Neural signatures of strategic types in a two-person bargaining game

Meghana A. Bhatt[a], Terry Lohrenz[a], Colin F. Camerer[b], and P. Read Montague[a,c,1]

[a]Department of Neuroscience, Baylor College of Medicine, Houston, TX 77030; [b]Divisions of the Humanities and Social Science and Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125; and [c]Department of Psychiatry, Baylor College of Medicine, Houston, TX 77030

Edited by Terrence J. Sejnowski, The Salk Institute for Biological Studies, La Jolla, CA, and approved September 28, 2010 (received for review July 6, 2010)

The management and manipulation of our own social image in the minds of others requires difficult and poorly understood computations. One computation useful in social image management is strategic deception: our ability and willingness to manipulate other people's beliefs about ourselves for gain. We used an interpersonal bargaining game to probe the capacity of players to manage their partner's beliefs about them. This probe parsed the group of subjects into three behavioral types according to their revealed level of strategic deception; these types were also distinguished by neural data measured during the game. The most deceptive subjects emitted behavioral signals that mimicked a more benign behavioral type, and their brains showed differential activation in right dorsolateral prefrontal cortex and left Brodmann area 10 at the time of this deception. In addition, strategic types showed a significant correlation between activation in the right temporoparietal junction and expected payoff that was absent in the other groups. The neurobehavioral types identified by the game raise the possibility of identifying quantitative biomarkers for the capacity to manipulate and maintain a social image in another person's mind.

decision making | individual differences | neuroeconomics

What do I think about you? What do I think you think about me? These basic assessments, underlying human social exchange, constitute crucial computations that all human brains must carry out if they are to navigate the complexities of social life. In larger-scale societies, survival and success hinge on the capacity to calibrate and monitor one's social image in the minds of others. Consequently, the question of how social signals manipulate the minds of others around us poses one of the central and most difficult computational problems underlying all social transactions. Until recently, quantitative neurobehavioral approaches to this problem have been absent.

The management of social image represents a refinement of the more thoroughly studied problem of "theory of mind" (1–4). Rather than simply modeling the goals and behavior of others, managing a social image requires that we understand that others also have a theory of our mind along a variety of cognitive dimensions, and therefore they maintain models of our own goals and behavior. The "second-order" belief problem of understanding others' perceptions of our self is at the heart of any strategic interaction, ranging from a simple game of cards to a complex business negotiation. To manipulate our own reputation in the mind of another agent requires that we estimate the depth to which our partner models our own mental state and the dimensions along which such modeling is likely to occur. These computations are particularly difficult because estimating another individual's model of oneself is an inherently recursive process—it requires keeping in mind my model of you, my model of your model of me, and so on (5–7).

One important case of a second-order belief computation is strategic deception, the manipulation of another person's beliefs about one's own goals or actions for the purpose of personal gain. For example, when someone bluffs in a poker game she attempts to change her opponents' beliefs by mimicking the behavior of someone with a winning hand. Note that bluffing inherently requires maintaining a belief about what is true (one's own cards) and what another player believes on the basis of your actions, even when the latter belief is hoped to be false. To carry this deception out successfully, the player must accurately model the way in which her behavior will affect her opponents' perception of the hidden variables in the game (the player's own cards), which in turn requires her to have a sufficiently accurate model of her opponent's model of her own behavior: how does this opponent believe you might act with a winning hand? Anyone who has played poker knows that this model will often include not only factors based on the structure of the game but also an understanding of how your own play in previous hands may have changed your opponent's beliefs about what type of player you are: are you the type to bluff? Although bluffing is one of the more clear-cut examples of the need to compute second-order beliefs, these computations are at the heart of almost any strategic interaction. However, not everyone is a good poker player. The ability to make accurate strategic computations and act upon them seems to vary greatly by person and context. What makes one person more strategic than another? Are there neural signatures of such differences?

A number of brain regions have been identified as parts of a possible "theory of mind" network. In addition, many areas known for more general tasks are often implicated in complex social decision making. For example, the dorsolateral prefrontal cortex (DLPFC), generally active in tasks involving cognitive control and complex decision making, has been implicated in social and theory of mind-related tasks (8, 9). The rostral prefrontal cortex [Brodmann area 10 (BA10)] has been implicated in a host of computations from mentalizing to goal maintenance (10–12). However, the precise computations executed at these loci remain under debate. Although medial prefrontal cortex seems to be modulated by deliberation and depth of theory of mind in certain games (13, 14), it notably fails to correlate with steps of reasoning in others (15). Similarly, the temporoparietal junction (TPJ) has been posited as a locus contributing to the maintenance and understanding of other people's beliefs (16–18) but has also been implicated in a host of nontheory of mind tasks, particularly those involving attentional reorienting (19, 20).

We take a formal approach to this problem by assaying the basic computational components of these second-order beliefs to probe their neural underpinnings. To do this, we use a modified version of a sender–receiver game (21): a bargaining task between two subjects. Strategic deception in this game requires the ability to maintain and update another person's beliefs, indicating the possible involvement of the TPJ. In addition, the execution of

Author contributions: M.A.B., C.F.C., and P.R.M. designed research; M.A.B. and T.L. performed research; P.R.M. contributed new reagents/analytic tools; M.A.B. and T.L. analyzed data; and M.A.B., T.L., C.F.C., and P.R.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

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

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1009625107/-/DCSupplemental.

sophisticated strategic deception also involves the execution of a long-term strategy, which requires prospective thinking, active goal maintenance, and cognitive control—suggesting the involvement of BA10 and DLPFC.

In this game, the "no feedback bargaining task," two players, a buyer and a seller, play 60 rounds of a bargaining task (Fig. 1). At the beginning of each round the "buyer" is informed of her private value v of a hypothetical object. She is then asked to "suggest a price" to the seller (values and prices are integers, 1–10). The seller then receives this suggestion and is asked to set a price p. If the seller's price is less than the private value v (which is known only to the buyer), the trade executes and the seller receives p, whereas the buyer receives $v - p$, the difference between the private value and the selling price. If the seller's price exceeds the buyer's value, the trade does not execute and both parties receive nothing. No feedback about whether the trade occurred is provided to either player.

The tradeable object has no value to either player if a trade does not occur. However, if a trade does occur, each player prefers a sale price that favors them. Buyers prefer lower prices and sellers prefer higher prices. This misalignment of incentives implies that the only equilibrium solution of the one-round version of this game is for no information transfer to occur (21). The buyer should "babble" and send suggestions with no informative relationship to her private value, and the seller should ignore this suggestion and set a price of either 5 or 6 (to maximize the expected revenue). However, this is the mutually optimal solution only if both players believe that the other is also playing in equilibrium. That is, babbling is only optimal if the seller is in fact ignoring buyer suggestions, and ignoring buyer suggestions is only optimal if they contain no meaningful information. In actuality, in these types of games players' beliefs about what others are actually likely to do are often not accurate (i.e., they are out of equilibrium). Therefore, descriptive models of belief formation and adjustment will be more involved than the simple equilibrium ones (5, 6, 21, 22). Models of this cognitive hierarchy type predict the existence of different behavioral types on the basis of the depth to which they model their opponent.

Results

Behavioral Results. Given these conditions, how do buyers actually play this game? Simple buyer strategies can be detected by regressing the buyer's private value v against their suggestion s sent to the seller. We restricted our analysis to the second half of the experiment to allow strategies to stabilize. From these linear regressions, we extracted two behavioral descriptors, the slope and the R^2 of the regression, and used these to cluster buyers into types (Fig. 2). The slope of this regression tracks roughly with the credibility of the buyer's suggestion (i.e., this slope tracks with how "good" the information contained in the suggestions is). If the slope is high, sellers should trust the information, and conversely if the slope is near zero, suggestions contain no information. In the interesting case in which the slope is negative, suggestions are actively misleading. Thus, we refer to this slope as a buyer's "information revelation" coefficient (IR).

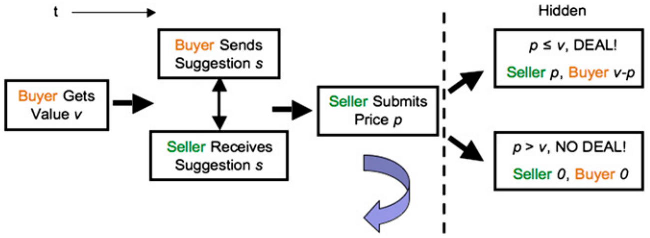

Fig. 1. Experimental task. At the beginning of each round the computer assigns a value for the widget to the buyer. The buyer "suggests a price" to the seller, who uses this information to set a final price for the object. The computer automates whether the deal occurs—if the price is less than or equal to the buyer's value, the seller receives the price, p, and the buyer receives the difference between the price and his private value, $v - p$. Otherwise, the deal fails and neither party receives anything. Neither party is informed of the outcome of the previous trial; payoffs are just added to a running tally of points kept by the computer.

The buyers fall into three distinct clusters: the "incrementalist" group ($n = 32$, blue) is characterized by a relatively high IR and high fit (large R^2). They are relatively honest with their price suggestions (the group mean slope is 0.57, consistent with suggesting prices equal to approximately half of the buyer's value, possibly to share the gains from trade equally). The "conservative" group ($n = 28$, green) generally show IRs close to zero and intermediate or low fit. Their suggestions may still contain information about the underlying value, but not much. Some of these actually had constant strategies and always sent a suggestion of 1. The third group, the strategic deceivers, or "strategists" ($n = 16$, red), are the most interesting. These buyers send suggestions that are negatively correlated with their private value. Because there is no feedback to either player after each trade, these strategic types have surmised that as long as they send a sequence of suggestions that mimics an incrementalist type, they should be able to make higher profits. For example, if they receive a value of 2, they will forego an immediate profit (which would be low at best) and send a high suggestion (for example, 8). Then when they get a high value, they can credibly send a low suggestion and reap a high profit from an unsuspecting seller. In addition to this data-driven clustering, we developed and estimated a model of belief formation, described in *SI Methods*, that predicts the existence of these three types of players independently, with each type possessing different depths or "levels" of theory of mind. The most sophisticated of these, the ones who reasoned most deeply about their opponents, should exhibit a negative IR. We designated these as "level-2" players. When we estimated this model on our subjects, we found that 14 of our 16 strategists were correctly classified as level-2 players. These level-2 subjects are designated by the triangles in Fig. 2*A*.

We assessed intelligence quotient (IQ) in 30 of our 76 subjects (11 incrementalists, 9 conservatives, and 10 strategists) and found that although incrementalists IQs were suggestively lower than conservative and strategist IQs, there were no significant differences among the three groups using a one-way ANOVA. More importantly, there was significant overlap among the three distributions, and there was no significant difference between conservative and strategist IQs. This shows that IQ alone does not account for the differences in behavior, and although above-average IQ seems to be a necessary condition for strategist behavior, it is not sufficient (Fig. 3*A*). We also assessed socioeconomic status in 65 of our 76 subjects and found no significant differences among the three groups (Fig. 3*B*). Overall, earnings were significantly lower in the incrementalists group than in the other two. Although there was no significant difference in mean earnings between conservatives and strategists, conservative earnings fell over a larger range, including many of the lowest and highest earnings overall (Fig. 3*C*). This is consistent with the cognitive hierarchy model because the superiority of the strategist or conservative approach to the game is dependent on the sophistication of the seller (i.e., the conservative approach does well against a credulous seller but does poorly against a more sophisticated, "level-1" seller).

Subject debriefing in the form of a free-response question administered after the experiment confirmed that the strategic deceivers were aware and deliberate in mimicking a distribution of suggestions that might be expected from a more truthful individual. One subject wrote "I tried to *throw off* [the] seller by saying the low things were high. . ." This comment, and other

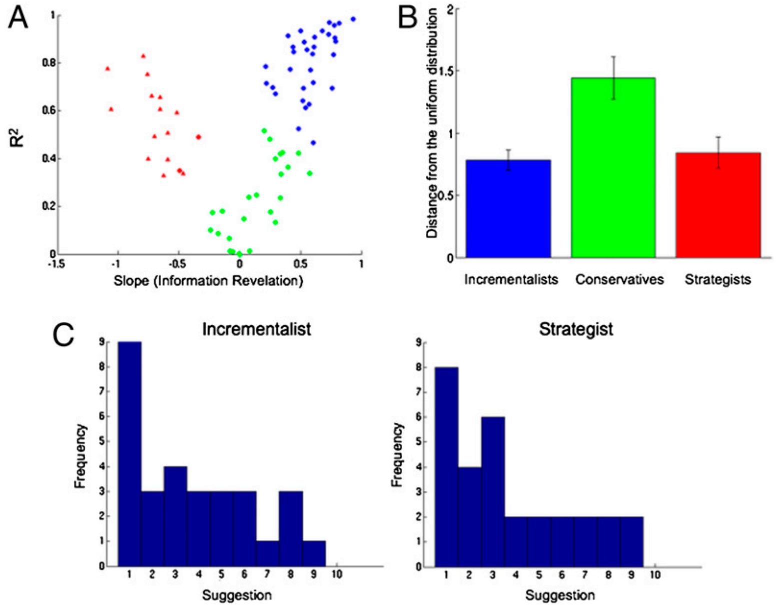

Fig. 2. Behavioral analysis. (*A*) Behavioral clustering in buyers. Incrementalists (blue) send suggestions that are highly correlated with their true value. Strategists (red) send suggestions that are negatively correlated with value. Strategists appear similar to incrementalists and thus reap the surplus from high-value trials. Conservative buyers (green) play closest to an economically rational actor and reveal no information about their value with their suggestions. Triangles indicate subjects who were classified as sophisticated "level-2" buyers according to a generative model. (*B*) Mean Kullback-Leibler (KL) distances of the players' choice distribution from the uniform distribution. Incrementalists and strategists are both significantly closer to the uniform distribution than conservatives but are not significantly different from each other. (*C*) Histograms showing suggestion frequencies for a single incrementalist (*Left*) and a single strategist (*Right*). Note that from the perspective of the seller the two are indistinguishable.

similar comments, showed a conscious and sophisticated model of how their suggestions might be processed by the seller over time. In contrast, the conservatives often simply stated that they always desired lower prices and therefore sent low suggestions, reflecting a simpler model of seller behavior. Finally, incrementalists tended to be more vague in their descriptions of their strategies than the other two groups.

Functional MRI Results. To probe the neural underpinnings of strategic behavior in buyers, we performed two sets of analyses. First, we performed between-group (defined by the behavioral clustering) comparisons of neural activity at various trial epochs. Second, we regressed buyers' neural activity against the buyer's IR coefficient at those same epochs. As with the behavioral data, we restricted our analysis of the functional MRI (fMRI) data to the second half of the experiment to allow strategies to stabilize.

Between-group comparisons over the individual subject first-level boxcar regressors over the entire trial (onset to decision) revealed two significant main effects of behavioral type that survive correction for multiple comparisons at the $P < 0.05$ level (either corrected for familywise error over gray matter at peak voxel, or cluster-level correction at $P < 0.001$, $k > 5$). First,

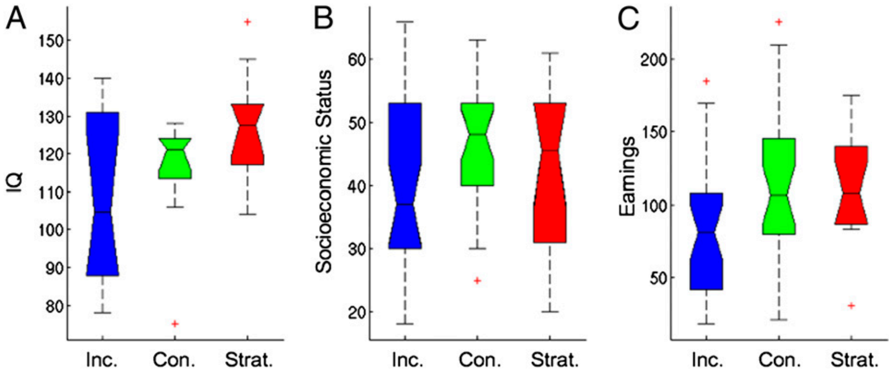

Fig. 3. Group differences were not explained by differences in IQ or socioeconomic status. (*A*) Although incrementalists had suggestively lower IQs than strategists or conservatives, this was not quite significant ($P = 0.07$, one-way ANOVA). However, both conservative and strategist IQs were significantly higher than average, whereas incrementalist IQs were not (11 incrementalists, 9 conservatives, and 10 strategists came back to take the IQ test); there was no significant difference between strategist and conservative IQs. (*B*) We also assessed socioeconomic status on 65 subjects using income, occupation, and education level and found there were no significant differences among the three groups according to one-way ANOVA. (*C*) Subject earnings by behavioral type: incrementalists had significantly lower final payments for the experiment ($P = 0.02$, one-way ANOVA); there was no significant difference in means between conservative and strategist earnings. Colored sections of the box-plots indicate the interquartile interval of the data, and whiskers show the total data extent excepting outliers, which are shown as red crosses. The black lines indicate the median data, and the notched section of the box gives a 95% confidence interval for the median.

strategists show a differentially higher activation in left rostral prefrontal cortex (BA10) shown in Fig. 4*A*, whereas incrementalists show differentially lower activation in the right DLPFC (rDLPFC) (Fig. 4*B*). In addition to these main effects, we found a significant ($P < 0.005$ at peak voxel, corrected for familywise error over gray matter) group × value (individual subject βs for the value at onset regressor) interaction in right TPJ (rTPJ) (Fig. 5). All three of these activations were significant at the $P < 0.001$ (uncorrected) level in the secondary analysis using IR as a continuous between-subject regressor. Both the rDLPFC and rTPJ activations also survived correction for multiple comparisons in this secondary analysis. Full statistics for both analyses are reported in *SI Methods*.

To further understand the nature of these activations we examined the time series of activity across the three groups in the identified regions. Confirming the whole-brain analysis, strategist activity in BA10 (Fig. 4*A*, *Right*) is significantly greater than for both conservatives and incrementalists, whereas conservative and incrementalist time courses were essentially identical. On the other hand, in the rDLPFC (Fig. 4*B*, *Right*), time course analysis reveals that although the area was characterized by decreased activation in the incrementalists, conservatives show an intermediate level of activation, between the activities of the incrementalists and the strategists. Finally, in the rTPJ (Fig. 5, *Right* panel), strategists showed a strong relationship between activation and value, which was absent in the other two groups.

Discussion

From a neural standpoint our understanding of social interactions is in its infancy. The present study attempts to shed light on an important aspect of social interaction: the understanding and manipulation of others' perceptions of us. Second-order belief formation is particularly interesting because there seems to be a wide range of abilities within the scope of normal human behavior. Indeed, in this simple bargaining task we used task behavior to uncover three distinct clusters of buyers. Further, fMRI revealed distinct neural correlates associated with these clusters.

To display strategic deception, subjects had to be able to consider the implications of current decisions on future payoffs and especially consider the counterfactual situation of what might happen if they chose the conservative strategy and engendered too much suspicion in the seller. This also requires the maintenance and continual updating of the "false beliefs" of their opponent, as well as cognitive control mechanisms necessary to inhibit the impulse to transfer information. In contrast, the conservatives only need to inhibit information transfer, and the incrementalists' naïve strategy simply anchors suggestions directly to the true value.

TPJ has been found repeatedly to be active in theory of mind tasks, particularly in the attribution of false or incongruent beliefs to another person (16, 19, 23). It is interesting that rTPJ activity is modulated by value rather than simply being more active in strategists, as might be expected. Saxe and Wexler (17) found that signals in the rTPJ were modulated by the degree of incongruence among multiple facts known about a target's mind. This finding shapes our interpretation of the modulation of activity in rTPJ by value in strategists: it is during the high-value trials that the strategists' bluff really matters. Even though strategists are deceiving during both high- and low-value trials, and their suggestions are always incongruent with their true value, the payoff only comes during the high-value trials. Additionally, strategists are effectively switching between two modes of behavior: reputation building, which occurs during low-value trials, and reward-collection, which occurs during high-value trials. This switch between attention to one's reputation and attention to one's actual payoffs is an example of attentional reorienting that has been associated with activity in TPJ (20). This activation also bears striking similarities to a nearby activation in superior temporal sulcus (STS) found by Hampton et al. [The reported peaks are close together, but do not appear to overlap: (52, −48, 20) for right TPJ in this study; (60, −54, 9) for STS in Hampton et al.] (13). In this article,

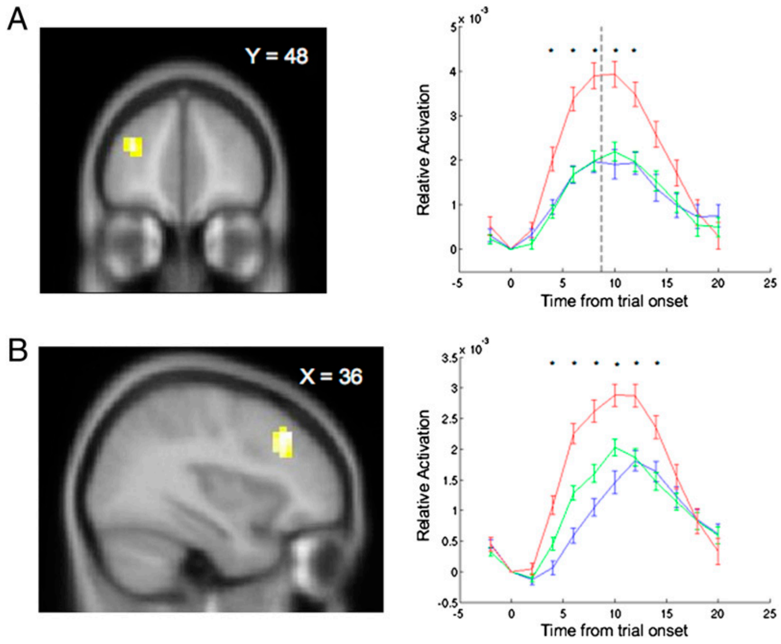

Fig. 4. Strategists differentially activate rDLPFC and left BA10. (*A*) Between-group analysis: strategist activation over the entire trial vs. other groups. *Left:* Left BA10. Peak voxel at (−32, 48, 20), $k = 14$, $t = 4.72$, $P = 0.049$ at peak voxel (corrected for familywise error over gray matter). *Right:* Time courses in BA10 by group. (*B*) Between-group analysis: incrementalist activation over the entire trial vs. other groups. *Left:* rDLPFC. Peak voxel at (36, 28, 36), $k = 27$, $t = 4.62$ at peak voxel, cluster-level $P = 0.044$ (corrected). *Right:* Time courses in rDLPFC by group. For both regions clusters are shown at $P < 0.001$, uncorrected. Cluster extents and cluster-level *P*s are reported at this threshold as well. Full statistics are reported in *SI Methods*. For time courses, all data are normalized to trial onset, dotted black line indicates average decision time, and asterisks indicate significance of the one-way ANOVA on activation at peristimilus time at the $P < 0.01$ level.

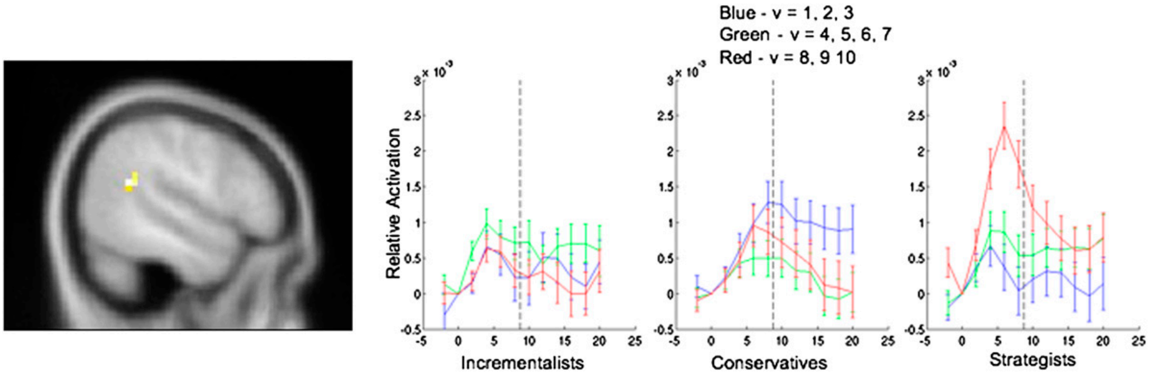

Fig. 5. Value modulates rTPJ activation only in the strategists. Between-group analysis: interaction of activation at trial onset with value, incrementalists vs. other groups. *Left:* rTPJ. Peak voxel at (52, −48, 20), $k = 10$, $t = 5.41$, $P = 0.004$ at peak voxel (corrected for familywise error over gray matter). Full statistics are reported in *SI Methods*. *Right:* Time courses in rTPJ by group.

bilateral STS activation was correlated with an "influence" parameter in a behavioral model. Interestingly, this parameter correlated with both expected reward and strategic switching in their task, much as high-value trials correlate to expected return and strategic switching in the strategists.

BA10 has been implicated in long-term goal maintenance and the use of prospective memory (10, 24), vital aspects of the strategists' forward-looking behavior. Burgess et al. propose a medial/lateral functional mapping of this region according to whether mental processing is stimulus oriented or stimulus independent, with the latter being associated with more lateral activations. This is consistent with our interpretation of the relatively lateral activation ($x = -32$) in BA10 as corresponding to the need for prospective thinking and goal maintenance in the strategist approach. As mentioned above, unlike incrementalists or conservatives, strategists have a distinct intermediate goal in the pursuit of reward: reputation building. The maintenance of the relatively long-term goal as a means to greater overall and future rewards is consistent with the stimulus-independent processing attributed to the area. The relative lack of activation at this locus for the other two groups (as well as the similarity of activity levels between the two other groups) accurately reflects the leap in task complexity between the conservative and strategist approaches.

On the other hand, rDLPFC shows a more continuous relationship between activation and strategic sophistication. Of the three areas highlighted here, it is the only one where the analysis using IR as a continuous between-subject regressor yielded a larger activation than the between-group analysis ($k = 31$ vs. $k = 27$; *SI Methods* provides full statistics on both analyses). The area has been consistently implicated in tasks involving working memory and cognitive control (25, 26). Both of these are functions that should be used by strategists, who must keep track of their previous suggestions to infer what their reputation is with the seller and inhibit the impulse to simply anchor their suggestion on their true value. The cognitive control function of rDLPFC in this task is further highlighted by the fact that conservatives also have elevated activity in the area as compared with incrementalists. One transcranial magnetic stimulation (TMS) study showing that disruption in rDLPFC reduces intertemporal building of a trustworthy reputation (but only when other people are highly trusting) is consistent with this cognitive control function of rDLPFC in strategizing (27).

The patterns of activity uncovered in the strategists uncover a set of regions involved in the successful manipulation of others' beliefs over time. BA10 is strongly recruited in strategists but not in incrementalists or conservatives. rDLPFC is recruited strongly by strategists, with some variance by conservatives, and weakly by incrementalists. Finally, the rTPJ, important for the attribution of false or incongruent beliefs to others and attentional reorienting, is strongly activated during strategist bluffing.

Human strategic thinking is both complicated and highly adaptive, driven by the coevolution of complex artificial socioeconomic environments and the mind's capacity to navigate those environments. Earlier studies have established neural correlates of the capacity to reason about other agents' likely behavior (14, 28) and the ability to learn from social information (29). Our study goes an important step further, by shedding light on how some agents make choices to manipulate other agents' perceptions of their own strategies. The pairing of the behavioral and neural data strongly suggests that strategists are guided in their deceit by the more schematic, forward-looking computations of BA10 and TPJ, in concert with heightened memory and control provided by DLPFC. It remains to be seen how a given individual finds herself in one group or the other: is strategic ability inherent, or can we train individuals to more easily identify strategic solutions by emphasizing the use of schematic representations and counterfactual analysis? Is strategic ability context dependent? Whatever the case, opportunities for strategic deception of this sort are possible only because of the existence, and in fact likely relative prevalence, of people with the tendency to be honest even when such honesty is not in their interest. However, in our admittedly circumscribed situation it is clear that there are three distinct classes of individuals who approach this strategic interaction in completely different ways and that these differences are manifest in qualitatively different neural signatures.

Our results suggest a method of understanding and quantifying individual differences—as clusters of behavior in an economic game (9)—and point to applications for the definition and diagnosis of mental disorders. Economic games can provide objective quantitative measures of strategic thinking (as in this study), social preferences (30, 31), risk preferences (32, 33), and a host of other potentially interesting characteristics. A better understanding of the range and joint distributions of these factors in the population could provide insight into those individuals who fall at the extremes of these distributions (i.e., those with mental disorders).

Methods

We regressed buyers' suggestions on their private values over the second half of the experiment, yielding three descriptive strategy parameters for each buyer—the slope, intercept, and fit (R^2). We normalized these three statistics across subjects by subtracting means and dividing by SDs. Clusters were identified using the k-means algorithm (34). The clusters did not change significantly when intercepts were excluded, therefore the results in the text are clustered using only slope and fit.

fMRI data were collected using 3-T Siemens scanners on 76 healthy subjects recruited in accordance with a protocol approved by the Baylor College of Medicine Institutional Review Board. High-resolution T1-weighted scans were acquired using a magnetization prepared rapid gradient echo sequence. Functional images were acquired with a repetition time of 2,000 ms and echo time of 25 ms. Thirty-seven 4-mm slices were acquired 30° off the anteroposterior commissural line, yielding functional voxels that were 3.4 mm × 3.4 mm × 4 mm.

Data were preprocessed using SPM2 algorithms (http://www.fil.ion.ucl.ac.uk/spm/software/spm2/) for slice–timing correction, motion correction, coregistration, gray/white matter segmentation, and normalization to the Montreal Neurological Institute template. Functional images were smoothed spatially using an 8-mm Gaussian kernel. All data were high-pass filtered (128 s); the regression error structure was assumed to be AR(1). Post-preprocessing voxels were 4 mm × 4 mm × 4 mm.

We considered two general linear models (GLMs). Key presses, head motion, and time derivatives were included as nuisance regressors in both models. The first model used separate point regressors at trial onset and decision, parameterized by value and suggestion, respectively. A nuisance regressor for selector appearance was also included. The second model used a boxcar regressor beginning at trial onset and ending at decision, parameterized by value and suggestion. Both analyses used separate regressors for events in the early (first 30) as opposed to late (second 30) trials of the experiment. Regressors were convolved with the standard hemodynamic response function. A gray matter mask was used for the second-level analysis, excluding voxels that were less than 40% likely to be gray matter.

After regions of interest were identified from the whole-brain GLMs time series were extracted in each cluster and averaged to produce time courses anchored to events of interest.

1. Frith C, Frith U (2005) Theory of mind. *Curr Biol* 15:R644–R646.
2. Amodio DM, Frith CD (2006) Meeting of minds: The medial frontal cortex and social cognition. *Nat Rev Neurosci* 7:268–277.
3. Keysers C, Gazzola V (2007) Integrating simulation and theory of mind: From self to social cognition. *Trends Cogn Sci* 11:194–196.
4. Tennie C, Frith U, Frith CD (2010) Reputation management in the age of the world-wide web. *Trends Cogn Sci*, 10.1016/j.tics.2010.07.003.
5. Camerer CF, Ho T, Chong J (2004) A cognitive hierarchy model of games. *Q J Econ* 119: 861–898.
6. Nagel R (1995) Unraveling in guessing games: An experimental study. *Am Econ Rev* 85:1313–1326.
7. Stahl D (1995) On players' models of other players: Theory and experimental evidence. *Games Econ Behav* 10:218–254.
8. Knoch D, Pascual-Leone A, Meyer K, Treyer V, Fehr E (2006) Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* 314:829–832.
9. Yoshida W, et al. (2010) Cooperation and heterogeneity of the autistic mind. *J Neurosci* 30:8815–8818.
10. Burgess PW, Dumontheil I, Gilbert SJ (2007) The gateway hypothesis of rostral prefrontal cortex (area 10) function. *Trends Cogn Sci* 11:290–298.
11. Burgess PW, Gilbert SJ, Dumontheil I (2007) Function and localization within rostral prefrontal cortex (area 10). *Philos Trans R Soc Lond B Biol Sci* 362:887–899.
12. Ramnani N, Owen AM (2004) Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. *Nat Rev Neurosci* 5:184–194.
13. Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proc Natl Acad Sci USA* 105:6741–6746.
14. Coricelli G, Nagel R (2009) Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proc Natl Acad Sci USA* 106:9163–9168.
15. Kuo WJ, Sjöström T, Chen YP, Wang YH, Huang CY (2009) Intuition and deliberation: Two systems for strategizing in the brain. *Science* 324:519–522.
16. Saxe R, Kanwisher N (2003) People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage* 19:1835–1842.
17. Saxe R, Wexler A (2005) Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia* 43:1391–1399.
18. Saxe R, Moran JM, Scholz J, Gabrieli J (2006) Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Soc Cogn Affect Neurosci* 1:229–234.
19. Decety J, Lamm C (2007) The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *Neuroscientist* 13:580–593.
20. Mitchell JP (2008) Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cereb Cortex* 18:262–271.
21. Crawford VP, Sobel J (1982) Strategic information transmission. *Econometrica* 50: 1431–1451.
22. Costa-Gomes M, Crawford VP, Broseta B (2001) Cognition and behavior in normal-form games: An experimental study. *Econometrica* 69:1193–1235.
23. Gobbini MI, Koralek AC, Bryan RE, Montgomery KJ, Haxby JV (2007) Two takes on the social brain: A comparison of theory of mind tasks. *J Cogn Neurosci* 19:1803–1814.
24. Reynolds JR, West R, Braver T (2009) Distinct neural circuits support transient and sustained processes in prospective memory and working memory. *Cereb Cortex* 19: 1208–1221.
25. MacDonald AW (2000) Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science* 288:1835–1838.
26. Mansouri FA, Tanaka K, Buckley MJ (2009) Conflict-induced behavioural adjustment: A clue to the executive functions of the prefrontal cortex. *Nat Rev Neurosci* 10: 141–152.
27. Knoch D, Schneider F, Schunk D, Hohmann M, Fehr E (2009) Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. *Proc Natl Acad Sci USA* 106:20895–20899.
28. Bhatt M, Camerer CF (2005) Self-referential thinking and equilibrium as states of mind in games: FMRI evidence. *Games Econ Behav* 52:424–459.
29. Behrens TE, Hunt LT, Woolrich MW, Rushworth MF (2008) Associative learning of social value. *Nature* 456:245–249.
30. Fehr E, Camerer CF (2007) Social neuroeconomics: The neural circuitry of social preferences. *Trends Cogn Sci* 11:419–427.
31. Henrich J, et al. (2005) "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behav Brain Sci*, 28:795–815, discussion 815–855.
32. Huettel SA, Stowe CJ, Gordon EM, Warner BT, Platt ML (2006) Neural signatures of economic preferences for risk and ambiguity. *Neuron* 49:765–775.
33. Hsu M, Bhatt M, Adolphs R, Tranel D, Camerer CF (2005) Neural systems responding to degrees of uncertainty in human decision-making. *Science* 310:1680–1683.
34. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. *Proc Fifth Berkeley Symp Math Stat Prob* 1:281–297.