Triangulating the neural, psychological & economic bases of guilt aversion

(May 12, 2011: publication day, *Neuron*)

Luke J. Chang,[1] Alec Smith,[2,5] Martin Dufwenberg,[2,6] Alan G. Sanfey[1,3,4]*

1 Department of Psychology, University of Arizona; 2 Department of Economics, University of Arizona; 3 Donders Institute for Brain, Mind & Behavior, Radboud University Nijmegen; 4 Behavioral Science Institute, Radboud University Nijmegen; 5 Humanities & Social Sciences, California Institute of Technology; 6 Department of Economics, University of Gothenburg

Email: ljchang@email.arizona.edu; acs@hss.caltech.edu; martind@eller.arizona.edu; asanfey@u.arizona.edu

Highlights

1) Guilt can be formally operationalized as failing to live up to another's expectations.
2) Guilt-aversion motivates cooperative behavior
3) Decisions which minimize future guilt are associated with insula, SMA, DLPFC, TPJ
4) Decisions which maximize financial reward are associated with vmPFC, NAcc, DMPFC

Abstract

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 (1759); 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, pro-social 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; but see also Ellingsen et al., 2010; Reuben et al., 2009) 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 J, 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 1 Panel A). 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 non-negative 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.

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

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 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., 2003; McCabe et al., 2000). 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 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).

Insert Figure 1 about here

Results

Behavioral Results

Insert Figure 2 about here

Our behavioral results demonstrated that participants behaved in a similar fashion to previous TG experiments (Camerer, 2003) (See 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 $b=0.85$, $se=0.06$, $t=15.20$, $p<0.001$ (see Figure 3, Panel A). 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, $b=0.90$, $se=0.04$, $t=21.32$, $p<0.001$ (see Figure 3, Panel B). Finally, participants reported that they would have felt more counterfactual guilt had they chosen to return less money then they actually did, $b=0.14$, $se=0.03$, $t=4.14$, $p<0.001$ (see Figure 3, Panel C). Taken together, these results suggest that participants behaved in a manner consistent with our model of guilt aversion.

Insert Figure 3 about here

Neuroimaging Results

We conducted several different analyses to examine the neural mechanisms underlying guilt aversion. Firstly, 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). Secondly, we explored processes that tracked parametrically with the predictions of the model. Thirdly, 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 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=3412.29ms, sd=1310.65) as compared to trials in which they returned less than their expectation (mean=3666.87ms, 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 methods). 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).

Insert Figure 4 about here

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 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, & 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 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 SI) 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).

Insert Figure 5 about here

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 Figure 4, Panel C, Figure S2 and

methods). 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.

Inter-regional 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 post-experimental 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 (a) motivated by a "warm glow" feeling and find cooperation inherently rewarding (Andreoni, 1990; Fehr and Camerer, 2007), (b) motivated to minimize the discrepancy between self and others' payoffs (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999), or (c) 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., 2009; 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 towards 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 & 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 then 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.

Methods

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 minutes including an extra 14 second fixation cross display at the beginning of the run to allow for T1 equilibrium, and another 21 second 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 1, Panel B. 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=11ms, TE=4 ms, matrix=256X256, slice thickness=1mm, gap=0mm) 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) (Stocker et al., 2006; Weiskopf et al., 2006) (TR=2000ms, TE=256ms, matrix=96X96, FOV=192mm, slice thickness=3.0mm, 42 axial slices).

Data Pre-Processing: Functional imaging data were preprocessed and analyzed using the FSL Software package 4.1.4 (FMRIB, Oxford, UK). The first 3 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 6-parameter rigid-body transformation. Images were spatially smoothed using a 5mm full width at half maximum Gaussian kernel. A high pass filter was used to cut off temporal periods longer than 66 seconds. All images were initially co-registered to the participant's high resolution structural scan and were then co-registered to the MNI 152 person 2mm 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 3-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 6 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 intra-run 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 participants' 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 2, Panel C 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 (+.99 -.33 -.33 -.33 for regressors 5-8). The results of this analysis can be seen in Figure 4 and Table 2.

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).

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. The Economic Journal *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. Journal of Memory and Language *59*, 340-412.

Bates, D., Maechler, M., and Dai, B. (2008). lme4: Linear mixed-effects models using s4 classes.

Battigalli, P., and Dufwenberg, M. (2007). Guilt in games. American Economic Review *97*, 170-176.

Battigalli, P., and Dufwenberg, M. (2009). Dynamic psychological games. Journal of Economic 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. (2009). 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. American Economic Review *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., and al., e. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. Nature Neuroscience *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 and Economic Behavior *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 *11*, 419-427.

Fehr, E., and Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. Quarterly Journal of Economics *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 U S A *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. Cognition and 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., Hytonen, K., Rijpkema, M., Smidts, A., and Fernandez, 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 U S A.

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 U S A *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., Rigdon, M.L., and Smith, V.L. (2003). Positive reciprocity and intentions in trust games. Journal of Economic Behavior and Organization *52*, 267-275.

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

Mellers, B., Schwartz, A., and Ritov, I. (1997). Elation and disappointment: Emotional responses to risky options. Psychological Science *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 multishot 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. Social cognitive and affective neuroscience *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.

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

Rabin, M. (1993). Incorporating fairness into game theory and economics. American Economic Review *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.

Reuben, E., Sapienza, P., and Zingales, L. (2009). Is mistrust self-fulfilling? Economics Letters *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 (1759)). The theory of moral sentiments (Indianapolis: Liberty Fund).

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 U S A *105*, 12569-12574.

Stocker, 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. Social cognitive and affective neuroscience *4*, 294-304.

Venables, W.N., and Ripley, B.D. (2002). Modern applied statistics with S., Fourth edn (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. The Economic Journal *111*, 295-321.

Acknowledgements

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 acknowlege funding from the National Institute of Aging (R21AG030768) to A.G.S., the National Institute of Mental Health (R03MH077058) to A.G.S. & (F31MH085465) to L.J.C., and the National Science Foundation to M.D.

Figure Legends

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).

Figure 2. Behavioral results. Panel A depicts a histogram of the Investor's Investment for all trials for all participants (mean=51.7%, sd=20.7%). Panel B depicts a 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%). Panel C depicts a 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%). Panel D depicts the percentage of the Investor's investment (multiplied by 4) that the Trustee decides to return (mean=38.37%, sd=7.80%).

Figure 3. Behavioral results. A) Investor's 1st order belief (E_1S_2) by the Trustee's 2nd order belief ($E_2E_1S_2$). B) The amount returned by the Trustee (S_2) by their 2nd 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).

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. 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 $Z>2.3$, $p<0.05$.

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.

Figures

Figure 1.

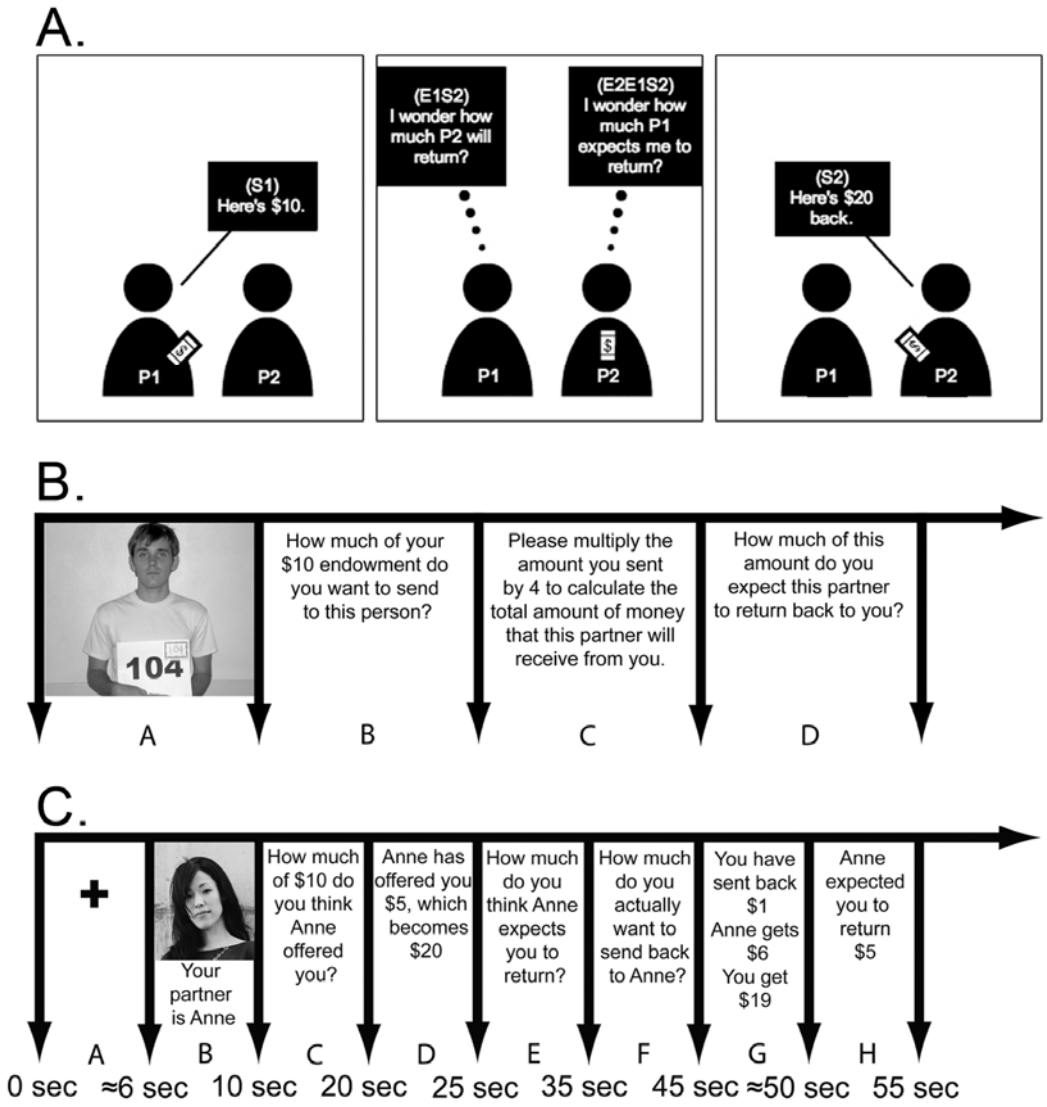

Figure 2.

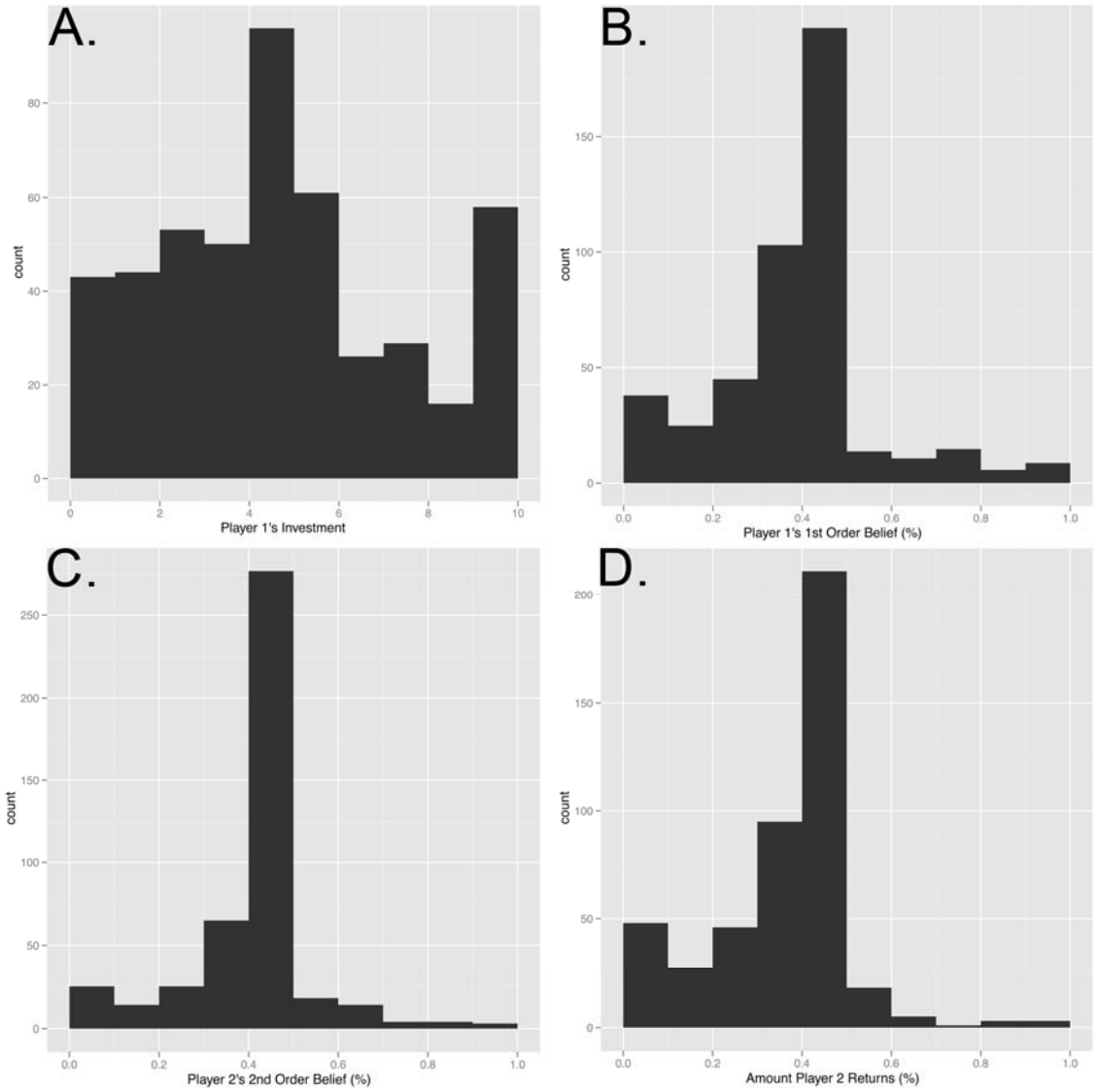
A.
count
Player 1's Investment
B.
count
Player 1's 1st Order Belief (%)
C.
count
Player 2's 2nd Order Belief (%)
D.
count
Amount Player 2 Returns (%)

Figure 3.

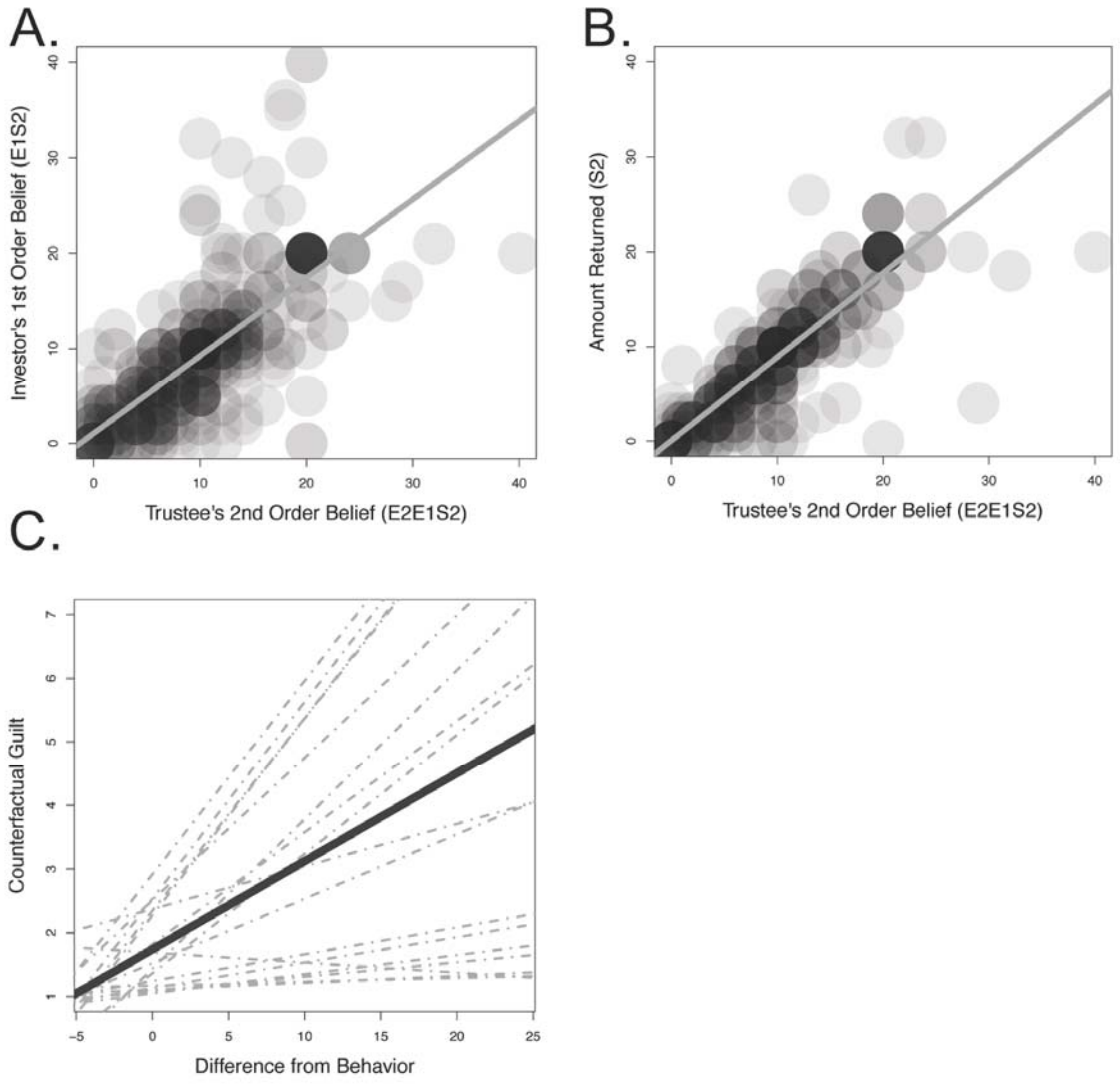
A.
Investor's 1st Order Belief (E1S2)
Trustee's 2nd Order Belief (E2E1S2)
B.
Amount Returned (S2)
Trustee's 2nd Order Belief (E2E1S2)
C.
Counterfactual Guilt
Difference from Behavior

Figure 4.

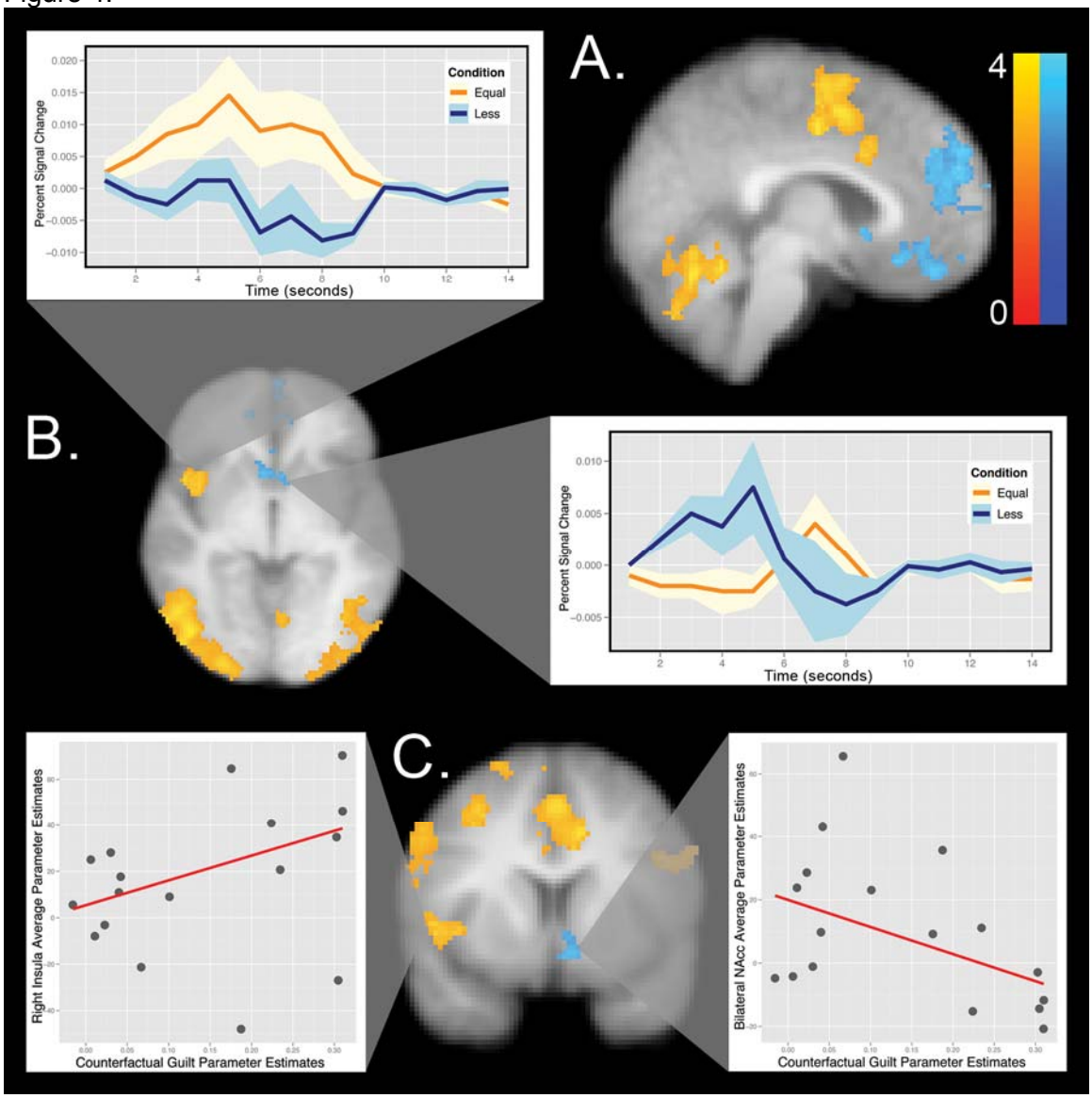
A.
4
0
Condition
Equal
Less
Percent Signal Change
Time (seconds)
B.
Condition
Equal
Less
Percent Signal Change
Time (seconds)
C.
Right Insula Average Parameter Estimates
Counterfactual Guilt Parameter Estimates
Bilateral NAcc Average Parameter Estimates
Counterfactual Guilt Parameter Estimates

Figure 5.