Neuron
Article

Triangulating the Neural, Psychological, and Economic Bases of Guilt Aversion

Luke J. Chang,[1] Alec Smith,[2,5] Martin Dufwenberg,[2,6] and Alan G. Sanfey[1,3,4,*]
[1]Department of Psychology
[2]Department of Economics
University of Arizona, 1503 E. University Boulevard, Tucson, AZ 85721, USA
[3]Donders Institute for Brain, Mind & Behavior
[4]Behavioral Science Institute
Radboud University Nijmegen, 6525EN Nijmegen, The Netherlands
[5]California Institute of Technology, Pasadena, CA 91125, USA
[6]University of Gothenburg, S-405 30 Gothenburg, Sweden
*Correspondence: asanfey@u.arizona.edu
DOI 10.1016/j.neuron.2011.02.056

SUMMARY

Why do people often choose to cooperate when they can better serve their interests by acting selfishly? One potential mechanism is that the anticipation of guilt can motivate cooperative behavior. We utilize a formal model of this process in conjunction with fMRI to identify brain regions that mediate cooperative behavior while participants decided whether or not to honor a partner's trust. We observed increased activation in the insula, supplementary motor area, dorsolateral prefrontal cortex (PFC), and temporal parietal junction when participants were behaving consistent with our model, and found increased activity in the ventromedial PFC, dorsomedial PFC, and nucleus accumbens when they chose to abuse trust and maximize their financial reward. This study demonstrates that a neural system previously implicated in expectation processing plays a critical role in assessing moral sentiments that in turn can sustain human cooperation in the face of temptation.

INTRODUCTION

Daily life confronts us on a regular basis with social situations in which we sometimes place trust in those around us or alternately are entrusted by others. Often, this takes the form of informal agreements, with the promise of benefits to all concerned if mutual trust is upheld. As an example, imagine we are in a coffee shop, and another customer asks us to watch over her laptop as she steps outside to make a phone call. Assuming we repay this trust and do indeed protect her laptop, it is clear what the benefit to her is. But what is in it for us? These everyday informal situations are a mainstay of our social life, but there is surprisingly little experimental research examining the question of what motivates this behavior. Indeed, although we may painstakingly deliberate the merits of entering a formal legal contract, we rarely give much thought to the psychological foundations of these more mundane arrangements. However, these decisions serve as the foundation for a safe (Sampson et al., 1997) and economically successful society (Smith, 1984; Zak and Knack, 2001), and thus increased knowledge of the neural structures that underlie these behaviors can provide valuable clues into the mechanisms that underlie these behaviors of trust and reciprocity.

Understanding the dynamic processes of strategic interactions has traditionally been under the purview of the field of economics. Classical models of human behavior have typically assumed that people maximize their own material self-interest; however, a host of experimental evidence demonstrates that people appear to care about the payoffs of others (Camerer, 2003). This insight has consequently resulted in the development of a number of models that emphasize other-regarding preferences. These models typically consider either the distribution of payoffs (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999) or other player's intentions (Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Rabin, 1993) and posit that cooperation occurs largely as the result of a positive, prosocial motivation (Fehr and Camerer, 2007).

An alternative mechanism underlying trust and reciprocity that has received considerably less empirical attention concerns the influence of affective state on interactive decision making, specifically the role of anticipated guilt in deciding to help others. Guilt can be conceptualized as a negative emotional state associated with the violation of a personal moral rule or a social standard (Haidt, 2003) and is particularly salient when one believes they have inflicted harm, loss, or distress on a relationship partner, for example when one fails to live up to the expectations of others (Baumeister et al., 1994). Acting to minimize guilt can thus be a powerful motivator in the decision-making process. According to this proposal, we may be particularly vigilant of our neighbor's laptop, not because of any prosocial feeling, but rather because we anticipate feeling terrible if anything happened when the owner expected us to care for it. Supporting this idea, some research has demonstrated that people are indeed guilt averse and in fact often do make decisions to minimize their anticipated guilt regarding a social interaction. While these studies have provided evidence that beliefs about others' expectations motivate cooperative behavior (Charness and Dufwenberg, 2006; Dufwenberg and Gneezy, 2000; Reuben

et al., 2009; but see also Ellingsen et al., 2010) and that specifically thinking about a guilty experience can promote greater levels of cooperation (Ketelaar and Au, 2003), no study to date has directly demonstrated that guilt avoidance is the mechanism that underlies these decisions to cooperate. However, sophisticated methods from neuroscience such as fMRI can provide important insights into the underlying mechanisms.

It is important to note that there is at present very limited understanding of how complex social emotions such as guilt are instantiated in the brain. The few previous studies investigating the neural underpinnings of this mechanism have employed methods which may not realistically evoke natural feelings of guilt, such as script-driven imagery (e.g., "remember a time when you felt guilt") (Shin et al., 2000) or imaginary vignettes (e.g., "I shoplifted a dress from the store") (Takahashi et al., 2004). Because we contend that that the anticipation of guilt can motivate prosocial behavior, it is critical to explore how guilt impacts decision making while participants are actually undergoing a real social interaction. According to our conceptualization of guilt, people balance how they would feel if they disappointed their relationship partner against what they have to gain by abusing their trust. It is possible that during this process people may even experience a preview of their future guilt at the time of the decision, which may be what ultimately motivates them to cooperate.

Therefore, the present study attempts to address these questions by integrating theory and methods from the diverse fields of psychology, economics, and neuroscience to understand the neural mechanisms that mediate cooperative behavior. We utilize a formal model of guilt aversion (Battigalli and Dufwenberg, 2007) developed within the context of Psychological Game Theory (PGT; Battigalli and Dufwenberg, 2009; Geanakoplos et al., 1989), which provides a mathematical framework to allow individual utility functions to encompass beliefs—a feature essential for modeling emotions. Importantly, using a formal model provides a precise quantification of the amount of guilt anticipated in each decision, and can be used to predict brain networks that track this signal. The use of computational models has been instrumental in understanding the neural systems underlying complex cognitive constructs involved in decision making such as prediction error (O'Doherty et al., 2004), uncertainty (Preuschoff et al., 2006), and mentalizing (Hampton and O'Doherty, 2007). This approach provides a principled method for both illuminating the neural responses to feelings of guilt and also exploring how they directly guide social decision making.

For example, consider how behavior might be modeled in the commonly-studied Trust Game (TG) (Berg et al., 1995) using a guilt-aversion model. In this game, a player (the Investor) must decide how much of an endowment to invest with a partner (the Trustee – see Figure 1A). Once transferred, this money is multiplied by some factor (often 3 or 4), and then the Trustee has the opportunity to return money back to the Investor. If the Trustee honors trust, and returns money, both players end up with a higher monetary payoff than originally endowed. However, if the Trustee abuses trust and keeps the entire amount, the Investor takes a loss. The standard economic solution to this game uses backward induction and predicts that a rational and selfish Trustee will never honor the trust given by the Investor, and the Investor realizing this, should never place trust in the first place, and will invest zero in the transaction. In contrast, our model of guilt aversion posits that a rational Trustee is interested in both maximizing their financial payoff (M_2) *and* minimizing their anticipated guilt associated with letting their partner down. Anticipated guilt can be operationalized as the nonnegative difference between the amount of money the Investor expects back (E_1S_2) and the amount that the Trustee actually returns (S_2). Because the Trustee typically does not know the Investor's true belief, their expectation of this belief, referred to as their second order belief ($E_2E_1S_2$), can be used as a proxy.

$$U_2 = M_2 - \Theta_{12}(E_2E_1S_2 - S_2)^+ \quad (1)$$

According to this model, the Trustee's anticipated guilt is thus based on their second order beliefs. The weight placed on anticipated guilt in the utility function is modulated by a guilt sensitivity parameter (Θ_{12}), which can vary for each partner the Trustee encounters. Participants make decisions, which maximize this utility function. If they are sufficiently guilt averse ($\Theta_{12} > 1$), then they will maximize their utility by returning the amount that they expect their partner will return, otherwise ($\Theta_{12} < 1$) they will receive the most utility from keeping all of the money (see Figure S1 available online for a simulation).

While a number of studies have investigated the neural systems underlying Investor's initial decisions to trust (Delgado et al., 2005; King-Casas et al., 2005; Krueger et al., 2007), there have been surprisingly few that have studied the Trustee's corresponding decisions to cooperate (Baumgartner et al., 2009; van den Bos et al., 2009). Previous work has found evidence that decisions to cooperate in an iterated Prisoner's Dilemma Game are associated with the ventral striatum (Rilling et al., 2002). However, it is important to note that decisions to cooperate in sequential games (i.e., the TG) may be fundamentally different from those in simultaneous-move games (i.e., Prisoner's Dilemma Game) because of the ability to visibly choose before the other player in the former (McCabe et al., 2000, 2003). Neuroscientific investigations of the TG have shown that decisions to abuse trust are associated with activity in the vmPFC and PCC (van den Bos et al., 2009). This study also observed interesting individual differences indicating that when making selfish decisions, trust abusers exhibit more activity in the ventral striatum and less activity in the insula, as compared to cooperators. These results suggest that decisions to betray trust by trust abusers may be motivated by reward-related regions such as the ventral striatum and vmPFC, while decisions to cooperate may be associated with the insula for cooperators. Another study of Trustee behavior has focused on honoring promises to reciprocate rather than cooperation per se (Baumgartner et al., 2009). Here, the authors found that dishonest participants had greater amygdala activation as compared to honest participants when deciding whether or not to reciprocate their partner's trust. While both of these studies examining Trustee behavior have provided important insights into their respective questions of interest, neither has provided evidence directly addressing the specific mechanism that underlies the decision to cooperate in these interactive scenarios.

The aim of the present study is to use a theory-driven approach to examine the neural processes associated with

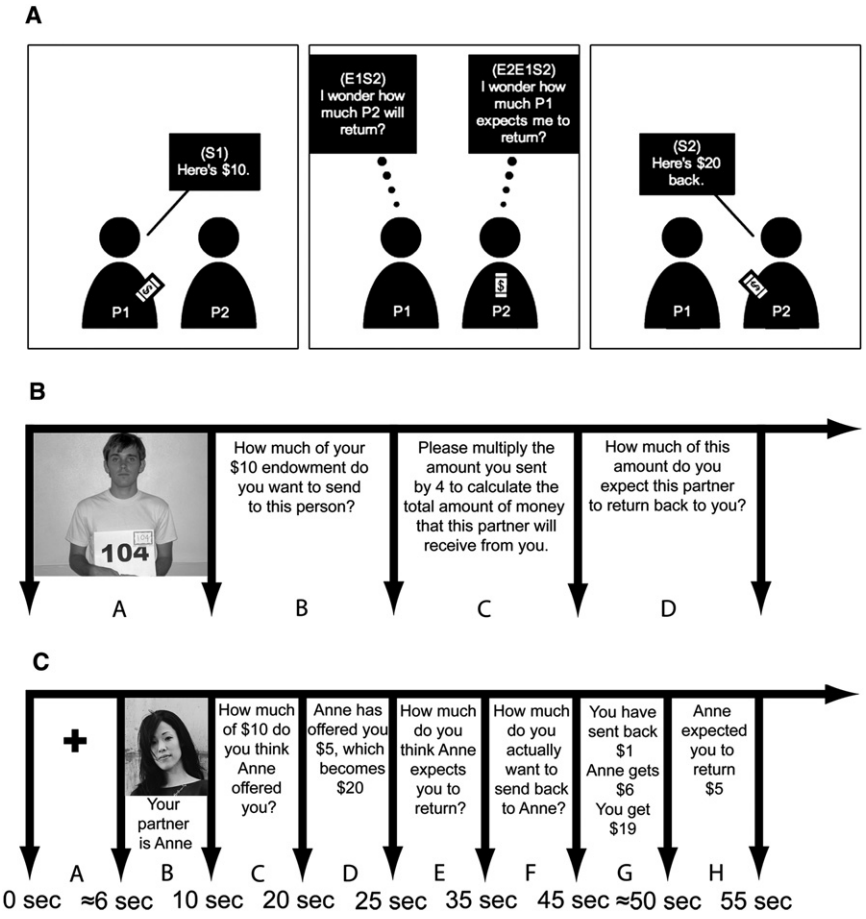


Figure 1. Trial Timeline

(A) Schematic of Trust Game (TG) with beliefs. Player 1 decides how much of their endowment they want to invest in Player 2 (S_1) and has an expectation about the amount of money that Player 2 will return (E_1S_2). The amount that Player 1 invests is multiplied by a factor of 4 by the experimenter. Player 2 has a belief about Player 1's expectation ($E_2E_1S_2$) and decides how much money to return back Player 1 (S_2).

(B) At session 1, all participants met as a group and played in the role of the Investor. After making an investment to every player, they were also asked how much of this amount (multiplied by 4) they believed their partner would return to them.

(C) Session 2 took place while the participants underwent functional magnetic resonance imaging and played in the role of Trustee. Participants first saw a fixation cross (A) and then a picture of their partner (B) on that round. Participants' beliefs about their partner's offer were then recorded (C) and then the actual offer was revealed (D). Next, participants' beliefs about the amount of money they believed their partner expected them to reciprocate were recorded (E) and they then decided how much they actually wanted to return (F). The final outcome was displayed (G) and then the partner's actual expectations were revealed (H).

guilt-motivated cooperation while the decision maker is immersed in a real, consequential interaction. As modeled by Equation 1, we elicit the participants' expectations and utilize them to isolate the neural systems involved in the anticipation of guilt. We predicted that the motivation to minimize anticipated guilt would induce participants to cooperate and that these cooperative decisions would therefore be associated with greater activity in the insula/acc and amygdala, based on previous studies of both guilt (Shin et al., 2000) and general negative affect (Calder et al., 2000; Damasio et al., 2000).

Thirty participants were recruited to play multiple single-shot rounds of a TG split over two sessions. Importantly, during this study we employed no deception, and therefore all participant interactions were both real and financially consequential. Use of this methodology allows us to examine actual interactions and also account for naturally occurring individual differences in both trust and reciprocity. During Session 1, all participants played as Investor and made an offer to every other participant in the experiment. In addition, we asked each participant to report the amount of money that they expected their partner to return (E_1S_2). Seventeen of these participants were recruited to play as the Trustee in a subsequent imaging session. During Session 2, each of these participants played 28 single-shot rounds of the TG as the Trustee while undergoing functional magnetic resonance imaging (fMRI). During the TG they received the actual offers made by each Investor during Session 1 (see Figure 1 for a trial timeline of both sessions). After learning about the amount of money player 1 sent, we first elicited the Trustee's second-order beliefs about the amount of money that they believed the Investor expected them to return ($E_2E_1S_2$). Participants could then return any amount of their multiplied investment in 10% increments (S_2). At the conclusion of Session 2, all participants were shown a recap of each round, and their subjective counterfactual guilt was assessed (see methods).

RESULTS

Behavioral Results

Our behavioral results demonstrated that participants behaved in a similar fashion to previous TG experiments (Camerer, 2003; Figure 2). The Investor usually sent some amount of their endowment to the Trustee, with the Trustee being quite accurate in predicting this investment (mixed effects regression, two-tailed; b = 0.15, se = 0.06, t = 2.29, p = 0.02). The Trustee was also generally accurate in predicting the Investors' expectations (b = 0.85, se = 0.06, t = 15.20, $p < 0.001$; Figure 3A). Supporting our model of guilt aversion, the Trustee used these expectations to guide their decision-making behavior, as they typically returned close to the amount of money that they believed their partner expected them to return (b = 0.90, se = 0.04, t = 21.32, $p < 0.001$; Figure 3B). Finally, participants reported that they

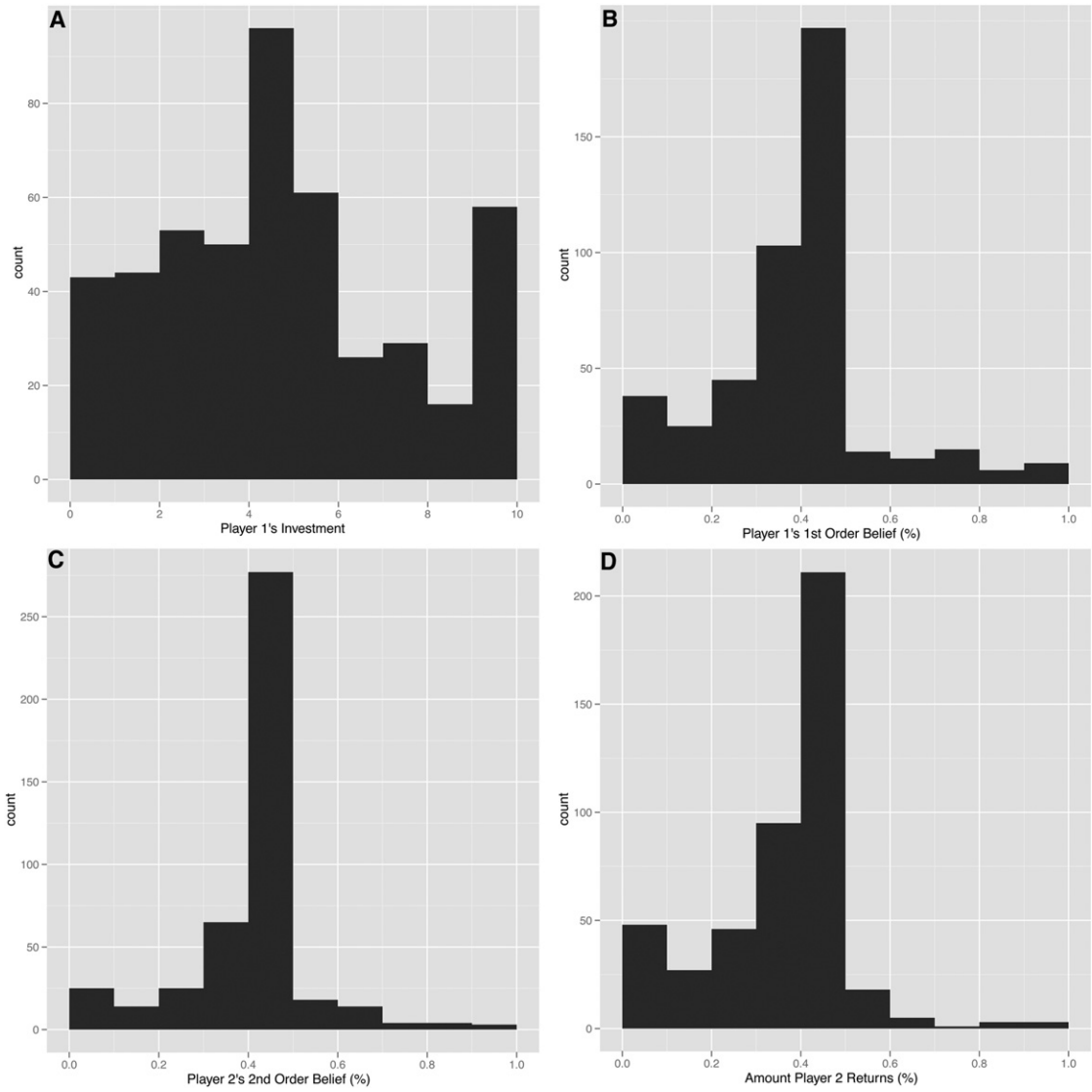


Figure 2. Behavioral Results

(A) Histogram of the Investor's Investment for all trials for all participants (mean = 51.7%, sd = 20.7%).

(B) Histogram of the percentage of their investment (multiplied by 4) that they expect the Trustee to return (1st Order Belief) (mean = 40.81%, sd = 10.44%).

(C) Histogram of the percentage of the Investor's investment (multiplied by 4) that the Trustee believes the Investor expects them to return (2nd Order Belief) (mean = 44.33%, sd = 3.52%).

(D) The percentage of the Investor's investment (multiplied by 4) that the Trustee decides to return (mean = 38.37%, sd = 7.80%).

would have felt more counterfactual guilt had they chosen to return less money than they actually did (b = 0.14, se = 0.03, t = 4.14, p < 0.001; Figure 3C). Taken together, these results suggest that participants behaved in a manner consistent with our model of guilt aversion.

Neuroimaging Results

We conducted several different analyses to examine the neural mechanisms underlying guilt aversion. First, a main contrast identified the neural processes underlying decisions that were consistent with the predictions of the guilt-aversion model (i.e., match expectations or not). Second, we explored processes that tracked parametrically with the predictions of the model. Third, we examined whether these processes could be explained by individual differences in guilt sensitivity estimated from their subjective counterfactual guilt ratings. Finally, we investigated the functional relationships between regions within the previously identified networks.

Main Contrast

To characterize the neural processes underlying the behavioral results, we attempted to isolate the two sources of value in Equation 1—the minimization of anticipated guilt and the maximization of financial reward. To do this, we compared trials during the decision phase in which participants returned the exact

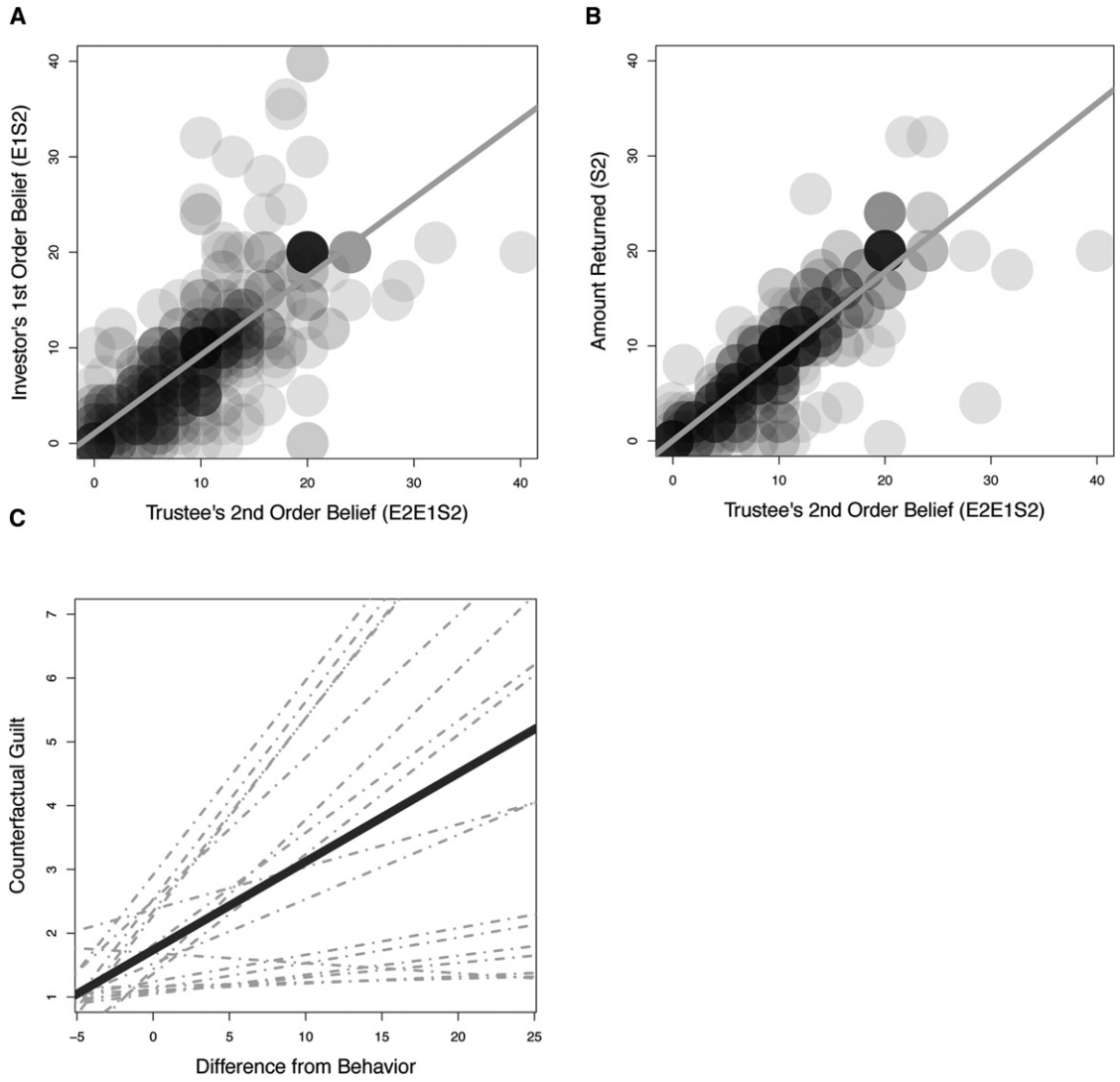


Figure 3. Behavioral Results

(A) Investor's first-order belief (E_1S_2) by the Trustee's second-order belief ($E_2E_1S_2$).

(B) The amount returned by the Trustee (S_2) by their second-order belief (see Table S1 for additional analyses).

(C) Participant's self-reported counterfactual guilt (the amount of guilt they would have felt had they returned less money) by the difference from their hypothetical choice from their actual behavior. The dotted lines represent participant's best linear unbiased predictors (BLUPs).

amount they believed their partner expected (i.e., minimized their anticipated guilt) to trials in which they returned less than they believed their partner expected (i.e., enhanced their financial reward). The duration of the decision phase was modeled as the time to decision. There was no significant difference in the response time between trials in which participants matched expectations (mean = 3412.29 ms, sd = 1310.65) as compared to trials in which they returned less than their expectation (mean = 3666.87 ms, sd = 1475.47; b = 0.25, se = 0.14, t = 1.80, p = 0.08). It is important to note that this response time is not particularly meaningful as participants were required to scroll through their choices and the starting point was random (see Experimental Procedures). The contrast, illustrated in Figure 4, revealed increased activity in the insula, supplementary motor area (SMA), dorsal anterior cingulate (DACC), dorsolateral prefrontal cortex (DLPFC), and parietal areas, including the temporal parietal junction (TPJ), when participants matched their second-order beliefs about their partner's expectations, thus minimizing guilt. Returning less than their second-order belief, and thereby increasing financial gain, was associated with greater activity in the ventromedial prefrontal cortex VMPFC, bilateral nucleus accumbens (NAcc), and dorsomedial prefrontal cortex (DMFPC) (See Table S2 for all identified regions).

Parametric Contrast

While the main contrast illustrates regions associated with minimizing expected guilt as compared to maximizing financial payoff, an additional question of interest is whether these activations change parametrically as a function of the actual deviation

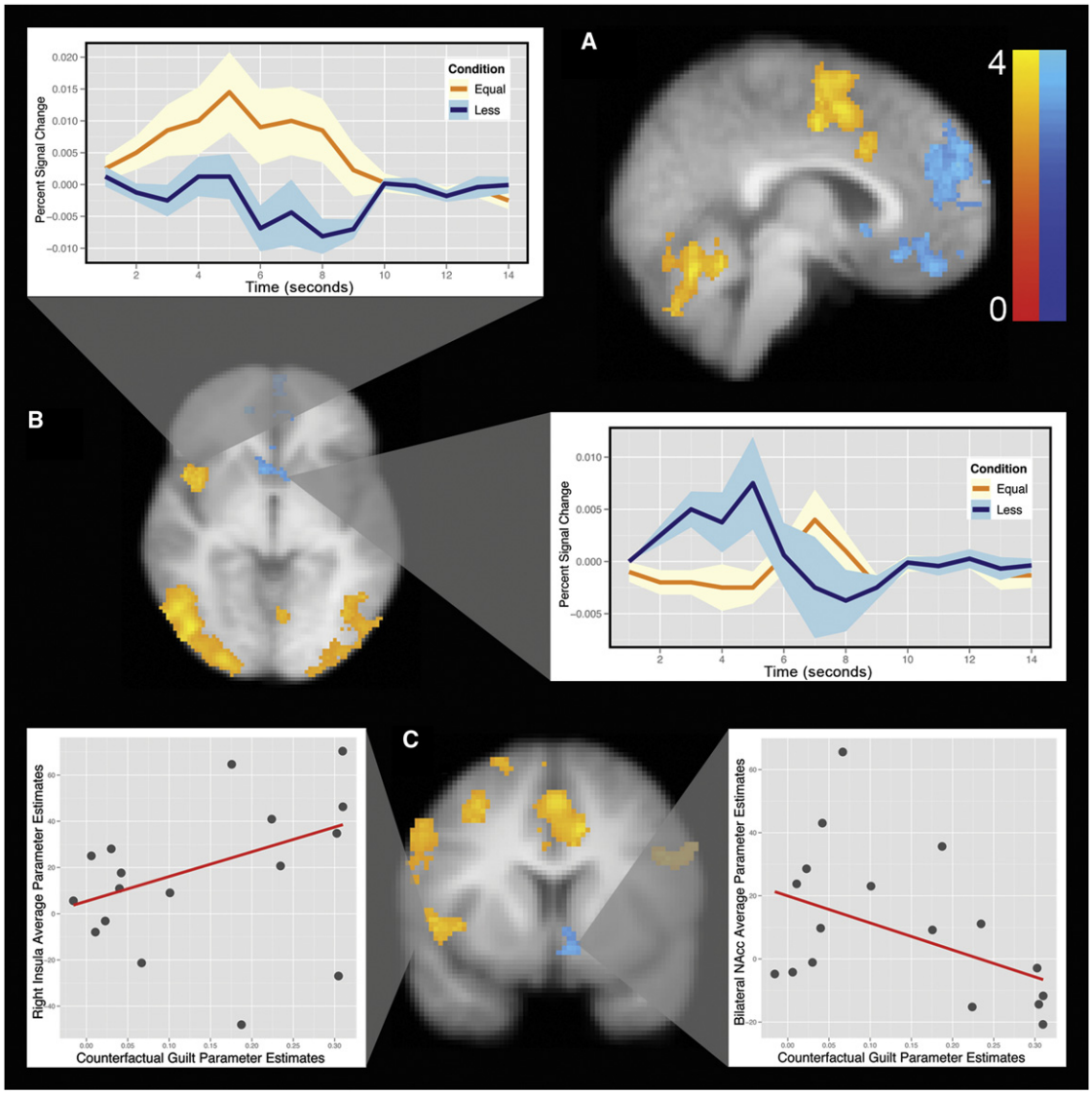


Figure 4. Minimizing Guilt Compared to Maximizing Financial Reward

(A) Increased activity (yellow) in the SMA, ACC, and cerebellum when matching expectations. Increased activity (blue) in the NAcc, VMPFC, and DMPFC can be seen when participants returned less than their second-order belief. The color map indicates Z values between 0 and 4. Error bars on the peristimulus plots reflect ± 1 standard error.

(B) Increased activity (yellow) in the insula when matching expectations and increased activity (blue) in the bilateral NAcc when returning less than their expectations.

(C) Increased activity in the insula, SMA, and right DLPFC (yellow) when matching expectations and increased activity (blue) in the left NAcc when returning less than expectations. The left blowup depicts the relationship between participant's counterfactual guilt sensitivity and their average activity for the insula. The right blowup depicts participant's estimated counterfactual guilt sensitivity and their average activity in the bilateral NAcc. See Figure S1 for a blowup of the SMA. Images are presented using radiological conventions (right = left) on the participant's average high resolution T1 image. The images are whole-brain thresholded using cluster correction $Z > 2.3$, $p < 0.05$.

See also Figure S2 and Table S2.

from matching expectations. To address this question we tested a parametric contrast that compared trials in which participants matched expectations to linear deviations from expectations (in 10% increments). Similar to the main contrast, matching expectations was associated with increased activity in the right insula, right DLPFC, SMA, ACC, and precuneous (see Figure 5 and Table S3). Returning incrementally less than expectations was associated with increased activity in the bilateral NAcc and MPFC (including VMPFC, DMPFC, and ACC).

However, participants systematically made slightly less money in trials in which they matched expectations (mean = \$12.28, sd = 5.88) compared to trials in which they returned less than they believed the other player expected (\$14.58, sd = 6.79; beta = −2.08, t = 2.53, $p < 0.05$). To address this potential

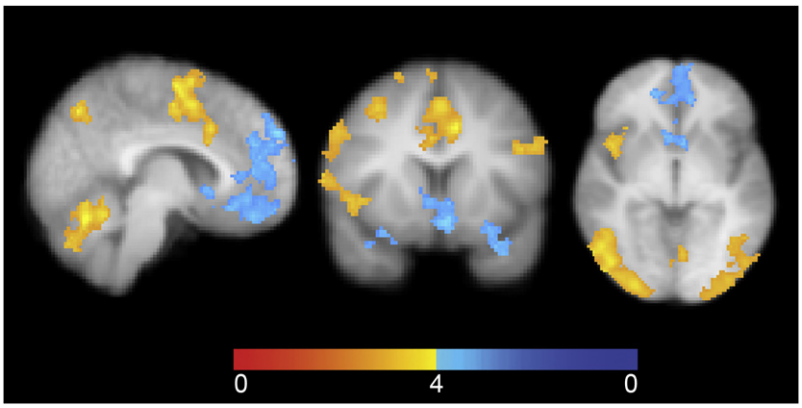


Figure 5. Parametric Contrast between Matching Expectations and Returning Less Than Second-Order Beliefs

This figure reflects the parametric contrast (+6 −1 −2 −3) of the regressors, indicating matching expectations, returning 10% less than expectations, returning 20% less than expectations, and returning +30% less than expectations. Images are displayed in radiological orientation (left = right) and are thresholded using whole brain cluster correction, $Z > 2.3$, $p < 0.05$. Color maps reflect Z values between 0 and 4. See also Figure S3 and Table S3.

confound and to rule out the possibility that the insula is simply tracking forgone financial payoffs rather than guilt aversion, we ran an additional analysis (see Supplemental Information) that allowed us to examine the effect of matching expectations while controlling for the amount of money that subjects return (i.e., their forgone financial payoff). Consistent with our interpretation, matching expectations was associated with increased activity in the insula, ACC, SMA, bilateral DLPFC, and TPJ. Regions associated with reward maximization (i.e., returning less than expectations) no longer survived cluster correction after controlling for forgone financial rewards, presumably as a consequence of high multicollinearity (see Figure S3 and Table S4).

Individual Differences

These data support the intriguing possibility suggested by our model that distinct networks may be processing competing motivations to either increase reward or decrease one's anticipated guilt. To examine this hypothesis further, we employed an individual differences approach in which we explored the relationship between differences in self-reported counterfactual guilt, assessed independently of the game, and our regions of interest across participants (see Figures 4C and S2; Experimental Procedures). Results from a robust regression (one-tailed) indicated that increased guilt sensitivity is positively related to increased activity in the insula and SMA ($b = 106.92$, $se = 50.44$, $p = 0.05$ and $b = 99.64$, $se = 46.49$, $p = 0.02$, respectively). That is, participants who reported that they would have felt more guilt had they returned less money showed increased insula and SMA activity when they matched expectations. In contrast, we observed a negative relationship between guilt sensitivity and the NAcc ($b = -89.17$, $se = 44.28$, $p = 0.03$), indicating that participants who reported that they would have experienced no change in guilt had they returned less money demonstrated increased activity in the NAcc when making a decision to maximize their financial reward. This effect is anatomically specific to these regions, as there were no significant relationships observed between guilt sensitivity and the right DLPFC, left DLPFC, VMPFC, or DMPFC.

Interregional Correlations

While we have primarily focused on disentangling the neural systems associated with the motivations underlying decision behavior, we also observed a network of regions that have previously been associated with an executive control system (e.g., DLPFC, parietal regions, and SMA) (Miller and Cohen, 2001) when participants matched expectations. Consistent with work that has suggested that the insula and SMA may comprise a distinct network which signals the need for executive control (Sridharan et al., 2008), we observed positive relationships between the insula and SMA across subjects ($r(16) = 0.64$, $p < 0.01$) and also between bilateral DLPFC and the SMA ($r(16) = 0.74$, $p < 0.001$), but no relationship between the insula and DLPFC (Pearson correlations, two-tailed). These relationships are concordant with previous conceptualizations of PFC functioning (Miller and Cohen, 2001) and suggest that the insula may recruit the dlPFC for increased self-control via the SMA. Finally, we also observed a significant negative relationship between activity in the insula and the NAcc across subjects ($r(16) = -0.56$, $p = 0.02$), hinting at a possible reciprocal relationship between these two systems, a relationship also predicted by our model.

DISCUSSION

Utilizing a formal game theoretic model of utility maximization involving guilt aversion (Battigalli and Dufwenberg, 2007), we find compelling evidence that moral sentiments aid in producing cooperative behavior in a consequential social exchange. Our model formalizes the psychological construct of guilt as a deviation from a perceived expectation of behavior and in turn posits that trust and cooperation may depend on avoidance of a predicted negative affective state. Congruent with our model's predictions, we observed evidence suggesting that when participants chose whether or not to honor an investment partner's trust distinct neural systems are involved in the assessment of anticipated guilt and in maximizing individual financial gain, respectively. These results provide converging psychological, economic, and neural evidence that a guilt-aversion mechanism underlies decisions to cooperate and demonstrate the utility of an interdisciplinary approach in assessing the motivations behind high-level decision-making.

Our experimental paradigm adds to the standard TG methodology by also eliciting participants' (second-order) beliefs, allowing us to test the predictions of the guilt-aversion model.

In addition, we did not employ deception, and all participant interactions were financially consequential, which importantly allows us to examine real interactions and also account for naturally occurring individual differences in both trust and reciprocity. Consistent with previous work (Charness and Dufwenberg, 2006; Dufwenberg and Gneezy, 2000), our results indicate that participants do indeed engage in mentalizing and are in fact able to accurately assess their partners' expectations. Further, as proposed by the model, participants use these expectations in their decisions and frequently choose to return the amount of money that they believe their partner expected them to return. Based on the postexperimental ratings that assess counterfactual guilt, we can infer that the motivation to match expectations is guilt aversion. Indeed, participants report that they would have felt more guilt had they returned less money in the game.

The guilt-aversion model explored here is distinct to other models of social preference as it posits that participants can mentalize about their partner's expectations and that they then use this information to avoid disappointing the partner. In contrast, other models conjecture that people are (1) motivated by a "warm glow" feeling and find cooperation inherently rewarding (Andreoni, 1990; Fehr and Camerer, 2007), (2) motivated to minimize the discrepancy between self and others' payoffs (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999), or (3) motivated to reciprocate good intentions and punish bad intentions (Dufwenberg and Kirchsteiger, 2004; Rabin, 1993). The guilt-aversion model thus provides a different psychological account of cooperation than other models because it incorporates both social reasoning and social emotional processing. The model also makes the interesting prediction that a social emotion is in effect an expectation error signal (Montague and Lohrenz, 2007), which functions to motivate people to behave consistent with shared social expectations. There is preliminary evidence indicating that these different motivations may be mediated by distinct neural systems. For example, altruism may be associated with areas associated with reward processing in the ventral striatum (Rilling et al., 2002). Inequity aversion may be associated with OFC (Tricomi et al., 2010), and intention-based reciprocity may be associated with a theory of mind network including the TPJ and the MPFC (van den Bos et al., 2009).

To understand the neural mechanisms underlying our model, we attempted to dissociate the competing motivations to either minimize guilt or maximize financial gain by comparing trials in which participants chose to match their partners' expectations to trials in which they returned less than they believed their partner expected. Participants exhibited increased activity in the insula, SMA, DACC, DLPFC, and parietal areas, including the TPJ, when they minimized their anticipated guilt by returning the amount of money that they believed their partner expected them to return. These results are consistent with another study which examined Trustee's decisions to cooperate (van den Bos et al., 2009), indicating that the belief elicitation procedure did not appear to alter the neural processing of cooperative decisions. The insula, SMA, and ACC have been implicated in a number of negative affective states such as guilt (Shin et al., 2000), anger (Damasio et al., 2000), and disgust (Calder et al., 2000) as well as physical pain, social distress (Eisenberger et al., 2003), and empathy for other's pain (Singer et al., 2004; see Craig, 2009, for a review). These studies support our conjecture that the prospect of not fulfilling the expectations of another can result in a negative affective state, which in turn ultimately motivates cooperative behavior. Finally, it is interesting to note that the neural systems involved in making decisions that minimize anticipated guilt are remarkably similar to those previously demonstrated to be involved in the decision to reject unfair offers in the Ultimatum Game (Sanfey et al., 2003), suggesting that at least one function of this network may be to motivate adherence to shared social expectations (Montague and Lohrenz, 2007). Recent work on decisions to conform to a perceived social norm has uncovered the same network (Berns et al., 2010; Klucharev et al., 2009), which indicates that perhaps the function of this frequently observed network is to track deviations from expectations and bias actions to maintain adherence to the expectation such as a moral rule or social norm. Sanfey et al., (2003) find that this network biases behavior to punish norm violators, while we observe here that this network biases behavior to be congruent with a socially shared expectation. This interpretation is consistent with a wealth of work on expectations in other domains of cognitive neuroscience such as novelty detection (Downar et al., 2000), placebo effects (Wager et al., 2004), and error monitoring (Miller and Cohen, 2001), suggesting that the network may be domain general (Dosenbach et al., 2006) and extend to social decision making.

An alternative interpretation of our results is that Trustees feel empathy toward the Investor and anticipate their partner's anticipated disappointment, which motivates them to cooperate. Empathy (like guilt) is another nebulous construct, though has yet to be formalized. Both empathy and guilt aversion require the ability to represent another's mental state (i.e., theory of mind) and directly relate to other's disappointment. However, one crucial distinction between the two constructs is that empathy posits that the Trustee feels the Investor's anticipated emotion (e.g., disappointment), while guilt-aversion contends that the act of disappointing a partner produces an emotion in the Trustee (e.g., guilt), which is qualitatively different from what the Investor is experiencing. Though our current design cannot parse apart these two interpretations, nor can our imaging results, as both of these constructs likely involve the insula (Singer et al., 2004), future work might attempt to differentiate between these two closely related constructs from both theoretical and empirical perspectives.

When participants returned less than their second-order belief and thereby increased their own financial gain, we found activation associated with greater activity in the VMPFC, bilateral NAcc, and DMFPC. These results became even more pronounced when we examined parametric deviations from expectation. Consistent with previous work that has examined decisions to abuse trust (van den Bos et al., 2009), we find increased activity in the VMPFC when participants return less than they believe their partner expected and predict that damage to this region would likely impair the ability to form accurate expectations, producing the guilt insensitive pattern of behavior observed in patient work (Krajbich et al., 2009). More broadly, however, these regions (i.e., NAcc and VMPFC) have received

attention for their role in computing value (Rangel et al., 2008) and the anticipation and processing of both primary and secondary reward (Dreher and Tremblay, 2009). In addition, we observed activity in the DMPFC, which has been implicated in mentalizing (Amodio and Frith, 2006) or simulating another's mental state. This signal may indicate that participants are engaging in reasoning about their partner's potential reaction to their decision. Together, these results suggest that maximizing one's utility involves a process of weighing the costs and benefits of letting a relationship partner down.

It is possible that the network associated with matching expectations is tracking forgone financial payoffs rather than guilt aversion per se. However, this interpretation is unlikely because we continue to observe activity in the insula when participants match expectations after controlling for the amount of money that participants chose to return. To provide further support for our interpretation that the competing motivations to maximize financial gain and minimize anticipated guilt are associated with distinct regions, we examined the relationship between the regions of interest (as defined by the group analyses) and independently assessed individual differences in guilt sensitivity. Consistent with our interpretation, we find that participants who report that they would have experienced more guilt had they returned less money demonstrated increased insula and SMA activation when they matched expectations. Conversely, participants who claimed that they would not have experienced any additional guilt had they returned less money showed increased activity in the NAcc when they in fact returned less than they believed their partner expected them to return. This implies that there is individual variability in the way in which anticipated guilt influences decisions. People who are more guilt sensitive have increased activity in the network associated with moral sentiments, while people with less guilt sensitivity have greater activity in those areas associated with reward and value.

Together, our results suggest that participants who are guilt sensitive may experience moral sentiments via the insula and SMA, which signals that they will feel guilty if they believe they let their investment partner down. This notion that feelings can be used as information in the decision-making process has been discussed in other domains of decision making such as risk (Damasio, 1994; Loewenstein et al., 2001; Mellers et al., 1997; Slovic et al., 2002) and regret (Coricelli et al., 2005). According to this framework, people generate anticipated emotions about how they might feel after choosing a particular outcome, which ultimately predicts their decision (Mellers et al., 1997). Interestingly, anticipatory feelings associated with risk have been reliably associated with the anterior insula (Critchley et al., 2001) and ACC (Coricelli et al., 2005), which provides further support for our argument that guilt aversion is generated by a sampling of the sentiment in question and is processed by the cingulo-insular network. Importantly, this extends the notion of anticipatory emotions from individual decision making to social contexts. These feelings originating in the insula may recruit the DLPFC to override the competing motivation to maximize financial gain and overall result in participants honoring their partner's trust and returning their initial investment. If this neural account is accurate, then we would predict that disrupting the DLPFC, insula, or ACC/SMA would result in participants choosing to return less money in the TG, as has indeed recently been demonstrated (Knoch et al., 2009). However, we make the divergent predictions that while disrupting all regions would reduce cooperative behavior, disrupting the DLPFC would still result in an affective response, while disrupting the insula or ACC/SMA would in contrast blunt the experience of guilt. Our results also predict that inaccurate expectations should also influence cooperative behavior. Overestimating partners' expectations would result in excessive guilt and enhanced associated insula/ACC/SMA activation, while underestimating partners' expectations would temper participant's guilt and insula/ACC/SMA activation and ultimately reduce their levels of cooperation, which is consistent with findings with patients with VMPFC damage (Krajbich et al., 2009).

This study demonstrates the synergistic effects of applying a neuroeconomic approach to the study of higher-level socio-cognitive-affective processes. Imprecise psychological constructs such as guilt can be formally operationalized using sophisticated economic models. In turn, the integration of psychological constructs into economic models can substantially improve their ability to predict actual decision-making behavior, in comparison to classical approaches. Finally, and most importantly, this interdisciplinary approach allows these mathematically quantified psychological constructs to be examined at the neural level in order to both better specify the theoretical models, as well as further understand the interactions between neural systems.

To return to our original example, our results suggest that one reason why we choose to stand guard over a stranger's possessions for no obvious reward is because signals originating in the insula and SMA remind us that allowing something bad to happen to the laptop, and thus deviating from the owner's expectations, would lead to strong feelings of guilt in the event of an untimely theft. Ultimately, gaining a greater mechanistic understanding of the microprocesses that can occur at a neural level can help facilitate greater understanding of emergent properties of macro-level interactive behavior that play a vital role in creating and maintaining a harmonious society.

EXPERIMENTAL PROCEDURES

Participants

Thirty participants (mean age = 18.5, female = 30%) were recruited from the University of Arizona campus, all of whom were screened for any significant health or neurological problems. The experiment was approved by the local Institutional Review Board and consisted of two separate sessions. From this sample, all participants that were eligible to enter the MRI environment (n = 17) were recruited from Session 1 to participate in Session 2 (mean age = 18.5, female = 53%). One participant from session 1 was excluded as a result of erratic responses, and some of one participant's fMRI data from the second session was lost due to technical reasons. Participants were assumed to be strangers.

Experimental Design

At session 1, all participants met as a group, were assigned an identification number, and had their individual pictures taken. After the instructions to the game were explained, all pictures were presented one at a time to the entire group. While the pictures were being presented, each participant played in the role of the Investor with the pictured participant and was endowed with

$10 for the round. After making an investment on the round, they were then asked how much of this amount (multiplied by 4) they believed their partner would return to them. At the end of the session, participants were paid $5 for their participation.

A subset of participants (n = 17) were recruited from Session 1 to participate in the second session, in which they played the TG in the role of the Trustee while being scanned using functional magnetic resonance imaging (fMRI). Each participant had an individually tailored paradigm, in which they decided how much money they wanted to return to the other participants in the experiment, based on these partners' actual proposals to them from Session 1. Each participant played a total of 28 rounds, distributed over four runs. Each run lasted exactly 7 min including an extra 14 s fixation cross display at the beginning of the run to allow for T1 equilibrium, and another 21 s fixation cross at the end of the run (210 volumes per run). The timeline of events in a typical round can be seen in Figure 1B. The stimuli were presented using E-Prime software via VisuaStim goggles (Resonance Technologies Inc, IL, USA), and participants indicated their answers by using a two-button fiber optic response box. Responses changed in 10% increments on each button press. These increments were randomly selected to either increase from $0 or decrease from the maximum amount of money for that round (which varied depending on how much had been sent by the partner), ensuring that the number of button presses was orthogonal to the amount of money selected, removing effects of any motor confounds. After participants selected their chosen amount of money, they used the second button to confirm this response.

After participants completed scanning, they rated their counterfactual guilt by indicating on a 7-point Likert scale the amount of guilt they believed they would have experienced had they returned a different amount of money, and were then paid a $20 participation fee. Finally, at the conclusion of the entire experiment, all participants were paid 50% of their earnings for one randomly selected trial. If participants participated in both sessions, they were paid for two separate trials. Participants in the first session that correctly predicted their partner's behavior for the trial selected received an additional $2 bonus (Charness and Dufwenberg, 2006; Dufwenberg and Gneezy, 2000). Only identification numbers were provided at the time of payment, thus ensuring that Trustees' responses were completely anonymous. No deception was employed in this study.

Data Acquisition

Each scanning session included a T1-weighted MPRAGE structural scan (TR = 11 ms, TE = 4 ms, matrix = 256 × 256, slice thickness = 1 mm, gap = 0 mm) and four functional runs. Functional scans were acquired in the axial plane using a 3-shot multiple echo planar imaging (MEPI) GRAPPA sequence which aided in reducing geometric distortions (Newbould et al., 2007). Parameters were optimized to maximize signal in regions associated with high susceptibility artifact (e.g., orbitofrontal cortex and medial temporal lobe) (Stöcker et al., 2006; Weiskopf et al., 2006) (TR = 2000 ms, TE = 256 ms, matrix = 96 × 96, FOV = 192 mm, slice thickness = 3.0 mm, 42 axial slices).

Data Preprocessing

Functional imaging data were preprocessed and analyzed using the FSL Software package 4.1.4 (FMRIB, Oxford, UK). The first three volumes of each functional run were discarded to account for T1 equilibrium effects. Images were corrected for slice scan time using an ascending interleaved procedure. Head motion was corrected using MCFLIRT using a six parameter rigid-body transformation. Images were spatially smoothed using a 5 mm full-width at half maximum Gaussian kernel. A high-pass filter was used to cut off temporal periods longer than 66 s. All images were initially coregistered to the participant's high-resolution structural scan and were then coregistered to the MNI 152 person 2 mm template using a 12 parameter affine transformation. All functional analyses are overlaid on the participants' average high-resolution structural scan in MNI space.

General Analysis Methods

A three-level mixed-effects general linear model (GLM) was used to analyze the imaging data. A first-level GLM was defined for each participant's functional run that included a boxcar regressor for each epoch of interest (e.g., decision phase) convolved with a canonical double-gamma hemodynamic response function (HRF). The duration of epochs in which participants submitted a response were modeled using the participant's reaction time (Grinband et al., 2008). To account for residual variance, we also included the temporal derivatives of each regressor of interest, the six estimated head movement parameters, and any missed trials as covariates of no interest. The resulting general linear model was corrected for temporal autocorrelations using a first-order autoregressive model. A second-level fixed effects model was fit for each subject to account for intrarun variability. For each participant, contrasts were calculated between parameter estimates for different regressors of interest at every voxel in the brain. A third-level mixed-effects model using FEAT with full Bayesian inference (Woolrich et al., 2004) was used to summarize group effects for every specified contrast. Statistical maps were corrected for multiple comparisons using whole-brain cluster correction based on Gaussian random field theory with an initial cluster threshold of $Z > 2.3$ and a Family Wise Error corrected threshold of $p < 0.05$ (Worsley et al., 1992). Peristimulus plots used functionally defined ROIs and were calculated by fitting a FIR model using fslroi 2.0 (Poldrack, 2007) and averaging within and then across participants.

Behavioral Analyses

All behavioral statistics were computed using the R statistical package (R Development Core Team, 2008). For regressions that included repeated observations, we used the lme4 mixed effects GLM package (Bates et al., 2008). Participants were treated as a random effect with varying intercepts and slopes. We report the regression coefficients (b), standard errors (SE), t values, and p values. Because there is no generally agreed upon method for calculating p values in mixed models, we used two separate methods. First, we calculated the degrees of freedom by subtracting the number of fixed effects from the total number of observations (Kliegl et al., 2007). Second, we generated confidence intervals from the posterior distribution of the parameter estimates using Markov Chain Monte Carlo methods (Baayen et al., 2008). These methods produced identical results. For robust regressions, we used the rlm function from the MASS package using MM estimation (Venables and Ripley, 2002).

Guilt Sensitivity Estimation

Our linear model of guilt aversion (Equation 1) makes sharp predictions about the amount of money that participants should return (see Figure S1 for a simulation). Our model allows for the guilt sensitivity parameter (θ_{12}) to vary for every Investor/Trustee interaction. There are two possible maxima of the utility function depending on θ_{12}. If participants are completely guilt averse ($\theta_{12} > 1$) then the model predicts they should always match their second-order belief. If they are completely guilt in-averse ($\theta_{12} < 1$) then they should always keep all of the money. Because all participants demonstrated some degree of guilt sensitivity, meaning that no subject always kept all of the money, all participants were classified as guilt averse and thus we observed no variability in Θ_{12}.

Counterfactual Guilt

To confirm that participants were actually motivated by anticipated guilt, we elicited their counterfactual guilt for each trial following the scanning session. After displaying a recap of each trial, we asked participants how much guilt they would have felt had they returned a different amount of money. This amount was randomly selected from all choices below and one choice above the amount they actually returned (choices increased or decreased in 10% increments). The deviation from the participant's actual choice was used to predict the amount of guilt that participants reported that would have felt had they returned that amount using a mixed effects regression. Thus, each participant's best linear unbiased predictions (BLUPs) (Pinheiro and Bates, 2000) represent their sensitivity to guilt. Larger slopes indicate that participants reported they would have felt more guilt had they returned less money, revealing a higher degree of guilt sensitivity, while smaller slopes reveal a low degree of guilt sensitivity with participants, indicating little change in the amount of guilt they would have experienced had they returned less money. The regression can be seen in Figure 2C along with each participant's BLUP.

Analysis 1, Main Contrast

To identify regions of the brain that are associated with anticipated guilt as predicted by our model, we examined trials during the return phase in which participants matched expectations by returning the amount of money that they believed their partner expected (n = 207), as compared to trials in which they returned less than they believed their partner expected (n = 183). This allowed us to identify neural systems associated with guilt aversion and also to see systems involved in maximizing financial payoffs. For this analysis, we excluded trials by modeling them as covariates of no interest where (1) the partner sent $0, and thus there was no decision for the participant to make (n = 33), (2) the participant returned more than their second order belief (n = 66), and (3) the participants either did not indicate their belief or the amount they wanted to return (n = 20). This model thus included the following 30 regressors:

(1) Face phase
(2) Prediction phase
(3) Investment phase
(4) Belief elicitation phase
(5) Decision phase when participants matched their partner's expectations (n = 207)
(6) Decision phase when participants returned 10% less than their partners' expectations (n = 99)
(7) Decision phase when participants returned 20% less than their partners' expectations (n = 46)
(8) Decision phase when participants returned 30%+ less than their partners' expectations (n = 38)
(9) Decision phase when participants returned more than their expectations (n = 66)
(10) Summary phase
(11) Handed-down-belief phase
(12) Missed trials
(13–24) Temporal derivatives of regressors 1–12
(25–30) Estimated head movement parameters (6)

We compared trials in which the participant matched their expectations to trials in which they returned less than their expectations (+0.99 −0.33 −0.33 −0.33 for regressors 5–8). The results of this analysis can be seen in Figure 4 and Table S2.

Analysis 2, Parametric Contrast

An additional question of interest is whether the activations found above change parametrically as a function of deviation from matching expectations. To address this, we tested a parametric contrast in which we compared trials in which participants matched expectations to a linear deviation in 10% increments Winsorized at 30%. Responses greater than or equal to 30% were grouped together, as these were relatively rare and this procedure ensured that the number of cases were balanced across regressors. This contrast specifically compared matching expectations to returning 10% less, 20% less, and 30+% less (+6 −1 −2 −3 for regressors 5–8) using the model from Analysis 1 .

Analysis 3, Counterfactual Guilt Correlations

To address the hypothesis that regions associated with guilt aversion should become more active as a function of guilt sensitivity, we extracted the average third-level parameter estimates from each of the regions of interest and examined their relationship with our measure of counterfactual guilt. We extracted the average values in the clusters located in the right and left DLPFC, insula, SMA, MOFC, and DMPFC by restricting to voxels that were located both in these clusters and in the respective anatomical masks taken from the Harvard-Oxford probabilistic atlas. Because of the small size of the nucleus accumbens, all voxels located in a bilateral anatomical mask were used regardless of statistical significance. We used the individual slopes (BLUPs) from the random effects component of the counterfactual guilt analysis as our metric of guilt sensitivity. Due to the noise of the two metrics (average beta values from a third-level imaging analysis and individual BLUPs from a mixed effects analysis) and non-Gaussian distribution, we used robust regression to estimate the effects using MM estimation (Venables and Ripley, 2002).

SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures, four tables, and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.neuron.2011.02.056.

ACKNOWLEDGMENTS

We thank Matt Kleinman for his help in collecting the data and Drs. Anouk Scheres, James Rilling, and Lynn Nadel for their helpful comments. We would like to acknowledge funding from the National Institute of Aging (R21AG030768) to A.G.S., the National Institute of Mental Health (R03MH077058) to A.G.S. and (F31MH085465) to L.J.C., and the National Science Foundation to M.D.

Accepted: February 25, 2011
Published: May 11, 2011

REFERENCES

Amodio, D.M., and Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. Nat. Rev. Neurosci. *7*, 268–277.

Andreoni, J. (1990). Impre altruism and donations to public goods: A theory of warm-glow giving. Econ. J. *100*, 464–477.

Baayen, R.H., Davidson, D.J., and Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. *59*, 340–412.

Bates, D., Maechler, M., and Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-39. http://CRAN.R-project.org/package=lme4.

Battigalli, P., and Dufwenberg, M. (2007). Guilt in games. Am. Econ. Rev. *97*, 170–176.

Battigalli, P., and Dufwenberg, M. (2009). Dynamic psychological games. J. Econ. Theory *144*, 1–35.

Baumeister, R.F., Stillwell, A.M., and Heatherton, T.F. (1994). Guilt: An interpersonal approach. Psychol. Bull. *115*, 243–267.

Baumgartner, T., Fischbacher, U., Feierabend, A., Lutz, K., and Fehr, E. (2009). The neural circuitry of a broken promise. Neuron *64*, 756–770.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. Games Econ. Behav. *10*, 122–142.

Berns, G.S., Capra, C.M., Moore, S., and Noussair, C. (2010). Neural mechanisms of the influence of popularity on adolescent ratings of music. Neuroimage *49*, 2687–2696.

Bolton, G.E., and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. Am. Econ. Rev. *90*, 166–193.

Calder, A.J., Keane, J., Manes, F., Antoun, N., and Young, A.W. (2000). Impaired recognition and experience of disgust following brain injury. Nat. Neurosci. *3*, 1077–1078.

Camerer, C.F. (2003). Behavioral Game Theory (New York: Russell Sage Foundation).

Charness, G., and Dufwenberg, M. (2006). Promises and partnership. Econometrica *74*, 1579–1601.

Coricelli, G., Critchley, H.D., Joffily, M., O'Doherty, J.P., Sirigu, A., and Dolan, R.J. (2005). Regret and its avoidance: A neuroimaging study of choice behavior. Nat. Neurosci. *8*, 1255–1262.

Craig, A.D. (2009). How do you feel—now? The anterior insula and human awareness. Nat. Rev. Neurosci. *10*, 59–70.

Critchley, H.D., Mathias, C.J., and Dolan, R.J. (2001). Neural activity in the human brain relating to uncertainty and arousal during anticipation. Neuron *29*, 537–545.

Damasio, A.R. (1994). Descartes' Error: Emotion, Reason, and the Human Brain (New York: Penguin Putnam).

Damasio, A.R., Grabowski, T.J., Bechara, A., Damasio, H., Ponto, L.L., Parvizi, J., and Hichwa, R.D. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. Nat. Neurosci. *3*, 1049–1056.

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

Dosenbach, N.U., Visscher, K.M., Palmer, E.D., Miezin, F.M., Wenger, K.K., Kang, H.C., Burgund, E.D., Grimes, A.L., Schlaggar, B.L., and Petersen, S.E. (2006). A core system for the implementation of task sets. Neuron *50*, 799–812.

Downar, J., Crawley, A.P., Mikulis, D.J., and Davis, K.D. (2000). A multimodal cortical network for the detection of changes in the sensory environment. Nat. Neurosci. *3*, 277–283.

Dreher, J.C., and Tremblay, L.K., eds. (2009). Handbook of Reward and Decision-Making (Burlington, MA: Academic Press).

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

Dufwenberg, M., and Kirchsteiger, G. (2004). A theory of sequential reciprocity. Games Econ. Behav. *47*, 268–298.

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

Ellingsen, T., Johannesson, M., Tjotta, S., and Torsvik, G. (2010). Testing guilt aversion. Games Econ. Behav. *68*, 95–107.

Falk, A., and Fischbacher, U. (2006). A theory of reciprocity. Games Econ. Behav. *54*, 293–315.

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

Fehr, E., and Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. Q. J. Econ. *114*, 817–868.

Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). Psychological games and sequential rationality. Games Econ. Behav. *1*, 60–79.

Grinband, J., Wager, T.D., Lindquist, M., Ferrera, V.P., and Hirsch, J. (2008). Detection of time-varying signals in event-related fMRI designs. Neuroimage *43*, 509–520.

Haidt, J. (2003). The moral emotions. In Handbook of Affective Sciences, R.J. Davidson, K.R. Scherer, and H.H. Goldsmith, eds. (Oxford: Oxford University Press), pp. 852–870.

Hampton, A.N., and O'Doherty, J.P. (2007). Decoding the neural substrates of reward-related decision making with functional MRI. Proc. Natl. Acad. Sci. USA *104*, 1377–1382.

Ketelaar, T., and Au, W.T. (2003). The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. Cogn. Emotion *17*, 429–453.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R., and Montague, P.R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. Science *308*, 78–83.

Kliegl, R., Risse, S., and Laubrock, J. (2007). Preview benefit and parafoveal-on-foveal effects from word n + 2. J. Exp. Psychol. Hum. Percept. Perform. *33*, 1250–1255.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., and Fernández, G. (2009). Reinforcement learning signal predicts social conformity. Neuron *61*, 140–151.

Knoch, D., Schneider, F., Schunk, D., Hohmann, M., and Fehr, E. (2009). Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. Proc. Natl. Acad. Sci. USA *106*, 20895–20899.

Krajbich, I., Adolphs, R., Tranel, D., Denburg, N.L., and Camerer, C.F. (2009). Economic games quantify diminished sense of guilt in patients with damage to the prefrontal cortex. J. Neurosci. *29*, 2188–2192.

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., Heinecke, A., and Grafman, J. (2007). Neural correlates of trust. Proc. Natl. Acad. Sci. USA *104*, 20084–20089.

Loewenstein, G.F., Weber, E.U., Hsee, C.K., and Welch, N. (2001). Risk as feelings. Psychol. Bull. *127*, 267–286.

McCabe, K.A., Smith, V.L., and LePore, M. (2000). Intentionality detection and "mindreading": Why does game form matter? Proc. Natl. Acad. Sci. USA *97*, 4404–4409.

McCabe, K., Rigdon, M.L., and Smith, V.L. (2003). Positive reciprocity and intentions in trust games. J. Econ. Behav. Organ. *52*, 267–275.

Mellers, B., Schwartz, A., and Ritov, I. (1997). Elation and disappointment: Emotional responses to risky options. Psychol. Sci. *8*, 423–429.

Miller, E.K., and Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. *24*, 167–202.

Montague, P.R., and Lohrenz, T. (2007). To detect and correct: Norm violations and their enforcement. Neuron *56*, 14–18.

Newbould, R.D., Skare, S.T., Jochimsen, T.H., Alley, M.T., Moseley, M.E., Albers, G.W., and Bammer, R. (2007). Perfusion mapping with multiecho multi-shot parallel imaging EPI. Magn. Reson. Med. *58*, 70–81.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R.J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science *304*, 452–454.

Pinheiro, J., and Bates, D. (2000). Mixed-Effects Models in S and S-Plus (New York, NY: Springer-Verlag).

Poldrack, R.A. (2007). Region of interest analysis for fMRI. Soc. Cogn. Affect. Neurosci. *2*, 67–70.

Preuschoff, K., Bossaerts, P., and Quartz, S.R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. Neuron *51*, 381–390.

Rabin, M. (1993). Incorporating fairness into game theory and economics. Am. Econ. Rev. *83*, 1281–1302.

Rangel, A., Camerer, C., and Montague, P.R. (2008). A framework for studying the neurobiology of value-based decision making. Nat. Rev. Neurosci. *9*, 545–556.

R Development Core Team (2008). R: A language and environment for statistical computing. (Vienna, Austria).

Reuben, E., Sapienza, P., and Zingales, L. (2009). Is mistrust self-fulfilling? Econ. Lett. *104*, 89–91.

Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., and Kilts, C. (2002). A neural basis for social cooperation. Neuron *35*, 395–405.

Sampson, R.J., Raudenbush, S.W., and Earls, F. (1997). Neighborhoods and violent crime: a multilevel study of collective efficacy. Science *277*, 918–924.

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

Shin, L.M., Dougherty, D.D., Orr, S.P., Pitman, R.K., Lasko, M., Macklin, M.L., Alpert, N.M., Fischman, A.J., and Rauch, S.L. (2000). Activation of anterior paralimbic structures during guilt-related script-driven imagery. Biol. Psychiatry *48*, 43–50.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R.J., and Frith, C.D. (2004). Empathy for pain involves the affective but not sensory components of pain. Science *303*, 1157–1162.

Slovic, P., Finucane, M.L., Peters, E., and MacGregor, D.G. (2002). The affect heuristic. In Heuristics and Biases, T. Gilovich, D. Griffin, and D. Kahneman, eds. (New York: Cambridge University Press), pp. 397–420.

Smith, A. (1984). The Theory of Moral Sentiments (Indianapolis: Liberty Fund). Originally published 1759.

Sridharan, D., Levitin, D.J., and Menon, V. (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. Proc. Natl. Acad. Sci. USA *105*, 12569–12574.

Stöcker, T., Kellermann, T., Schneider, F., Habel, U., Amunts, K., Pieperhoff, P., Zilles, K., and Shah, N.J. (2006). Dependence of amygdala activation on echo time: Results from olfactory fMRI experiments. Neuroimage *30*, 151–159.

Takahashi, H., Yahata, N., Koeda, M., Matsuda, T., Asai, K., and Okubo, Y. (2004). Brain activation associated with evaluative processes of guilt and embarrassment: An fMRI study. Neuroimage *23*, 967–974.

Tricomi, E., Rangel, A., Camerer, C.F., and O'Doherty, J.P. (2010). Neural evidence for inequality-averse social preferences. Nature *463*, 1089–1091.

van den Bos, W., van Dijk, E., Westenberg, M., Rombouts, S.A.R.B., and Crone, E.A. (2009). What motivates repayment? Neural correlates of reciprocity in the Trust Game. Soc. Cogn. Affect. Neurosci. *4*, 294–304.

Venables, W.N., and Ripley, B.D. (2002). Modern Applied Statistics with S, Fourth Edition (New York: Springer).

Wager, T.D., Rilling, J.K., Smith, E.E., Sokolik, A., Casey, K.L., Davidson, R.J., Kosslyn, S.M., Rose, R.M., and Cohen, J.D. (2004). Placebo-induced changes in FMRI in the anticipation and experience of pain. Science *303*, 1162–1167.

Weiskopf, N., Hutton, C., Josephs, O., and Deichmann, R. (2006). Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: A whole-brain analysis at 3 T and 1.5 T. Neuroimage *33*, 493–504.

Woolrich, M.W., Behrens, T.E., Beckmann, C.F., Jenkinson, M., and Smith, S.M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. Neuroimage *21*, 1732–1747.

Worsley, K.J., Evans, A.C., Marrett, S., and Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab. *12*, 900–918.

Zak, P.J., and Knack, S. (2001). Trust and Growth. Econ. J. *111*, 295–321.

Neuron, Volume *70*

Supplemental Information

Triangulating the Neural, Psychological, and Economic Bases of Guilt Aversion

Luke J. Chang, Alec Smith, Martin Dufwenberg, and Alan G. Sanfey

Supplemental Information Inventory

Figures

Figure S1 related to Equation 1. Simulation of Guilt-Aversion Model
The guilt-aversion model makes two behavioral predictions depending on Θ_{12}. If $\Theta_{12} < 1$, then the optimal choice (S_2) that maximizes the utility function (U_2) is to keep all of the money. If $\Theta_{12} > 1$, then the optimal choice which maximizes U_2 is to match expectations and return the amount that player 2 believes that player 1 expects them to return. Here we plot the behavioral predictions for Player 2's choice if Player 1 invests \$10 (which become \$40) and Player 2 believes that Player 1 expects them to return \$20 for varying Θ_{12}s (e.g., 0.5 or 1.5).

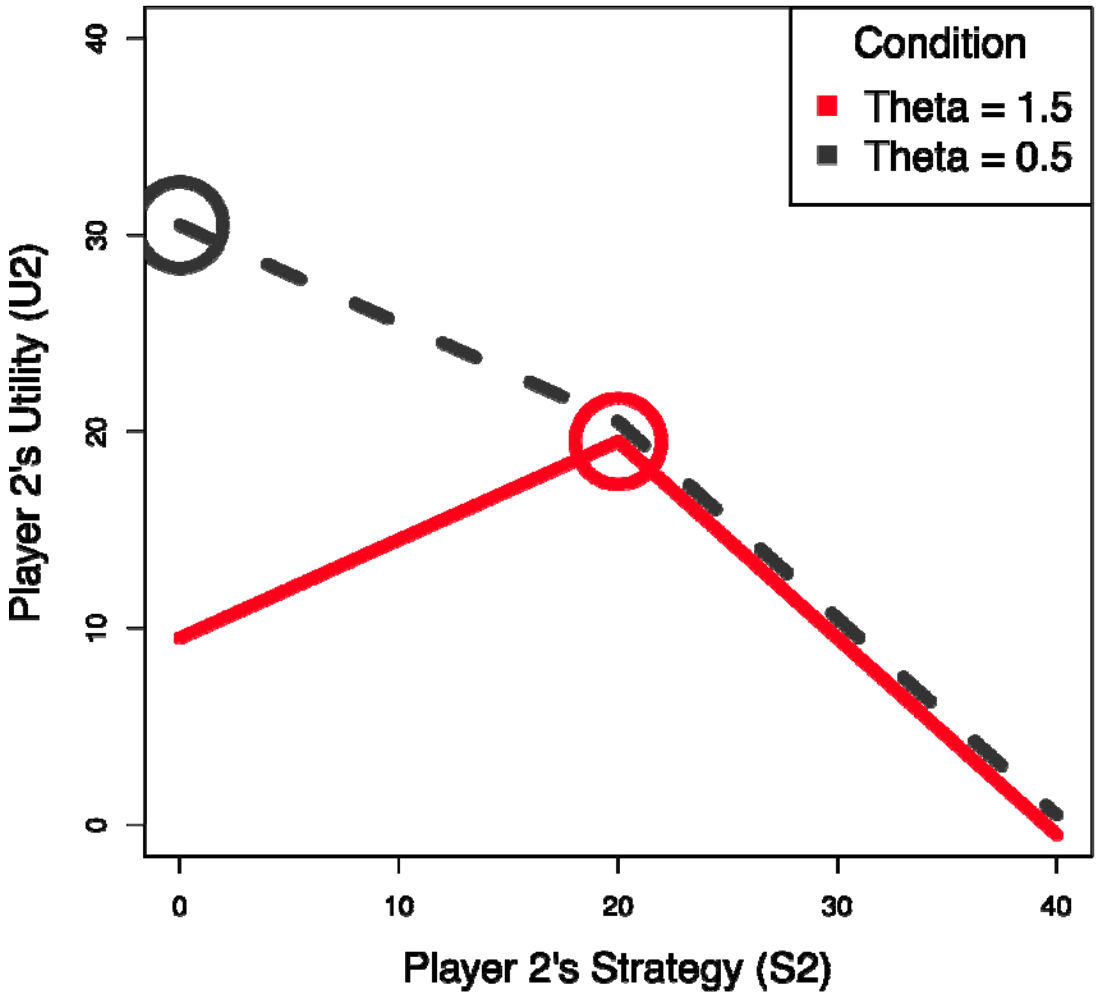


Figure S2 related to Figure 4. Relationship between SMA and Guilt Sensitivity. Participant's best linear unbiased predictors (BLUPs) from the counterfactual guilt analysis predict the average parameter estimate of voxels in the SMA ROI using robust regression.

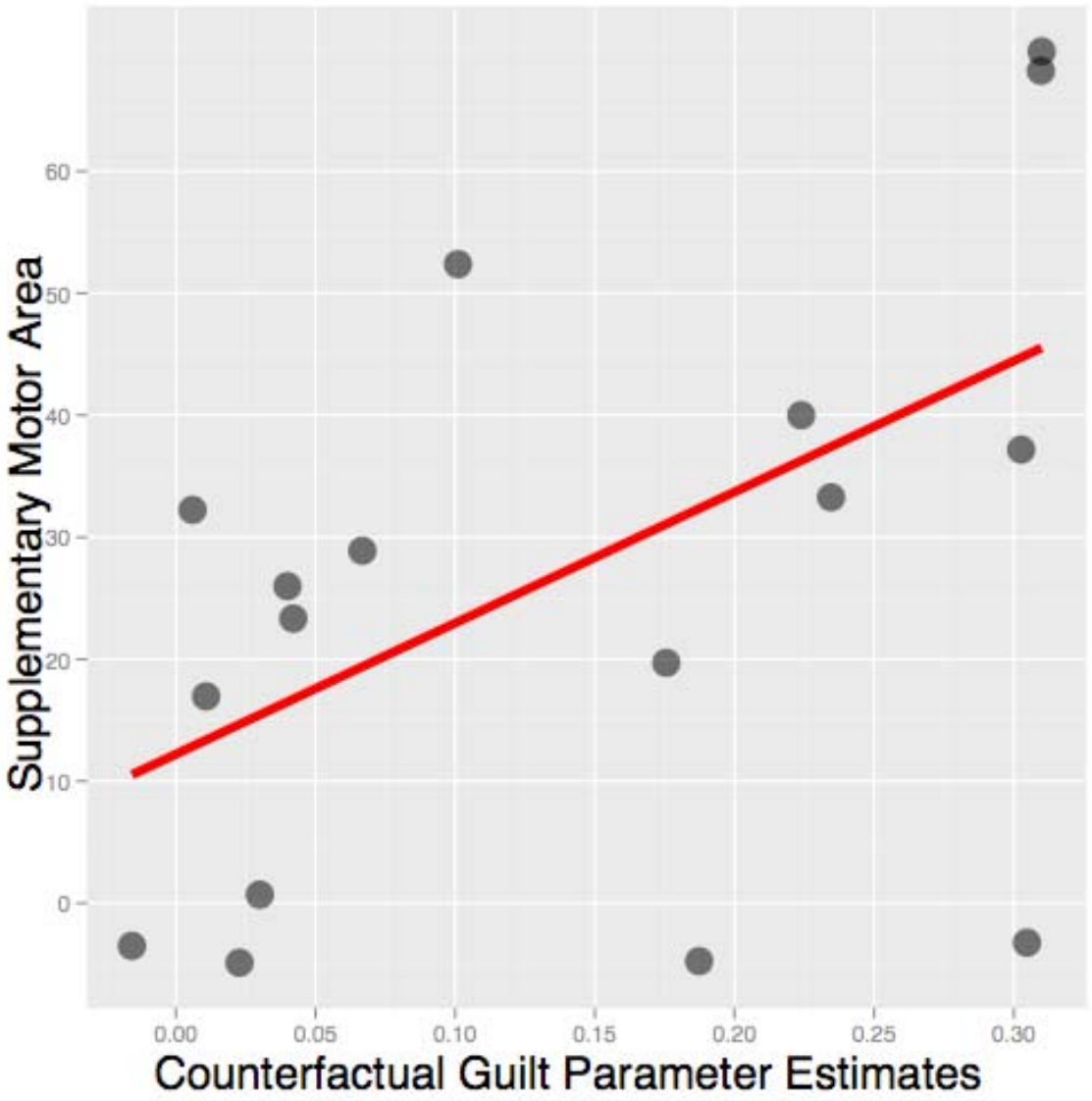
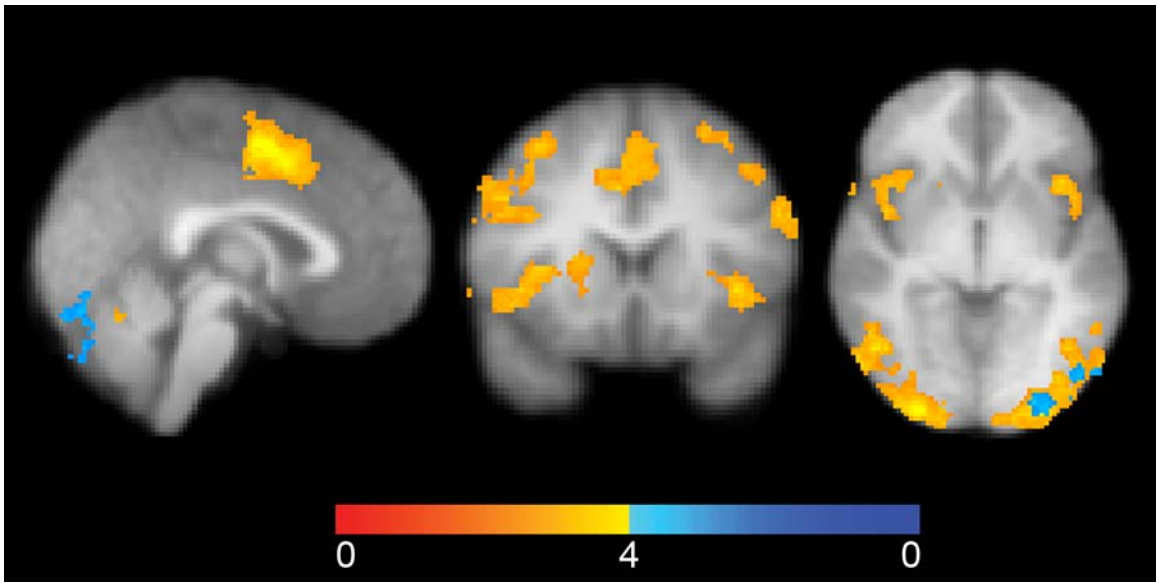


Figure S3 related to Figure 5. Guilt-Aversion Controlling for Player 2's Choice. This figure depicts activity associated with matching expectations (orange) and returning less than participants believed their partner expected them to return (blue). Images are displayed in radiological orientation (left=right) and are thresholded using whole brain cluster correction, $Z > 2.3$, $p < 0.05$. Color maps reflect Z values between 0 and 4.



Tables

Table S1 (related to figure 3). The Effect of Beliefs on Trustee Behavior. This table illustrates the results of a mixed effects regression analysis, in which Trustee's 2nd order beliefs (E2E1S2) significantly predict the amount of money that they chose to return to the Investor (S2) controlling for the size of the initial investment (Offer Amount). Participants were treated as a random effect with varying intercepts (s^2=0.004, SD=0.06). All variables were normalized to [0,1]. There was a correlation of 0.34 between the two fixed effects predictors. These results indicate that there is a significant effect of expectations on behavior controlling for the subgame.

Predictor	Parameter Estimate	SE	t - Value	p - Value
Intercept	0.09	0.03	3.54	< .001
Offer Amount	0.17	0.03	6.82	< .001
E2E1S2	0.46	0.04	10.83	< .001

Table S2 (related to Figure 4). Brain activations for Matching Compared to Returning Less then Expectations Contrast. This table reflects the contrast matching expectations (i.e. 2nd order beliefs) compared to returning less than expectations and shows the local maxima of clusters surviving cluster correction $Z > 2.3$, $p < 0.05$ in MNI space. Cortical and subcortical regions were identified using the Harvard-Oxford Probabilistic Anatomical Atlas, while the cerebellar regions were identified using a probabilistic cerebellar atlas (Diedrichsen et al., 2009). Abbreviations: DMPFC=dorsomedial prefrontal cortex, DLPFC=dorsolateral prefrontal cortex, TPJ = temporal-parietal junction, SMA = supplementary motor area, OFC = orbitofrontal cortex, ACC = anterior cingulate cortex.

	Hemisphere	Region	BA	Z Value	X	Y	Z
Less > Equal							
	L	dmPFC (Frontal Pole)	10	3.61	-12	64	24
	L	Nucleus Accumbens	25	2.99	-8	8	-12
	L	Paracingulate	10	3.41	-6	54	6
	L	Superior Frontal Gyrus (DMPFC)	9	3.76	-12	46	50
	L	Medial OFC	11	3.16	-8	26	-16
	R	Caudate	25	3.11	14	20	4
	R	Medial OFC	11	3.1	4	28	-16
	R	Nucleus Accumbens	25	2.92	6	14	-4
	R	Rostral ACC	10	3.39	14	46	0
	R	Sub Genual ACC	25	3.28	0	14	-14
	R	Superior Frontal Gyrus (DMPFC)	10	3.51	4	56	26
Equal > Less							
	L	ACC	24	3.51	-4	18	34
	L	Cerebellum (Left Crus I)	19	3.65	-42	-74	-26
	L	Middle Frontal Gyrus (DLFPC)	45	3.49	-48	30	26
	L	Fusiform	19	3.53	-44	-70	-20
	L	Lateral Occipital Cortex	37	3.49	-56	-64	6
	L	Posterior Cingulate Cortex	NA	3.4	-12	-28	42
	L	Postcentral gyrus	3	4.27	-34	-30	58
	L	Precentral Gyrus	6	4.17	-34	-4	48
	L	Precentral Gyrus	43	3.96	-58	2	28
	L	SMA	NA	3.75	-4	-2	48
	L	Superior Parietal Lobule	40	4.22	-34	-40	50
	L	Supramarginal Gyrus (TPJ)	2	3.6	-54	-36	36
	R	ACC	24	3.32	6	16	36
	R	Cerebellum (Right Crus I)	37	3.9	36	-58	-30
	R	Cerebellum (Right VI)	NA	3.92	30	-46	-38
	R	Cerebellum (Vermis VI)	NA	3.75	6	-66	-22
	R	Middle Frontal Gyrus (DLPFC)	46	3.5	36	42	28
	R	Inferior Temporal gyrus	37	4.12	52	-50	-22
	R	Insula	48	3.42	42	8	0
	R	Lateral Occiptal Cortex, Inferior division	19	4.18	50	-72	-8
	R	Lateral Occiptal Cortex, Superior division	7	3.82	26	-62	42
	R	Occipital Pole	18	3.63	30	-94	-10
	R	SMA	NA	3.64	2	2	50
	R	Superior Parietal Lobule	40	3.88	36	-46	46
	R	Supramarginal Gyrus (TPJ)	2	3.48	50	-28	36

Table S3 (related to Figure 5). Brain activations for Parametric Contrast of Matching Compared to Returning Less then Expectations. This table reflects the parametric contrast matching expectations (i.e. 2nd order beliefs) compared to returning less than expectations (i.e. 10%, 20%, +30%) and shows the local maxima of clusters surviving cluster correction $Z > 2.3$, $p < 0.05$ in MNI space. Cortical and subcortical regions were identified using the Harvard-Oxford Probabilistic Anatomical Atlas, while the cerebellar regions were identified using a probabilistic cerebellar atlas (Diedrichsen et al., 2009). Abbreviations: DMPFC=dorsomedial prefrontal cortex, DLPFC=dorsolateral prefrontal cortex, TPJ = temporal-parietal junction, SMA = supplementary motor area, OFC = orbitofrontal cortex, ACC = anterior cingulate cortex.

	Hemisphere	Region	BA	Z Value	X	Y	Z
Less > Equal							
	L	Caudate	25	3.08	-10	18	4
	L	Lateral OFC Cortex	47	4.06	-36	32	-18
	L	Middle Frontal Gyrus	9	3.52	-28	24	42
	L	Nucleus Accumbens	25	3.3	-6	10	-8
	L	Paracingulate Gyrus	9	3.7	-6	54	6
	L	Rostral ACC	32	3.18	-10	40	20
	L	Sub Genual ACC	25	3.54	-4	14	-16
	L	Superior Frontal Gyrus (DMPFC)	9	4.11	-12	46	50
	L	Superior Frontal Gyrus (DMPFC)	8	3.69	-20	26	58
	L	Superior Frontal Gyrus (DMPFC)	10	3.92	-12	64	24
	L	Temporal Pole	38	3.51	-42	20	-30
	R	ACC	11	3.38	0	30	-6
	R	Caudate	25	3.19	12	20	4
	R	Lateral OFC Cortex	11	3.25	22	36	-16
	R	Medial OFC Cortex	11	3.51	4	44	-20
	R	Nucleus Accumbens	25	3.08	8	16	-4
	R	Paracingulate Gyrus	10	4.05	14	46	0
	R	Posterior Insula	48	3.06	40	0	-14
	R	Rostral ACC	11	3.21	8	36	2
	R	Superior Frontal Gyrus (DMPFC)	10	3.94	2	56	32
	R	Temporal Pole	28	3.35	26	10	-26
	R	Temporal Pole	38	3.24	42	20	-20
Equal > Less							
	L	Angular Gyrus	19	3.45	-44	-56	44
	L	Cerebellum (Crus I)	NA	3.61	-40	-74	-26
	L	Dorsal ACC	32	4.02	-8	14	38
	L	Fusiform	19	3.57	-44	-70	-20
	L	Lateral Occipital Cortex, Inferior Division	37	3.91	-58	-64	8
	L	Lateral Occiptal Cortex, Inferior Division	19	3.81	-52	-74	-14
	L	Middle Frontal Gyrus (DLPFC)	45	3.88	-44	28	24
	L	Middle Frontal Gyrus (DLPFC)	46	3.08	-32	40	24
	L	Postcentral Gyrus	40	4.74	-38	-34	40
	L	Precentral Gyrus	6	4.48	-34	-6	48
	L	Precentral Gyrus	4	4.43	-58	0	30
	L	SMA	NA	3.85	-4	-2	48
	L	Superior Parietal Lobule	40	4.05	-34	-52	58
	L	Supramarginal Gyrus	2	3.88	-58	-28	44
	L	Supramarginal Gyrus (TPJ)	40	4.01	-46	-34	38
	R	Central Opercular Cortex	48	3.58	48	-2	10

R	Cerebellum (I-IV)	NA	3.71	4	-52	-16
R	Cerebellum (Right Crus I)	NA	4.1	36	-58	-30
R	Cerebellum (Vermis VI)	NA	4.22	6	-64	-22
R	Cerebellum (VI)	NA	4.17	30	-46	-38
R	Cerebellum (X)	NA	3.62	30	-36	-44
R	DLPFC	46	3.64	36	42	28
R	Dorsal ACC	24	3.53	6	16	36
R	Fusiform	37	3.68	42	-48	-20
R	Inferior Temporal Gyrus	37	4.09	48	-60	-12
R	Insula	48	3.73	46	14	-2
R	Lateral Occipital Cortex, Inferior Division	19	4.21	50	-72	-8
R	Lateral Occipital Cortex, Superior Division	7	4.12	26	-62	42
R	Occipital Pole	18	3.97	30	-92	-8
R	Precentral Gyrus	6	4.07	46	-4	56
R	Precuneous	7	3.81	6	-70	48
R	SMA	6	3.94	6	-6	64
R	Supramarginal gyrus	40	4.19	44	-40	48

Table S4 (related to figure S3). Brain Activations for Guilt-Aversion Controlling for Player 2's Choice. This table reflects the activity associated with matching expectations (i.e., (E2E1S2-S2)=0) and returning less than participants believed their partner expected them to return (i.e., $(E2E1S2-S2)^{+}$) controlling for Player 2's choice (i.e., S2) and shows the local maxima of clusters surviving cluster correction $Z > 2.3$, $p < 0.05$ in MNI space. Cortical and subcortical regions were identified using the Harvard-Oxford Probabilistic Anatomical Atlas, while the cerebellar regions were identified using a probabilistic cerebellar atlas (Diedrichsen et al., 2009). Abbreviations: DLPFC=dorsolateral prefrontal cortex, TPJ = temporal-parietal junction, SMA = supplementary motor area, OFC = orbitofrontal cortex, ACC = anterior cingulate cortex, STS = superior temporal sulcus.

	Hemisphere	Region	BA	Z Value	X	Y	Z
Return Less than Expectations							
	L	Cerebellum (Crus I)	NA	3.47	-38	-76	-28
	L	Cerebellum (Crus II)	NA	3.32	-6	-80	-30
	L	Lateral Occipital Cortex, Inferior Division	18	3.53	-32	-88	-16
	L	Lateral Occipital Cortex, Superior Division	7	3.32	-22	-66	40
	L	Middle Temporal Gyrus	20	3.23	-60	-32	-18
	R	Superior Parietal Lobule	7	3.38	32	-54	54
Match Expectations							
	L	DLPFC (Middle Frontal Gyrus)	46	3.42	-38	36	28
	L	Insula (Anterior)	48	3.69	-38	14	-2
	L	Insula (Posterior)	48	3.9	-42	-2	6
	L	Lateral Occipital Cortex, Inferior Division	19	4.35	-44	-76	-16
	L	Lateral Occipital Cortex, Superior Division	19	4.41	-26	-70	26
	L	Postcentral Gyrus	4	4.85	-48	-18	52
	L	Precentral Gyrus	6	4.02	-56	6	32
	L	SMA	NA	4.32	-2	-2	54
	L	Supramarginal Gyurs (TPJ)	2	4.11	-48	-30	36
	R	ACC	24	3.39	6	16	36
	R	Caudate	NA	3	18	14	8
	R	Cerebellum (V)	37	3.53	14	-52	-22
	R	Cerebellum (Crus I)	19	3.51	38	-76	-24
	R	Cerebellum (VI)	NA	3.79	16	-58	-28
	R	DLPFC (middle frontal gyrus)	46	3.88	34	36	28
	R	Insula (Anterior)	47	3.41	42	18	-6
	R	Insula (Middle)	48	3.27	44	4	-2
	R	Lateral Occipital Cortex, Inferior Division	19	4.45	34	-86	-8
	R	Lateral Occipital Cortex, Inferior Division	37	4.32	52	-64	-8
	R	Lateral Occipital Cortex, Superior Division	7	4.12	34	-60	44
	R	Parietal Operculum Cortex (TPJ)	48	3.23	56	-22	16
	R	Precentral Gyrus	6	4.27	60	10	32
	R	Precentral Gyrus	6	3.92	28	-8	48
	R	Supramarginal Gyrus, Posterior Division	40	4.09	42	-44	46

R	Supramarginal Gyrus, Posterior Division (TPJ)	22	3.93	66	-44	18
R	Temporal Occipital Fusiform Cortex	37	3.86	38	-48	-24

Supplemental Experimental Procedures

Guilt-Aversion Controlling for Player 2's Behavior: As a consequence of our design, participants make systematically less money in trials in which they match expectations compared to trials in which they return less than they believe the other player expected them to return. Presumably, participants choose to return more money to Player 1 because they are more motivated by minimizing guilt aversion then maximizing financial payoff. However, to rule out the possibility that the insula is simply tracking forgone financial payoffs rather than guilt-aversion, we ran an analysis which allowed us to examine the effect of matching expectations (i.e., guilt-aversion=0 in eq(1)), while controlling for the amount of money that they choose to return (i.e., their forgone financial payoff or S2). This model included the following regressors:

1) Return phase
2) Guilt (i.e., E2E1S2-S2)+
3) Match Trials (i.e., Guilt = 0)
4) Player 2's Choice (i.e., S2)
5) Over Match Trials (i.e., E2E1S2-S2)-
6) Face phase
7) Prediction phase
8) Investment phase
9) Belief elicitation phase
10) Summary phase
11) Handed-down-belief phase
12) Missed trials
13-24) Temporal derivatives of regressors 1 – 12
25-31) Estimated head movement parameters (n=6)

We report the results for the independent variance associated with matching expectations (regressor 3) and linear deviations of returning less money than participants believed Player 1 expected (regressor 2), while controlling for all of the other regressors in Figure S3 and Table S4. We were forced to exclude 8/66 runs due to a lack of variability in either regressor 2 or 3.

Supplemental References

Diedrichsen, J., Balsters, J.H., Flavell, J., Cussans, E., and Ramnani, N. (2009). A probabilistic MR atlas of the human cerebellum. Neuroimage *46*, 39-46.