Neural basis of egalitarian behavior

Christopher T. Dawes[a,1], Peter John Loewen[b], Darren Schreiber[c], Alan N. Simmons[d,e], Taru Flagan[d,f], Richard McElreath[g], Scott E. Bokemper[h], James H. Fowler[c,i], and Martin P. Paulus[d,e,f]

[a]Department of Politics, New York University, New York, NY 10012; [b]Department of Political Science, University of Toronto, Toronto, ON, Canada M5S 3G3; [c]Department of Political Science, [d]Department of Psychiatry, [f]Laboratory of Biological Dynamics and Theoretical Medicine, and [i]School of Medicine, University of California at San Diego, La Jolla, CA 92103; [e]Psychiatry Service, Veterans Affairs San Diego Healthcare System, San Diego, CA 92161; [g]Department of Anthropology, University of California, Davis, CA 95616; and [h]Department of Political Science, University of Nebraska, Lincoln, NE 68588

Edited by Terrence J. Sejnowski, The Salk Institute for Biological Studies, La Jolla, CA, and approved March 2, 2012 (received for review November 18, 2011)

Individuals are willing to sacrifice their own resources to promote equality in groups. These costly choices promote equality and are associated with behavior that supports cooperation in humans, but little is known about the brain processes involved. We use functional MRI to study egalitarian preferences based on behavior observed in the "random income game." In this game, subjects decide whether to pay a cost to alter group members' randomly allocated incomes. We specifically examine whether egalitarian behavior is associated with neural activity in the ventromedial prefrontal cortex and the insular cortex, two regions that have been shown to be related to social preferences. Consistent with previous studies, we find significant activation in both regions; however, only the insular cortex activations are significantly associated with measures of revealed and expressed egalitarian preferences elicited outside the scanner. These results are consistent with the notion that brain mechanisms involved in experiencing the emotional states of others underlie egalitarian behavior in humans.

behavioral economics | egalitarianism

An intriguing aspect of human behavior that has long puzzled scholars (1) is that individuals are willing to sacrifice their own resources to promote equality in groups. For example, when dividing resources between oneself and others, people make fair divisions, at a cost to themselves, when the interaction is anonymous and the opportunity for reciprocity does not exist (2). Individuals reject unequal divisions offered by others, even if rejection means neither party receives anything (3); they reject payment for a task after having observed another receiving a higher payment for the same task, even when they had accepted the lower payment before observing others' payoffs (4); and they voluntarily pay a personal cost to increase the resources of the poorest members of their groups and decrease the resources of the richest members, even when no reputational benefits or reciprocity can be expected (5).

Recent neuroscience studies using functional MRI (fMRI) have begun to identify the neural mechanisms underlying other-oriented behavior. For example, reward-related mechanisms play an important role when individuals engage in costly punishment of people who choose not to contribute to a mutual effort (6), during both mandatory and volitional giving to others (7), and when contemplating more equal divisions of resources (8). However, the focus of these experiments has been on dyadic instead of group interactions. Moreover, although these studies identify brain areas active during valuation, they do not identify which activations can be used to predict which individuals will actually engage in egalitarian behavior outside of a scanner.

To better understand what neural mechanisms might underlie egalitarian behavior in groups, we use a procedure called the "random-income" game. In this game, participants in a group are arbitrarily assigned a level of income and the group is assigned to one of three specific levels of income distribution by a computer (see *SI Methods*). Past analyses of the random-income game have shown that participants are willing to pay to take from the rich and to give to the poor, even in circumstances where the participants have no control of their initial income level and the group's initial income distribution. This willingness to pay is driven in part by emotional responses to unequal outcomes (5). Importantly, those with the greatest sensitivity to inequality in the random-income game are also those most likely to engage in costly punishment of noncontributors in a separate public-goods game (9), a behavior that has been shown to promote cooperation in humans (10).

We conducted an fMRI experiment on 20 subjects (10 male, 10 female) to identify brain regions involved in different stages of a random-income game paradigm (Fig. 1). In this experiment, each scanned subject was placed in a group with three subjects outside the scanner in each round, and all subjects made decisions in 20 rounds, giving each scanner subject 60 opportunities to decide whether to send positive or negative tokens to group members of different incomes. Consistent with other studies conducted outside the scanner (5), the behavioral results show that participants were increasingly willing to pay to take money from those with the highest incomes and to give money to those with the lowest (Fig. 1).

Results

Based on results of other studies involving social inequality (11, 12), we focused on activations within the ventromedial prefrontal cortex (vmPFC), which includes the medial orbitofrontal cortex (13). Consistent with this earlier work (8), we find that activation in this area is associated with decision-making in the random-income game. However, we did not find a significant relationship between activation and measures of expressed and revealed egalitarian preferences elicited outside the scanner (Fig. 2).

In addition to the vmPFC, we focused on the insular cortex as a region of interest (ROI) because it has been shown to be an important neural substrate in a diverse set of experiments that involve the relationship between the individual and others. For example, unfair treatment by others in the ultimatum game appears to trigger insular activity (3), with higher activations corresponding to more frequent rejection of the unfair offers (12). Inequality aversion within different contextual situations has also been associated with activation in the insula in other ultimatum-game scenarios (14). Others have reported that individuals with personality disorders exhibit dysfunction in the insula that is associated with impaired trust and social cooperation (15). These and other results suggest that the insular cortex in general is critical for the perception of internal states (16), and that the anterior insula is particularly important for awareness of our own feelings as they relate to others (17), which may be especially relevant for empathy in decision-making situations (18). Insula activation has also been linked to volitional prosocial behavior induced by em-

Author contributions: C.T.D., P.J.L., D.S., A.N.S., T.F., R.M., J.H.F., and M.P.P. designed research; C.T.D., P.J.L., D.S., A.N.S., T.F., J.H.F., and M.P.P. performed research; C.T.D., P.J.L., D.S., A.N.S., J.H.F., and M.P.P. analyzed data; and C.T.D., P.J.L., D.S., A.N.S., S.E.B., J.H.F., and M.P.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: cdawes@nyu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1118653109/-/DCSupplemental.

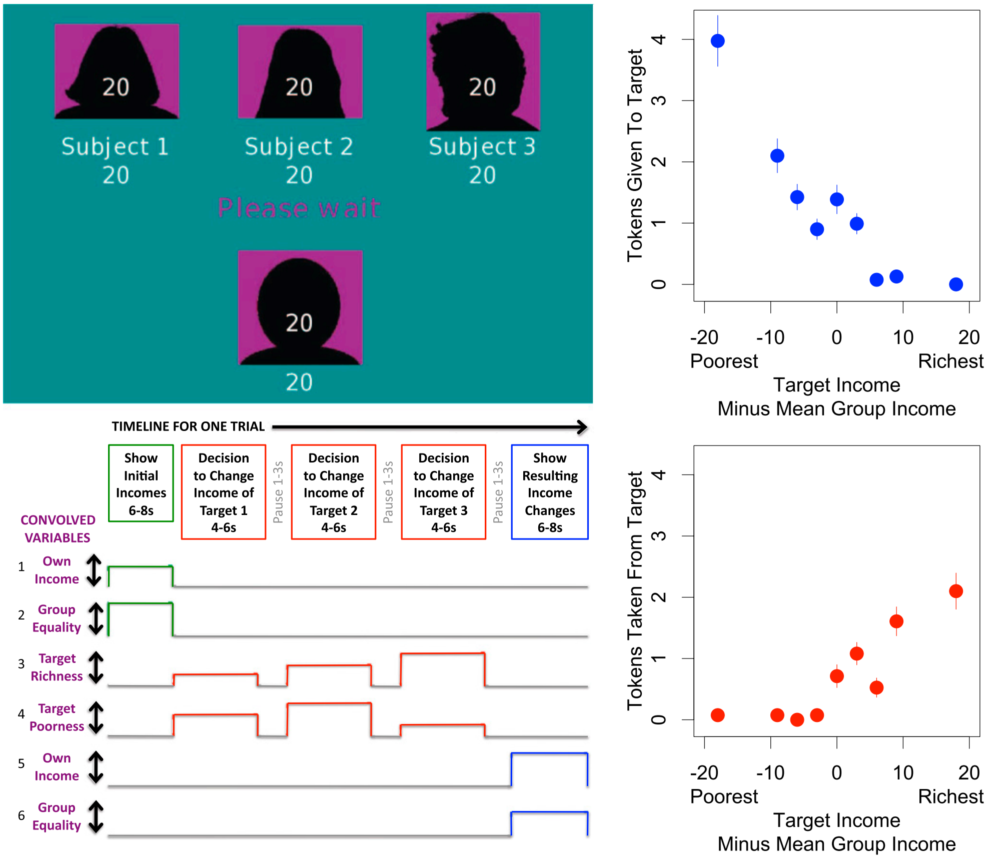

Fig. 1. Sample initial screenshot (*Upper Left*) shows own income at bottom and income of each other group member at top. During decision phases, income beneath silhouette changes to reflect incomes that will result from the subject's actions. The timeline for each trial (*Lower Left*) included an initial phase where randomly drawn incomes for each group member were shown, three decision phases in which the subject sees a target group member's income and then chooses whether or not to give or take tokens from the target, and an outcome phase where resulting incomes for each group member are shown. To identify ROIs, we regressed voxel activations on the convolution of six time-varying covariates with heights depending on initial own income, initial group equality, target "richness" (target income minus mean group income among those who earn more than average), target "poorness" (mean group income minus target income among those who earn less than average), change in one's own income, and change in group equality. Consistent with previously published work (1), scanner subjects tend to give tokens to low earners (*Upper Right*) and take away tokens from high earners (*Lower Right*). Vertical lines indicate SEMs.

pathy toward other's emotional or physical pain (19). Thus, we expected that this brain region would be a likely candidate for involvement in egalitarian behaviors.

Our neuroimaging results support the role of the insular cortex in egalitarian behavior. Specifically, we identified a region within the transition area of lateral inferior frontal gyrus and anterior insula (Fig. 2), where activations correlated significantly with changes in group equality when resulting incomes were shown at the end of each trial. These activations were also significantly associated with two additional measures of egalitarianism. First, the activations correlated with a widely used index of self-reported egalitarianism (Huber regression, $P = 0.009$) (Fig. 2, *Top*, *Center*). Second, the activations were significantly associated with egalitarian behavior measured by a series of dictator games (*SI Methods*) played after the scan session (Huber regression, $P = 0.05$) (Fig. 2, *Top*, *Right*).

Discussion

To summarize, this experiment shows that some parts of the brain are more active during egalitarian outcomes, and these activations are correlated with egalitarian behavior inside the scanner. However, a more crucial result is that the activations are also correlated with behavior outside the scanner, including self-reported preferences for egalitarian outcomes and game behavior that reveals how willing subjects are to use their own resources to obtain egalitarian outcomes within their groups. Taken together, the evidence suggests that the anterior insular cortex plays a critical role in egalitarian behavior in humans. This conclusion is consistent with a broader view of the insular cortex as a neural substrate (17) that processes the relationship of the individual with respect to his or her environment (19). The predominately left-lateralized activation may point toward the possibility of a positive valence or energy-preserving mode- (17) related processing during egalitarian behavior (i.e., individuals may see the group as a greater good that is worth preserving). The fact that the insula is directly involved in physiological, food, and pain-related processing supports the general notion that prosocial behavior, which is important for survival of both the individual and the group/species, is implemented on a fundamental physiological level similar to breathing, heartbeat, hunger, and pain.

Adam Smith (1) contended that moral sentiments like egalitarianism derived from a "fellow-feeling" that would increase with our level of sympathy for others, predicting not merely aversion to inequity, but also our propensity to engage in egalitarian behaviors. The evidence here supports such an interpretation. Although individuals may experience internal rewards when punishing antisocial behavior (6) and may have preferences for social equality (8), our results suggest that it is the brain mechanisms involved in experiencing the emotional and social states of self and others (17–21) that appear to be driving egalitarian behaviors.

Our results have important implications for theories of the evolution of prosocial behavior that suggest culturally transmitted "leveling mechanisms"—for example, food sharing and monogamy—

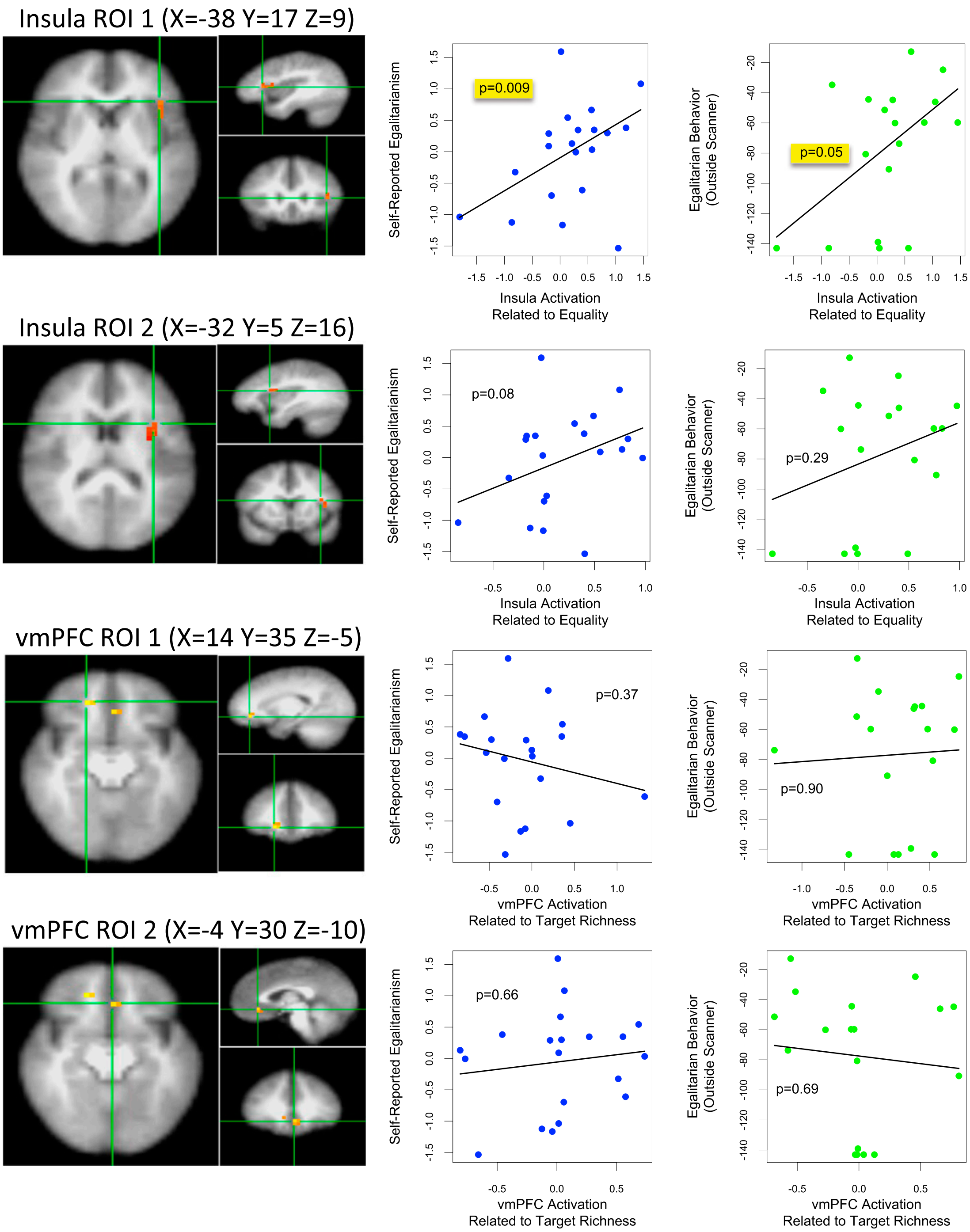

Fig. 2. ROIs in the insular cortex where activations correlated significantly with change in group equality when resulting incomes were shown at the end of each trial. Between subject activations in Insula ROI 1 predict a self-reported measure of egalitarianism (first row, *Center*). They also predict egalitarian behavior as measured in a series of dictator games with different multipliers (first row, *Right*). Insula ROI 2 (second row) shows the same pattern, but the relationships are not significant. Two ROIs in the vmPFC (third and fourth rows) show no relationship with egalitarianism measured outside the scanner. Lines and *P* values based on Huber regression.

stifle within-group competition and create circumstances in which intergroup antagonism generates selective pressure for altruistic behaviors (22). A concern for equality may have originally evolved because it fostered the conditions necessary for early human groups to maintain a high level of cooperation (23). Future research should focus on the interconnectivity of regions of the brain involved in egalitarianism and altruism to better understand how these two behaviors may have coevolved.

Methods

Stimulus/Task. To measure egalitarian behavior we use the random-income game (5). In this game, subjects are divided into groups of four anonymous members each. Each player receives a sum of monetary units (MUs) randomly assigned by a computer and each MU equals US $0.05. To maximize differences in initial group inequality we created three kinds of groups: a purely equal group (MUs for each group member = 20, 20, 20, and 20), a low-inequality group (MUs = 11, 17, 23, and 29), and a high-inequality group (MUs = 2, 14, 26, 38). Subjects are shown the payoffs of other group members for that round and are then provided an opportunity to give up to 10 "negative tokens" or "positive tokens" to other players. Each negative token costs 1 MU and decreases the payoff of a targeted individual by 3 MUs; positive tokens also cost 1 MU, but increase the targeted individual's payoff by 3 MUs. Groups are randomized after each round to prevent reputation from influencing decisions; interactions between players are strictly anonymous and subjects know this. Furthermore, by allowing participants more than one behavioral alternative, the experiment eliminates possible demand effects (24).

Subjects also completed a self-report survey that included a battery of items, including a standard set of questions that measure egalitarian preferences. At the conclusion of the study, subjects completed five rounds of the modified dictator game (25).

Subjects. All subjects were undergraduate social science majors at the University of California at San Diego and provided informed consent. Based on a pilot study, we chose individuals who bought more positive or negative tokens in relatively unequal conditions of the random-income game to complete an fMRI screening questionnaire. As a result, the subjects selected tend to be more egalitarian than those in the overall pilot sample. Although this selection process potentially makes the sample less representative overall, it increases the number of observations of egalitarian behavior and therefore improves the power of statistical tests. Subjects who were deemed eligible based on the screening questionnaire were invited to participate in the neuroimaging phase of the experiment. The final fMRI sample consisted of 10 males and 10 females that were all very similar in age.

Image Acquisition. fMRI data were collected at the University of California at San Diego on a 3T GE CXK4 Magnet with an eight-channel brain array coil to axially acquire T2*-weighted echo-planar images (EPI) (field-of-view 230 mm, 64 × 64 matrix; 30 2.6-mm thick slices; 1.4 mm gap; TR = 2,000 ms, TE = 32 ms, flip angle = 90°). The basic structural and functional image processing were conducted with the Analysis of Functional NeuroImages (AFNI) software package (26). All EPI images were aligned to the high-resolution anatomical images and resampled to a voxel size of 4 × 4 × 4 mm (from the original 3.7 × 3.7 × 4 mm). Data were temporally smoothed, spatially blurred with a 6-mm FWHM spatial filter, and normalized to Talairach space (via AFNI's auto Talairach program, followed by visual inspection of each structural image).

Data Analysis. A regression model was constructed for a contrast against the hypothesized voxel activation, based on a blood-oxygen level-dependent hemodynamic response function with 4- to 6-s peaks (Fig. 1). The model is composed of six time-varying convolved regressors with heights that vary based on certain aspects of the trial (Fig. 1). The parameters of the model were varied by round based on the different characteristics of each respective trial of the random-income game. The coefficient for each regressor was established using a standard general linear model.

The regression model consists of three parts that are designed to measure activation during each phase of the trial:

Introduction phase regressors. An "own income" regressor was computed as the initial income assigned to the subject by the computer, and a "group inequality" regressor was calculated as the initial SD of all four incomes in the group.

Decision phase regressors. A "target richness" regressor was defined as the target's income minus the group average income for the trial. Negative values for this regressor (where the target was poorer than average) were assigned a 0. Similarly a "target poorness" variable was defined as the group average income for the trial minus the target's income. Negative values (where the target was richer than average) were assigned a 0. These two regressors index the degree to which, during decision-making, the target is either doing better or worse than the group as a whole.

Outcome phase regressors. These regressors were used to indicate the change in individual and group incomes during the trial resulting from token purchases by all four group members. An "own income change" regressor measured the subject's final income minus his or her initial income, and a "group equality change" regressor measured the final SD of group incomes minus the initial SD.

A statistical model was used to relate changes in EPI intensity to differences in task characteristics (27). All slices of the EPI scans were temporally aligned following registration to ensure that different relationships with the regressors were not a result of the acquisition of different slices at different times during the repetition interval. EPIs were coregistered using a 3D-coregistration algorithm (28) that has been developed to minimize the amount of image translation and rotation relative to all other images. Six motion parameters were obtained for each subject. Three motion regressors (roll, pitch, and yaw), a linear trend, and constant were included in the model. Based on the fit to the regressors of interest, percent signal changes were calculated on the spatial blurred (6-mm Gaussian blur) normalized brains. Specifically, we used six convolved regressors, two for each phase of the task (Fig. 1). A threshold adjustment method based on Monte-Carlo simulations was applied to prevent identification of false-positive areas of activation (29). Based on these simulations, it was determined that a voxel-wise a priori probability of 0.01 (t = 2.552) would result in a corrected cluster-wise activation probability of 0.01 (one-sided) if a minimum volume 320 μL and a connectivity radius of 4.0 mm were to be considered.

Following the individual analysis, we determined how differences in behavior across subjects were related to brain activation differences that are indexed by coefficients on one of the six regressors. Specifically, we considered the behavior during the 20 rounds × 3 group members (60 total) possible decisions and computed two measures of egalitarian behavior: (*i*) a "take from rich" measure equal to the Pearson correlation between "target richness" and the amount of the change in the target income resulting from the subject's purchases of positive and negative tokens; (*ii*) a "give to poor" measure equal to the Pearson correlation between "target poorness" and the amount of the change in the target income resulting from the subject's purchases of positive and negative tokens. To identify ROIs, these measures were used in a multiple regression analysis with the behavior as the independent measures and the first level analysis as the dependent measure. These results were analyzed in R (30) for group effects and correlations. Using the "take from rich" measure, we identified four ROIs using "natomically constrained functional regions of interest" (31). For the insular cortex we constructed a probability mask. The probability mask was modeled after a well-documented atlas of various neurological structures. Briefly, to extract a mask for the insular cortex, we used Individual Brain Atlases using Statistical Parametric Mapping software (32), a toolbox for segmenting structural MR images. All programs in this toolbox are developed in MATLAB (33), based on a widely used neuroimaging software package, SPM (Wellcome Trust Centre for Neuroimaging, London, United Kingdom). This package uses the nonlinear registration and gray matter segmentation processes performed through SPM5 subroutines. The group insula mask was obtained by averaging across the individual insular masks and requiring that for the common insula mask each voxel had to cover the insula gray matter in at least 50% of all subjects.

Second-Level Analysis. All second-level analyses were conducted using the statistical programming language R (30) and with SPSS software, version 10. Specifically, a mixed-model analysis was conducted with the R procedure *lme*, which is part of the *nlme* library. The fixed effects were decision type, the random effects were subjects. Moreover, we conducted voxel-wise multiple linear regression analyses with performance during the task as independent measures, and the percent signal change during the different decision-making phases as the dependent measure using the *lm* procedure in R. Finally, we conducted a between-subjects analysis of activations in the identified ROIs (as measured by coefficients on the regressors) and measures of self-reported and behavioral egalitarianism outside the scanner using the *rlm* procedure in R.

ACKNOWLEDGMENTS. J.H.F. thanks the University of California at San Diego Committee on Research for generous research support.

1. Smith A (1759) *The Theory of Moral Sentiments* (Printed for A. Millar and A. Kincaid and J. Bell, London).
2. Henrich JP, et al. (2004) *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies* (Oxford Univ Press, Oxford, New York).
3. Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the Ultimatum Game. *Science* 300:1755–1758.
4. Brosnan SF, De Waal FB (2003) Monkeys reject unequal pay. *Nature* 425:297–299.
5. Dawes CT, Fowler JH, Johnson T, McElreath R, Smirnov O (2007) Egalitarian motives in humans. *Nature* 446:794–796.
6. de Quervain DJ, et al. (2004) The neural basis of altruistic punishment. *Science* 305: 1254–1258.
7. Harbaugh WT, Mayr U, Burghart DR (2007) Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science* 316:1622–1625.
8. Tricomi E, Rangel A, Camerer CF, O'Doherty JP (2010) Neural evidence for inequality-averse social preferences. *Nature* 463:1089–1091.
9. Johnson T, et al. (2009) The role of egalitarian motives in altruistic punishment. *Econ Lett* 102:192.
10. Fehr E, Gächter S (2002) Altruistic punishment in humans. *Nature* 415:137–140.
11. Fliessbach K, et al. (2007) Social comparison affects reward-related brain activity in the human ventral striatum. *Science* 318:1305–1308.
12. Tabibnia G, Satpute AB, Lieberman D (2008) The sunny side of fairness: Preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychol Sci* 19:339–347.

13. Lacerda AL, Hardan AY, Yorbik O, Keshavan MS (2003) Measurement of the orbitofrontal cortex: A validation study of a new method. *Neuroimage* 19:665–673.
14. Wright ND, Symmonds M, Fleming SM, Dolan RJ (2011) Neural segregation of objective and contextual aspects of fairness. *J Neurosci* 31:5244–5252.
15. King-Casas B, et al. (2008) The rupture and repair of cooperation in borderline personality disorder. *Science* 321:806–810.
16. Craig AD (2002) How do you feel? Interoception: The sense of the physiological condition of the body. *Nat Rev Neurosci* 3:655–666.
17. Craig AD (2009) How do you feel—now? The anterior insula and human awareness. *Nat Rev Neurosci* 10:59–70.
18. Singer T, Critchley HD, Preuschoff K (2009) A common role of insula in feelings, empathy and uncertainty. *Trends Cogn Sci* 13:334–340.
19. Masten CL, Morelli SA, Eisenberger NI (2011) An fMRI investigation of empathy for 'social pain' and subsequent prosocial behavior. *Neuroimage* 55:381–388.
20. Dewall CN, et al. (2010) Acetaminophen reduces social pain: Behavioral and neural evidence. *Psychol Sci* 21:931–937.
21. Immordino-Yang MH, McColl A, Damasio H, Damasio A (2009) Neural correlates of admiration and compassion. *Proc Natl Acad Sci USA* 106:8021–8026.
22. Bowles S (2006) Group competition, reproductive leveling, and the evolution of human altruism. *Science* 314:1569–1572.
23. Boehm C (1999) *Hierarchy in the Forest* (Harvard, Cambridge).
24. Orne MT (1962) On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *Am Psychol* 17:776.
25. Andreoni J, Miller J (2002) Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica* 70:737—753.
26. Cox RW (1996) AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res* 29:162–173.
27. Haxby JV, Hoffman EA, Gobbini MI (2000) The distributed human neural system for face perception. *Trends Cogn Sci* 4:223–233.
28. Eddy WF, Fitzgerald M, Noll DC (1996) Improved image registration by using Fourier interpolation. *Magn Reson Med* 36:923–931.
29. Forman SD, et al. (1995) Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magn Reson Med* 33:636–647.
30. R Development Core Team (2011) R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna). Available at http://cran.r-project.org/.
31. Johnstone T, et al. (2005) Stability of amygdala BOLD response to fearful faces over multiple scan sessions. *Neuroimage* 25:1112–1123.
32. Alemán-Gómez Y, Melie-García L, Valdés-Hernandez, P (2006) IBASPM: Toolbox for automatic parcellation of brain structures. Available on CD-Rom in *NeuroImage* 31(Suppl 1). Available at http://www.thomaskoenig.ch/Lester/ibaspm.htm.
33. MATLAB version 7.10.0. (The MathWorks Inc., Natick, MA). Available at http://www.mathworks.com.