Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Games and Economic Behavior 52 (2005) 424–459
GAMES and Economic Behavior
www.elsevier.com/locate/geb

Self-referential thinking and equilibrium as states of mind in games: fMRI evidence ☆

Meghana Bhatt, Colin F. Camerer *

Division of Social Sciences 228-77, California Institute of Technology, Pasadena, CA 91125, USA

Received 3 February 2005

Available online 17 May 2005

Abstract

Sixteen subjects' brain activity were scanned using fMRI as they made choices, expressed beliefs, and expressed iterated 2nd-order beliefs (what they think others believe they will do) in eight games. Cingulate cortex and prefrontal areas (active in "theory of mind" and social reasoning) are differentially activated in making choices versus expressing beliefs. Forming self-referential 2nd-order beliefs about what others think you will do seems to be a mixture of processes used to make choices and form beliefs. In equilibrium, there is little difference in neural activity across choice and belief tasks; there is a purely neural definition of equilibrium as a "state of mind." "Strategic IQ," actual earnings from choices and accurate beliefs, is negatively correlated with activity in the insula, suggesting poor strategic thinkers are too self-focused, and is positively correlated with ventral striatal activity (suggesting that high IQ subjects are spending more mental energy predicting rewards).

JEL classification: C70; C91

☆ This research was supported by a Packard Foundation grant to Steven Quartz, and an internal Caltech grant.
* Corresponding author.
E-mail addresses: mbhatt@tanis.hss.caltech.edu (M. Bhatt), camerer@hss.caltech.edu (C.F. Camerer).

0899-8256/$ – see front matter
doi:10.1016/j.geb.2005.03.007

1. Introduction

Game theory has become a basic paradigm in economics and is spreading rapidly in political science, biology, and anthropology. Because games occur at many levels of detail (from genes to nations), game theory has some promise for unifying biological and social sciences (Gintis, 2003).

The essence of game theory is the possibility of *strategic thinking*: Players in a game can form beliefs about what other players are likely to do, based on the information players have about the prospective moves and payoffs of others (which constitute the structure of the game). Strategic thinking is central to game theory, but is also important in market-level phenomena like signaling, commodity and asset market information aggregation, and macroeconomic models of policy setting.

Despite the rapid spread of game theory as an analytical tool at many social levels, very little is known about how the human brain operates when thinking strategically in games. This paper investigates some neural aspects of strategic thinking using fMRI imaging. Our eventual goal is to build up a behavioral game theory that predicts how players choose and the neural processes that occur as they play. The data can also aid neuroscientific investigations of how people reason about other people and in complex strategic tasks.

In our experiments, subjects' brain activity is imaged while they play eight 2-player matrix games which are "dominance-solvable"[1]—that is, iterated deletion of dominated strategies (explained further below) leads to a unique "equilibrium" in which players' beliefs about what other players will do are accurate and players best respond to their beliefs. (In equilibrium, nobody is surprised about what others actually do, or what others believe, because strategies and beliefs are synchronized, presumably due to introspection, communication or learning.)

The subjects perform three tasks in random orders: They make choices of strategies (task C); they guess what another player will choose ("beliefs," task B); and they guess what other players think *they* will choose ("2nd-order beliefs," task 2B). Every player being scanned plays for money with another subject who is outside of the scanner.

In a game-theoretic "equilibrium," beliefs are correct, and choices are optimal given beliefs. One way for the brain to reach equilibrium is for neural activity in the C, B, and $2B$ tasks to be similar, since at equilibrium all three tasks "contain" the others, i.e. choice is a best response to belief, so the choice task invokes a belief formation. Any difference in activation across the three conditions is suggestive that different processes are being used to form choices and beliefs. In fact, as we show below, in experimental trials in which choices and beliefs are in equilibrium, there is little difference in activity in making a choice and expressing a belief; so this provides a purely neural definition of equilibrium (as a "state of mind"). Differences in activity across the three tasks might help us understand why players are out of equilibrium, so these differences are the foci of most of our analyses.

The first focus is the difference between making a choice and expressing a belief (i.e., the comparison between behavior and fMRI activation in the C and B conditions).

[1] In a dominance-solvable games, if players do not play dominated strategies, and guess that others will not, iteratively, then the result is an equilibrium configuration of strategy choices by players, and beliefs about what others will do, which are mutually consistent.

If choices are best-responses to beliefs, then the thinking processes underlying choice and belief formation should highly overlap; choice and belief are like opposite sides of the same coin. (Put differently, if you were going to build brain circuitry to make choices and form beliefs, and wanted to economize on parts, then the two circuits would use many shared components.)

In contrast, disequilibrium behavioral theories that assume limited strategic thinking allow players to choose without forming a belief, per se, so that C and B activity can differ more significantly. For example, Camerer et al. (2004a, 2004b) present a theory of limited strategic thinking in a cognitive hierarchy (building on earlier approaches[2]). In their theory some "0-step" players just choose randomly, or use some algorithm which is thoughtful but generates random choice—in any case, they will spend more energy on choice than belief. "One-step" thinkers act as if they are playing 0-step players, so they compute a choice but do not think deeply while forming a belief (e.g., they do not need to look at the other player's payoffs at all since they do not use these to refine their guess about what others will do). Two-step players think they are playing a mixture of 0- and 1-step players; they work harder at forming a belief, look at other players' payoffs, and use their belief to pick an optimal choice. Models of this sort are precise (more statistically precise than equilibrium theories) and fit most experimental data sets from the first period of a game (before learning occurs) better than Nash equilibrium does (Camerer et al., 2004a). These limited-thinking theories allow larger differences in cognitive activity between the acts of *choosing* a strategy and *expressing a belief* about another player's strategy than equilibrium theories do. A 1-step player, for example, will look at all of her own payoffs and calculate the highest average payoff when making a choice, but when guessing what strategy another player will choose she can just guess randomly. Such a player will do more thinking when choosing than when stating a belief. This possible difference in processing motivates our analysis of differential brain activity during the C and B tasks.[3]

The second focus of the analysis is on the difference in activity while forming beliefs in the B task and 2nd-order beliefs in the $2B$ task. One way agents might form 2nd-order beliefs is to use general circuitry for forming beliefs, but apply that circuitry as if they were the other player (put themselves in the "other player's brain"). Another method is self-referential: Think about what they would like to choose, and ask themselves if the other player will guess their choice or not. These two possibilities suggest, respectively, that the B and $2B$ conditions will activate similar regions, or that the C and $2B$ regions will activate similar regions.

Besides contributing to behavioral game theory (see Camerer, 2003), imaging the brain while subjects are playing games can also contribute to basic social neuroscience (e.g.,

[2] See Nagel, 1995; Stahl and Wilson, 1994; Costa-Gomes et al., 2001; Hedden and Zhang, 2002 and Cai and Wang, 2004.

[3] An ideal test would compare activity of subjects who are capable of performing different thinking steps across games of different complexity. For example, a low-step thinker should show similar activity in simple and complex games (because they lack the skill to think deeply about complex games). A high-step thinker would stop at a low-level choice in a simple game (where k and higher steps of thinking prescribe the same choice) but would do more thinking in complex games. Unfortunately, we have not found a solid psychometric basis to "type-cast" players reliably into steps of thinking; when we can do so, the comparison above will provide a useful test.

Adolphs, 2003). Cognitive social neuroscientists are interested in spectrum disorders[4] like autism, in which people lack a normal understanding of what other people want and think. The phrase "theory of mind" (ToM) describes neural circuitry that enables people to make guesses about what other people think and desire (sometimes called "mind-reading" or "mentalizing"; e.g., Siegal and Varley, 2002; Gallagher and Frith, 2003; Singer and Fehr, 2005).

Using game theory to inform designs and generate sharp predictions can also provide neuroscientists interested in ToM and related topics with some new tools which make clear behavioral predictions and link tasks to a long history of careful theory about how rational thinking relates to behavior.

In this spirit, our study extends ToM tasks to include simple matrix games. While there has been extensive research into first order beliefs: the simple consideration of another person's beliefs, there has been very little investigation of 2nd-order beliefs, especially when they are self-referential—i.e., what goes on in a person's brain when they are trying to guess what another person thinks *they* will do?

1.1. Why study choices, beliefs and 2nd order beliefs?

Figure 1 shows the exact display of a matrix game (our game 3) that row players saw in the scanner, in the 2*B* task where they are asked what the column player thinks they will do.[5] The row and column players' payoffs are separated onto the left and right halves of the screen (in contrast to the usual presentation).[6] Row payoffs are in a submatrix on the left; column player payoffs are in a submatrix on the right (which was, of course, explained to subjects).

The Fig. 1 game can be "solved" (that is, a Nash equilibrium can be computed) by three steps of iterated deletion of dominated strategies.[7] The row player's strategy *C* is dominated by strategy *B* (i.e., regardless of what the column player does, *B* gives a higher payoff than *C*); if the row player prefers earning more she will never choose *C*. If the column player guesses that row will never play *C* (the dominated strategy is "deleted," in game theory language—i.e., the column player thinks *C* will never be played by an earnings-maximizing row player), then strategy *BB* becomes a dominant strategy for the column

[4] A "spectrum" disorder is one which spans a wide range of deficits (inabilities) and symptoms—it has relatively continuous gradation. This suggests a wide range of neural circuits or developmental slowdowns contribute to the disorder, rather than a single cognitive function.

[5] The placeholder letter "*x*" is placed in cells and rows which are inactive in an effort to create similar amounts of visual activity across trials, since matrices had different numbers of entries.

[6] The split-matrix format was innovated by Costa-Gomes et al. (2001), who used it to separate eye movements when players look at their own payoffs or the payoffs of others, in order to judge what decision rules players were using (see also Camerer et al., 1994). The matrices are more complex than many fMRI stimuli but we chose to use affine transformations of the CGCB matrices to permit precise comparability of our choice data to theirs. Our current study did not track eye movements but it would be simple to use this paradigm to link eye movement to fMRI activity, or to other temporally-fine measures of neural activity.

[7] A *strictly* dominated strategy is one that has a lower payoff than another strategy, for *every* possible move by one's opponent; A *weakly* dominated strategy has weakly lower payoffs than another strategy against all strategies and strictly lower payoffs against at least one of the opponent's strategies. A dominant strategy is one that gives the highest possible payoff against all of the opponent's strategies.

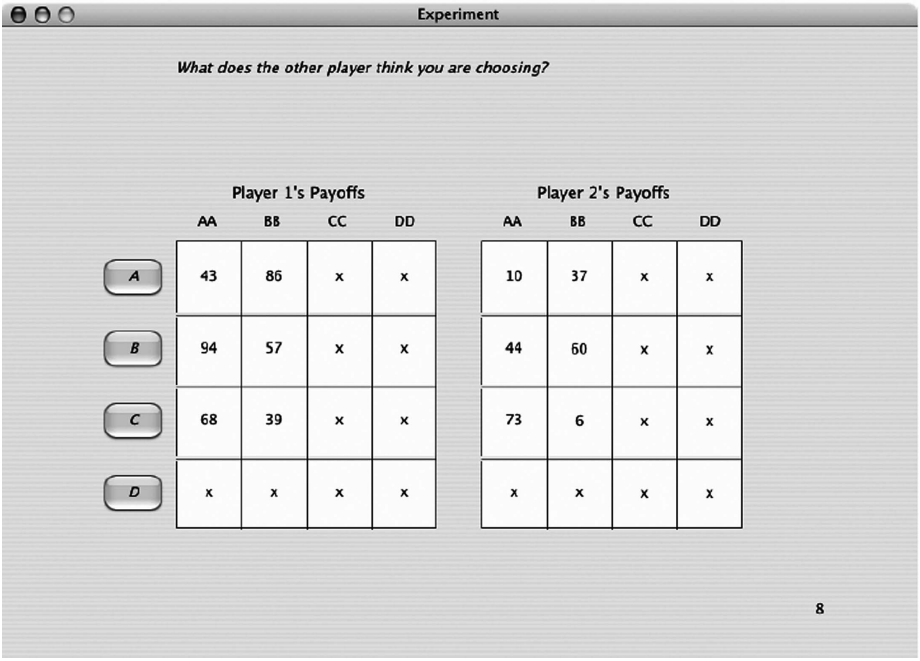


Fig. 1. A three-step game used in the experiment, as presented in the scanner (game 3). C is dominated. Deleting C makes AA dominated. Deleting AA and C makes A dominant. The unique Nash equilibrium is therefore (A, BB). Only 31% and 61% (respectively) chose these strategies (see Appendix A). The Camerer–Ho CH model (see text) with $\tau = 1.5$ predicts 7% and 55%.

player. If the row player guesses that the column player guesses she (the row player) will never play C, and the row player infers that the column player will respond with BB, then strategy A becomes dominant for the row player. Of course, this is a long chain of reasoning which presumes many steps of mutual rationality.

Putting aside the fMRI evidence in our study, simply comparing choices, beliefs and iterated beliefs as we do could be interesting in game theory for a couple of reasons. A common intuition is that higher-order beliefs do not matter. But Weinstein and Yildiz (2004) show that in games which are not dominance-solvable, outcomes depend sensitively on higher-order beliefs (if they are not restricted through a common knowledge assumption à la Harsanyi). Empirically, their theorems imply that knowing more about higher-order beliefs is necessary to guess what will happen in a game.

Goeree and Holt's (2004) "theory of noisy introspection" assumes that higher-order beliefs are characterized by higher levels of randomness or uncertainty. Increased uncertainty might appear as lower levels of overall brain activity (or higher, if they are thinking harder) for 2nd-order beliefs compared to beliefs and choices. Furthermore, increased uncertainty should be manifested by poorer behavioral accuracy for higher-order beliefs.

Second-order beliefs also play a central role in games involving deception. By definition, a successful deception requires a would-be deceiver to know she will make one choice, A, but also believe the other player thinks she will make a *different* choice, B.

The capacity for deception therefore requires a player to hold "false 2nd-order beliefs" in mind—that is, to plan choices which are different from what (you think) others think you will do.[8]

Finally, second-order beliefs also play an important role in models of social preferences, when a player's utility depends directly on whether they have lived up to the expectations of others (see Rabin, 1993). Dufwenberg and Gneezy (2000) studied trust games in which players could pass up a sure amount x and hope that a second player gave them a larger amount y from a larger sum available to divide. They found that the amount the second player actually gave was modestly correlated (0.44) with the amount the second player thought the first player expected (i.e., the second player's 2nd-order belief). The second player apparently felt some obligation to give enough to match player 1's expectations.[9] These kinds of emotions require 2nd-order beliefs as an input.

Trying to discern what another person believes about *you* is also important in games with asymmetric information, when players have private information that they know others know they have, and in games where a "social image" might be important, when people care what others think about them (in dictator and public goods games, among others).

1.2. Neuroeconomics, and what it is good for

This paper is a contribution to "neuroeconomics," a rapidly-emerging synthesis (and subject of this special issue) which grounds details of basic economic processes in facts about neural circuitry (Camerer et al., 2004c, 2005; Zak, 2005; Glimcher and Rustichini, 2004).

Neuroeconomics is an extension of *behavioral economics*, which uses evidence of limits on rationality, willpower and self-interest to reform economic theory; neural imaging is just a new type of evidence. Neuroeconomics is also a new part of *experimental economics*, because it extends experimental methods which emphasize paying subjects according to performance, and tying predictions to theory, to include studies with animals, lesion patients (and "temporary lesions" created by TMS), single-neuron recording, EEG and MEG, psychophysiological recording of heart rate, skin conductance, pupil dilation, tracking eye movements, and PET and fMRI imaging (McCabe and Smith, 2001). Neuroeconomics is also part of *cognitive neuroscience*, since these studies extend the scope of what neuroscientists understand to include "higher-order cognition" and complex tasks involving social cognition, exchange, strategic thinking, and market trading that have been the focus of microeconomics for a long time.

[8] Whether or not a person can understand false beliefs is a key component of theory of mind and is also a test used to diagnose autism. In a classic "Sally–Anne" task, a subject is told that Sally places a marble in her basket and leaves the room. Anne then moves the marble from the basket to a box and also leaves the room. Sally re-enters the room. The subject is then asked where Sally will look for her marble. Since the child believes that the marble is in the box, she must be able to properly represent Sally's different belief—a *false* belief—to answer correctly, that Sally will look in the basket. Most children switch from guessing that Sally will look for the marble in the box (a selfreferentially-grounded mistake) to guessing that she will be looking in the basket at around 4 years old. Autistic children make this switch later or not at all. See Gallagher and Frith (2003) for more detail.

[9] However, about a third of the player 2's gave less than they thought others expected.

One reaction to the idea of neuroeconomics is that economic models do not need to include neural detail to make good predictions, because they are agnostically silent about whether their basic assumptions are actually satisfied, or simply lead to outcomes "as if" the assumptions were true.[10] As a result, one can take a conservative or radical view of how empirical studies like ours should interact with conventional game theory.

The conservative view is that neural data are just a new type of evidence. Theories should get extra credit if they are consistent with these data, but should not be penalized if they are silent about neural underpinnings.

The radical view is that all theories, eventually, will commit to precisely how the brain (or some institutional aggregation, as in a firm or nation-state's actions) carries out the computations that are necessary to make the theory work. Theories that make accurate behavioral predictions and also account for neural detail should be privileged over others which are neurally implausible.

Our view leans toward the radical. It cannot be bad to have theories which predict choices from observable structural parameters and which *also* specify precise details of how the brain creates those choices. (If we could snap our fingers and have such theories for free, we would.) So the only debatable question is whether the cognitive and neural data available *now* are good enough to enable us to *begin* to use neural feasibility as a central way to judge the plausibility of as-if theories of choice.

We think this is a reasonable time to begin using neural activation to judge plausibility of theories because there are many theories of choice in decision theory and game theory, and relatively few data to sharply separate those theories. Virtually all theories appeal vaguely to plausibility, intuition, or anecdotal evidence, but these are not scientific standards. Without more empirical constraint, it is hard to see how progress can be made when there are many theories. Neural data certainly provide more empirical constraint.

Furthermore, in many domains current theories *do not* make good behavioral predictions. For example, equilibrium game theories clearly explain many kinds of experimental data poorly (e.g., Camerer, 2003). Studying cognitive detail, including brain imaging, will inevitably be useful for developing new concepts to make *better* predictions.[11]

An argument for the imminent value of neural data comes by historical analogy to recent studies which track eye movements when subjects play games Camerer et al. (1994); Costa-Gomes et al. (2001) (CGCB); Johnson et al. (2002); Costa-Gomes and Crawford (2004); Johnson and Camerer (2004). When payoffs are placed on a computer screen, different algorithms for making choices can be tested as joint restrictions on the choices implied by

[10] The "as if" mantra in economics is familiar to cognitive scientists in the form of David Marr's influential idea that theories can work at three levels—"computational" (what an economist might call functional or as-if); "algorithmic" or "representational" (what steps perform the computation); and "implementation" or hardware (see Glimcher, 2003 for a particularly clear discussion). Ironically, Marr's three-level idea licensed cognitive scientists to model behavior at the highest level. We invoke it to encourage economists who operate exclusively at the highest level, to commit game theory to an algorithmic view, to use evidence of brain activity to make guesses about algorithms and to therefore discipline ideas about highest-level computation.

[11] Furthermore, neuroeconomics will get done whether economists endorse it or not, by smart neuroscientists who ambitiously explore higher-order cognition carefully but without the benefit of decades of training about how delicate theoretical nuances might matter and which can guide design. Engaging with the energetic neuroscientists is therefore worthwhile for both sides.

those algorithms, *and* whether players look at the payoff numbers they need to execute an algorithm.

Eye tracking has been used in three published studies to separate theories which make similar behavioral predictions. Camerer et al. (1994) and Johnson et al. (2002) studied three-period bargaining games in which empirical offers are somewhere between an equal split and the subgame perfect self-interest equilibrium (which requires subjects to "look ahead" to future payoffs if bargaining breaks down in early periods; see Camerer, 2003, Chapter 4). They found that in 10–20% of the games subjects literally did not glance at the possible payoff in a future period, so their offers could not be generated by subgame perfect equilibrium. Johnson and Camerer (2004) found that the failure to look backward, at the possible payoffs of other players in previous nodes of a game, helped explain deviations from "forward induction." CGCB found that two different decision rules, with very similar behavioral predictions about chosen strategies, appeared to be used about equally often, when only choices were used to infer what rules were used. But when lookup information was used, one rule was inferred to be much more likely. If CGCB had only used choices to infer rules, *they would have drawn the wrong conclusion about what rules people were using*.

Those are three examples of how inferences from choices alone do not separate theories nearly as well as inferences from both choices *and* cognitive data. Perhaps neural activity can have similar power as attentional measures, as evidence accumulates and begins to make sense.

The hard part is creating designs that link neural measures to underlying latent variables. Our work is guided by the "design triangle" illustrated in Fig. 2. The triangle shows experimental stimuli (on the top of the triangle) which produce measured output—brain activation, skin conductance, eye movements, and so on (lower left)—which can, ideally, be interpreted as expressions of underlying variables or algorithms which are not directly observable (lower right). For the experiments reported in this paper, the underlying constructs which are illuminated by brain activity are hypotheses about the decision processes players are using to generate choices and beliefs.

Keep in mind that while brain pictures like those shown below highlight regions of activation, we are generally interested not just in regions but in neural *circuitry*—that is, how various regions collaborate in making decisions. Understanding circuitry requires a variety of methods. fMRI methods are visually impressive but place subjects in an unnatural (loud, claustrophobic) environment and the signals are weak so many trials are needed to average across. Neuroscience benefits from many tools. For example, looking at tissue in primate brains helps establish links between different regions ("connectivity"). Other methods include psychophysiological measurement (skin conductance, pupil dilation, etc.), studies of patients with specialized brain damage, animal studies, and so forth. Neuroscience is like detective work on difficult cases: There is rarely a single piece of evidence that is definitive. Instead, the simplest theory that is consistent with the most different types of evidence is the one that gets provisionally accepted, and subject to further scrutiny. This paper should be read in this spirit, as extremely tentative evidence which will eventually be combined with many new studies to provide a clear picture.

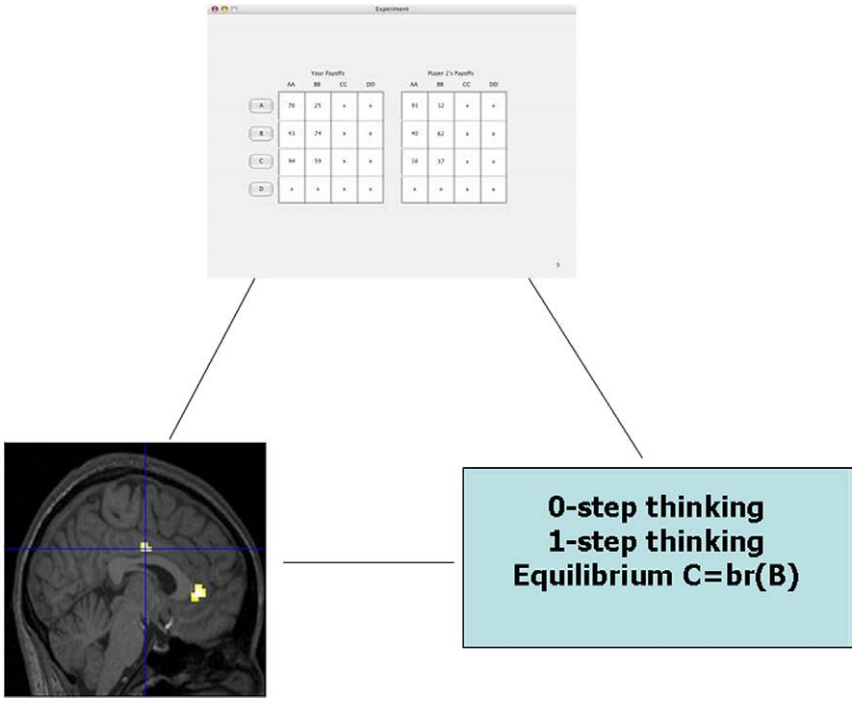


Fig. 2. Neureconomics design: Designs relate stimuli (top) to latent variables or algorithms (right) which generate interpretable activation (left). Experimental economics studies link stimuli (top) and variables (right). Many neuroscience studies just report links between stimuli (top) and activation (left). The neuroeconomics challenge is to make all 3 fit.

2. Neural correlates of strategic thinking

2.1. Methods

Sixteen subjects were scanned,[12] one at a time, in a 3T Siemens Trio scanner at Caltech (Broad Imaging Center) as they performed C, B and $2B$ tasks across each of eight games. The games and order of the three tasks were fixed across subjects. Appendix A shows the games (which are transformations of games in CGCB), the instructions, and give some methodological details.

In keeping with healthy experimental economics convention, both players were financially rewarded for one task and game that was chosen at random after they came out of the scanner. If a choice task was chosen, then the choices of both players determined their payoffs ($.30 times experimental points). If a belief or 2nd-order belief task was chosen for payment, a player earned $15 if her belief B matched the other player's choice, or $15 if her 2nd-order belief $2B$ matched the other player's belief.

[12] To experimental social scientists, 16 seems like a small sample. But for most fMRI studies this is usually an adequate sample to establish a result because adding more subjects does not alter the conclusions much.

Pairs of subjects were recruited on campus at Caltech through SSEL lab recruiting software.[13] One subject performed the tasks in the scanner, as the row player, while the other performed them in an adjacent room, as the column player.

We give only a quick sketch of fMRI technique here. Methods of measurement and analysis are extremely complex and still evolving. Appendix A has more detail (or see, e.g., Huettel et al., 2004).

Each subject first has their brain "structurally scanned" (as in medical applications) to establish a sharper picture of the details of brain anatomy for six minutes. Then each subject proceeds through a series of screens (like Fig. 1) one at a time, at their own pace (response times averaged 8–25 seconds; see Appendix A). They make choices and express beliefs by pressing buttons on a box they hold in their hand. After each response is recorded, there is a random lag from 6–10 seconds with a "fixation cross" on a blank screen to hold their visual attention in the center of the screen and allow blood flow to die down. The entire set of tasks took from 7 to 15 minutes.

The scanner records 32–34 "slices" of brain activity every 2 seconds (one "TR"). Each slice shows blood flow in thousands of three-dimensional "voxels" which are $3 \times 3 \times 3$ millimeters in size. Our analysis is "event-related," which means we ask which voxels are unusually active when a particular stimulus is on the screen. The analysis is a simple linear regression where dummy variables are "on" when a stimulus is on the screen and "off" otherwise. This "boxcar" regression is convolved with a particular function that is well-known to track the hemodynamic response of blood flow. The regression coefficients of activity in the BOLD (blood-oxygenation level dependent) signal in each voxel tell us which voxels are unusually active. Data from all subjects are then combined in a random effects analysis. We report activity which is significantly different from chance at a p-value < 0.001 (a typical threshold for these studies), and for clusters of at least 5 adjacent voxels where activity is significant (with exceptions noted below).

2.2. Behavioral data

Before turning to brain activity, we first describe some properties of the choices and expressed beliefs. Appendix A shows the relative frequencies of subject choices, expressed beliefs, and expressed 2nd-order beliefs, in each game.

Table 1 shows the percentages of trials, for games solvable in different numbers of steps of deletion of dominated strategies, in which players made equilibrium choices. The table includes the choice data from CGCB's original study using these games. First note that the percentages of subjects making the equilibrium strategy choice in our study is similar for row and column players, who are respectively, in and out of the scanner. (None of the row–column percentages are significantly different.) However, equilibrium play in our games is

[13] Since Caltech students are selected for remarkable analytical skill, they are hardly a random sample. Instead, their behavior is likely to overstate the average amount of strategic thinking in a random population. This is useful, however, in establishing differential activation of regions for higher-order strategic thinking since the subjects are likely to be capable of higher-order thinking in games that demand it.

Table 1
Percentages of equilibrium play across games and player type

Type of game	Row player (in scanner)	Column player (out of scanner)	Row + column mean	CGCB mean	New data − CGCB z-statistic
2×2, row has a dominant decision	0.75	0.61	0.68	0.93	−3.21*
2×4, row has a dominant decision	0.56	0.72	0.65	0.96	−3.24*
2×2, column has a dominant decision	0.50	0.61	0.56	0.80	−2.46*
2×4, column has a dominant decision	0.63	0.56	0.59	0.70	−0.94
2×3, 2 rounds of iterated dominance	0.47	0.58	0.53	0.69	−1.49
3×2, 3 rounds of iterated dominance	0.22	0.22	0.22	0.22	−0.02

less frequent than in CGCB's experiment, significantly so in the simplest games.[14] Since the frequencies of equilibrium play by the in-scanner row player and the out-of-the-scanner column player are similar, the lower percentage of equilibrium play in our experiments is probably due to some factor other than scanning.[15]

Table 2 reports the frequency of trials in which $C = br(B)$ (where $br(B)$ denotes the best response to belief B), $B = br(2B)$, $C = 2B$, and in which all three of those conditions are met simultaneously (our stringent working definition of "an equilibrium trial" hereafter).

Equilibrium trials are generally rare (23%). Comparing the match of beliefs and choices across categories, a natural intuition is that as players reason further up the hierarchy from choices, to beliefs, to iterated beliefs, their beliefs become less certain. Therefore, 2nd-order beliefs should be less consistent with beliefs than beliefs are with choices, and 2nd-order beliefs and choices should be least consistent (Goeree and Holt, 2004). (In terms of the Table 2 statistics, the three rightmost column figures should decline from left to

[14] Of course, eliciting choices, beliefs, and 2nd-order beliefs in consecutive trials might affect the process of choice, perhaps promoting equilibration. But the close match of our observed $C = br(B)$ rate to the Costa-Gomes and Weizsäcker's (2004) rate, and the lower rate of equilibrium choices compared to CGCB's subjects (who only made choices) suggests the opposite. Also keep in mind that our subjects report a single strategy as a belief, and are rewarded if their guess is exactly right, which induces them to report the mode of their distribution. (For example, if they think *AA* has a p chance and *BB* has a $1 - p$ chance they should say *AA* if $p > 0.5$.) Costa-Gomes and Weizsäcker elicited a probability distribution of probability across all possible choices. Their method is more informative but we did not implement it in the scanner because it requires a more complex response which is difficult and time-consuming using button presses.

[15] The difference between our rate of conformity to equilibrium choice and CGCB's may be due to the fact that beliefs are elicited, although one would think that procedure would increase depth of reasoning and hence conformity to equilibrium. We think it is more likely to result from a small number players who appeared to act altruistically, trying to make choices which maximize the total payoff for both players (which often leads to dominance violation—e.g., cooperation in prisoners' dilemma games). Since this kind of altruism is surprisingly difficult to pin down carefully, we continue to use all the data rather than to try to separate out the altruistically-minded trials.

Table 2
Frequencies of choice and belief matching for the row player

Type of game	Equilibrium (all 3 conditions hold)	$C = br(B)$	$B = br(2B)$	$C = 2B$
Row has dominant strategy	0.31	0.66	0.59	0.69
Column has dominant strategy	0.44	0.75	0.75	0.88
2×3 game with two steps of dominance	0.13	0.63	0.66	0.69
3×2 game with three steps of dominance	0.06	0.59	0.53	0.75
Overall	0.23	0.66	0.63	0.75

right.) That intuition is wrong for these data. The fractions of trials in which $C = br(B)$, and $B = br(2B)$ are about the same. The number of subjects who make optimal choices given their belief ($C = br(B)$) is only 66%. This number may seem low, but it is similar to statistics reported by Costa-Gomes and Weizsäcker (2004) (who also measured beliefs more precisely than we did).

More interestingly—and foreshadowing brain activity we will see later—the frequency with which choices match 2nd-order beliefs ($C = 2B$) is actually *higher*, for all classes of games, than the frequency with which $B = br(2B)$ (75 versus 63% overall). This is a hint that the process of generating a self-referential iterated belief might be similar to the process of generating a choice, rather than simply iterating a process of forming beliefs to guess what another player believes about oneself.

Given these results, and the success of parametric models of iterated strategic thinking (e.g., Camerer et al., 2004a), an obvious analysis is to sort subjects or trials into 0, 1, 2 or more steps of thinking and compare activity. But the current study was not optimally designed for this analysis, so analyses of this type are not insightful.[16]

2.3. *Differential neural activity in choice* (C) *and belief* (B) *tasks*

In cognitive and neural terms, 0- and 1-step players do not *need* to use the same neural circuitry to make choices and to express beliefs. Thus, any difference in neural activation

[16] Comparing trials sorted into low-steps of thinking (0 or 1) and high steps shows very little differential activation of high relative to low in either choice or belief tasks, and substantial activation of low relative to high in cingulate and some other regions. The a priori guess is that higher thinking steps produce more cingulate (conflict) activation, so we do not think the sorting into apparent 0- and 1-step trials is accurate enough to permit good inferences at this stage. A design tailored for this sort of "typecasting" analysis could be used in future research. There are many handicaps from the current design for linking inferred thinking steps to brain activity. One problem is that in many games, choices of higher-step thinkers coincide. Another problem is that it is difficult to weed out altruistic choices, so they are typically misclassified in terms of steps of thinking which adds noise. A cross-subject analysis (trying to identify the typical number of thinking steps for each subject) did not work because individual subject classification is noisy with only eight games (see also Chong et al., 2005). It is also likely that these highly skilled subjects did not vary enough in their thinking steps to create enough variation in behavior to pick up weak behavior-activation links.

in the two conditions (C and B) is a clue that some players, on some trials, are making choices without forming beliefs of the sort that require any deep processing about what other players will do, so that belief elicitation is actually a completely different sort of neural activity than choice.[17] Therefore, the first comparison we focus on is between row players *choosing* strategies and *expressing* beliefs about what column players will do.

Figure 3 shows brain "sections" which reveal four significantly higher activations in the choice (C) condition compared to the belief (B) condition (i.e., the "$C > B$ subtraction") which have 10 or more adjacent voxels ($k > 10$).[18] The differentially active regions are the posterior cingulate cortex (PCC),[19] the anterior cingulate cortex (ACC), the transitional cortex between the orbitofrontal cortex (OFC) and the agranular insula (which we call frontal insula, FI),[20] and the dorsolateral prefrontal cortex (DLPFC). The sections each show differential activity using a color scale to show statistical significance. A 3-dimensional coordinate system is used which locates the middle of the brain at $x = y = z = 0$. The upper left section (a) is "sagittal," it fixes a value of $X = -3$ (that is 3 mm to the left of the zero point on the left-right dimension). The upper right section (b) is "coronal" at $Y = +48$ (48 mm frontal or "anterior" of the $Y = 0$ point). The lower left section (c) is "transverse" (or "axial") at $Z = -18$, 18 mm below the zero line.

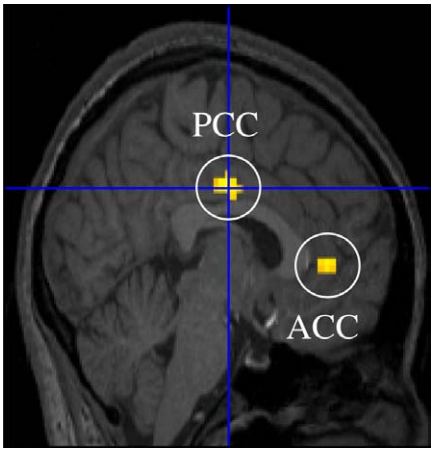
Figure 4 shows the time courses of raw BOLD signals on the y-axis (in normalized percentage increases in activity) in the PCC region identified above (left, or superior, in the upper left section Fig. 3(a)), for the C (thick line), B (thin line) and $2B$ (dotted line) tasks. These pictures show how relative brain activity increases or decreases in a particular area

[17] An important caveat is that different tasks, and game complexities, will produce different patterns of eye movement. Since we do not have a complete map of brain areas that participate in eye movements for the purpose of decision (though see Glimcher, 2003), some of what we might see might be part of general circuitry for eye movement, information acquisition, etc., rather than for strategic thinking per se. The best way to tackle this is to record eye tracking simultaneously with fMRI and try to use both types of data to help construct a complete picture.

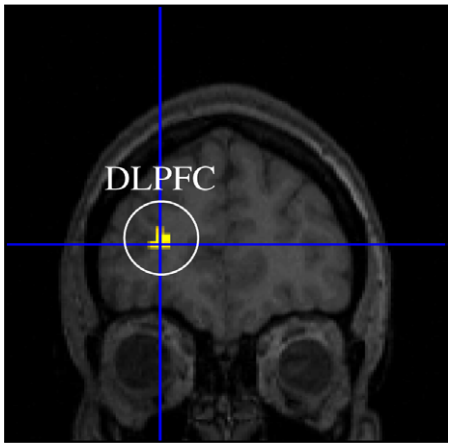
[18] A very large fifth region not shown in Fig. 3 is in R occipital cortex (9, −78, 9, $k = 202$, $t = 6.77$). When we use a smaller k-voxel filter, $k = 5$ (used in Fig. 3) there are four additional active regions besides the R occipital and those shown in Fig. 3 (see Table A.4 in Appendix A) which are not especially interpretable in terms of strategic thinking.

[19] We use the following conventions to report locations and activity: The vector (−3, −9, 33, $k = 5$, positive in 14 of 16 subjects) means that the voxel with peak activation in the cluster has coordinates $x = -3$, $y = -9$, $z = 33$. The coordinates x, y, and z respectively measure distance from the left to the right of the brain, from front ("anterior") to back ("posterior"), and bottom ("inferior") to top ("superior"). The figure $k = 5$ means the cluster has 5 voxels of 3 cubic millimeters each. The number of subjects with positive regression coefficients is an indication of the uniformity of the activation across subjects. Table A.4 in Appendix A shows coordinates for all regions mentioned in this paper, and some regions that are not discussed in the text.

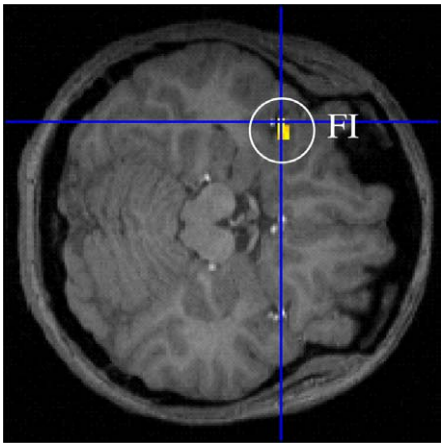
[20] FI and ACC are the two regions of the brain known to contain spindle cells. Spindle cells are large elongated neurons which are highly "arborized" (like a tree with many branches, they project very widely, and draw in information and project information to many parts of the brain) that are particular to humans and higher primate kin, especially bonobos and chimpanzees (Allman et al., 2002). It is unlikely that any of these brain areas are solely responsible for our ability to reason about others. In fact it seems that the pathologies where individual do not have these abilities, namely Autism and Asperger's syndrome, do not involve lesions of any specific areas of the brain, but rather more generalized developmental problems including a decreased population of spindle cells (Allman, Caltech seminar), decreased connectivity to the superior temporal sulcus (Castelli et al., 2002), and defects in the circuitry of the amygdala (Siegal and Varley, 2002).



(a) $X = -3$



(b) $Y = +48$



(c) $Z = -18$

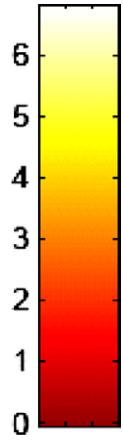


Fig. 3. Areas of significantly differential activity in choice minus belief conditions, all trials, at $p < 0.001$ (uncorrected). (a) Top area is posterior cingulate cortex, PCC (-3, -12, 33, $k = 24$, $t = 5.12$; 14 of 16 subjects positive); right area is anterior cingulate cortex/genu ACC (6, 42, 0; $k = 33$, $t = 4.62$; 15 of 16 subjects positive). (b) dorsolateral prefrontal cortex DLPFC (-27, 48, 9; $k = 14$, $t = 4.74$; 15 of 16 subjects positive). (c) transition cortex/FI (-42, 12, -18; $k = 31$, $t = 4.60$, 14 of 16 subjects positive).

over time, for different tasks. The time courses also show standard error bars from pooling across trials; when the standard bars from two lines do not overlap, that indicates statistically significant patterns of activation. The 0 time on the x-axis is when the task stimulus is first presented (i.e., the game matrix appears). The x-axis is the number of scanning cycles (TRs). Each TR is 2 seconds, so a number 4 on the x-axis is 8 seconds of clock time. Perhaps surprisingly, when the stimulus is presented the ACC actually *deactivates* during these tasks (the signal falls). Since blood flow takes one or two TR cycles to show up in imaging (about 3–5 seconds), the important part of the time sequence is in the middle of

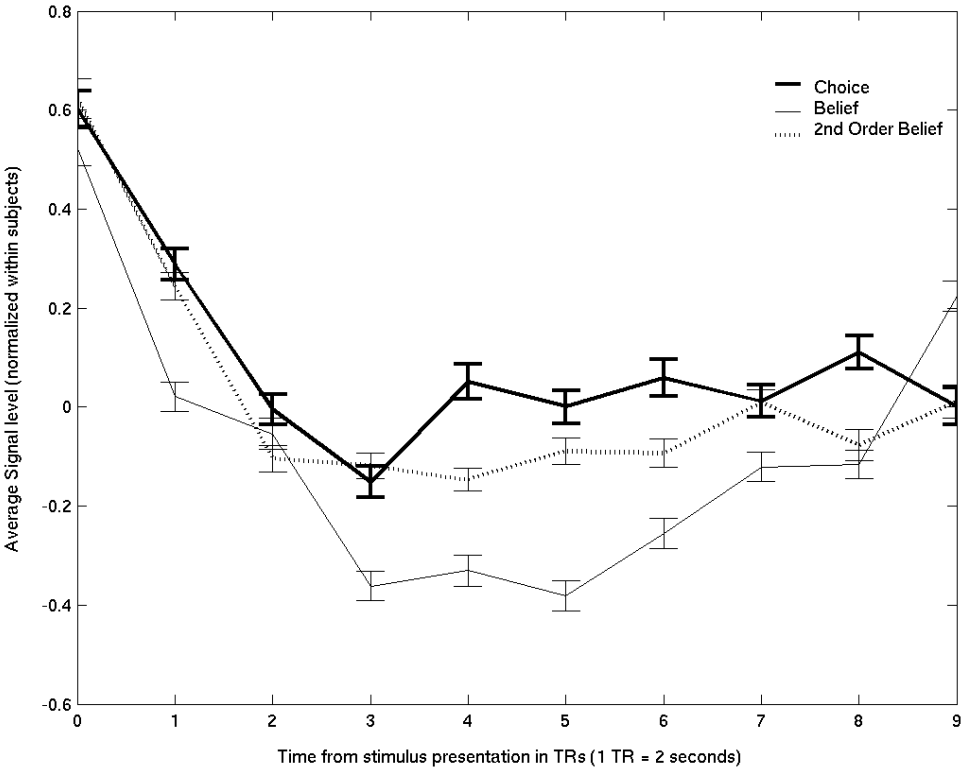


Fig. 4. Time course of activity in posterior cingulate $(-3, -12, 33)$ in choice (C, thick line), belief (B, thin line) and 2nd-order belief ($2B$, dotted line) tasks.

the graph, between 3 TRs and 8 TRs (when most of the responses are made, since they typically take 8–10 seconds; see Appendix A for details).

The important point is that during the choice task (thick line), PCC deactivation is higher than in the $2B$ and B tasks—hence the differential activation in C minus B shown in the previous Figure 3(a). Most importantly, note that the $2B$ task activity lies between the C and B activity. This is a clue that guessing what someone thinks you will do $(2B)$ is a mixture of a guessing process (B), and choosing what you will do (C). This basic pattern—$2B$ is between C and B—also shows up in time courses of activity for all the other areas highlighted in the brain sections in Fig. 3.

Figure 5 shows the location of anterior cingulate cortex (ACC, in yellow) and orbitofrontal cortex (pink). The cingulate cortex is thought to be important in conflict resolution and "executive function" (e.g. Miller and Cohen, 2001). The ACC and PCC regions that are differentially active in choosing rather than forming beliefs have both been implicated in ToM and in other social reasoning processes. The PCC is differentially active in moral judgments that involve personal versus impersonal involvement and many other kinds of processing that involve emotional and cognitive conflict (e.g., Greene and Haidt, 2002). D. Tomlin (personal communication) has found relative activation in the very most

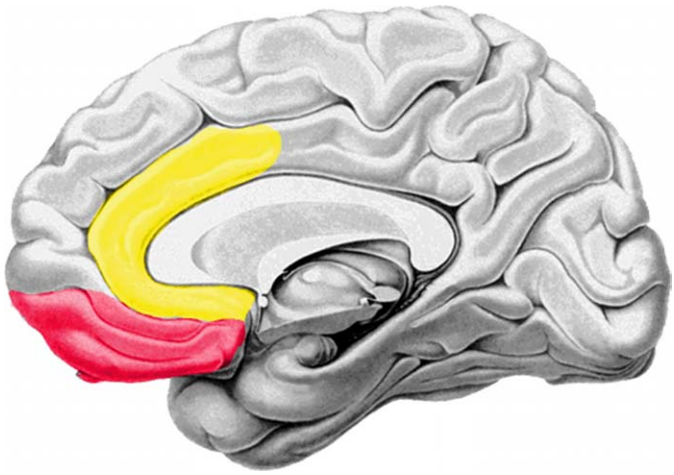


Fig. 5. A brain drawing showing anterior cingulate cortex (ACC, yellow) and orbitofrontal cortex (OFC, pink). The front of the brain (anterior) is to the left. Reprinted with permission of Ralph Adolphs.

anterior (front) and posterior (back) cingulate regions that are shown in Fig. 3 in repeated trust games with a very large sample (almost 100 pairs of players), after another player's decision is revealed.[21] Since their subjects are playing repeatedly, presentation of what another player actually does provides information on how he may behave in the next trial, it is possible that this evidence is immediately used to start making the players next decision.

The fact that all these regions are more active when people are making choices, compared to expressing beliefs, suggests that a very simple neural equation of forming a belief and choosing is leaving out some differences in neural activity that are clues to how the processes may differ.

The FI region we identify is close to an area noted by Gallagher et al. (2002) (38, 24, −20) in the inferior frontal cortex. Their study compared people playing a mixed-equilibrium (rock, paper, scissors) game against human opponents versus computerized opponents. The identification of a region differentially activated by playing people, which is nearby to our region is a clue that this inferior frontal/FI region might be part of some circuitry for making choices in games against other players.

Differential activation in frontal insula (FI) is notable because this area is activated when people are deciding how to bet in ambiguous situations relative to risky ones, in the sense of Ellsberg or Knight (Hsu et al., 2005). This suggests choice in a game is treated like an ambiguous gamble while expressing a belief is a risky (all-or-none) gamble. This

[21] Tomlin et al. reported a "self-other" map of the cingulate which includes the most anterior and posterior regions we see in Fig. 3. They studied brain activation during repeated partner trust games. When the other player's behavior was shown on a screen, the most anterior (front of the brain) region was active, independent of the player role. When one's own behavior was shown, more middle cingulate regions were activated. The most posterior (back) regions were activated when either screen was shown. The brain often "maps" external parts of the world (retinotopic visual mapping) or body (somatosensory cortex). The cingulate map suggests a similar kind of "sociotopic" mapping in the cingulate.

interpretation is consistent with 0- and 1-step thinking, in which evaluating strategies and likely payoffs occurs with a shallow consideration of what other players will do, which seems more ambiguous than forming a belief.

2.4. Equilibrium as a state of mind: Choice and belief in- and out-of-equilibrium

The evidence and discussion above suggests that the processes of making a strategic choice and forming a belief are *not* opposite sides of a neural coin. Interesting evidence about this neural-equivalence hypothesis emerges when the trials are separated into those in which all choices and beliefs are in equilibrium (i.e., $C = br(B)$, $B = br(2B)$ and $C = 2B$) and those which are out of equilibrium (one or more of the previous three parenthetical conditions does not hold).

Figure 6 shows sections of differential activity in the C and B tasks during equilibrium trials. This is "your brain in equilibrium": There is only one area actively different (at $p < 0.001$) in the entire brain. This suggests that equilibrium can be interpreted not only

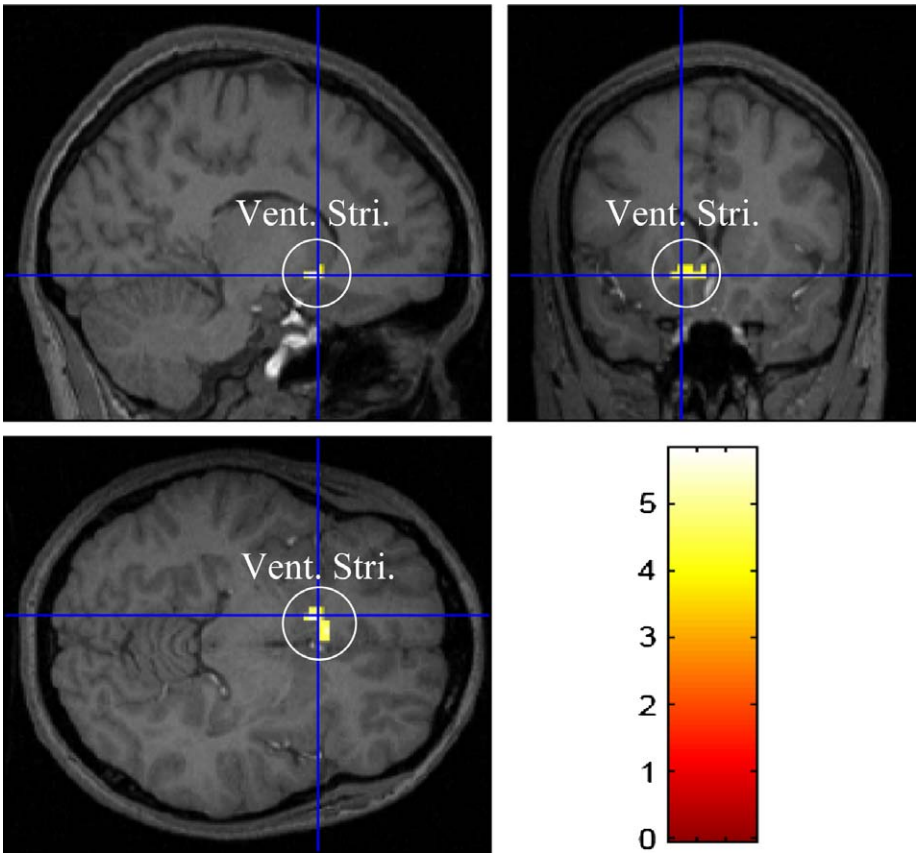


Fig. 6. This is your brain in equilibrium: Area of significant differential activation in $C > B$ for in-equilibrium trials. The only significant area at $p < 0.001$ (-3, 21, -3; $k = 20$, $t = 5.80$) is ventral striatum.

as a behavioral condition in which choices are optimal and beliefs rational, but also can be interpreted neurally as a *state of mind*: When choices, beliefs and 2nd-order beliefs all match up accurately, and are mutual best responses, there is only a minimal difference in activation between choice and belief, which means the mechanisms performing those tasks are highly overlapping.[22]

Figure 6 does show one important differential activation, however, in the ventral striatum. This region is involved in encoding reward value of stimuli and predicting reward (e.g., Schultz, 2000. This area is also differentially activated when we compare choice to the 2nd order belief task, t-statistic > 4 in several overlapping voxels). This difference could be due to the difference in rewards in the choice and belief tasks. Note that activation in FI is not significantly different between the C and B tasks in equilibrium (cf. Fig. 3), which is a clue that perceived ambiguity from choosing is lower when choices and beliefs are in equilibrium.

Figure 7 shows the C minus B differential activation in trials when choices and beliefs are out of equilibrium. Here we see some areas of activation similar to those in the overall C minus B subtraction.[23] The novel activity here is in the paracingulate frontal cortex region (Brodmann area *BA* 8/9; Fig. 7, upper left section). This region has appeared in mentalizing tasks in two studies. One is the Gallagher et al. (2002) study of "rock, paper, scissors"; a paracingulate area just anterior to the one in Fig. 7 is differentially active when subjects played human opponents compared to computerized algorithms.[24] McCabe et al. (2001) also found significant differential activations in the same area among subjects who were above the median in cooperativeness in a series of trust-like games, when they played humans versus computers.

In our tasks, of course, choosing and expressing belief are both done with another opponent in mind (in theory). Activation of the paracingulate region in our non-equilibrium $C > B$ subtraction and in Gallagher et al.'s and McCabe et al.'s human–computer difference suggests that people are reasoning more thoughtfully about their human opponent

[22] The difference between in- and out-of-equilibrium $C > B$ activity does not simply reflect the complexity of the games which enter the two samples, because separating the trials into easy (solvable by dominance for row or column) and hard (solvable in 2–3 steps) does not yield a picture parallel to Figs. 6–7. The difference is also not due to lower test power (there are fewer in-equilibrium than out-of-equilibrium trials) because the strategic areas active in Fig. 7 are not significantly activated in the in-equilibrium $C > B$ subtraction (paracingulate $t = 0.36$; dorsolateral prefrontal, $t = 1.34$).

[23] Note that the Fig. 3 activations, which pool all trials, do not look like a mixture of the Fig. 6 (in-equilibrium trials) and Fig. 7 (out-of-equilibrium trials) activities. However, the areas which are differentially active below the $p < 0.001$ threshold when all trials are pooled do tend to have activation in the in- and out-of-equilibrium subsamples, but activation is more weakly significant in the subsamples and vice versa. In the $C > B$ subtraction for out-of-equilibrium trials, the PCC is active at $p < 0.01$ and the ACC at $p < 0.005$. The dorsolateral prefrontal region (see Fig. 7) at $(-30, 30, 6, k = 14)$ which is active $(p < 0.001)$ in the out-of-equilibrium trials is just inferior to the region active in all trials $(-27, 48, 9, k = 14)$.

[24] In both conditions the subjects were actually playing against randomly chosen strategies (which is the Nash equilibrium for this game). The occasional practice of deception in economics experiments conducted by neuroscientists raises a scientific question of whether it might be useful to agree on a no-deception standard in this emerging field, as has been the stubborn and useful norm in experimental economics to protect the public good of experimenter credibility.

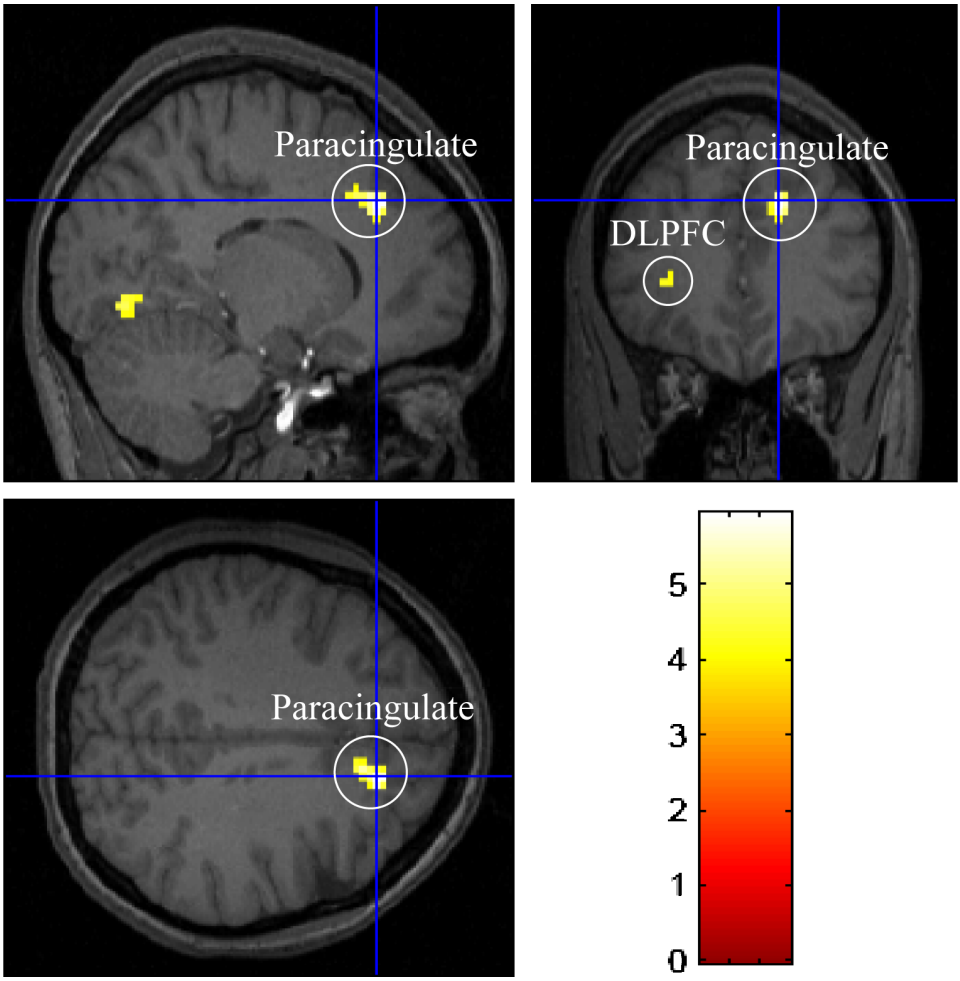


Fig. 7. This is your brain out-of-equilibrium: Areas of significant differential activation in $C > B$ for out-of-equilibrium trials. Largest area (15, 36, 33; $k = 39$; $t = 5.93$, 12 of 13 positive) is paracingulate cortex (BA 9), visible in all three sections. Posterior area in the sagittal section (left in upper left section) is occipital cortex (12, -75, -6; $k = 19$, $t = 4.84$). Ventral area in the coronal section (leftmost activity in the upper right section) is dorsolateral prefrontal cortex (-30, 30, 6, $k = 14$, $t = 4.85$).

when choosing rather than believing. This pattern is consistent with low-level strategic thinking in which players do not spend much time thinking about what others will do in forming beliefs, when they are out of equilibrium.

The difference we observe in brain activity in- and out-of-equilibrium is similar to Grether et al.'s (2004) fMRI study of bidding in the incentive-compatible Vickrey second-price auction. After players were taught they should bid their values (a dominant strategy), activity in the ACC was diminished.

2.5. Self-referential iterated strategic thinking: 2nd-order beliefs versus beliefs

The second comparison we focus on is differential activity in the brain when row players are asked what they think the column players think *they* (the row players) will do—their 2nd-order beliefs—compared to brain activity when they are just asked to state beliefs about what column players will do.

Figure 8 shows differential activity in the $2B$ condition, compared to B, in those trials where players were out of equilibrium.[25] The large ($k = 35$ at $p = 0.005$) voxel area is the anterior insula (a smaller subset of these voxels, $k = 3$, are still significant at $p = 0.001$).

The insula is the region in the brain responsible for monitoring body state and is an important area for emotional processing (see Fig. 9 for a picture of where the insula is). Parts of the insula project to frontal cortex, amygdala, cingulate, and ventral striatum. The insula is hyperactive among epileptics who feel emotional symptoms from seizures (fear, crying, uneasiness; Dupont et al., 2003), and in normal subjects when they feel pain, disgust and social anxiety. Sanfey et al. (2003) found that the insula was activated when subjects received low offers during the ultimatum game. Eisenberger et al. (2003) found the area was activated when subjects were made to feel socially excluded from a computerized game of catch. Importantly for us, the insula is also active when players have a sense of self-causality from driving a cursor around a screen (compared to watch equivalent cursor movement created by others; Farrer and Frith, 2001), or recall autobiographical memories (Fink et al., 1996). These studies suggest that insula activation is part of a sense of "agency" or self-causation, a feeling to which bodily states surely contribute. Our region overlaps with the area found by Farrer and Frith.

The insula activation in creating 2nd-order beliefs supports the hypothesis that 2nd order belief formation is not simply an iteration of belief formation applied to imagine how what other players believe about you. Rather, it is a combination of belief-formation and choice-like processes. We call this the self-referential strategic thinking hypothesis. The basic facts that C and $2B$ activations tend to be very similar, C and $2B$ choices often match up (Table 2), and that activations in the C and $2B$ tasks both tend to be different from B in similar ways,[26] supports this hypothesis too.

2.6. Individual differences: Brain areas that are correlated with strategic IQ

All the analyses above pool across trials and subjects (assuming random effects). Another way to approach the data is to treat each subject as a unit of analysis, and ask how activation is correlated with behavioral differences in skill, across subjects.

To do this we first calculate a measure of "strategic IQ" for each subject. Remember that subjects actually had a human opponent in these games. Since subjects did not receive any feedback until they came out of the scanner (and one of each of the C, B and $2B$ trials

[25] This $2B > B$ subtraction for the in-equilibrium trials yields no significant regions at $p < 0.001$. As noted earlier, this shows that being in equilibrium can be interpreted as a state of mind in which forming beliefs and 2nd-order beliefs are neurally-similar activities.

[26] Differential $C > B$ activation in the same insula region observed in the $2B > B$ subtraction is marginally significant ($t = 2.78$), and is positive for 10 out of the 13 subjects in the sample.

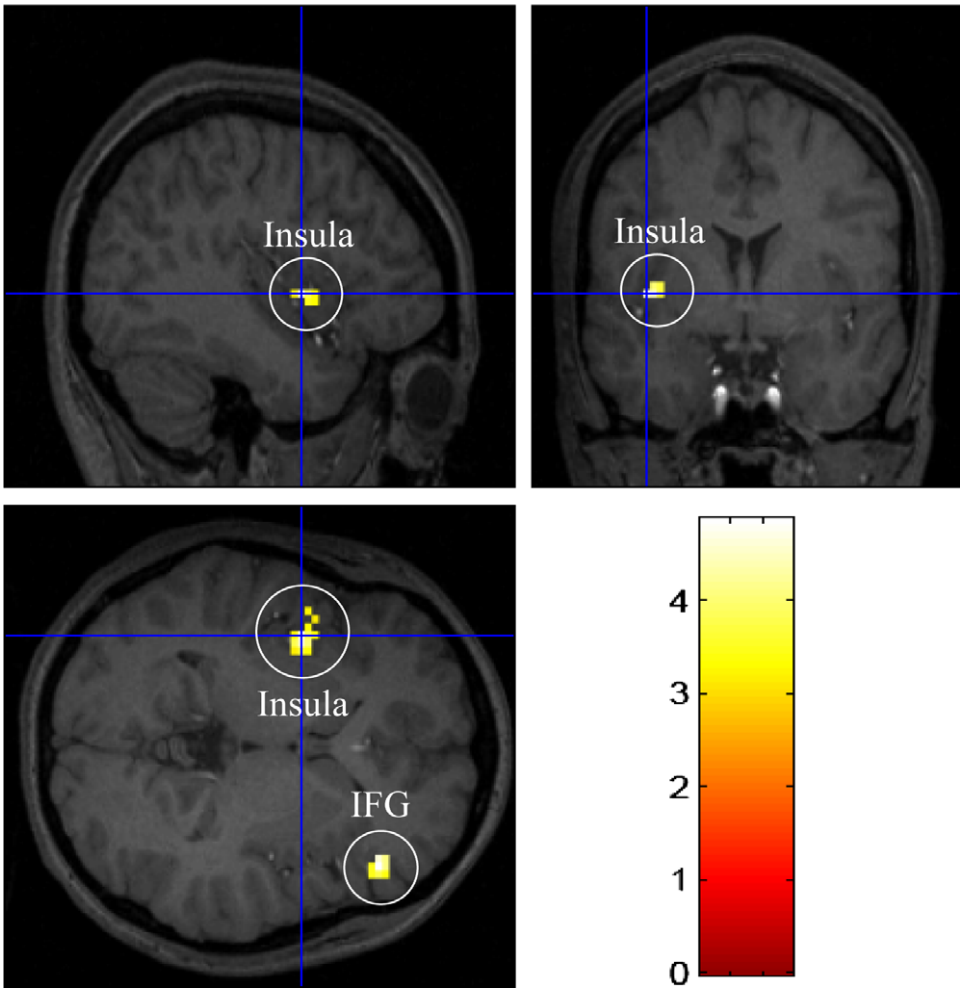


Fig. 8. Differential activity in iterated belief ($2B$) minus belief (B) conditions, out-of-equilibrium trials only. Significance level $p < 0.005$ (uncorrected). $N = 13$ because some subjects did not have enough non-Nash trials to include. Area visible in all three sections is left insula (-42, 0, 0, $k = 35$, $t = 4.44$, 12 of 13 positive). This area is still active but smaller in cluster size at lower p-values ($k = 9$ at $p = 0.002$, $k = 3$ at $p = 0.001$). The other active region in the transverse slice (lower left) is inferior frontal gyrus (45, 33, 0; $k = 13$, $t = 4.85$).

was chosen randomly for actual payment), it makes sense to judge the *expected* payoffs from their choices, and the accuracy of their beliefs, by comparing each row subject with the population average of *all* the column players.[27] We use this method to calculate the expected earnings for each subject from their choices, and from accuracy of their beliefs

[27] This is sometimes called a "mean matching" protocol. It smoothes out the high variance which results from matching each in-scanner subject with just one other subject outside the scanner.

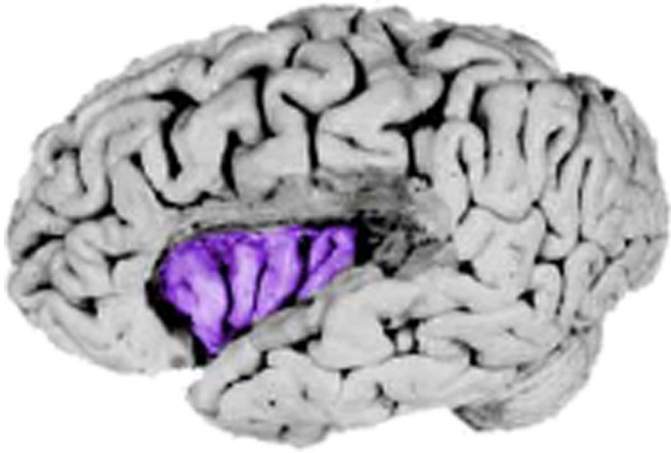


Fig. 9. A brain drawing showing insula cortex (in purple), as it would appear with the temporal lobe peeled back at the Sylvian fissure. The front of the brain (anterior) points to the left. Drawing reprinted with permission of Ralph Adolphs.

(i.e., how closely did their beliefs about column players' choices match what the column players *actually did*?) and similarly for 2nd-order beliefs. Their earnings in each of the three tasks are then standardized (subtracting the task-specific mean and dividing by the standard deviation). Adding these three standardized earnings numbers across the C, B and $2B$ tasks gives each subject's strategic IQ relative to other subjects. (The three numbers are only weakly correlated, about 0.20, across the three tasks, as is typical in psychometric studies.)

We then regressed activation during the choice task on these strategic IQs. The idea is to see which regions have activity that is correlated with strategic IQ.

We expected to find that players with higher strategic IQ might have, for example, stronger activation in ToM areas like cingulate cortex or the frontal pole BA 10. However, we found no correlations with strategic IQ in areas most often linked to ToM. Positive and negative effects of skill on activation in these areas might be canceling out. That is, players who are skilled at strategic thinking might be more likely to think carefully about others, which activates mentalizing regions. However, they may also do so more effortlessly or automatically, which means activity in those regions could be lower (or their responses more rapid).[28]

[28] The identification problem here is familiar in labor economics, where there is unobserved skill. If you run a regression on output (y) against time worked (t) across many workers, for example, it might be negative because the most skilled workers are so much more productive per unit time that they can produce more total output in a shorter time than slow workers, who take longer to produce less. Similarly, Chong et al. (2005) recorded response times of subjects and then inferred the number of steps of thinking the subjects were doing from their choices. Surprisingly, they found that the number of thinking steps was negatively correlated with response time. This puzzle can be explained if the higher-step thinkers are much faster at doing each step of thinking. It might also mean, as noted in footnote 14, that subjects classified as 0-step thinkers are actually doing something cognitively

However, choice-task activity in a $k = 13$ voxel cluster in the precuneus and a $k = 11$ voxel cluster in the caudate (dorsal striatum), are positively correlated with SIQ ($p < 0.001$ and $p < 0.05$ respectively), as shown in Fig. 10. The precuneus neighbors the posterior cingulate (PCC) and is implicated in "integration of emotion, imagery, and memory" (Greene and Haidt, 2002). Perhaps high-SIQ players are better at imagining what others will do, and this imaginative process in our simple matrix games uses all-purpose circuitry that is generally used in creating empathy or doing emotional forecasting involving others. The SIQ-caudate correlation shown in Fig. 7 is naturally interpreted as reflecting the greater certainty of rewards for the high SIQ subjects. This shows a sensible link between actual success at choosing and guessing in the games (experimental earnings) and the brain's *internal* sense of reward in the striatum.

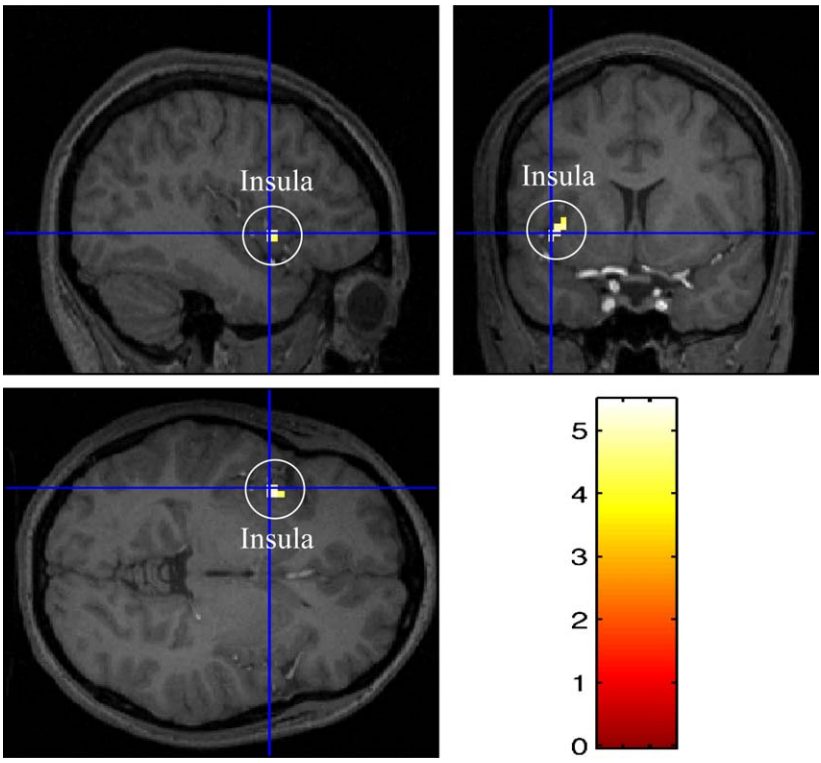
We also find interesting *negative* correlations between strategic IQ and brain activity during the choice task. Figure 11 shows the strong negative correlation between SIQ and activity in the left anterior insula (-39, 6, -3, $k = 25$) in the choice task, relative to a baseline of all other tasks, and also shows the insula region of interest in a sagittal slice.[29] Note that the low-SIQ players have an *increase* in activation relative to baseline (i.e., the y-axis values for those with negative standardized SIQ are positive), while the high-SIQ players have a decrease (negative y-axis values).

As noted above, the region of anterior insula in Fig. 11 which is correlated with SIQ is also differentially active in the $2B$ task relative to the B task. We interpret this as evidence that subjects are self-focused when forming self-referential iterated beliefs. The increase in insula activity might be an indication that too much self-focus in making a choice is a mistake—subjects who are more self-focussed do not think enough about the other player and make poorer choices and less accurate guesses. An alternative explanation is that subjects who are struggling with the tasks, and earn less, feel a sense of unease, or even fatigue from thinking hard while lying in the scanner (remember that the insula is activated by bodily discomfort). The higher insula activation for lower strategic IQ players may be the body's way of expressing strategic uncertainty to the brain. The fact that there is *de*activation in the choice task for higher SIQ players suggests a different explanation for them—e.g., by concentrating harder on the games they "lose themselves" or forget about body discomfort.

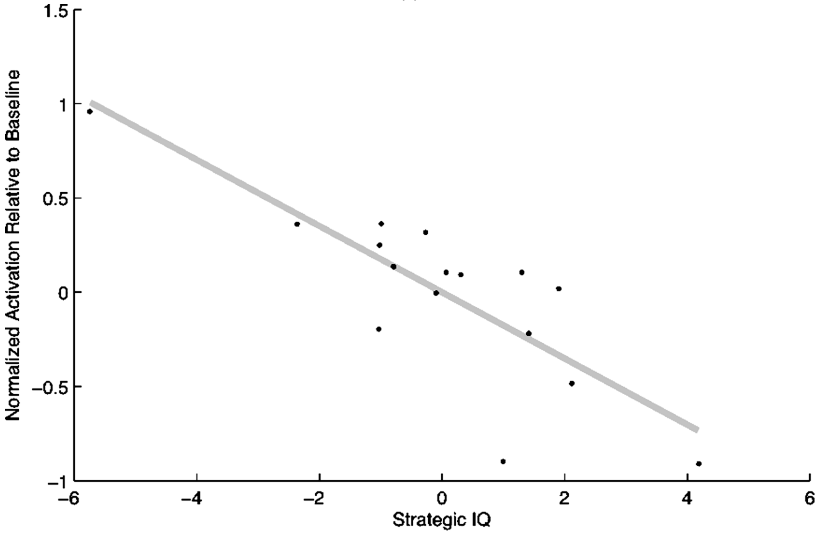
The fact that insula activity is negatively correlated with strategic IQ suggests that self-focus may be harmful to playing games profitably. A natural followup study to explore this phenomenon is to compare self-referential iterated beliefs of the form "what does subject A think that B thinks I (i.e., A) will do" with "what does *someone else* (C) think

sophisticated which the model cannot classify as higher-level thinking. (In some games, this even includes Nash equilibrium choices.)

[29] The y-axis is the regression coefficient in normalized signal strength (%) for each subject from a boxcar regression which has an independent dummy variable of $+1$ when the choice task stimulus is on the screen—from screen onset to the time that the subject made a decision with a button press—and 0 otherwise. The activation is scaled for each subject separately in percentage terms, so the results do not merely reflect differences in overall activation between subjects. The rank-order correlation corresponding to the correlation in Fig. 8(a) is -0.81 ($t = 5.08$) so it is not simply driven by outliers.

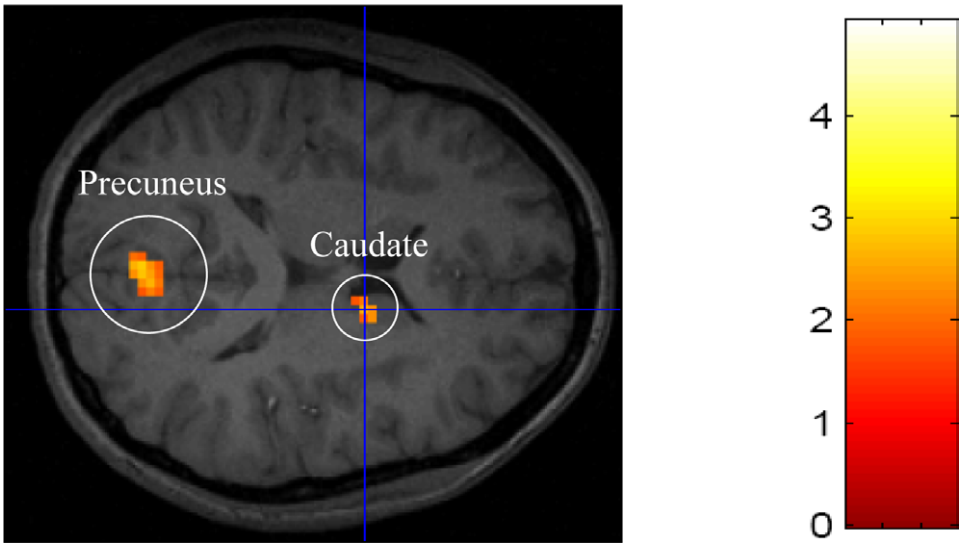


(a)

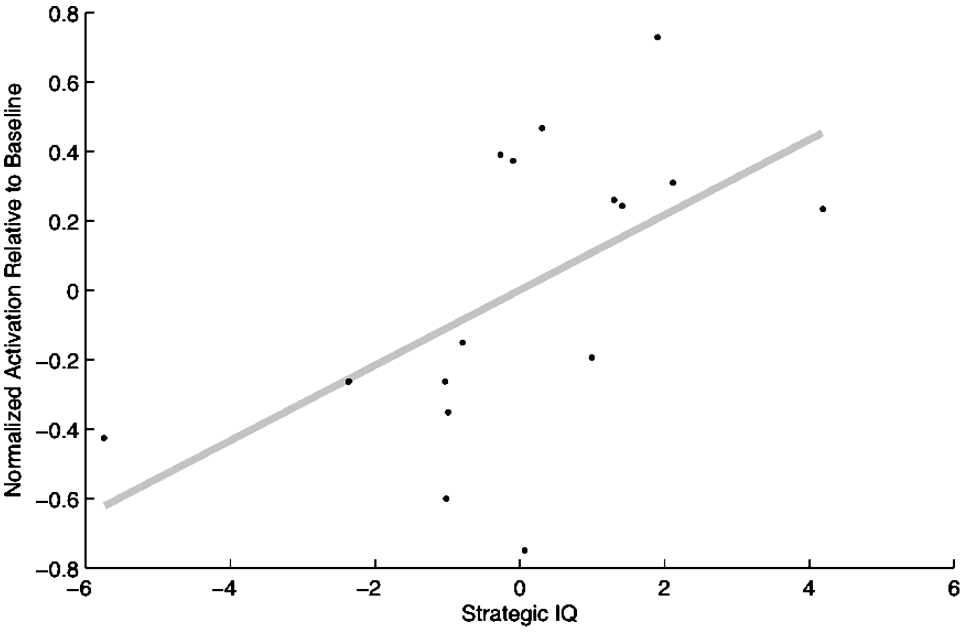


(b)

Fig. 10. (a) Sagittal slice showing L insula (-42, 6, -3, $k = 12$, $t = 5.34$), $p < 0.0005$. (b) Cross-subject correlation between L insula relative activity (y-axis) and relative SIQ (x-axis) ($r = -0.82$, $p < 0.0001$; rank-order correlation $= -0.81$).



(a)



(b)

Fig. 11. (a) Areas positively correlated with SIQ ($p < 0.05$): Precuneus (on left, 3, −66, 24, $k = 312$, $t = 4.90$), caudate (dorsal striatum) (12, 0, 15, $k = 11$, $t = 2.52$). (b) Cross-subject correlation between relative caudate activity (y-axis) and relative SIQ (x-axis) ($r = 0.56$, $p < 0.025$; rank-order correlation = 0.60).

B thinks *A* will do" (a non-self-referential 2nd order belief task). If self-focus harms the ability to guess accurately what *B* thinks you (*A*) will do, a third party (*C*) may be more accurate about guessing *B*'s beliefs about *A*'s move than *A* is. This possibility is related to psychology experiments on "transparency illusions" (Gilovich and Medvec, 1998) and "curse of knowledge" (Camerer et al., 1989; Loewenstein et al., 2003). In these experiments, subjects find it hard to imagine that other people do not know what they the subjects themselves know.

At this point, we do not know empirically if non-self-referential 2nd-order beliefs are more accurate than self-referential 2nd-order beliefs. The key point is that we would never have thought to ask this question until the neuroeconomic method suggested a link between insula activity, self-reference, and low strategic IQ. This is one illustration of the capacity of neural evidence to inspire new hypotheses.

3. Discussion and conclusion

Our discussion has two parts. We first mention some earlier findings on neuroscientific correlates of strategic thinking. Then we will summarize our central findings, and briefly conclude about how to proceed.

3.1. Other neuroscientific evidence on strategic thinking

An irony of neuroeconomics is that neuroscientists often find the most basic principles of rationality *useful* in explaining human choice, while neuroeconomists like ourselves hope to use neuroscience to help us understand *limits* of rationality in complex decision making (usually by suggesting how to weaken rationality axioms in biologically-realistic ways).[30] As a result the simplest studies of strategic thinking by neuroscientists focus on finding brain regions that are specially adapted to do the simplest kind of strategic thinking—reacting differently to humans compared to nonhuman computerized algorithms. As noted earlier, when subjects played mixed-equilibrium and trust games, respectively, against humans rather than computerized opponents, Gallagher et al. (2002) found activation in inferior frontal areas and paracingulate areas, and McCabe et al. (2001) found activity in the frontal pole (BA10), parietal, middle frontal gyrus and thalamic areas.

[30] The same irony occurs in models of risky choice where strategic thinking plays no role. Glimcher (2003) shows beautifully how simple expected value models clarified whether parietal neurons encode attention, intention or—the winner—something else (expected reward). At the same time, decision theorists imagine that neural circuitry might provide a foundation in human decision making for theories showing how choices violate simple rationality axioms—viz., that evaluations are reference-dependent, probabilities are weighted nonlinearly, and emotional factors like attention and optimism play a central role in risky decision making. A way to reconcile these views is to accept that simple rationality principles guide highly-evolved pan-species systems necessary for survival (reward, food, sex, violence) but that complex modern choices are made by a pastiche of previously-evolved systems and are unlikely to have evolved to satisfy rationality axioms only discovered in recent decades. Understanding such modern decisions forces us to become amateur neuoroscientists and learn about the brain, and talk to those who know the most about it.

A few other studies have focused on reward and emotional regions in games. Rilling et al. (2002) found striatal activation in response to mutual cooperation in a PD, which they interpret as a rewarding "warm glow" that sustains cooperation. De Quervain et al. (2004) find nucleus accumbens activation when third-party players sanction players who betrayed the trust of another player, showing a "sweet taste of revenge" (which is also price-sensitive, revealed by prefrontal cortical activity). The Sanfey et al. (2003) study on ultimatum games showed differential insula, ACC, and dorsolateral prefrontal activation for low offers. Singer et al. (2004) found that merely seeing the faces of players who had cooperated activated reward areas (striatum), as well as the insula. The latter finding suggests where game-theoretic concepts of a person's "reputation" are encoded in the brain and are linked to expected reward. Tomlin et al. (personal communication) find that the most anterior and posterior cingulate regions are active when players are processing what other players have done in a repeated trust games.

Many of these regions are also active in our study. The insula, active in evaluating low ultimatum offers and upon presentation of cooperating partners, is also active in creating 2nd-order beliefs in our study. The cingulate regions in Tomlin et al. are also prominent when players are choosing strategies, compared to guessing what other players will do.

Special subject pools are particularly informative in game theory, where stylized models assume players are both self-interested (almost sociopathic) and capable of great foresight and calculation. Hill and Sally (2002) compared autistic children and adults playing ultimatum games. About a quarter of their autistic adults offered nothing in the ultimatum game, which is consistent with an inability to imagine why others would regard an offer of zero as unfair and reject it. Offers of those adult autistics who offer more than zero cluster more strongly around 50% than the autistic childrens' offers, which are sprinkled throughout the range of offers. The child–adult difference suggests that socialization has given the adults a rule or "workaround" which tells them how much to offer, even if they cannot derive an offer from the more natural method of emotionally forecasting what others are likely to accept and reject. Gunnthorsdottir et al. (2002) found that subjects high on psychometric "Machiavellianism" ("sociopathy lite") were twice as likely to defect in one-shot PD games than low-Mach subjects.

A sharp implication in games with mixed equilibria is that all strategies that are played with positive probability should have equal expected reward. Platt and Glimcher (1999) found neurons in monkey parietal cortex that have this property. Their parietal neurons, and dorsolateral prefrontal neurons in monkeys measured by Barraclough et al. (2004), appear to track reinforcement histories of choices, and have parametric properties that are consistent with Camerer and Ho's (1999) dual-process EWA theory, which tracks learning in many different games with human subjects.[31]

[31] In the Camerer–Ho theory, learning depends on two processes: (1) A process of reinforcement of actual choices, probably driven by activity in the limbic system (striatum), and (2) a potentially separate process of reinforcing unchosen strategies according to what they would have paid (which probably involves a frontal process of counterfactual simulation similar to that involved in regret). A parameter δ represents the relative weight on the counterfactual reinforcement relative to direct reinforcement. Estimates by Barraclough et al. (2004) from activity in monkey prefrontal cortex support the two-process theory. They estimate two reinforcements: When the monkeys choose and win (reinforcement by Δ_1), and when they choose and lose (Δ_2). In their two-strategy

Still other studies have focussed on coarse biological variables rather than detailed brain processes. In sequential trust games, Zak et al. (2003) find a link between levels of oxytocin—a hormone which rises during social bonding (such as intimate contact and breast-feeding)—and trust. Gonzalez and Loewenstein (2004) found that circadian rhythms (whether you're a night or morning person) affected behavior in repeated trust (centipede) games—players who are "off peak" tended to cooperate less.

3.2. What we have learned

In this paper, we scanned subjects' brain activity using fMRI as they made choices, expressed beliefs, and expressed iterated "2nd-order" beliefs. There are three central empirical findings from our study:

- A natural starting point for translating game theory into hypotheses about neural circuitry is that most of the processes in making choices and forming beliefs should overlap when players are in equilibrium. Indeed, in trials where choices and beliefs are in equilibrium, this hypothesis is true—the only region of differential activation between choice and belief tasks is the striatum, perhaps reflecting the higher "reward activity" from making a choice compared to guessing. In general, however, making a choice (rather than making a guess) differentially activates posterior and anterior cingulate regions, frontal insula, and dorsolateral prefrontal cortex. Some of these regions are part of "theory of mind" circuitry, used to guess what others believe and intend to do. The cingulate activity suggests that brains are working harder to resolve cognitive-emotional conflicts in order to choose strategies.
- Forming self-referential 2nd-order beliefs—guessing what others think you will do—compared to forming beliefs, activates the anterior insula. This area is also activated by a sense of agency or self-causation (as well as by bodily sensations like disgust and pain). Combined with behavioral data and study of the time courses of activation, this suggests that guessing what others think you will do is a mixture of forming beliefs and making choices. For example, this pattern of activity is consistent with people anchoring on their own likely choice and then guessing whether other players will figure out what they will do, when forming a self-referential 2nd-order belief.
- Since subjects actually play other subjects, we can calculate how much they earn from their choices and beliefs—their "strategic IQ." When they make choices, subjects with higher strategic IQ have stronger activation in the caudate region (an internal signal of predicted reward which correlates with actual earnings) and precuneus (an area thought to integrate emotion, imagery and memory, suggesting that good strategic thinking may use circuitry adapted for guessing how other people feel and what they might do). Strategic IQ is negatively correlated with activity in insula, which suggests that

games, the model is mathematically equivalent to one in which monkeys are not reinforced for losing, but the unchosen strategy is reinforced by Δ_2. The fact that Δ_2 is usually less than Δ_1 in magnitude (see also Lee et al., 2004) is equivalent to $\delta < 1$ in the Camerer–Ho theory (less reinforcement in the second process from unchosen strategies), which corresponds to parametric measures from many experimental games with humans (see Camerer et al., 2004b).

too much self-focus harms good strategic thinking, or that poor choices are neurally expressed by bodily discomfort.

It is too early to know how these data knit together into a picture of brain activity during strategic thinking. However, activity in cingulate cortex (posterior, neighboring precuneus, anterior, and paracingulate) all appear to be important in strategic thinking, as does activity in dorsolateral prefrontal cortex, the insula region and in reward areas in the striatum. The most novel finding is that activity in creating self-referential 2nd-order beliefs activates insula regions implicated in a sense of self-causation. That interpretation, along with the fact that 2nd-order beliefs are highly correlated with choices, is a clue that higher-order belief formation is not a simple iteration of belief formation. Furthermore, the link between self-focus suggested by insula activity and its negative correlation with low strategic IQ suggests that third-party 2nd-order beliefs (*C* guessing what *B* thinks *A* will do) might be more accurate than self-referential 2nd-order beliefs (*A* guessing what *B* thinks *A* will do). This novel prediction shows how neural evidence can inspire a fresh idea that would not have emerged from standard theory.

Note that the study of brain activation is not really intended to confirm or refute the basic predictions in game theory; that kind of evaluation can be done just by using choices (see Camerer, 2003). Instead, our results provide some suggestions about a *neural* basis for game theory which goes beyond standard theories that are silent about neural mechanisms. Neural game theories will consist of specifications of decision rules and predictions about *both* the neural circuitry that produces those choices and its biological correlates (e.g., pupil dilation, eye movements, etc.). These theories should also say something about how behavior varies across players who differ in strategic IQ, expertise, autism, Machiavellianism, and so forth. Linking brain activity to more careful measurements of steps of strategic thinking is the next obvious step in the creation of neural game theory.

Acknowledgments

We got help from an understanding referee, special issue editor Aldo Rustichini, and audiences at ESA, FUR, Neuroeconomics 2004 (Kiawah), Iowa, NYU, and UCLA (both the economics audience and Marco Iacobonni's group). Advice and assistance from Ralph Adolphs, Cedric Anen, Sayuri Desai, Paul Glimcher, Ming Hsu, Galen Loram, Read Montague, John O'Doherty, Kerstin Preuschoff, Antonio Rangel, Michael Spezio, Damon Tomlin and Joseph Wang were also helpful. Steve Flaherty and Mike Tyszka's technical support was also very helpful.

Appendix A. Order of games and tasks, raw choice data in games fMRI regions in texts scans, methods, and instructions

In the Table A.1: In the "CGCB transform" column, in notation $Gx\ (r, c; Y - Z)$, Gx denotes name and letter, r and c are constants added to original CGCB payoffs to transform them to experimental currency payoffs we used, and $Y - Z$ denotes original rows or

Table A.1
Order of games, transformation from original CGCB games, and order of tasks for each game

Game	CGCB transform	Task order	Game type
1	$2A$ $(-10, -5; AA - BB)$	$C, B, 2B$	Row player has dominant strategy
2	$3A$ $(-20, +10)$	$2B, C, B$	Column player has dominant strategy
3	$5A$ $(+15, -13; A - C)$	$2B, B, C$	3×2 Game, 3 steps of dominance for row player
4	$5B$ $(-7, +11, B - C)$	$B, 2B, C$	3×2 Game, 3 steps of dominance for row player
5	$6A$ $(-17, -3; AA - BB)$	$C, 2B, B$	2×3 Game, 2 steps of dominance for row player
6	$6B$ $(+7, +0; AA - CC)$	$B, C, 2B$	2×3 Game, 2 steps of dominance for row player
7	$9A$ $(+19, +19; A - C)$	$C, B, 2B$	Row player has a dominant strategy
8	$9B$ $(0, 0)$	$B, C, 2B$	Column player has dominant strategy

Table A.2
Frequency of strategy choices $A - D$ and $AA - DD$ in our study vs. Costa-Gomes et al. (2001) data. (CGCB data denoted "C"; "n/a." denotes strategies that did not exist in a particular game)

	A		*B*		*C*		*D*		*AA*		*BB*		*CC*		*DD*	
#	New	*C*	New	*C*	New	*C*	New	*C*	New	*C*	New	*C*	New	*C*	New	*C*
1	.25	.21	.75	.79	n/a	n/a	n/a	n/a	.61	.69	.39	.31	n/a	n/a	n/a	n/a
2	.50	.86	.50	.14	n/a	n/a	n/a	n/a	.61	.92	.39	.08	n/a	n/a	n/a	n/a
3	.31	.21	.56	.79	.13	.00	n/a	n/a	.39	.23	.61	.77	n/a	n/a	n/a	n/a
4	.25	.14	.63	.71	.13	.14	n/a	n/a	.44	.46	.56	.54	n/a	n/a	n/a	n/a
5	.44	.79	.56	.21	n/a	n/a	n/a	n/a	.22	.38	.17	.00	.61	.62	n/a	n/a
6	.50	.36	.50	.64	n/a	n/a	n/a	n/a	.56	.77	.22	.08	.22	.15	n/a	n/a
7	.38	.08	.00	.00	.06	.00	.56	.92	.56	.46	.44	.54	n/a	n/a	n/a	n/a
8	.38	.07	.63	.93	n/a	n/a	n/a	n/a	.11	.08	.00	.00	.17	.00	.72	.92

columns that are switched to create our matrices. Example: Our game 3 (see text, Fig. 1) is CGCB game $5A$ with 15 added to all row payoffs, 13 subtracted from all column payoffs, and rows A and C switched. In game 6 there was a math error in one cell: for (B, AA) in our game we added 6 instead of 7 to the corresponding cell in CGCB, this did not change the strategic structure of the game.

A.1. Methodological details

Pairs of subjects were recruited on campus at Caltech through SSEL lab recruiting software.[32] One subject performed the tasks in the scanner, as the row player, while the other performed them in an adjacent room, as the column player. These three tasks were given in a random order for each game to control for order effects.

In the scanner each subject proceeds through a series of screens (like Fig. 1) one at a time, at their own pace. They press buttons on a box with 4 buttons to record their responses (choosing a row strategy in C and $2B$ tasks, and a column strategy across the bottom of

[32] Since Caltech students are selected by the admissions committee, for their unusual analytical skill, they are hardly a random sample. Instead, their behavior is likely to overstate the average amount of strategic thinking in a random population. This is useful, however, in establishing differential activation of regions for higher-order strategic thinking since the subjects are likely to be capable of higher-order thinking in games that demand it.

Table A.3
Distributions of free response times (25th, 50th—median—and 75th percentiles) in seconds across tasks and games

	Choice (C) median			Belief (B) median			2nd order ($2B$) median		
	25%	50%	75%	25%	50%	75%	25%	50%	75%
Game 1	11.4	20.4*	26.2	11.3	12.5	18.3	5.78	8.58	13.7
Game 2	8.87	11	20.9	6.58	7.75	13.5	14.5	22.3*	25.5
Game 3	8.58	10.7	16.3	9.61	11.2	20.2	16.8	25*	42.8
Game 4	2.91	7.83	15	11.4	16.6∗	32.9	6.08	10.8	23.9
Game 5	18.6	24.9*	37.3	6.55	11.6	16.7	7.92	10.1	23.9
Game 6	8.1	9.5	13.4	19.6	25.2*	42.8	4.61	6.54	15.1
Game 7	17.6	25.5*	42	6.08	9.23	14.1	6.58	10	17.3
Game 8	6.17	8.05	12.2	15.8	20.9*	26	5.67	11.1	13.8

Note: Response times are typically about twice as long for the first task presented.
* Denotes task which was presented first (e.g., the $2B$ task was first in game 3).

the screen in B tasks). After each response is recorded, there is a random lag from 6–10 seconds with a "fixation cross" to hold their visual attention in the center of the screen. The entire set of tasks took from 7 to 15 minutes.

At the end of the experiment 1 of the 24 tasks was chosen at random and subjects were paid according to their payoffs in the games at a rate of \$0.30 a point, if a choice task was picked, or were given \$15 for a correct answer to the belief tasks. All payments were in addition to a \$5 show-up fee.

Subjects in the scanner were debriefed after the experiment to control for any difficulties in the scanner and to get self-descriptions as to their strategies. The most common strategy described was a hybrid between cooperation and self-interest where they acted largely to maximize their own payoffs, but would cooperate if a small loss to herself would result in a large gain to the other player.[33] Some subjects seemed empirically more cooperative than others, but we treated all subjects similarly in our analysis.

To do the scanning, we first acquired a T1-weighted anatomical image from all row players. (This is a sharper-resolution image than the functional images taken during behavior so that we can map areas of activation onto a sharper image to see which brain areas are active.) Functional images were then acquired while subjects in the scanner played with subjects outside the scanner. They were acquired with a Siemens 3T MRI scanner using a T2-weighted EPI (TR = 2000 msec TE = 62 ms, 34 (32 for smaller heads) 3 mm slices), 32–34 slices depending on brain size. The slice acquisition order was $(2, 4, 6, \ldots, 1, 3, 5, \ldots)$. Data was acquired with one functional run per subject.

Data were analyzed using SPM2. Data were first corrected for time of acquisition, motion-corrected, coregistered to the T1-weighted anatomical image, normalized to the MNI brain and smoothed with an 8 mm kernel. The data were then detrended using a high-pass filter of periods greater that 128 seconds and an AR(1) correction.

[33] Subjects reporting this strategy included some who'd taken one or more classes in game theory and were familiar with the concept of Nash equilibrium.

Table A.4
Coordinates (x, y, z), cluster sizes (k), and t-statistics for subtractions and activity-behavior correlations reported in the text

Comparison	Signif. threshold	Area	x	y	z	Cluster size k	T-stat.
Choice > Belief (all games, all subjects)	$p = 0.001$	R Occipital Lobe	9	−78	9	202	6.77
		Cingulate Gyrus	−3	−12	33	24	5.12
		L Dorsolateral	−27	48	9	14	4.74
		ACC	6	42	0	33	4.62
		Frontal Insula	−42	12	−18	31	4.60
		R Cerebellum	9	−42	−27	17	4.49
		R Insula	36	12	−3	6	4.10
2nd order Belief > Belief (out of equilibrium games only)	$p = 0.001$	L Insula	−42	3	0	3	4.44
		Inferior Frontal Gyrus	45	33	0	8	4.85
	$p = 0.002$	L Insula	−42	3	0	9	4.44
		Inferior Frontal Gyrus	45	33	0	13	4.85
Choice-task activity *negatively* correlated with SIQ (games w/dominant strategies excluded)	$p = 0.0005$	Left Insula	−42	6	−3	12	5.34
		BA 11	−24	45	−15	6	5.47
		R Cerebellum	9	−78	−18	6	5.28
Choice-task activity *positively* correlated with SIQ (games w/dominant strategies excluded)	$p = 0.001$	Precuneus	3	−66	24	13	4.90
	$p = 0.05$	Caudate	12	0	15	11	2.52
		Precuneus	3	−66	24	312	4.90
		R Occipital/ Cerebellum	18	−87	−21	33	3.61
		Precentral Gyrus	−42	−18	42	45	2.90
		Occipital Gyrus	−27	−63	−12	12	2.35
		L Occipital	−36	−84	−15	6	2.28
		R Occipital	48	−69	36	13	2.24
Choice > Belief (in equil.)	$p = 0.001$	Ventral Striatum	−3	21	−3	20	5.80
Choice > Belief (out of equil.)	$p = 0.01$	Cingulate/BA 24	−3	−12	33	n.a.*	2.76
	$p = 0.005$	ACC	6	42	0	13	3.17
		ACC	15	42	0	13	3.33
	$p = 0.001$	Paracingulate	15	36	33	39	5.93
		L Dorsolateral	−30	30	6	14	4.85
		R Occipital	12	−75	−6	19	4.84
		R Occipital	30	−60	9	12	4.73

Note: R and L denote right and left hemispheres, respectively.

* Cluster size is not reported for this voxel since at this p-value there is so much activity that clusters overlap significantly. In this instance we do not feel that the cluster size is particularly informative, we report the t-statistic merely to show that there is some activity in the Choice > Belief (out of equil.) contrast that overlaps with what we see in the overall Choice > Belief contrast.

For each analysis the general linear model was constructed by creating dummy variables that were "on" from the stimulus onset time until the decision. These dummy variables were convolved with the standard hemodynamic response function. Standard t-tests were used to determine whether coefficient on one dummy variable is greater than that on another. Data from all the subjects were combined using a random-effects model. The cross-subject regressions regress regression coefficients of treatment affects across voxels against behavioral measures of strategic IQ.

A.2. Instructions to subjects

This is an experiment on decision-making. The decisions you make will determine a sum of money you will receive at the end of this experiment. If you read these instructions carefully, you stand to earn a substantial sum of money.

The questions in this experiment will all involve playing "matrix games." For the duration of the experiment Player 1 will be the "row player" and Player 2 will be the "column player." You will be shown a series of game that look something like this:

	Player 1's payoff			Player 2's payoffs		
	AA	*BB*	*CC*	*AA*	*BB*	*CC*
A	15	16	35	6	20	7
B	10	20	30	7	23	10
C	20	17	36	0	7	3

In these games the row player chooses a row and the column player chooses a column. Above, the row player would choose *A*, *B* or *C* and the column player would choose *AA*, *BB*, or *CC*. You will both make these decisions simultaneously and the cell that is determined by your choices determines your payoff. For example: If in the above example the row player had chosen *B* and the column player had chosen *CC*—The row player: Player 1, would receive 30 points and the column player: Player 2, would receive 10 points. If on the other hand Player 1 had selected *C* and Player 2 had selected *BB* the payoffs would be 17 for Player 1 and 7 for Player 2.

In addition to playing the games you will be asked some questions about the games during the course of the experiment. You will be asked what you think the other player will choose, and what you think the other player believes *you* will choose. These questions will be mixed in with the games in a random order so pay close attention to the question at the top of the screen. If you are Player 2 (outside the scanner) you may not go back and forth among the questions.

Payment

In addition to playing the games you will be asked some questions about the games during the course of the experiment. At the end of the experiment we will select one game or question and award you for your performance on that game or question. You will earn $15 for a correct answer to a question, or $0.30 a point for points earned in the game. In addition you be given a $5.00 show-up fee.

Questions:

1) What is your age?
2) What is you sex? (F/M)
3) Are you left handed or right handed?
4) Have you taken any courses in Economics and/or Game Theory. If so, please list these below.
   a.
   b.
   c.
   d.
   e.
5) In game a. below, if the row player chooses *C* and the column player chooses *AA*, what are both players' payoffs?
6) Practice games—If you're Player 1, choose a row. If you're Player 2, choose a column.

a.	Player 1's Payoffs			Player 2's Payoffs		
	AA	*BB*	*CC*	*AA*	*BB*	*CC*
A	10	12	48	20	19	12
B	5	30	25	78	42	60
C	20	13	0	50	7	9
D	43	16	27	15	10	13

b.	Player 1's Payoffs			Player 2's Payoffs		
	AA	*BB*	*CC*	*AA*	*BB*	*CC*
A	0	−1	1	0	1	−1
B	1	0	−1	−1	0	1
C	−1	1	0	1	−1	0

If you have any further question about how to play these games, ask the experimenter now.

References

Adolphs, R., 2003. Cognitive neuroscience of human social behavior. Nature Rev. Neurosci. 1, 165–178.

Allman, J., Hakeem, A., Watson, K., 2002. Two phylogenetic specializations in the human brain. Neuroscientist 8, 335–346.

Barraclough, D., Conroy, M.L., Lee, D., 2004. Prefrontal cortex and decision making in a mixed-strategy game. Nature Neuro 7, 404–410.

Cai, H., Wang, J.T.-Y., 2004. Over-communication and bounded rationality in strategic information transmission games: An experimental investigation. Games Econ. Behav. In press.

Camerer, C.F., 2003. Behavioral Game Theory: Experiments on Strategic Interaction. Princeton Univ. Press, Princeton.

Camerer, C.F., Ho, T., 1999. Experience-weighted attraction (EWA) learning in normal-form games. Econometrica 67, 827–874.

Camerer, C.F., Ho, T.-H., Chong, J.K., 2004a. A cognitive hierarchy theory of one-shot games. Quart. J. Econ. 119, 861–898.

Camerer, C.F., Ho, T., Chong, J.K., 2004b. Behavioral game theory thinking, learning, teaching. In: Hück, S. (Ed.), Advances in Understanding Strategic Behaviour: Game Theory, Experiments, and Bounded Rationality. Essays in Honour of Werner Güth. Palgrave Press, Basingstoke.

Camerer, C.F., Johnson, E., Sen, S., Rymon , T., 1994. Cognition and framing in sequential bargaining for gains and losses. In: Binmore, K., Kirman, A., Tani, P. (Eds.), Frontiers of Game Theory. MIT Press, Cambridge.

Camerer, C.F., Loewenstein, G., Prelec, D., 2004c. Neuroeconomics: Why economics needs brains. Scand. J. Econ. 106, 555–580.

Camerer, C.F., Loewenstein, G., Prelec D., 2005. Neuroeconomics: How neuroscience can inform economics. J. Econ. Lit. In press.

Camerer, C.F., Loewenstein, G., Weber, M., 1989. The curse of knowledge in economic settings. J. Polit. Economy 97, 1232–1254.

Castelli, F., Frith, C., Happé, F., Frith, U., 2002. Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. Brain 125, 838–849.

Chong, K., Camerer, C.F., Ho, T., 2005. A cognitive hierarchy theory of games and experimental analysis. In: Zwick, R. (Ed.), Experimental Business Research, vol. II. Kluwer Academic. In press.

Costa-Gomes, M.V., Crawford, V.P., 2004. Cognition and behavior in two-person guessing games: An experimental study. Working paper. University of California-San Diego, Department of Economics, April.

Costa-Gomes, M., Weizsäcker, G., 2004. Stated beliefs and play in normal form games. Working paper. University of York, December 7. Available from: http://econ.tau.ac.il/papers/game/mcggw11.pdf.

Costa-Gomes, M., Crawford, V., Broseta, B., 2001. Cognition and behavior in normal-form games: An experimental study. Econometrica 69, 1193–1235.

Dufwenberg, M., Gneezy, U., 2000. Measuring beliefs in an experimental lost wallet game. Games Econ. Behav. 30, 163–182.

Dupont, S., Bouilleret, V., Hasboun, D., Semah, F., Baulac, M., 2003. Functional anatomy of the insula: New insights from imaging. Surg. Radiol. Anat. 25, 113–119.

Eisenberger, N.I., Lieberman, M.D., Williams, K.D., 2003. Does rejection hurt? An fMRI study of social exclusion. Science 302, 290–292.

Farrer, C., Frith, C.D., 2001. Experiencing oneself vs. another person as being the cause of an action: The neural correlates of the experience of agency. Neuroimage 15, 596–603.

Fink, G.R., Markowitsch, H.J., Reinkemeier, M., Bruckbauer, T., Kessler, J., Heiss, W.D., 1996. Cerebral representation of one's own past: Neural networks involved in autobiographical memory. J. Neurosci. 16, 4275–4282.

Gallagher, H., Frith, C., 2003. Functional imaging of 'theory of mind.' Trends Cogn. Sci. 7, 77–83.

Gallagher, H., Anthony, J., Roepstorff, A., Frith, C., 2002. Imaging the intentional stance in a competitive game. Neuroimage 16, 814–821.

Gilovich, T., Medvec, V., 1998. The illusion of transparency: Biased assessments of others' ability to read emotional states. J. Personality Soc. Psych. 75, 332–346.

Gintis, H. 2003. Towards a unity of the human behavioral sciences. Working paper. Santa Fe Institute. Available from: http://www.umass.edu/preferen/gintis/unity.pdf.

Glimcher, P.W., 2003. Decisions, Uncertainty and the Brain: The New Science of Neuroeconomics. MIT Press, Cambridge.

Glimcher, P.W., Rustichini, A., 2004. Neuroeconomics: The consilience of brain and decision. Science 306, 447–452.

Goeree, J., Holt, C., 2004. A model of noisy introspection. Games Econ. Behav. 46, 365–382.

Gonzalez, R., Loewenstein, G., 2004. Effects of circadian rhythm on cooperation in an experimental game. Working paper. Carnegie Mellon.

Greene, J., Haidt, J., 2002. How (and where) does moral judgment work? Trends Cogn. Sci. 6, 517–523.

Grether, D., Plott, C.R., Rowe, D., Sereno, M., Allman, J.M., 2004. An fMRI study of selling strategy in second price auctions. Working paper No. 1189. Caltech. Available from: http://www.hss.caltech.edu/SSPapers/wp1189.pdf.

Gunnthorsdottir, A., McCabe, K., Smith, V.L., 2002. Using the Machiavellianism instrument to predict trustworthiness in a bargaining game. J. Econ. Psych. 23, 49–66.

Hedden, T., Zhang, J., 2002. What do you think I think you think? Theory of mind and strategic reasoning in matrix games. Cognition 85, 1–36.

Hill, E., Sally, D., 2002. Dilemmas and bargains: Theory-of-mind cooperation and fairness. Working paper. University College London.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., Camerer, C.F., 2005. Known and unknown probability: fMRI and lesion-patient evidence. Working paper. Caltech.

Huettel, S., Song, A., McCarthy, G., 2004. Functional Magnetic Resonance Imaging. Sinauer Associates, Inc.

Johnson, E.J., Camerer, C.F., 2004. Thinking backward and forward in games. In: Brocas, I., Castillo, J. (Eds.), The Psychology of Economic Decisions, vol. 2: Reasons and Choices. Oxford Univ. Press, Oxford.

Johnson, E.J., Camerer, C.F., Sen, S., & Rymon, T., 2002. Detecting failures of backward induction: Monitoring information search in sequential bargaining. J. Econ. Theory 104, 16–47.

Lee, D., Conroy, M.L., McGreevy, B.P., Barraclough, G.J., 2004. Reinforcement learning and decision making in monkeys during a competitive game. Cogn. Brain Res. 22, 45–58.

Loewenstein, G., Moore, D., Weber, R., 2003. Paying $1 to lose $2: Misperceptions of the value of information in predicting the performance of others. Working paper. Carnegie Mellon Department of Social and Decision Sciences. Available from: http://www.andrew.cmu.edu/user/rweber/Mispredicting_information.pdf.

McCabe, K., Smith, V.L., 2001. Neuroeconomics. In: Nadel, L. (Ed.), Encyclopedia of Cognitive Sciences. MIT Press, Cambridge.

McCabe, K., Houser, D., Ryan, L., Smith, V.L., Trouard, T., 2001. A functional imaging study of cooperation in two-person reciprocal exchange. Proc. Nat. Acad. Sci. 98, 11832–11835.

Miller, E.K., Cohen, J.D., 2001. An integrative theory of prefrontal cortex function. Ann. Rev. Neurosci. 24, 167–202.
Nagel, R., 1995. Unraveling in guessing games: An experimental study. Amer. Econ. Rev. 85, 1313–1326.
Platt, M.L., Glimcher, P.W., 1999. Neural correlates of decision variables in parietal cortex. Nature 400, 233–238.
de Quervain, D., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., Fehr, E., 2004. The neural basis of altruistic punishment. Science 27, 1254–1258.
Rabin, M., 1993. Incorporating fairness into game theory and economics. Amer. Econ. Rev. 83, 1281–1302.
Rilling, J.K., Gutman, D.A., Zeh, T.R., Pagnoni, G., Berns, G.S., Kilts, C.D., 2002. A neural basis for social cooperation. Neuron 35, 395–405.
Sanfey, A.G., Rilling, J.K., Aaronson, J.A., Nystrom, L.E., Cohen, J.D., 2003. The neural basis of economic decision-making in the ultimatum game. Science 300, 1755–1758.
Schultz, W., 2000. Multiple reward signals in the brain. Nature Rev. Neuro. 1, 199–207.
Siegal, M., Varley, R., 2002. Neural systems involved in 'theory of mind'. Nature Rev. Neuro. 3, 463–471.
Singer, T. Fehr, E., 2005. The neuroeconomics of mind reading and empathy. Amer. Econ. Rev. Pap. Proc., May.
Singer, T., Kiebel, S., Winston, J., Dolan, R., Frith, C., 2004. Brain responses to the acquired moral status of faces. Neuron 41, 653–662.
Stahl, D., Wilson, P., 1994. Experimental evidence on players' models of other players. J. Econ. Behav. Organ. 25, 309–327.
Weinstein, J., Yildiz, M., 2004. Finite-order implications of any equilibrium. Working paper. MIT. Available from: http://econ-www.mit.edu/graduate/candidates/research.htm?student=jonw.
Zak, P., 2005. Neuroeconomics. Philosophical Transactions Roy. Soc. B. In press.
Zak, Paul, Kurzban R., Matzner W., 2003. Oxytocin is associated with interpersonal trust in humans. Working paper. Claremont Graduate School Department of Economics.