# Inventory of Supplemental Information

Supplemental Figure 1, related to Figure 2. This figure shows the relationship between estimated contribution and interpersonal ratings of friendliness and dominance across conditions, for comparison to the relationship shown in Figure 2 between estimated contribution and liking.

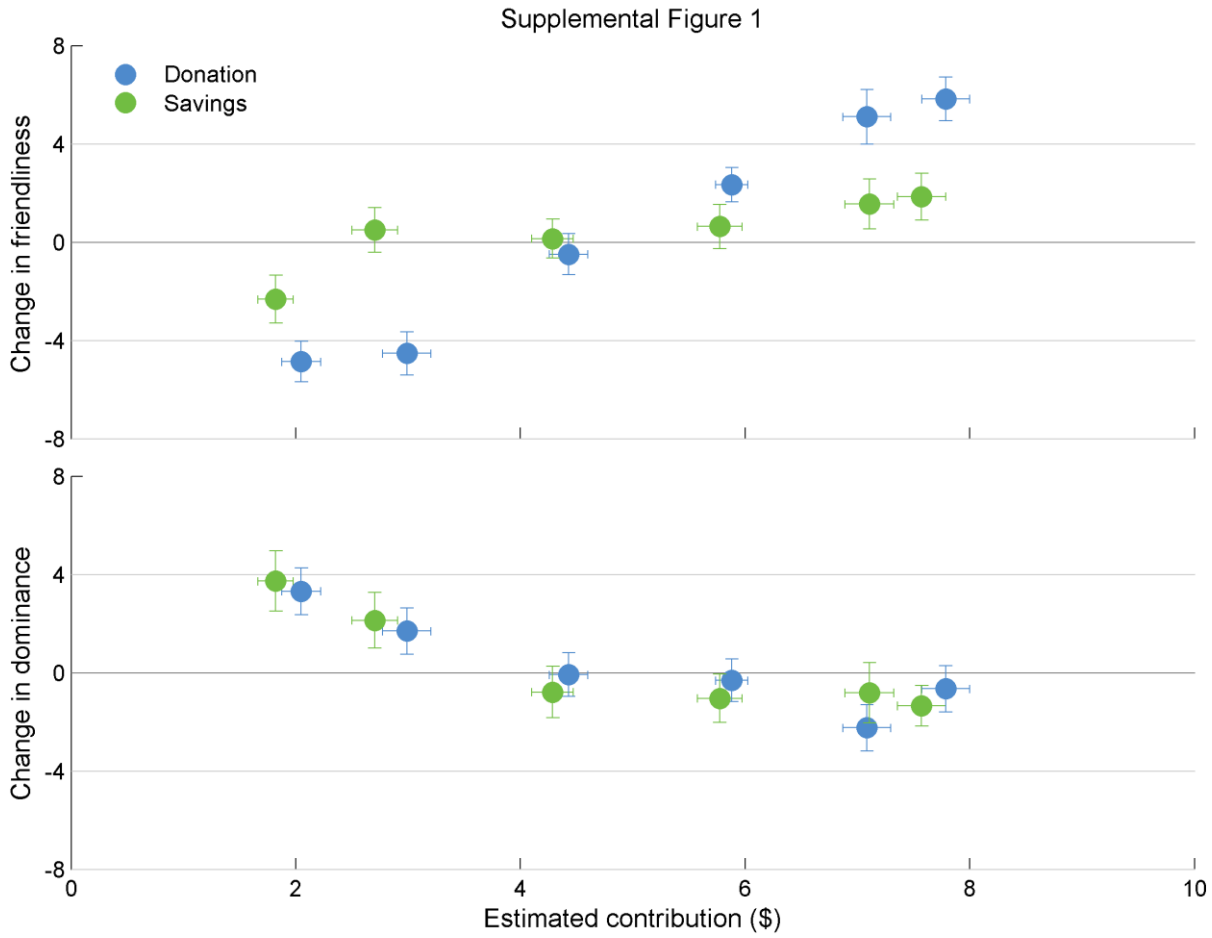
Supplemental Table 1. Performance and reaction time in behavioral study, related to Table 1. This table provides inference accuracy and reaction time data for the behavioral study, for comparison to the same measures in the FMRI study presented in Table 1.

Supplemental Table 2. Activation during inference phase, related to Table 2. This table provides activation data for the interaction of inferred contribution and condition, analyzed using the reinforcement learning model (as compared to the same interaction analyzed with the standard model, shown in Table 2). This model contains additional control regressors for inferential certainty and estimated contribution sum. This table also provides activation data for the inference phase within conditions using the standard model (as compared to the between-condition contrasts presented in Table 2).

Supplemental Table 3. Activation correlated with inferential errors during feedback phase within conditions (reinforcement learning model), related to Table 3. This table provides activation data for the feedback phase within conditions (as compared to the between-condition contrasts presented in Table 3).

Supplemental Experimental Procedures. This section explains the statistical analysis of interpersonal ratings in the behavioral study, explains the imaging analysis in greater detail than the main text, and provides details for the specific reinforcement learning algorithm used in this analysis.

Supplemental References.




**Supplemental Figure 1, related to Figure 2. Framing of estimated contributions selectively affects ratings of friendliness.** Top: Points represent change in friendliness ratings from before to after the task for each player, plotted against the estimated average contribution for that player. Bottom: Points represent change in dominance ratings from before to after the task for each player, plotted against the estimated average contribution for that player. Participants in the Savings condition saw savings amounts ($10 – contributions); contributions are displayed here for clarity. Error bars are standard errors across participants.

Supplemental Table 1

*Performance and reaction time in behavioral study, related to Table 1*

| Block of trials | % correct (*SEM*) | | Reaction time, ms (*SEM*) | |
|---|---|---|---|---|
| | Donation | Savings | Donation | Savings |
| 1st | 57.94 (2.16) | 59.21 (2.13) | 4252.91 (216.25) | 4669.40 (291.77) |
| 2nd | 64.60 (2.34) | 64.76 (2.24) | 4541.02 (293.38) | 4323.22 (289.07) |
| 3rd | 67.78 (1.86) | 67.46 (2.05) | 4283.20 (266.92) | 4023.51 (243.73) |
| 4th | 67.30 (1.65) | 68.89 (1.99) | 4202.01 (242.82) | 4181.57 (249.66) |

*Note.* $n$ = 84 (42 in Donation condition, 42 in Savings condition). Blocks are 15 trials long.

Standard errors of the mean (*SEM*) are calculated within block and condition.

Supplemental Table 2

*Activation during inference phase, related to Table 2*

| Region | Peak Z-score | X | Y | Z | Cluster size (vox) |
|---|---|---|---|---|---|
| Reinforcement learning model | | | | | |
| *High > Low (Donation > Savings)* | | | | | |
| Posterior inferior frontal gyrus | 4.09 | 52 | -10 | 22 | 14 |
| Medial precuneus | 3.86 | 14 | -54 | 62 | 26 |
| Inferior temporal cortex | 3.58 | -40 | -2 | -28 | 18 |
| Posterior superior temporal gyrus | 3.56 | -50 | -48 | 10 | 11 |
| Middle frontal gyrus | 3.49 | -42 | 30 | 10 | 11 |
| Ventromedial PFC | 3.48 | 0 | 42 | -6 | 10 |
| Standard model | | | | | |
| *Donation condition only: High > Low* | | | | | |
| Ventromedial PFC | 4.20 | -2 | 46 | -4 | 84* |
| Superior temporal sulcus | 3.87 | 60 | -18 | -10 | 17 |
| Angular gyrus | 3.87 | -52 | -62 | 36 | 17 |
| Rostromedial PFC | 3.78 | -12 | 60 | 18 | 39 |
| Rostromedial PFC | 3.74 | 14 | 58 | 12 | 21 |
| Ventromedial PFC | 3.66 | 12 | 52 | -2 | 12 |
| Medial parietal cortex | 3.53 | 16 | -56 | 62 | 10 |

| | | | | | |
|---|---|---|---|---|---|
| *Donation condition only: Low > High* | | | | | |
| Lingual gyrus | 4.34 | 24 | -70 | 8 | 22 |
| Dorsomedial PFC | 3.73 | -6 | 16 | 44 | 13 |
| *Savings condition only: High > Low* | | | | | |
| No clusters active at this threshold. | | | | | |
| *Savings condition only: Low > High* | | | | | |
| Precentral gyrus | 4.47 | 22 | -24 | 74 | 25 |
| Precentral gyrus | 4.21 | 62 | -6 | 22 | 18 |
| Thalamus | 4.16 | -30 | -28 | 6 | 19 |
| Superior frontal gyrus | 4.08 | 16 | 26 | 62 | 21 |
| Superior frontal gyrus | 3.97 | 24 | 6 | 68 | 22 |
| Medial frontal gyrus | 3.86 | 10 | -10 | 74 | 17 |
| Dorsolateral PFC | 3.67 | 24 | 24 | 44 | 10 |
| Inferior frontal gyrus | 3.64 | 46 | 28 | -10 | 12 |
| Putamen | 3.62 | 36 | -16 | -6 | 10 |
| Dorsomedial PFC | 3.59 | 0 | 12 | 46 | 12 |

*Note.* [a] indicates subpeaks within a cluster. PFC = prefrontal cortex. * = cluster size $p < 0.05$ corrected for multiple comparisons across the whole brain. Activations in table were thresholded voxelwise at $p < 0.001$ and with a cluster size ≥ 10 voxels (whole-brain corrected cluster-size threshold = 59 voxels [reinforcement learning model ], 51 voxels [standard model, Donation condition], 41 voxels [standard model, Savings condition]). T-statistics were converted to Z-scores for reporting. Coordinates are reported in MNI/ICBM152 coordinates, as in SPM5. Resampled voxel size was 2 x 2 x 2 mm.

Supplemental Table 3

*Activation correlated with inferential errors during feedback phase within conditions (reinforcement learning model), related to Table 3*

| Region | Peak Z-score | X | Y | Z | Cluster size (vox) |
|---|---|---|---|---|---|
| *Donation condition only: positive correlations with model* | | | | | |
| Putamen | 4.05 | -24 | 16 | 2 | 14 |
| Parahippocampal gyrus | 4.03 | 34 | -36 | -16 | 13 |
| *Donation condition only: negative correlations with model* | | | | | |
| Inferior frontal gyrus | 4.14 | -30 | 22 | 20 | 19 |
| *Savings condition only: positive correlations with model* | | | | | |
| Rostromedial PFC | 3.85 | 0 | 58 | 28 | 14 |
| *Savings condition only: negative correlations with model* | | | | | |
| Lingual gyrus | 3.86 | 2 | -70 | 12 | 46* |
| Posterior cingulate | 3.83 | 34 | -64 | 12 | 14 |
| Dorsomedial PFC | 3.73 | -10 | 12 | 58 | 25 |

*Note.* PFC = prefrontal cortex. * = cluster size $p < 0.05$ corrected for multiple comparisons across the whole brain. Activations in table were thresholded voxelwise at $p < 0.001$ and with a cluster size $\geq 10$ voxels (whole-brain corrected cluster-size threshold = 63 voxels [Donation condition], 41 voxels [Savings condition]). T-statistics were converted to Z-scores for reporting. Coordinates are reported in MNI/ICBM152 coordinates, as in SPM5. Resampled voxel size was 2 x 2 x 2 mm.

Supplemental Experimental Procedures

*Interpersonal ratings*

In the behavioral study, player judgment questionnaires before and after the task included ratings for eight interpersonal traits: friendly, outgoing, assertive, cunning, antisocial, introverted, submissive, and undemanding. Each trait was framed as a question (e.g., "How friendly does this person seem?"), and participants indicated their answers on a 9-point Likert-type scale anchored at 1 by *not at all*, at 5 by *somewhat*, and at 9 by *very*. The trait adjectives were chosen from earlier studies to span the interpersonal circumplex (Knutson, 1996; Wiggins, 1979) and provide robust measures of two key interpersonal dimensions: dominance and friendliness (cf. Fiske et al., 2007; Oosterhof and Todorov, 2008).

Pre-task dominance ratings for each player were constructed from the pre-task trait ratings according to the formula:

Assertive – Submissive + (√2/2)*Outgoing + (√2/2)*Cunning – (√2/2)*Undemanding – (√2/2)*Introverted

corresponding to each adjective's geometric location on the circumplex. Similarly, pre-task friendliness ratings were constructed according to the formula:

Friendly – Antisocial + (√2/2)*Outgoing + (√2/2)*Undemanding – (√2/2)*Cunning – (√2/2)*Introverted

Post-task dominance and friendliness ratings were constructed identically, using the post-task trait ratings. As with liking, dominance and friendliness ratings were then analyzed as changes from before to after the task, including the pre-task rating as a covariate of no interest.

*Imaging analysis*

During spatial preprocessing, functional images were first corrected for slice timing to the middle slice and realigned to the first image in the run. Next, in-plane and high-resolution anatomical images were coregistered to the mean functional image using normalized mutual information and normalized to the MNI template brain using standard options (8-mm source smoothing and affine transformation, followed by 16 nonlinear iterations of discrete cosine basis warping, using a frequency cutoff of 25 mm). Functional images were then normalized with the in-plane parameters, resampled to 2 x 2 x 2 mm, and smoothed with a 4-mm FWHM Gaussian filter.

The standard model examined experimental effects across task events. Six regressors (one for each player) modeled the appearance of players in the face phase. Four regressors modeled the onset of the inference phase, one each for High and Low inferences crossed with subsequently correct or incorrect inferences. Four regressors modeled the onset of the feedback phase, one each for High and Low feedback crossed with correct or incorrect feedback. Effects were modeled with stick functions of 0 duration (1 at onset and 0 otherwise). For each inference regressor, a first-order parametric modulator weighted with trial-by-trial reaction time was also included.

The reinforcement learning model examined effects correlated with trial-by-trial learning. The output of a reinforcement learning algorithm (see below) was used to estimate each participant's learning of the average contribution associated with each player over time, which was then used to calculate regressors for each participant's imaging data. Three regressors modeled the face phase: a stick function for the appearance of any player, a parametric modulator for each player's estimated average contribution level on that trial, and a parametric

modulator for certainty (the absolute difference between the player's estimated contribution and the mean estimated contribution over all players and trials). Five regressors modeled the inference phase: a stick function for the onset of any inference phase, a contrast function for High vs. Low inferences (1 at the onset of High inferences, -1 at the onset of Low inferences, and 0 elsewhere), and parametric modulators for reaction time, inferential certainty, and the sum of estimated contributions. Inferential certainty was calculated as the absolute difference between the sum of estimated contributions on that trial and 19.5. Twelve regressors modeled the feedback phase, two for each player: a stick function for feedback for that player, and a parametric modulator for the player-specific inferential error on that trial.

All parametric modulators in both models were first-order, and all were mean-centered before convolution with the hemodynamic response function.

*Reinforcement learning algorithm*

The reinforcement learning model used a modified Rescorla-Wagner learning rule (Rescorla and Wagner, 1972) to parameterize how participants learned and updated estimated contribution levels. In each participant's model, initial estimated contribution levels were set identically for all players and then updated for each player at each feedback phase by a fraction of the difference between the actual and estimated contributions.

Initial fitting suggested that participants learned faster at the beginning of the experiment than at the end – a rational response to a stable environment. This stronger weighting of early information could be modeled with an exponential decay in the learning rate over time. The final model thus fit three parameters for each participant - an initial estimate $\theta$, a learning rate $\alpha$, and a

decay rate β. Each participant's estimated contribution level $E_{p,t}$ for player *p* at appearance *t* could be expressed as a function of the parameters and the actual contributions $C_{p,t}$:

$$E_{p,0} = \theta$$

$$E_{p,t+1} = E_{p,t} + (\alpha / t^{\beta})$$

The probability of inferring High or Low on each trial was then assumed to be a logistic function of the summed estimated contribution levels for that trial.

The *nlmefit* function in MATLAB 7.7 was used to estimate mixed-level models, treating inferences as the lower-level unit within participants. Models were estimated using maximum likelihood and diagonal covariance, and parameters that yielded impossible inferences (greater than $10 or less than $0) were not allowed. Missing inferences were removed before model fitting.

Classification rates indicated reasonable fits to all participants' inference data. Fits ranged from 56.67 – 86.67% in the Donation condition (*M* = 71.50%, *SEM* = 1.74%) and were marginally higher in the Savings condition, ranging from 61.67 – 85.00% (*M* = 75.93%, *SEM* = 1.69%; $t(36) = 1.82$, $p < 0.08$).

## Supplemental References

Fiske, S.T., Cuddy, A.J.C., and Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. Trends Cogn. Sci. *11*, 77-83.

Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. J. Nonverbal Beh. *20*, 165-182.

Oosterhof, N.N., and Todorov, A. (2008). The functional basis of face evaluation. Proc. Natl. Acad. Sci. USA *105*, 11087-11092.

Rescorla, R.A., and Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Classical conditioning II: Current research and theory, A.H. Black, and W.F. Prokasy, eds. (New York: Appleton Century Crofts), pp. 65-99.

Wiggins, J.S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. J. Pers. Soc. Psychol. *37*, 395-412.