\ (1-p)^{k-10}p & \text{if } k \geq 10 \end{cases}$

where $p$ = 0.25. Thus after a minimum block length of 10 trials, additional trials could be triggered randomly until such a consensus was reached, or the block ended. This feature of the design is important because it means that the participants therefore cannot easily exploit knowledge about how many trials would be left in a block and thus cannot use this information

to alter their strategy accordingly.

At the beginning of each trial, each participant was asked to make a choice between the pair of items by pressing a button with their right hand (index finger for the left item; and middle finger for the right) with no time constraint (*Decision* phase; Figure 1B). The two items were randomly positioned left or right of the fixation point in every trial. The chosen item was immediately highlighted by a gray frame, initiating the *ISI* (inter-stimulus-interval) phase. After a waiting time for the other group-members' decisions and a jittered interval (1.5-4.5s), the others' choices were revealed to the participant by red dots (*Outcome* phase, 2s). Notably, participants were not able to identify each of the other group members; they were informed only about the distribution of the red dots (i.e., the number of participants choosing each of the two items). If all the dots were located below the image of the item the participant had chosen, i.e., consensus, the participant was informed that she/he obtained the item (*Instruction* phase, 3s) and moved to the next block after the jittered *ITI* (2-6s). Otherwise, *Decision* phase on the next trial in the same block was initiated following the *ITI*.

In the control experiment, settings were almost the same as those in the main experiment, except for that each participant tried to build a consensus with computer algorithms instead of other human participants. In this experiment, the instructions were carefully designed in order to make the task as "impersonal" as possible so as to avoid participants attributing agency to the computer algorithms. We thus did not use any suggestion of human-likeness in the computer algorithms. Specifically, we instructed to participants, "*You will get the item if all the red dots are located below the image of the item you choose within one block of trials. Otherwise, you will get nothing … At the initial trial in a block, the placement of each dot under a particular*

*item is determined randomly by the computer. On subsequent trials, it is possible that a dot will switch position from one item to the other. The probability of an individual dot moving from one item to the other depends on where the other dots are located, and the choice you yourself make*". In actuality, the computer algorithm to determine the location of each red dot was designed to mimic the not-scanned participants' actual choice behavior in the main experiment, both in terms of the tendency to follow the majority's choice and reaction times (Figure S1C).

***Reward payment***

Participants received a participation fee of $60. Furthermore, at the end of the experiment we created a pool for the participant's choices of all the three tasks. The computer selected one choice at random from the pool, and the selected choice was actually implemented. The specific process was as follows: (i) the computer randomly selected pre-, post-scanning BDM auction task or Consensus decision-making task; (ii) if pre- or post-scanning BDM auction task was selected, one trial in the task was selected at random and actually implemented; (iii) if Consensus decision-making task was selected, one block in the task was selected at random. If the participant reached a consensus with the other group-members in the block, she/he got the item; otherwise, she/he did not get anything. Since participants did not know which choice was selected, they should have treated every choice as if it were the only one. Also note that for this reason they did not need to consider the possibility of receiving the same item twice.

**Computational models**

To determine the key computational variables involved in the Consensus decision-making task, we constructed a family of computational models and fit those models to the participants' actual choice behaviors.

***Full model.*** A decision value for one item, say *A*, is constructed by

$$Q(A) = w_1 P(A) + w_2 G(A) + w_3 S(A), \quad (1)$$

where $P(A)$ is the participant's preference for *A*, $G(A)$ is the percentage of group-members who chose *A* on the previous trial, $S(A)$ is the estimated stickiness of *A* (see below for the derivation), and $w_1$, $w_2$ and $w_3$ denote the decision weights for the three variables. The value for the other item, *B*, is similarly constructed. The decision values govern the participant's choice probability (of the item *A*) as follows:

$$q(A) = f(Q(A) - Q(B))$$
$$= f(w_1(P(A) - P(B)) + w_2(G(A) - G(B)) + w_3(S(A) - S(B))), \quad (2)$$

where $f(z) = 1/[1 + \exp(-z)]$ is a sigmoidal function.

The hidden stickiness is inferred by a simple Bayesian learning algorithm (see Figure 3A for the graphical description of the inference). The Bayesian-learner estimates the hidden variable, *S*, from observable variables: the others' choices on the current trial, *Y*, and the group members' choices on the previous trial, *G*. Here, let's assume the stickiness, *S*, reflects the others' aggregated preference for each item without loss of generality. That is, the stickiness of an item results from the others' strong preference for the item. Precisely, $S(A)$ reflects the others' aggregated *relative* preference for *A*, $\bar{P}_O(A) \equiv P_O(A) - P_O(B)$ where $P_O$ denotes the others' preference for each item. Likewise, $S(B)$ reflects $\bar{P}_O(B) \equiv P_O(B) - P_O(A)$, and so

$\bar{P}_O(B) = -\bar{P}_O(A)$. Then, the Bayesian-learner's goal is to update estimates of the others' aggregated relative preference, $\bar{P}_O$, from their current choices, *Y*, and the group-members' prior choices, *G*,

$$p'(\bar{P}_O) \propto p(\bar{P}_O)\, p(Y|\bar{P}_O, G). \quad (3)$$

Note that a variable the Bayesian-learner estimates is the "aggregated" preference because in our task participants could not identify each of the other group-members. Moreover, it is worth mentioning that the estimated variable is the "relative" preference, as inference about the absolute preference participants have for each item is unlikely to be determinable from the choice data.

Assuming the Bayesian-learner believes that the others' choices *Y* are generated by their own preference $\bar{P}_O$ and the group members' prior choice *G*, likelihood of *Y* given $\bar{P}_O$ and *G* can be computed as a product of the others' choice probabilities,

$$\begin{aligned} p(Y|\bar{P}_O, G) &= q(A|\bar{P}_O, G)^{M_A}\, q(B|\bar{P}_O, G)^{M_B} \\ &= f(w_1\bar{P}_O(A) + w_2(G(A) - G(B)))^{M_A} \times \\ &\quad f(w_1\bar{P}_O(B) + w_2(G(B) - G(A)))^{M_B}, \end{aligned} \quad (4)$$

where $M_A$ and $M_B$ represent the number of others who choose *A* and *B* respectively ($M_A + M_B = M - 1$; *M* is a group size, *four* for *four*-person blocks and *six* for *six*-person blocks), and *f* is a sigmoidal function (as in equation (2)). The assumption about the learner's belief is reasonable, given that as shown in Figure 2C participants' choices were actually guided by their own preference and the group-members' prior choice in this task.

On the first trial in each block, the Bayesian learner has a *normally distributed* prior belief about the others' aggregated relative preference. This prior belief has the basis on the

assumption that the learner believes each of the others' relative preference is uniformly distributed in [$-4, $4] and so the aggregate preference over the $M-1$ other group-members is, by the central limit theorem, approximately normally-distributed with mean *zero* and variance $64/[12(M-1)]$.

In this paper, we term the expectation of the others' aggregated relative preference as stickiness, i.e., $S(A) \equiv \int \bar{P}_O(A) p\left(\bar{P}_O(A)\right) d\bar{P}_O(A)$. Participants utilize the estimated stickiness, *S*, at the time of decision to make a choice (see equation (1)).

To fit the model to the participants' actual choices, we employed a hierarchical modeling approach (Figure S4A) (Daw, 2011). In this approach, each participant's individual-level decision parameters are assumed to be drawn from common population-level normal distributions: for participant *i*, $w_1^{(i)} \sim N(\mu_1, \sigma_1)$, $w_2^{(i)} \sim N(\mu_2, \sigma_2)$ and $w_3^{(i)} \sim N(\mu_3, \sigma_3)$. Then, the likelihood of the participant *i*'s choices is,

$$p(\boldsymbol{c}_i|\mu_1,\sigma_1,\mu_2,\sigma_2,\mu_3,\sigma_3)$$
$$= \iiint p\left(\boldsymbol{c}_i|w_1^{(i)},w_2^{(i)},w_3^{(i)}\right) p\left(w_1^{(i)}|\mu_1,\sigma_1\right) p\left(w_2^{(i)}|\mu_2,\sigma_2\right) p\left(w_3^{(i)}|\mu_3,\sigma_3\right) dw_1^{(i)} dw_2^{(i)} dw_3^{(i)} \quad (5)$$

where $p\left(\boldsymbol{c}_i|w_1^{(i)},w_2^{(i)},w_3^{(i)}\right)$ can be derived by equation (2). The likelihood of the full choice data set from all participants (1,…,*N*) is simply a product over the participants:

$$p(\boldsymbol{c}_1 \cdots \boldsymbol{c}_N|\mu_1,\sigma_1,\mu_2,\sigma_2,\mu_3,\sigma_3) = \prod_{i=1}^{N} p(\boldsymbol{c}_i|\mu_1,\sigma_1,\mu_2,\sigma_2,\mu_3,\sigma_3). \quad (6)$$

We estimated the population-level parameters using Maximum likelihood estimation of equation (6) (minimizing the negative log likelihood, $-\log p(\boldsymbol{c}_1 \cdots \boldsymbol{c}_N|\mu_1,\sigma_1,\mu_2,\sigma_2,\mu_3,\sigma_3)$, by using *fminsearch* in Matlab R2013b, MathWorks). For model comparisons (Figure 3G, and

Figures S1G and S4CD), the obtained negative log likelihood was converted to Bayesian information criterion (BIC) that takes into account the numbers of free parameters (smaller values indicate better fit).

Once the population parameters are determined, we can recover the distribution of each participant's individual-level parameters as follows:

$$p\left(w_1^{(i)}, w_2^{(i)}, w_3^{(i)}\right) \propto p\left(\boldsymbol{c}_i | w_1^{(i)}, w_2^{(i)}, w_3^{(i)}\right) p\left(w_1^{(i)}, w_2^{(i)}, w_3^{(i)} | \mu_1, \sigma_1, \mu_2, \sigma_2, \mu_3, \sigma_3\right). \quad (7)$$

Then, the expected value of each parameter can be derived by, e.g., for $w_1^{(i)}$,

$$\widehat{w_1^{(i)}} = \int w_1^{(i)} \iint p\left(w_1^{(i)}, w_2^{(i)}, w_3^{(i)}\right) dw_2^{(i)} dw_3^{(i)}\, dw_1^{(i)}. \quad (8)$$

We used the expected values of $w_1^{(i)}$, $w_2^{(i)}$ and $w_3^{(i)}$ as each participant's decision weights for the three computational variables in the behavioral and fMRI analyses.

In a supplemental analysis, we also employed an individual modeling approach (Figure S4B) in which the individual-level parameters are estimated for each individual participant separately without the hierarchical structure. In this analysis, each model's goodness of fit was assessed by Model evidence (approximated by BIC (Daw, 2011)) and then entered into a Bayesian model selection procedure (BMS) (Stephan et al., 2009) for the model comparisons (Figures S2C and S4E).

***Alternative partial models.*** We also implemented a family of partial models that included only *one* or *two* of the three decision variables used in the full model above to compute the decision value (a separate model was implemented to capture each possible combination of decision variables). For example, a model named "P + G" in Figure 3G constructs a decision value based

only on the participant's own preference and the group members' prior choice, i.e., $Q(A) = w_1 P(A) + w_2 G(A)$.

***Additional alternative models.*** We also considered some variants of the P + G model, in which the decision weight for the group-members' prior choice, $w_2$, is modulated trial-by-trial, as proposed by a theoretical study (Couzin et al., 2005).

In a model, named "P + $G_{\text{modulated I}}$" in Figure S4, the weight is updated as proposed in (Couzin et al., 2005):

$$w_{2,t+1} = \begin{cases} w_{2,t} + \eta & if\ C = preferred\ item \\ w_{2,t} - \eta & if\ C \neq preferred\ item, \end{cases} \quad (9)$$

where $C$ denotes an item chosen by the participant on the trial $t$. Here, the initial weight and the increment $\eta$ are individual-level free-parameters.

In another model, named "P + $G_{\text{modulated II}}$" in Figure S4,

$$w_{2,t+1} = \begin{cases} w_{2,t} + \eta & if\ C = the\ majority's\ choice \\ w_{2,t} - \eta & if\ C \neq the\ majority's\ choice. \end{cases} \quad (10)$$

In the other model, named "P + $G_{\text{modulated III}}$" in Figure S4,

$$w_{2,t+1} = \begin{cases} w_{2,t} + \eta & if\ preferred\ item = the\ majority's\ choice \\ w_{2,t} - \eta & if\ preferred\ item \neq the\ majority's\ choice. \end{cases} \quad (11)$$

Furthermore, we considered another alternative model, named "P x G" in Figure S4, that constructs a decision value based on the *expected reward* of the item, i.e., preference value multiplied by the likelihood of consensus on the item. Assuming that the likelihood is approximated with the percentage of other group-members' who chose the item on the previous trial, $G_O$, a decision value can be represented by $Q(A) = w_1 P(A)^{\alpha} \times G_O(A)$, where $\alpha$ governs the participant's risk-preference.

***Social conformity model.*** We constructed a model reflecting social conformity effects that conflict between participants' own preference and observed group members' choice alters their preference. In this model, when participants' preference is incongruent with the majority's choice, their preference is updated in proportion to the percentage of group-members' who chose the item on the *current* trial. That is,

$$P_{t+1}(A) = P_t(A) + \eta[G_t(A) - 50],$$

where the learning rate $\eta$ is an individual-level free-parameter and the updated preference carry over across blocks (preference for the other item, *B*, is similarly updated). A decision value at trial *t* + 1 is thus computed as $Q_{t+1}(A) = w_1 P_{t+1}(A)$.

**Additional behavioral analyses**

**Stability of preferences and effects of social conformity.** We checked if participants' preference for each item was altered during the experiment. Contrasting the preference measured before and after the experiment, we found the participants' preference was overall quite stable (Figure S1D). There was no significant change ($p > 0.40$, two-tailed, for all items) irrespective of item categories ($p > 0.9$, two-tailed, for foods; and $p > 0.4$, two-tailed, for trinkets) or if they reached consensus on the item or not ($p > 0.1$, two-tailed, for consensus items; and $p > 0.7$, two-tailed, for no-consensus items).

A closer examination, however, revealed a small but significant change in the participants' preference in a manner consistent with previous studies on social conformity (Campbell-Meiklejohn et al., 2010; Charpentier et al., 2014; Klucharev et al., 2009). When participants reached a consensus on the non-preferred item by compromise, their preference for

the consensus/non-preferred item increased while that for the paired no-consensus/preferred item decreased (Figure S1E), implying that conflict between participants' own preference and observed group members' choice altered their preference (Izuma, 2013). Here, it is worth noting that the preference change is unlikely to be strongly confounded with the effect of group members' prior choices in our model. This is because there was no significant correlation between the social conformity effect (degree of the preference change) and the weight of the group-members' prior choices estimated by the model fitting ($p > 0.4$, two-tailed, for all cases; Figure S1F). Thus, the effects of social conformity on preference and strategic consideration of group members' prior choices are two distinct effects. Moreover, the social conformity account has little explanatory power to predict participants' actual choice behavior in our experiment. To test this, we constructed a computational model reflecting purely social conformity effects, but without any strategic components (see Social conformity model), and compared the goodness of fit with our other models that do incorporate strategic considerations as described in the previous section. Formal model comparison showed the social conformity model provided a worse fit compared to the other models (Figure S1G). Taken together, although there is a slight effect of social conformity on individual preferences in the study, social conformity effects per se do not appear to be playing a major role in explaining participants' choice behavior.

**Individual behavior in the *main* experiment: additional alternative explanations.** In this experiment, participants in principle could employ more complicated strategies. One possibility might be that participants learned some aspects of the task-structure or about others' preferences *across* blocks, which would result in a decrease in the number of trials taken to reach consensus

in late blocks. Another possibility might be that participants engaged in temporal coordination/collusion. An example of this in the context of our experiment would be if participants reached a consensus on the item that was preferred by one participant, and then in the next block they reached a consensus on the item preferred by another participant. This type of temporal coordination, so-called "alternating reciprocity", has been observed in previous studies of *two*-person battle of sexes games (Rapoport et al., 1976; Sonsino and Sirota, 2003). We reasoned that if it was the case in our experiment, a sequence of whether the participant won or lost (win: consensus on the preferred item; lose: otherwise) would exhibit an oscillatory pattern at least to some degree.

To examine these possibilities, we looked into the temporal property of the number of trials and sequences of wins/losses. Inconsistent with the possibility that participants adopted complicated strategies, we did not find any significant temporal-trends in the number of trials (ANOVA, $p > 0.4$; Jonckheere's trend test on the individual participant level, $p < 0.05$, two-tailed, only in *two* of the 20 participants) or auto-correlations in the sequence of wins/losses ($r = 0.047 \pm 0.040$, $p > 0.25$, two-tailed; on the individual participant level, $p < 0.05$, two-tailed, only in *one* of the 20 participants). Moreover, it is worth noting that in this experiment the same pair of items was never presented again; and that each participant couldn't uniquely identify each of the other individuals (see Experimental Tasks). Given these features of the experiment, as well as the results of the additional analyses, we believe that it is unlikely that participants utilized these complex strategies.

**fMRI Data Acquisition**

The fMRI images were collected using a 3T Siemens (Erlangen) Trio scanner located at the Caltech Brain Imaging Center (Pasadena, CA) with a 32-channel radio frequency coil. The BOLD signal was measured using a one-shot T2*-weighted echo planar imaging sequence (Volume TR = 2780 ms, TE = 30 ms, FA = 80°). Forty-four oblique slices (thickness = 3.0 mm, gap = 0 mm, FOV = 192 × 192 mm, matrix = 64 × 64) were acquired per volume. The slices were aligned 30° to the AC–PC plane to reduce signal dropout in the orbitofrontal area (Deichmann et al., 2003). After the four functional sessions, high-resolution (1 $mm^3$) anatomical images were acquired using a standard MPRAGE pulse sequence (TR = 1500 ms, TE = 2.63 ms, FA = 10°).

**fMRI Data Analysis**

***Preprocessing.*** We used the SPM8 software (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London) for image processing and statistical analysis. fMRI images for each participant were preprocessed using the standard procedure in SPM8: after slice timing correction, the images were realigned to the first volume to correct for participants' motion, spatially normalized, and spatially smoothed using an 8 mm FWHM Gaussian kernel. High-pass temporal filtering (using a filter width of 128 s) was also applied to the data.

***General linear model I.*** A separate general linear model (GLM) was defined for each participant. The GLM contained parametric regressors representing the three key computational variables at the onset of decision (Figure 1B): the participant's preference for the chosen item,

the percentage of group-members who had previously selected the item that was chosen by the participant on the current trial, and the estimated stickiness of the chosen item.

Specifically, the participant-specific design matrices contained the following regressors: (1) two stick functions at the onset of decision phase separately for *four*-person and *six*-person blocks, and two boxcar functions in the period of the decision phase for *four*-person and *six*-person blocks; (2) two stick functions at the timing of motor response for *four*-person and *six*-person blocks; (3) two stick functions at the onset of outcome phase, and two boxcar functions in the period of the phase; and (4) two stick functions at the onset of instruction phase, and two boxcar functions in the period of the phase. We also included the parametric modulators of the stick functions at the onset of decision for *four*-person and *six*-person blocks, encoding the participant's preference for the chosen item, the percentage of group-members who had previously selected the item that was chosen by the participant on the current trial, and the estimated stickiness of the chosen item, respectively. Furthermore, to control for nuisance effects of outcome-related neural signals, we included two parametric modulators at the onset of outcome: the group members' current choice (the percentage of group-members who chose the same item as the participant did), and the prediction error about the stickiness (posterior minus prior mean of the stickiness for the chosen item). All the regressors were convolved with a canonical hemodynamic response function. In addition, six motion-correction parameters were included as regressors of no-interest to account for motion-related artifacts. Notably, in the model specification procedure, we concatenated four sessions into one GLM because of the small number of trials in each session (note: constant regressors coding for each session were included); and serial orthogonalization of parametric modulators was turned off.

We defined contrasts of interest for the three key computational variables at the time of decision, independent of group size, as a [1 1] contrast of corresponding regressors for *four*-person and *six*-person blocks. For each participant, the contrasts were estimated at every voxel of the whole-brain and entered into a random-effects analysis.

***General linear model II.*** The GLM contained parametric regressors representing the choice probability assigned by the computational model to the participant's chosen item at the onset of decision phase for *four*-person and *six*-person blocks, in addition to the regressors included in GLM I. Note that the choice probability regressors were orthogonalized to the other parametric regressors at the same timing.

***Psychophysiological interaction (PPI) analysis.*** The PPI analysis was performed by the standard procedure in SPM8. We first extract BOLD signals from independent ROIs (see below): vmPFC, right pSTS/TPJ and bilateral IPS (combination of the right and left IPS regions); and then created PPI regressors by forming interactions of the BOLD signals (physiological factors) and the stick function regressors at the onset of decision phase for *four*-person and *six*-person blocks (psychological factors). The GLM for the PPI analysis therefore contained the following regressors: (1) physiological factors, BOLD signals from the three ROIs; (2) psychological factors, the stick function regressors at the onset of decision phase for *four*-person and *six*-person blocks; and (3) PPI factors, interaction terms of the psychological and physiological factors, as well as the other regressors included in GLM I.

***Whole-brain analysis.*** We set our significance threshold at $p < 0.05$ whole-brain corrected for multiple comparisons at cluster-level. The minimum spatial extent, $k = 63$, for the threshold was

estimated based on the underlying voxel-wise *p*-value, $p < 0.005$ uncorrected, by using the AlphaSim program in Analysis of Functional NeuroImages (AFNI) (Cox, 1996).

***ROI analysis.*** ROI analyses were performed by the Marsbar software. Each ROI or small-volume was defined as a 10mm sphere centered on the coordinates extracted from the prior studies on social and non-social decision-making. For vmPFC (Figure 4), we used the coordinates [-5 37 -10] extracted from a meta-analysis (Clithero and Rangel, 2013); for pSTS/TPJ (Figure 5), we used the coordinates [55 -45 13] which are the average co-ordinates extracted from the four most relevant computational studies of STS/TPJ function (Behrens et al., 2008; Boorman et al., 2013; Hampton et al., 2008; Haruno and Kawato, 2009); for right and left IPS (Figure 6), we used the coordinates [41 -51 39] and [-37 -56 42] which are averaged co-ordinates extracted from the five most relevant prior computational fMRI studies involving the IPS (d'Acremont et al., 2013; Daw et al., 2006; Glascher et al., 2010; Hare et al., 2011; Suzuki et al., 2012).

## Supplemental References

Campbell-Meiklejohn, D.K., Bach, D.R., Roepstorff, A., Dolan, R.J., and Frith, C.D. (2010). How the opinion of others affects our valuation of objects. Curr Biol *20*, 1165–1170.

Charpentier, C.J., Moutsiana, C., Garrett, N., and Sharot, T. (2014). The brain's temporal dynamics from a collective decision to individual action. Journal of Neuroscience *34*, 5816–5823.

Clithero, J.A., and Rangel, A. (2013). Informatic parcellation of the network involved in the computation of subjective value. Social Cognitive and Affective Neuroscience *9*, 1289–1302.

Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. *29*, 162–173.

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., and Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. Nature *441*, 876–879.

Deichmann, R., Gottfried, J., Hutton, C., and Turner, R. (2003). Optimized EPI for fMRI studies of the orbitofrontal cortex. NeuroImage *19*, 430–441.

Haruno, M., and Kawato, M. (2009). Activity in the Superior Temporal Sulcus Highlights Learning Competence in an Interaction Game. Journal of Neuroscience *29*, 4542–4547.

Izuma, K. (2013). The neural basis of social influence and attitude change. Curr Opin Neurobiol *23*, 456–462.

Klucharev, V., Hytonen, K., Rijpkema, M., Smidts, A., and Fernandez, G. (2009). Reinforcement Learning Signal Predicts Social Conformity. Neuron *61*, 140–151.

Rapoport, A., Gordon, D.G., and Guyer, M.J. (1976). The 2x2 Game (University of Michigan Press).

Russell, M.T., and Karol, D. (2002). The 16PF Fifth Edition administrator's manual (Champaign, IL: Institute for Personality and Ability Testing).

Sonsino, D., and Sirota, J. (2003). Strategic pattern recognition—experimental evidence. Games and Economic Behavior *44*, 390–411.

Wheelwright, S., Baron-Cohen, S., Goldenfeld, N., Delaney, J., Fine, D., Smith, R., Weil, L., and Wakabayashi, A. (2006). Predicting Autism Spectrum Quotient (AQ) from the Systemizing Quotient-Revised (SQ-R) and Empathy Quotient (EQ). Brain Res. *1079*, 47–56.