DECONSTRUCTING THE ROLE OF EXPECTATIONS IN COOPERATIVE BEHAVIOR WITH DECISION NEUROSCIENCE

by

Luke Joseph Chang

Copyright © Luke Joseph Chang 2012

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF PSYCHOLOGY

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2012
As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Luke J. Chang entitled “Deconstructing the role of expectations in cooperative behavior with decision neuroscience” and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

________________________________________ Date: 5/17/11
Alan G. Sanfey

________________________________________ Date: 5/17/11
John J.B. Allen

________________________________________ Date: 5/17/11
Dave Sbarra

________________________________________ Date: 5/17/11
Lee Ryan

Final approval and acceptance of this dissertation is contingent upon the candidate’s submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

________________________________________ Date: 5/17/11
Dissertation Director: Alan G. Sanfey
STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Luke J Chang
ACKNOWLEDGMENTS

I would like to acknowledge all of the people who have contributed in various ways to my development as a scientist, particularly those who have inspired me to become excited by new ideas, take risks, work harder, and find success in failure. I am incredibly grateful for all of the wonderful mentorship, support, and guidance I have received from various faculty in the department including Dave Sbarra, John Allen, Varda Shoham, Michael Rohrbaugh, Lee Ryan, and Catherine Shisslak. I’m especially appreciative of all the special help I received from Michael Frank in learning computational modeling and Martin Dufwenberg in learning Game Theory. Most importantly, I would like to thank my advisor, Alan Sanfey for his guidance, support, friendship, and remarkable mentorship, both in action and by example. I am incredibly appreciative of the level of trust he placed in me and the amount of independence he allowed in my work. I am also grateful for all the times he helped bail me out when I got in over my head.

I would also like to extend my gratitude to all of my lab mates, Trevor Kvaran, Katia Harle, Filippo Rossi, and Mascha van’t Wout, for all of their help over the years. You have all been wonderful friends, teachers, mentors, and colleagues. I’d also like to thank my friends and colleagues Brad Doll, Jim Cavanagh, Alec Smith, Mike X Cohen, and Tal Yarkoni for both teaching me and inspiring my work.

I would also like to thank all the Research Assistants who helped with data collection on the various projects including: Matt Kleinman, Carly Furgersen, Nico Warner, Alex Wilkins, Tess Gemberling, and Katie Martin.

I would also like to acknowledge the financial support that helped make this work possible from the National Institute of Mental Health to LJC (F31MH085465) and to AGS (R03MH077058), the National Institute of Aging to AGS (R21AG030768), the University of Arizona Technology Research Initiative Fund to LJC, and the University of Arizona Graduate College to LJC.

Finally, none of this would have been possible without my wonderful wife, Eunice. She has been there for me every step along the way from helping me decide to pursue a PhD, to helping me realize that it’s time to finish it, and every other step in between. I am forever grateful for all of your sacrifices, support, encouragement, and patience.
DEDICATION

To my parents,

Nancy Susan Chang and Franklin Minto Chang

For letting believe that I could do anything I wanted,

And for helping make that possible.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>8</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>9</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>10</td>
</tr>
<tr>
<td><em>Game Theory</em></td>
<td>11</td>
</tr>
<tr>
<td><em>Modeling</em></td>
<td>13</td>
</tr>
<tr>
<td><em>Expectations and Cooperation</em></td>
<td>15</td>
</tr>
<tr>
<td><em>Expectations, Cooperation, and the Brain</em></td>
<td>20</td>
</tr>
<tr>
<td>Specific Aims</td>
<td>22</td>
</tr>
<tr>
<td>Author Contributions</td>
<td>23</td>
</tr>
<tr>
<td>References</td>
<td>43</td>
</tr>
<tr>
<td>PRESENT PROJECT</td>
<td>25</td>
</tr>
<tr>
<td>Study 1 (Appendix A)</td>
<td>25</td>
</tr>
<tr>
<td>Study 2 (Appendices B &amp; C)</td>
<td>26</td>
</tr>
<tr>
<td>Study 3 (Appendix D)</td>
<td>27</td>
</tr>
<tr>
<td>Study 4 (Chapter 1)</td>
<td>28</td>
</tr>
<tr>
<td>Study 5 (Appendix E)</td>
<td>29</td>
</tr>
<tr>
<td>Conclusion</td>
<td>30</td>
</tr>
<tr>
<td>CHAPTER 1: MANIPULATING THE SOCIAL NORM (STUDY 4)</td>
<td>31</td>
</tr>
<tr>
<td>Introduction</td>
<td>31</td>
</tr>
<tr>
<td>Methods</td>
<td>32</td>
</tr>
<tr>
<td>Results</td>
<td>35</td>
</tr>
<tr>
<td>Discussion</td>
<td>41</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

FIGURE 1. PROPOSERS’ EXPECTATIONS..............................................36
FIGURE 2. PROPOSERS’ BEHAVIOR....................................................37
FIGURE 3. RESPONDERS’ EXPECTATIONS..........................................39
FIGURE 4. RESPONDERS’ BEHAVIOR...............................................40
ABSTRACT

This project attempts to understand the role of expectations in cooperative behavior using the interdisciplinary approach of Decision Neuroscience. While cooperation provides the foundation for a successful society, the underlying bio-psycho-social mechanisms remain surprisingly poorly understood. This investigation deconstructs cooperation into the specific behaviors of trust, reciprocation, and norm enforcement using the Trust and Ultimatum Games from behavioral economics and combines formal modeling and functional magnetic resonance imaging to understand the neurocomputational role of expectations in these behaviors. The results indicate that people appear to use context specific shared expectations when making social decisions. These beliefs are malleable and appear to be dynamically updated after an interaction. Emotions such as guilt and anger can be formally operationalized in terms of others’ expectations and appear to be processed by a specific neural system involving the anterior insula, anterior cingulate cortex, and supplemental motor cortex. Importantly, these neural signals appear to motivate people to not only behave consistent with these expectations, but also to help others update their beliefs when these expectations are violated. Further, violations of social expectations appear to promote enhanced memory for norm violators. This work demonstrates the neural and computational basis of moral sentiments.
INTRODUCTION

Daily life confronts us on a regular basis with social situations in which we choose to cooperate with those around us. Often, this takes the form of informal agreements, with the promise of benefits to all concerned if mutual cooperation is upheld. 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, Raudenbush, & Earls, 1997) and economically successful society (Smith, 1984 (1759); Zak & Knack, 2001), and thus it is essential to understand the processes that underlie these behaviors. A promising new approach, termed Neuroeconomics (Glimcher, Camerer, Poldrack, & Fehr, 2009) or Decision Neuroscience (Sanfey, 2007) attempts to integrate theory and methods from the diverse fields of psychology, economics, and neuroscience to understand the bio-psycho-social processes that underlie complex behaviors such as cooperation. The strength of this approach lies in its ability to use mathematical models of behavior to quantify complex psychological constructs in order to better characterize and understand the underlying neural processes. While a few preliminary studies have indicated that cooperation may lead to increased positive moods and increases in activity in regions of the brain associated with reward processing (Krueger, et al., 2007; Rilling, et al., 2002; Tabibnia & Lieberman, 2007), there has not yet been a systematic study of the
mechanisms underlying cooperation. One potential reason is there is currently no overarching theoretical conceptualization that has permitted a methodical investigation. Cooperation can be operationally defined as the act of multiple agents working together to attain a mutually beneficial goal. This act is comprised of the dissociable actions of trust and reciprocation, which are contingent upon beliefs. For example, trust is conditional on the belief that a partner will reciprocate and reciprocation is dependent on the belief that the partner trusts them to reciprocate. I contend that these expectations are dynamic and continuously updated and result in an emotional response when they are violated. Based on this framework, I predict that cooperation may be indirectly sustained via both norm enforcement and memory of expectation violations. The aim of this project, thus, is to provide a novel conceptual framework based on expectations to systematically understand the processes that facilitate cooperative behavior using the methodological approach of Decision Neuroscience.

*Game Theory*

The study of cooperation has traditionally been under the purview of economics and social psychology, and has been largely dominated by the theoretical approach of Game Theory. Game Theory attempts to use mathematical models to prescriptively determine the optimal behavior that maximizes individual utility within the context of an economic game (von Neumann & Morgenstern, 1944).
Games are stripped down examples of strategic situations that attempt to model specific kinds of social interactions involving multiple rational agents. Game Theory has spawned a multitude of different individual games, all of which have been used to examine slightly different aspects of strategic interactions. For example, one game commonly used to study cooperative behavior in the context of financial investment is the Trust Game (TG: Berg, Dickhaut, & McCabe, 1995). This game is typically played with two players – an investor and a trustee. The investor is endowed with a sum of money and can choose to invest any amount of that endowment in the trustee. The amount that they choose to invest is typically multiplied by the experimenter by a factor of 3 or 4 and the trustee then decides how much money they want to send back. The trustee can choose to reciprocate the investor's trust by returning more money then was initially invested, or abuse their trust by keeping all or most of the money. Trust is operationally defined as the amount of money that a player invests in their partner and trustworthiness is defined as the likelihood that the trustee will reciprocate trust (Camerer, 2003). The classic game theoretic solution is for the trustee to keep all of the money. The rational investor realizing that their partner is not going to reciprocate maximizes their payoff by keeping all their endowment. Most people, however, are willing to invest about 50% of their endowment and most trustees reciprocate about 95% of what was invested 50% of the time (Berg, et al., 1995).
Another game commonly used to study cooperative behavior in the context of economic bargaining is the Ultimatum Game (UG: Guth, Schmittberger, & Schwarze, 1982). This game is typically played with two players, one of whom is charged with splitting a sum of money, say $10, between the two. The responder must either accept or reject the offer put forth by the proposer. If the responder accepts the offer, each player is paid accordingly. If the responder rejects, however, then neither player receives anything. The standard game theoretic solution is for the proposer to offer the least amount of money possible, and that this will be accepted, on the grounds that something is better than nothing. However, countless experiments have demonstrated that proposers typically offer about half of their endowment, and that responders reject offers of 20% of the pot or less about 50% of the time (Camerer, 2003).

Modeling

There have been a number of theories proposed in response to classical economic theory’s failure to adequately predict behavior in these games. All of these theories continue to rely on the assumption that individuals are interested in maximizing their individual utility. However, these newer theories posit that individuals possess social preferences, which can be incorporated into their utility functions (Fehr & Camerer, 2007). Models provide a precise mathematical operationalization of a theory, which is useful for comparing theories, testing
specific hypotheses, and also for generating novel predictions. For example, one popular model has proposed that people possess distributional preferences and are motivated to minimize inequity in payoff outcomes between players (Bolton & Ockenfels, 2000; Dawes, Fowler, Johnson, McElreath, & Smirnov, 2007; Fehr & Schmidt, 1999). Others have argued that people value intentions more than simply minimizing inequality in the payoffs (Dufwenberg & Kirchsteiger, 2004; Falk & Fischbacher, 2006; Rabin, 1993) and there is some evidence that intention-based models can provide a better account of behavior than outcome-based models (Falk, Fehr, & Fischbacher, 2003; McCabe, Rigdon, & Smith, 2003). However, there are a number of findings that cannot be accounted for by either class of models (Henrich, et al., 2001; Henrich, et al., 2006; Xiao & Houser, 2005; Yamagishi, et al., 2009). These studies suggest that participants appear to have context specific expectations about the normative behavior and are motivated to behave consistent with these expectations (Bohnet & Zeckhauser, 2004; Krupka & Weber, 2009; Sanfey, 2009). This notion has recently been formalized using Psychological Game Theory (Battigalli & Dufwenberg, 2009; Geanakoplos, Pearce, & Stacchetti, 1989), which is a mathematical framework to allow beliefs to be incorporated into utility functions. These theories posit that people experience emotional reactions in response to expectation violations, which biases behavior to be consistent with the expectation (Battigalli & Dufwenberg, 2007; Dufwenberg, 2002; Dufwenberg & Gneezy, 2000; Smith, Working Paper). This development has allowed the psychological construct of
emotion to be formally operationalized and has resulted in a new class of economic models that are contingent upon expectations. Other classes of models developed within the domain of computer science use expectations to model trial-by-trial learning. These models are referred to as reinforcement learning (Sutton & Barto, 1998) and posit that learning occurs via prediction error signals that result when feedback differs from expectations (Rescorla & Wagner, 1972). In summary, there are a number of different classes of models that attempt to understand the individual processes underlying cooperative behavior. These models are important because they both allow specific hypotheses about the computational process underlying behavior to be systematically tested and also generate trial-by-trial predictions which can be used to uncover neural systems that parametrically track these processes (O'Doherty, Hampton, & Kim, 2007). The use of formal models to investigate the dynamics of neural systems is a recent development that possesses extraordinary potential, particularly for studying complex phenomena such as social cooperation.

**Expectations and Cooperation**

The success of human civilization can largely be attributed to our remarkable ability to cooperate with others. This seemingly innate capacity to cooperate has likely been accompanied by an instinctive ability to subscribe to a system of rules. At its most basic level, cooperation can only occur if two agents share a similar
belief. This means that two agents can only coordinate on a mutually beneficial outcome if they share the expectation that a particular outcome will be beneficial. This simple notion serves as the foundation of social contract theory, which posits that the legitimacy of a rule is contingent upon the consent of the governed (Rousseau, 1968 (1762)). Social norms describe shared beliefs about the appropriate behavior for a given context (Bicchieri, 2006; Elster, 1989; Fehr & Fischbacher, 2004) and serve as the basis for conventions and laws. Social psychologists have demonstrated that people are motivated to conform to social norms in a variety of contexts (Asch, 1956; Cialdini, Reno, & Kallgren, 1990; Milgram, 1963; Prentice & Miller, 1993).

Some social norms are only followed provisional on the belief that others will follow it. These beliefs are referred to as conditional norms and are often accompanied by competing motivations to either maximize utility for the individual or the group. For example, there is evidence that cooperation in a public goods game is conditional on the belief that others will cooperate (Fischbacher, Gachter, & Fehr, 2001). However, if participants do not believe that others will contribute to the public good, they are likely to free ride. Similarly, the act of trust is also contingent on the belief that the partner will reciprocate (King-Casas, et al., 2005). Players invest more money in the Trust Game when they believe there is a high probability their partner will honor their trust and decrease their likelihood of reciprocating when their trust has been abused.
Thus, one aspect of cooperation, trust, appears to be based on a conditional expectation, which changes as a function of experience.

Reciprocation is another aspect of cooperation that can be framed in terms of expectations. There is some evidence that reciprocating trust is based on inferring the expectations of a partner (Dufwenberg & Gneezy, 2000; McCabe, et al., 2003). Participants are more likely to honor trust if they believe that their partner expected them to reciprocate (Dufwenberg & Gneezy, 2000), which may be conveyed by a promise (Charness & Dufwenberg, 2006) or the size of the investment (McCabe, et al., 2003). The motivation for this behavior may be to avoid feeling guilt (Battigalli & Dufwenberg, 2007). Guilt refers to the aversive emotional state experienced after believing that a relationship partner was let down (Baumeister, Stillwell, & Heatherton, 1994) and has been argued to be a prosocial emotion (Tangney, Stuewig, & Mashek, 2007) because it prompts taking reparative action (Carlsmith & Gross, 1969; Darlington & Macker, 1966) to ameliorate the aversive state (Ketelaar & Au, 2003; Regan, 1971). In the context of cooperation, guilt has been formally operationalized as a counterfactual emotion based on beliefs about a partner’s expectations (i.e. their second order beliefs) and can be successfully minimized by taking action consistent with these beliefs (Battigalli & Dufwenberg, 2007; Charness & Dufwenberg, 2006; Dufwenberg, 2002; Dufwenberg & Gneezy, 2000).
Norm enforcement is another aspect of cooperation that is based on expectations. People appear to form beliefs about a social norm and are motivated to adhere to it and punish violators. For example, a credible threat of a sanction will promote increased adherence to the social norm (Fehr & Gachter, 2000; Yamagishi, 1986). Additional evidence suggests that independent observers will often incur a cost to punish a norm violator, even when decisions to do so will never affect their payoff (Fehr & Gachter, 2002). This has been referred to as altruistic punishment and appears to scale with deviation from the norm. Like reciprocation, this behavior appears to be motivated by an emotion. Social transgressions in the Ultimatum game elicit negative emotions of anger (Pillutla & Murnighan, 1996) and disgust (Chapman, Kim, Susskind, & Anderson, 2009). Third party observers report higher levels of anger when the deviation between a free rider’s contribution and the rest of the group was large compared to when the deviation was small (Fehr & Gachter, 2000). These self-reported emotions in reaction to unfair offers have also been associated with physiological responses (Ben-Shakhar, Bornstein, Hopfensitz, & van Winden, 2007; Chapman, et al., 2009; van ’t Wout, Kahn, Sanfey, & Aleman, 2006). Similar to guilt, anger has been formalized as the deviation between expectations about the contextual norm and biases behavior to enforce the norm (Smith, Working Paper).
Taken together, these findings indicate that expectations play a critical role in the various components of cooperative behavior. The act of trust appears to be based on a conditional norm, while the acts of reciprocation and norm-enforcement appear to be based on a social norm. The motivation to adhere to and enforce a social norm seems to arise from emotional responses of guilt and anger respectively. An additional question is how these processes influence subsequent interactions. Repeated Trust Games indicate that the conditional beliefs underlying decisions to trust are updated after every trial (Delgado, et al., 2005; King-Casas, et al., 2005). However, less is known about how these interactions affect other cognitive systems such as memory. One prominent theory has proposed that humans may have evolved a specific ability to remember “cheaters” – individuals who benefit themselves by violating a social contract (Cosmides & Tooby, 1992). However, support for this theory has been mixed with some studies reporting enhanced memory of cheaters (Chiappe, et al., 2004; Mealy, Daood, & Krage, 1996; Oda, 1997), while others find either no effect (Barclay & Lalumiere, 2006; Mehl & Buchner, 2008) or support for enhanced confidence in remembering altruists (Barclay & Lalumiere, 2006). An alternative explanation that can potentially reconcile these discrepancies in the literature is that people are sensitive to expectation violations and there is some supporting evidence indicating that people may have enhanced memory for partners that behave contrary to social conventions, regardless of their behavior (Barclay, 2008).
Growing interest in social cognitive neuroscience has yielded a number of studies investigating the neural processes associated with cooperation. However, a failure to actually test specific theories, inconsistent methodologies, and a lack of a unifying interpretive framework has unfortunately made it difficult to integrate these findings. Expectations may provide a useful framework to understand these findings and to generate novel predictions. Work on expectations has yielded a consistent neural system which includes the anterior insula and anterior cingulate cortex (ACC)/supplementary motor area (SMA). This system, often referred to as the salience detection network (Seeley, et al., 2007) or goal-directed network (Dosenbach, et al., 2006), has been thought to be involved in interoception (Craig, 2002) and subjective awareness (Craig, 2009) and has been demonstrated to be functionally involved in switching between the default and executive control networks (Sridharan, Levitin, & Menon, 2008). The salience detection network has been observed when anticipating an aversive event such as a shock (Butler, et al., 2007; Phelps, et al., 2001) and may potentially mediate the placebo effect (Wager, et al., 2004; Wager, Scott, & Zubieta, 2007). In addition, there is overwhelming evidence implicating this system when an unexpected event occurs (Downar, Crawley, Mikulis, & Davis, 2000; Huettel, Mack, & McCarthy, 2002) or when detecting an error (Debener, et al., 2007; Ullsperger, et al., 2007).
al., 2005; Klein, et al., 2007). Consistent with this work, this same network as been associated with conflict with a social norm (Klucharev, Hytonen, Rijpkema, Smidts, & Fernandez, 2009) and also appears to be involved with actually conforming behavior to the norm (Berns, Capra, Moore, & Noussair, 2009).

Numerous studies have investigated the neural systems associated with Investors' decisions to trust in a repeated game and have reliably implicated the ACC/MPFC and the striatum (Delgado, et al., 2005; King-Casas, et al., 2005; Krueger, et al., 2007; McCabe, Houser, Ryan, Smith, & Trouard, 2001; Rilling, et al., 2002; Tomlin, et al., 2006). The MPFC findings are typically interpreted as evidence of theory of mind processing (Amodio & Frith, 2006) while the striatal results are taken as evidence that conditional beliefs are being updated via prediction error (Delgado, et al., 2005; King-Casas, et al., 2005). The sole study investigating decisions to reciprocate in the Trust Game found activity in the insula and ACC/SMA (Van Den Bos, Van Dijk, Westenberg, Rombouts, & Crone, 2009). These results suggest that conditional beliefs may be updated by a prediction error signal in a similar manner as basic learning and that reciprocation involves similar processes as conforming behavior to a norm.

There have been a number of studies investigating norm enforcement using the UG (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Tabibnia & Lieberman, 2007) and third party punishment (de Quervain, et al., 2004). These studies
have revealed two distinct neural systems. Rejecting unfair offers is associated with increased activity in the insula, ACC/SMA, & DLPFC and has been interpreted as evidence of an emotional response (Sanfey, et al., 2003). Striatal activity has been associated both with accepting fair offers and third-party punishment (de Quervain, et al., 2004). Interestingly, dividing money when there is a threat of sanction promotes greater conformity to the norm and is associated with insula activity (Spitzer, Fischbacher, Herrnberger, Gron, & Fehr, 2007). These results suggest that enforcing a norm may be associated with the insula/ACC/SMA system or a striatal system.

*Specific Aims*

The aim of the present project is thus to investigate the role of expectations in cooperation using the framework of Decision Neuroscience. Cooperation will be operationalized as the dissociable behaviors of (1) trust, (2) reciprocation, and (3) norm enforcement. In addition, the influence of expectations on social memory will also be examined. Importantly, expectations will be formally modeled using mathematical tools from psychological game theory and reinforcement learning, which will facilitate greater interpretability of the associated neural processes. Utilizing this framework, I predict that the various aspects of cooperation will be associated with a brain network that has been reliably associated with expectations (e.g. insula, ACC/SMA). This study has the potential to provide a
novel contribution to the fields of Decision Neuroscience and Social Cognitive Neuroscience by both introducing an overarching theoretical framework and demonstrating how formal models can be used to operationalize nebulous constructs.

*Author Contributions*

The work presented in this dissertation has been conducted over the past 4 years in collaboration with a number of colleagues. While these colleagues have provided some conceptual and minor technical assistance to this work, I am primarily responsible for the experimental design, data collection and analysis and writing. My doctoral advisor, Dr. Alan Sanfey was instrumental in supervising this process and provided considerable assistance in editing this work.

The work on trust (Chang, Doll, van ’t Wout, Frank, & Sanfey, 2010) was conducted in collaboration with Brad Doll, Drs. Mascha van’t Wout, Michael Frank, and Alan Sanfey and published in *Cognitive Psychology* in 2010. The conceptual framework was a follow up to Dr. van’t Wout’s study on perceptions of facial trustworthiness and behavioral trust (van ’t Wout & Sanfey, 2008) and utilized a few models developed initially by Brad Doll and Dr. Frank (Doll, Jacobs, Sanfey, & Frank, 2009). I developed the main model tested in this work under
the supervision of Dr. Frank and designed, collected and analyzed the data, and wrote up the manuscript under the supervision of Dr. Sanfey.

The work on reciprocation (Chang, Smith, Dufwenberg, & Sanfey, In Press) was conducted in collaboration with Drs. Alec Smith, Martin Dufwenberg, and Alan Sanfey and was accepted for publication in *Neuron* in 2011. The theoretical framework for guilt-aversion was initially developed by Dr. Dufwenberg (Battigalli & Dufwenberg, 2009; Charness & Dufwenberg, 2006; Dufwenberg, 2002). The formal modeling work was conducted under the supervision of Dr. Smith. I designed the study, collected and analyzed the data, and wrote the manuscript under the supervision of Dr. Sanfey.

The work on norm enforcement (Chang & Sanfey, Under Review) was conducted in collaboration with Dr. Alan Sanfey and has been submitted for publication. I designed the study, collected and analyzed the data, and wrote up the results under his supervision.

The work on social memory (Chang & Sanfey, 2009) was conducted in collaboration with Dr. Alan Sanfey and has been published in *Frontiers in Behavioral Neuroscience* in 2009. Dr. Sanfey and I designed the study and wrote up the results together, and I collected and analyzed the data under his supervision.
PRESENT PROJECT

Using the framework of Decision Neuroscience, this project investigates the role of expectations in cooperation in 5 studies. Cooperation was deconstructed into the distinct constructs of trust (Study 1), reciprocation (Study 2), and norm enforcement (Studies 3 & 4). In addition, the influence of expectations on social memory was examined (Study 5). The specific details of this project are presented in the papers appended to this dissertation. The only project that has not yet been submitted for publication is described in chapter 1 (Study 4). The following is a summary of the most important findings.

Study 1 (Appendix A)

Recent efforts to understand the mechanisms underlying human cooperation have focused on the notion of trust, with research illustrating that both initial impressions (Delgado, et al., 2005; van 't Wout & Sanfey, 2008) and previous interactions (King-Casas, et al., 2005) impact the amount of trust people place in a partner. Less is known, however, about how these two types of information interact in iterated exchanges. Study 1 examined how implicit initial trustworthiness information interacts with experienced trustworthiness in a repeated Trust Game. Consistent with my hypotheses, these two factors reliably influence behavior both independently and synergistically, in terms of how much money players were willing to entrust to their partner and also in their post-game
subjective ratings of trustworthiness. To further understand this interaction, I used Reinforcement Learning models to test several distinct processing hypotheses. These results suggest that trustworthiness is a belief about probability of reciprocation based initially on implicit judgments, and then dynamically updated based on experiences. Importantly, my model was not only able to account for player’s behavior in the game, but also was able to predict their post-game trustworthiness ratings.

Study 2 (Appendices B & C)

An additional open question is why do people often choose to reciprocate other’s trust when they can better serve their interests by acting selfishly? Early theoretical work suggested that people find cooperation intrinsically rewarding (Rilling, et al., 2002) and may choose to reciprocate to experience a warm glow” (Andreoni, 1990). However, an alternative hypothesis is that people dislike disappointing others’ expectations as it is associated with a negative affective state such as guilt. Anticipating this negative state may motivate cooperative behavior. In Study 2, I 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 their partner’s trust. The results indicate that Trustees were able to accurately infer the Investor’s expectations and, consistent with the model of guilt, appeared to use these expectations to
make their decisions to cooperate. In the imaging analyses, I observed increased activation in the insula, supplementary motor area, dorsolateral prefrontal cortex (PFC), and temporal parietal junction when participants were behaving consistent with the guilt 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. These results appeared to be modulated by individual differences in guilt sensitivity. 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.

*Study 3 (Appendix D)*

While the previous study demonstrated that participants were motivated to cooperate to avoid violating others’ expectations, I was also interested in whether people would take action and enforce a shared expectation when violated. This notion of expectation violation is different from the widely accepted theory of fairness, which is understood in terms of equity of payoffs. In Study 3, I outline a novel, expectation-based, neurocomputational model of social preferences, and find that it outperforms the standard inequity-aversion model in explaining decision behavior in a social interactive bargaining task. This is supported by fMRI findings showing that the tracking of social expectation violations is
processed by anterior cingulate cortex, extending previous computational conceptualizations of this region to the social domain.

Study 4 (Chapter 1)

The two previous studies (Chang & Sanfey, Under Review; Chang, et al., In Press) demonstrated that participants’ expectations about others' beliefs and behaviors played an important computational role in how players made their decisions. However, the results of these studies were based on correlating expectations with behavior. My formal model of expectations in the Ultimatum Game (Chang & Sanfey, Under Review) predicts that manipulating expectations will change player's strategies. Consistent with the predictions of the model, there is preliminary evidence that Responders who expected low proposals in the Ultimatum Game are more likely to accept more inequitable offers (Kleinman, Chang, & Sanfey, In Preparation; Sanfey, 2009). However, it remains an open question as to how both players will respond when they believe that the normative behavior for Proposers is to offer more than half of their endowment. In Study 4, I tested this precise hypothesis and examined both players behavior after instructing players that the social norm is for Proposers to offer more than half of their endowment. In accordance with the predictions of the model, I found that our manipulation resulted in Proposers offering more than half of their endowment and Responders increasing their rejection rates of unfair offers
compared to control conditions.

Study 5 (Appendix E)

Finally, I was interested in how these violations of expectations in the game impacted subsequent memory for these interactions. In Study 5, I investigated this question by using functional magnetic resonance imaging to scan participants as they viewed photographs of people they had either recently played an Ultimatum Game with in the role of Responder, or that they had never seen before. Based on previous work that has investigated “cheater detection”, I was interested in whether participants demonstrated a relative enhanced memory for partners that made either fair or unfair proposals. I found no evidence, either behaviorally or neurally, supporting enhanced memory based on the amount of money offered by the Proposer. However, participants’ initial expectations about the offers they would experience in the game did appear to influence their memory. Participants demonstrated relatively enhanced subjective memory for partners that made proposals that were contradictory to their initial expectations. In addition, I observed two distinct brain systems that were associated with partners that either offered more or less than the participants’ expectations. Viewing pictures of partners that exceeded initial expectations was associated with the bilateral anterior insula, anterior cingulate cortex/premotor area, striatum, and bilateral posterior hippocampi, while viewing partners that
offered less than initial expectations was associated with bilateral temporal-parietal junction, right STS, bilateral posterior insula, and precuneus. These results suggest that memory for social interaction may not be guided by a specific cheater detection system, but rather a more general expectation violation system.

Conclusion

Taken together, these results provide compelling evidence for the role of expectations in cooperative behavior. People appear to use shared expectations of normative behavior for specific contexts when making social decisions. These beliefs are malleable and are updated after receiving feedback from each interaction. Emotions processed by a specific neural system involving the anterior insula, ACC, and SMA appear to motivate people to not only behave consistent with these expectations, but also to help others update their beliefs when these expectations are violated. This work demonstrates how the interdisciplinary decision neuroscience framework can be used to investigate the neural, psychological, and economic mechanisms underlying amorphous and complex phenomena such as cooperation.
CHAPTER 1: MANIPULATING THE SOCIAL NORM (STUDY 4)

Introduction

We have previously provided evidence suggesting that expectations play an important role for decision-making in the Ultimatum Game (Chang & Sanfey, Under Review). Participants appear to generate an expectation of a context specific social norm and are more likely to reject offers that deviate from this expectation. This notion has been supported by other findings. Proposers are more likely to make a fair offer if they believe that (1) most people make fair offers (Krupka & Weber, 2009), (2) they might be punished if they don’t (Spitzer, et al., 2007), and (3) the Responder can express their emotional reaction to their offer (Xiao & Houser, 2009). There is also evidence that Responders are more likely to accept inequitable offers, if they believe that to be the norm (Sanfey, 2009). Finally, there is a hint of evidence that inequitable offers in the Responder’s favor (i.e. hyperfair) may also be rejected with a higher likelihood, albeit to a lesser extent than inequitable offers in the Proposer’s favor (Henrich, et al., 2006; Yamagishi, et al., 2009). The expectation hypothesis posits that people are interested in conforming to and enforcing the norm. Two counterintuitive predictions of this hypothesis are that (1) Proposers should be more likely to make a hyperfair offer (i.e. offer more than 50% of their endowment) if they believe that to be the norm and (2) Responders should be more likely to reject unfair offers if they believe the norm to be hyperfair offers.
Thus, the goal of Study 4 is to test these two novel predictions behaviorally in the context of an Ultimatum Game.

Methods

Participants

One hundred and six students (mean age = 19.37, sd = 2.76, female = 50%) were recruited from the University of Arizona Psychology subject pool to participate in this experiment. All participants gave informed consent according to procedures approved by the University of Arizona's Institutional Review Board.

Procedure

This study employed a between-subjects design to examine the role of hyper-fair expectations in Proposer and Responder behavior in the Ultimatum Game (Guth, et al., 1982). Proposers were randomly assigned to one of three conditions: high expectations (n=22), low expectations (n=21), and no expectations (n=21) and met in large groups. Participants first signed both experimental and photo consent forms, and then each had their photo taken holding their participant identification number. Participants then read the instructions for the game and answered questions to make sure they understood how the game worked. Next, participants read a very subtle manipulation, which indicated the normative
proposer behavior. For example, participants in the high expectation group read, “Just to give you some information about how the game is typically played by college students, in general, the most common offers made are quite high, that is, an offer of $7 or $8 to your partner when dividing a $10 pot.” Participants in the low expectation read, “Just to give you some information about how the game is typically played by college students, in general, the most common offers made are quite low, that is, an offer of $1 or $2 to your partner when dividing a $10 pot. Finally, participants in the control condition did not receive any information about the normative behavior. Participants then reported their expectations about the types of offers they believed most people would offer and made their actual offer. Afterwards, participants reported their expectations about the likelihood of responders accepting different offer amounts and then filled out questionnaires. Participants played for real money and were paid based on how Player 2 responded to their offer at the end of the experiment several weeks later.

Responders were randomly assigned to two conditions: high expectations (n=22) or no expectations (n=20) and met individually in the laboratory. Responders received a manipulation that was identical to the Proposers, but responded to offers on a computer running Eprime. Responders encountered two repetitions of offers ranging from $1-$9 (18 in total). Most of these offers were made by actual Proposers and were randomly selected from the distribution of offers made by all Proposers in the three expectation conditions. However, for offers
that were never actually made (e.g., $1, $2, and $9), Proposer faces were randomly assigned to these “fake” offers to ensure that participants encountered all possible nonzero offers. Afterwards participants rated the proposers faces on 6 dimensions (e.g., attractiveness, aggressiveness, competence, likeability, trustworthiness, and fairness) on a 5 point likert scale all of the faces that they played with and faces that they didn’t and filled out questionnaires. Participants were paid based on their decision for one randomly selected trial at the end of the experiment.

Expectations

Prior to making their decisions, we elicited participants’ beliefs about the kinds of offers they expected to encounter, with participants being asked the number of people out of 100 that they believed would make each offer from the set [$1, $9]. These elicited expectations were used to create a distribution of the frequency of offers that they expected to encounter. The weighted mean of this distribution was used to represent each participant’s initial expectation (Sanfey, 2009).

Analyses

All analyses were conducted using the R statistical framework (R_Development_Core_Team). The Proposer data was analyzed using a one-
way ANOVA and t-tests in the context of a general linear model using the LM function. Responder data was analyzed using a mixed logit model with varying intercepts using the GLMER function from the LME4 package (Bates, Maechler, & Dai, 2008).

Results

Proposers

We found a significant effect of our manipulation on Proposers’ expectations, F(2,60)=4.10, p=0.02 (see Figure 1). Participants who were told that most people make high offers expected most Proposers to give significantly larger offers (mean=4.49, sd=1.07) compared to the control condition (mean=3.39, sd=1.36), b=1.10, se=0.39, t=2.85, p=0.006. There was no significant difference between the low expectation condition (mean=4.04, sd=1.30) and the control or high expectation conditions, b=0.65, se=0.39, t=1.64, p=0.16 and b=-0.45, se=0.38, t=-1.18, p=0.24 respectively.
Consistent with our hypothesis, we observed a significant effect of our expectation manipulation on the amount of money the Proposers offered (see Figure 2), $F(2,60)=6.68$, $p=0.002$. Proposers who believed most people would
make hyper fair offers gave significantly more money (mean=5.63, sd=1.43) than participants in the control condition (mean=4.38, sd=1.20), b=1.26, se=0.36, t=3.47, p<0.01. We did not observe a significant difference between the amount of money offered in the low expectation condition (mean=4.65, sd=0.81) compared to the control condition, b=0.27, se=0.37, t=0.73, p=0.47.

Figure 2. Proposers Behavior

Note: Error Bars reflect ± 1 SE.
Responders

We did not observe a significant effect of our manipulation on Responders’ Expectations (see Figure 3), $F(1,40)=0.75$, $p=0.39$. However, participants in the high expectation condition did expect Proposers to make slightly higher offers (mean=4.86, sd=1.40) compared to the control condition (mean=4.51, sd=1.17).
Consistent with previous research, we found that the size of the offer was significant in predicting Responders' decisions, with larger offers increasing the likelihood of accepting the offer, odds ratio = 5.09, log-odds = 1.63, se = 0.20,
z=8.00, p<0.001. We also observed a significant interaction between the expectation manipulation and offer amount for accepting offers, odds ratio=0.44, log-odds=-.82, se=0.22, p<0.001. This interaction is depicted in Figure 4 and indicates that participants in the high expectation condition (n=20) were less likely to accept low offers than participants in the control condition (n=22). We did not observe a significant main effect of condition, odds ratio=4.23, log-odds=1.55, se=0.94, p=0.12.

Figure 4. Responder Behavior

Note: Error Bars reflect ± 1 SE.
Discussion

In this study we were interested in examining the effect of manipulating expectations about normative Proposer behavior on subsequent Proposer and Responder behavior in the Ultimatum Game. We were specifically interested in how a hyper-fair norm (i.e., expecting Proposers to give more than 50% of their endowment) would impact both players behavior. Consistent with the predictions of our expectation model, our manipulation increased the average amount offered by Proposers and decreased the likelihood of responders accepting low offers.

These results provide further empirical support for our formal model of expectations (Chang & Sanfey, Under Review). Importantly, these findings cannot be explained by the competing inequity aversion account as this framework predicts that participants only care about minimizing discrepancies in payoff outcomes expectations and thus expectations should not affect player’s decisions. In fact, our Proposer data suggest that our manipulation increased inequity in payoffs in the other player’s favor.

Unfortunately, we did not find a significant effect of low expectations on Proposer behavior. This provides a challenge to our theory, but can potentially be explained by the fact that participants wanted to ensure that they would make some money. Also, it’s possible that participants in this condition found the
manipulation to be implausible. Unfortunately we did not collect post-experiment feedback. While we manipulated expectations about what most Proposers offer, this may be an independent expectation of the amount that they believe most responders will accept. This specific hypothesis should be explored in future research.

Taken together with previous work (Chang & Sanfey, Under Review; Kleinman, et al., In Preparation; Sanfey, 2009), these results provide compelling support for our theory of expectations. Participants appear to generate context specific beliefs about the social norm and behave consistent with these beliefs. Players also appear to be motivated to maintain the norm by punishing violators. These shared expectations appear to be malleable as a subtle manipulation about whether the norm is high or low can change how people make decisions. It remains unclear if these norms are simply serving as an anchor point or are actually influencing beliefs. We have preliminary work which indicates that these beliefs can be updated based on experience (Kleinman, et al., In Preparation), but this question should be explored more thoroughly in future work.
APPENDIX A: SEEING IS BELIEVING: TRUSTWORTHINESS AS A DYNAMIC BELIEF (STUDY 1)

Originally published in *Cognitive Psychology*. Reprinted with permission from the authors.

Seeing is believing: Trustworthiness as a dynamic belief

Luke J. Chang\textsuperscript{a}, Bradley B. Doll\textsuperscript{b}, Mascha van ’t Wout\textsuperscript{b}, Michael J. Frank\textsuperscript{b}, Alan G. Sanfey\textsuperscript{a,}\textsuperscript{*}

\textsuperscript{a} Department of Psychology, University of Arizona, 1503 E. University Blvd, Tucson, AZ 85721, United States
\textsuperscript{b} Departments of Cognitive & Linguistic Sciences and Psychology, Brown University, 190 Thayer St, Providence, RI 02912-1978, United States

\textbf{A R T I C L E   I N F O}

Article history:
Accepted 22 March 2010

Keywords:
Decision-making
Trust
Social learning
Trustworthiness
Reinforcement learning
Face perception

\textbf{A B S T R A C T}

Recent efforts to understand the mechanisms underlying human cooperation have focused on the notion of trust, with research illustrating that both initial impressions and previous interactions impact the amount of trust people place in a partner. Less is known, however, about how these two types of information interact in iterated exchanges. The present study examined how implicit initial trustworthiness information interacts with experienced trustworthiness in a repeated Trust Game. Consistent with our hypotheses, these two factors reliably influence behavior both independently and synergistically, in terms of how much money players were willing to entrust to their partner and also in their post-game subjective ratings of trustworthiness. To further understand this interaction, we used Reinforcement Learning models to test several distinct processing hypotheses. These results suggest that trustworthiness is a belief about probability of reciprocation based initially on implicit judgments, and then dynamically updated based on experiences. This study provides a novel quantitative framework to conceptualize the notion of trustworthiness.

\textcopyright{} 2010 Elsevier Inc. All rights reserved.

1. Introduction

The success of human civilizations can be largely attributed to our remarkable ability to cooperate with other agents. Cooperative relationships, in which individuals often endure considerable risk, are built on the foundation of trust – a nebulous construct that is nevertheless intimately tied to both...
interpersonal (Rempel, Holmes, & Zanna, 1985) and economic prosperity (Zak & Knack, 2001). However, as anyone who has ever purchased a used car can attest, not everyone turns out to actually be trustworthy. Thus, the accurate inference of an individual’s level of trustworthiness is crucial for the development of a successful relationship. How then does one actually assess trustworthiness? Business people often attest to the importance of looking a future partner in the eye and physically shaking their hand before signing a contract. When physical meetings are not an option, people frequently rely on reputation, which in the world of online commerce has taken the form of buyer testimonials about a seller’s prior transactions on sites such as eBay. This is consistent with the notion that the best predictor of an individual’s level of trustworthiness is their behavior in previous interactions with us (Axelrod & Hamilton, 1981; King-Casas et al., 2005). We are more likely to invest trust in someone previously shown to be trustworthy than someone who has previously betrayed us. Therefore, one useful model for inferring the trustworthiness of another is to make an initial assessment based on available information, and then update this judgment based on subsequent interactions.

Because trust is an amorphous construct, it is often difficult to measure and operationalize. From a psychological perspective, trust can be considered the degree to which an individual believes that a relationship partner will assist in attaining a specific interdependent goal (Simpson, 2007). While this definition can apply to any number of social interactions, consider an example of confiding in a colleague. Alex has been privately deliberating a decision and seeks feedback from Trevor. Alex is interested in Trevor’s perspective, but does not want David to know. Trust, in this example is Alex’s belief that Trevor will not divulge this sensitive information to David. Of course, trust is a broad concept that extends to many different aspects of social interaction and may depend on the assessment of a variety of factors, including, honesty, competence, competitiveness, and greed. However, in order to study trust experimentally, it is necessary to have a good operationalization of it, even though this may be limiting.

The Trust Game is a task that has been developed by Behavioral Economists to serve as a proxy for everyday situations involving trust like in our example. The Trust Game explicitly measures our Psychological operationalization of trust, and assesses the degree to which an individual is willing to incur a financial risk with a partner (Berg, Dickhaut, & McCabe, 1995). This simple game involves two players, A and B. Player A is endowed with an initial amount of money, say $10, and can choose to invest any amount of this endowment with B. The amount that Player A invests is multiplied by the experimenter by some factor, usually 3 or 4, and then Player B decides how much of this enlarged endowment, if any, they would like to return to Player A. The partner can choose to repay the investor’s trust by returning more money than was initially invested, or abuse their trust by keeping all (or most) of the money. In this game, trust is operationally defined as the amount of money that a player invests in their partner, and trustworthiness is defined as the likelihood that the partner will reciprocate trust. Evidence from empirical work has shown that most investors are willing to transfer about half of their endowment. In turn, when the investment is multiplied by a factor of 3, partners are usually willing to reciprocate trust so that both partners end up with approximately equal payoffs (Berg et al., 1995). This simple game provides a useful behavioral operationalization of trust, and also demonstrates that in general players exhibit both trust and trustworthiness, contrary to the standard predictions of economic Game Theory (Camerer, 2003). Additionally, this game can serve as a framework for experimental manipulations of social signals. Two types of social signals we will investigate in this study are the degree to which both the initial judgments of a partner and previous experience with that partner can alter decisions of trust and reciprocity.

Research has demonstrated that trustworthiness is often rapidly inferred from social signals, and in turn influence behavior in the Trust Game. Trustworthiness judgments are influenced by brief social interactions (Frank, Gilovich, & Regan, 1993), as well as information about an individual’s moral character (Delgado, Frank, & Phelps, 2005). Even more subtly, signals of trustworthiness can be detected from simply viewing faces (Winston, Strange, O’Doherty, & Dolan, 2002). Facial expressions can be processed outside of conscious awareness (Morris et al., 1998), and indeed, competence judgments about an individual can be made within 100 ms (Willis & Todorov, 2006) and affective judgments about an individual can be made as quickly as 140 ms (Pizzagalli et al., 2002). Individuals who are attractive or who appear happy are also more likely to be viewed as trustworthy (Scharlemann, Eckel, Kacelnik, & Wilson, 2001). Our group has recently investigated how initial
impressions can influence trust, and demonstrated that implicit judgments of facial trustworthiness can predict the amount of financial risk a person is willing to take in a Trust Game (van 't Wout & Sanfey, 2008). In this study, normed ratings of player trustworthiness (as assessed by briefly viewing a photograph of each player) were a significant predictor of how much money these players were given in a standard one-shot Trust Game. These set of studies support the notion that both explicit (e.g., information about a partner’s moral character) and implicit social signals (e.g., facial trustworthiness) can influence initial judgments of trustworthiness, and that these judgments can in turn impact the degree to which people actually place trust in other individuals in a meaningful social interaction.

Social signals can also be inferred from repeated interactions. The best predictor of whether a person will place trust in their partner in a given Trust Game round is whether or not this partner previously reciprocated trust (King-Casas et al., 2005). The process of placing trust when it has previously been reciprocated, but stopping once trust is abused, is often referred to as a tit-for-tat strategy, and has been demonstrated to be the optimal strategy for repeated interactions (Axelrod & Hamilton, 1981). Repeated interactions have also been shown to influence subjective ratings of moral character in a Prisoner’s Dilemma game (Singer, Kiebel, Winston, Dolan, & Frith, 2004) and in a Trust Game (Delgado et al., 2005). These findings suggest that in a repeated interaction, trustworthiness can be learned based on the history of a partner’s behavior.

One model of investor behavior in the context of a Trust Game therefore involves an initial judgment of trustworthiness based on available information, which is then updated based on subsequent interactions with that partner. One question that currently remains unanswered is the interaction between this initial assessment and the subsequent updating, for example, the degree to which each signal may contribute to the final trust decision, and how concordant (a trustworthy face engaged in reciprocal behavior) and conflicting (e.g., a trustworthy face who does not reciprocate) information is handled. No study has directly investigated this question in the context of a Trust Game, though one experiment has provided preliminary evidence suggesting that initial judgments may influence the way information from repeated interactions is updated (Delgado et al., 2005). In this study, participants played a repeated Trust Game with three fictional characters. Prior to interacting with these purported partners, participants were given a short vignette describing the moral character of each partner. One character was depicted as “good”, one “neutral”, and a third as “bad”. The investigators observed that participants rated the “good” character as more trustworthy at the start of the game, and were in turn more likely to trust them. However, because all partners reciprocated 50% of the time, participants learned to trust the “good” partners less over time, and in fact began to “match” the 50% reinforcement probability (Herrnstein, 1961). At the conclusion of the game, even though participants trusted the “good” partner less than they did at the beginning, they still placed more trust in them than either of the other partners, and were still investing more than 60% of the time. There are two possible interpretations of this finding from a reinforcement learning framework. First, as suggested by the authors, the positive moral information may have biased the participants to ignore negative feedback, meaning they were unable to update the value of the partner after they were betrayed. An alternative interpretation is that the positive moral information increased the initial trust evaluation of the partner, but did not influence the way the participant interpreted the feedback. According to this interpretation, if given enough trials, the participant would have eventually learned the 50% reinforcement rate, though it would have taken longer compared to the neutral and negative partners. However, the design employed in this study makes it difficult to assess which hypothesis is more likely.

A useful method to examine the question of how initial judgment and experience interact is to employ mathematical models of behavior. Reinforcement learning (RL) is concerned with understanding how people learn from feedback in repeated interactions with the environment (Sutton & Barto, 1998). Assuming that the decision-maker is attempting to maximize his or her reward on each trial, one strategy is to predict the value of an environmental state, and then update these predictions based on the actual feedback received. One method for updating predicted values is to use the simple Rescorla–Wagner delta rule (Rescorla & Wagner, 1972), which quantifies on each trial the difference between the predicted value $V_S(t)$ and actual reward $r$, with this difference referred to as prediction error.

\[
\delta = r_S - V_S(t)
\]
The most straightforward way to learn the value of the relevant stimulus $s_t$ is to update its predicted value in proportion to the current prediction error $\delta_t$. The degree to which the prediction error influences the new value is scaled by a learning rate $\alpha$, where $0 < \alpha < 1$.

$$V_s(t + 1) = V_s(t) + \alpha\delta$$

Thus, receiving rewards greater than expected will lead one to increase the value associated with a given stimulus. Conversely, receiving rewards that were less than expected will cause a decrease in that value. Using a RL approach, all stimuli have an initial starting reward value, which is updated via a learning rule. Because of its simplicity, this framework not only provides a very powerful way to understand how people learn from feedback, but also provides a principled way to understand how social signals influence learning in a repeated Trust Game.

Use of this framework to understand how people learn in a social context encourages very specific hypothesis testing, and has the potential to provide insight into the subtle processes involved in social learning. To date, relatively few studies have attempted to study social learning from an RL perspective (Behrens, Hunt, Woolrich, & Rushworth, 2008; King-Casas et al., 2005). However, some recent studies have begun to use modeling in conjunction with behavior to better understand how social decision-making develops. For example, one experiment (Hampton, Bossaerts, & O’Doherty, 2008) used computational modeling to provide insight into the process of mentalizing about another player’s strategy in a game known as the Inspection Game. Additionally, Apesteguia, Huck, and Oechssler (2007) demonstrated that when given the opportunity to view other player’s behavior in a game, people will often imitate the strategy that provides the highest payoff. Greater discrepancies between an individual’s payoff and another player’s payoff result in an increased likelihood of switching to the other strategy.

Finally, a few recent studies have utilized computational approaches to study how social advice can impact learning (Biele, Rieskamp, & Gonzalez, 2009; Doll, Jacobs, Sanfey, & Frank, 2009). In these studies, prior to a standard learning task, participants are given information (termed as advice or instructions) from either another participant or from the experimenter about the optimal choice. These experiments have found evidence supporting the notion that social information leads to learning biases, namely that accurate information helps participants learn better, while inaccurate information impairs learning. Biele et al. (2009) found support for a model that assigned greater weight to outcomes consistent with the advice than to the same outcomes on unadvised choices. Doll et al. (2009) found that the best fit of the behavioral data was produced by a model that initialized constructed stimuli to a higher than normal starting value and reduced the impact of instruction inconsistent outcomes while increasing the impact of instruction consistent outcomes. These studies suggest that explicit information such as advice or moral information can impact not only initial expectations but also how people learn from feedback. Information consistent with the prior information is weighted higher in the value update, and information inconsistent with the advice is weighted lower.

However, no study to date has examined how implicit information impacts learning in an interactive social decision scenario.

The present study adapted the design of van ‘t Wout and Sanfey (2008) to examine how implicit initial trustworthiness information (i.e. facial features) interacts with experienced trustworthiness (i.e. the probability of reciprocation) in a repeated Trust Game. First, we expected to replicate our previous finding that facial trustworthiness influences initial financial risk-taking in a social context (van ‘t Wout & Sanfey, 2008). Second, we expected to replicate other work, which has demonstrated that previous experiences also influence behavior (Axelrod & Hamilton, 1981; King-Casas et al., 2005). Finally, and most importantly, we predicted that these two processes, facial trustworthiness and experienced trustworthiness, would interact such that partners that both look trustworthy and reciprocate frequently will be entrusted with the most money. To increase our construct validity, we employed multiple measurements of trustworthiness, which included behavior in the Trust Game as well as subjective ratings. To further characterize our behavioral findings, we used RL models to test three distinct processing hypotheses – (1) initialization, (2) confirmation bias, and (3) dynamic belief. The Initialization models (GL initialization & trust decay) posit that the implicit trustworthiness judgments influence behavior at the beginning of the game, but are eventually overridden by the player’s actual experiences (i.e. whether or not trust is reciprocated). The Confirmation Bias model proposes
that initial implicit trustworthiness judgments influence the way feedback (i.e. non-reciprocated trust) is updated throughout the interactions (Biele et al., 2009; Delgado et al., 2005; Doll et al., 2009). This model assumes that learning is biased in the direction of the initial impressions. Finally, the Dynamic Belief model proposes that the facial trustworthiness judgment serves as an initial trustworthiness belief, which is continuously updated based on the player’s experience in the game. These beliefs, in turn, influence learning. This model equally emphasizes the initial judgment and experience and predicts that players will learn to give more money to partners that are trustworthy, and less money to partners that betray trust. By explicitly formalizing the potential mechanisms via these models, this study can increase our understanding of how trust is placed in social economic exchanges.

2. Methods

2.1. Participants

Sixty-four undergraduates were recruited from the psychology participant pool at the University of Arizona and received course credit for their participation in the experiment. Three participants were excluded after indicating during debriefing that they did not understand the experiment, leaving a total of 61 participants (mean age = 18.67, sd = 1.38, female = 79%). All participants gave informed consent, and the study was approved by the local Institutional Review Board.

2.2. Trust Game

Participants played a repeated Trust Game in the role of Player A as described in the introduction. We employed a 2 × 2 within-subjects design in which partner’s level of facial trustworthiness (high or low) was crossed with partner’s level of experienced trustworthiness (high or low). Each Player B represented one of the experimental conditions. Level of facial trustworthiness was assessed via independent ratings of partner photographs (see below). We defined level of experienced trustworthiness as a high (80%) or low (20%) probability of reciprocating an offer. Money invested by the participant was multiplied by a factor of 4. If an offer was reciprocated, Player B always reciprocated 50% of the total multiplied amount sent by Player A. When trust was not reciprocated, Player B did not return any of the multiplied amount of money back to Player A. Participants played 15 randomly ordered interspersed rounds with each partner (60 trials total, plus 30 slot machine gambles). On each trial, participants were endowed with $10. Each trial lasted 16 s and began with a short fixation cross (1000 ms) followed by a picture of the partner (3500 ms). Participants then decided how much money they wanted to invest by scrolling through offers that randomly increased or decreased in $1 increments. After submitting their investment, participants were shown the amount of money they chose to invest (multiplied by 4). Player B’s decision to either keep all of the money or reciprocate was then revealed along with a summary of the payoffs to each player. If the participant did not submit an offer in time (8000 ms), they forfeited all of their money for the round. As a non-social control, participants also had an opportunity to “gamble” with two different slot machines. Just like the Trust Game trials, participants were allowed to invest any amount of their $10 endowment in $1 increments. One slot machine paid out twice the investment with an 80% probability and the other slot machine paid out at a 20% probability. Thus, the slot machine trials were identical to the Trust Game trials except that they offered a non-social context.

2.3. Stimuli

The stimuli for the human partners were selected from the Winston stimuli set (Adolphs, Tranel, & Damasio, 1998) based on trustworthiness ratings from a previous study (van ‘t Wout & Sanfey, 2008). Two sets of four faces were selected that were matched for trustworthiness and attractiveness. Each set consisted of one male and one female that were previously rated high on trustworthiness (mean = 4.10) and low on trustworthiness (mean = 3.10) on a 7-point Likert scale.
Participants were randomly assigned to play with one of the picture sets, with the other set serving as a control. For each participant the high and low trustworthy pictures were randomly assigned to an experienced trustworthiness condition (high vs. low). The four cells were balanced across subjects, $\chi^2(3) = 0.56, p = 0.91$. This ensured that any observed effects could not be attributable to a single picture. At the end of the game participants rated both sets of stimuli – those they played against and those they did not – on trustworthiness, attractiveness, competence, aggressiveness, and likeability using a 5-point Likert scale. Participants were also asked to estimate the percentage of the time that each of the players they played with reciprocated their offers, from 0% to 100% in 10% increments. All stimuli were presented on a laptop via EPrime software (Psychology Software Tools, Inc., Pittsburgh, PA).

3. Results

3.1. Behavioral data

Overall, as expected, we found a main effect of reciprocity, where participants gave more money overall to partners who reciprocated 80% of the time (mean = 5.64, se = 0.18) as compared to partners who only reciprocated 20% of the time (mean = 3.42, se = 0.20) using a repeated measures ANOVA $F(1, 60) = 125.70, p < 0.001, \eta^2 = 0.68$. There was a trend for partner type that approached significance, $F(2, 120) = 2.41, p = 0.09$ where participants tended to invest more money in participants who looked more trustworthy (mean = 4.76, se = 0.17) as compared to both those who looked untrustworthy (mean = 4.44, se = 0.17) and computer controls (mean = 4.38, se = 0.23). There was a significant partner by reciprocity interaction $F(2, 120) = 5.34, p = 0.006, \eta^2 = 0.08$, such that participants invested the most money in high trustworthy partners that reciprocated 80% of the time (mean = 6.11, se = 0.23) and the least amount of money in low trustworthy partners that reciprocated 20% of the time (mean = 3.26, se = 0.22) (see Fig. 1).

Additionally, we successfully replicated our previous finding (van’t Wout & Sanfey, 2008), in which facial trustworthiness impacted the amount of money invested on the first trial of each pairing, $F(2, 120) = 3.22, p = 0.04, \eta^2 = 0.05$. In their first interaction, participants invested significantly more money in high trustworthy looking partners (mean = 5.1, se = 0.23) as compared to low trustworthy looking partners (mean = 4.37, se = 0.23), $p = 0.03$ (see Fig. 2, Panel A).

Finally, the interaction that we observed in our main behavioral results was not completely driven by a strong effect of facial trustworthiness at the beginning of the experiment. We observed a significant interaction between facial trustworthiness and probability of reciprocation on the last trial of the experiment $F(2, 120) = 4.13, p = 0.02, \eta^2 = 0.06$, suggesting that the effect persists throughout all 15 trials (see Fig. 2, Panel B). These data were log transformed to account for negative skew in the data.

3.2. Post-experiment ratings

As a manipulation check to ensure that the pictures used in each condition were viewed appropriately, we compared all of the ratings for each participant’s control picture set. These picture sets were counterbalanced across participants, so that control sets served as the experimental pictures for other participants. Mixed effects regression revealed that participants rated those partners that were selected to look more trustworthy as actually being more trustworthy $b = 0.44$ (se = 0.17), $t = 2.66, p < 0.05$, more attractive $b = 0.60$ (se = 0.11), $t = 5.25, p < 0.05$, more likeable $b = 0.56$ (se = 0.14), $t = 3.90, p < 0.05$, more competent $b = 0.39$ (se = 0.12), $t = 3.12, p < 0.05$, and less aggressive $b = -0.83$ (se = 0.13), $t = -6.20, p < 0.05$ then those that were selected to look untrustworthy.

To determine the effect of the repeated interaction on post-game trustworthiness judgments, we examined the main effects of the initial trustworthiness judgment, the probability of reciprocation, and their interaction. All predictions were supported (see Fig. 3). There was a significant main effect of partner type, $F(2, 120) = 7.05, p < 0.001, \eta^2 = 0.10$. Partners that were selected to look more trustworthy based on ratings from an independent sample (mean = 3.10, se = 0.09) were rated as more trustworthy as compared to partners that were selected to look untrustworthy (mean = 2.66,
Partners that reciprocated more frequently (mean = 3.61, se = 0.07) were rated as more trustworthy than those that reciprocated infrequently (mean = 2.18, se = 0.08), F(1, 60) = 183.16, p < 0.001, $\eta^2 = 0.75$. Finally, there was a significant interaction between initial trustworthiness and probability of reciprocation, F(2, 120) = 8.58, p < 0.001, $\eta^2 = 0.13$, such that partners that both looked trustworthy and reciprocated frequently were rated the most trustworthy (mean = 4.02, se = 0.11) and partners that both looked untrustworthy and reciprocated infrequently were rated the least trustworthy (mean = 1.93, se = 0.12). Also, it is interesting to note that partners that looked trustworthy and reciprocated frequently were rated as more trustworthy (mean = 4.02, se = 0.11) in comparison to both novel faces matched for trustworthiness (mean = 3.43, se = 0.08) t(60) = 4.20, p < 0.001, and for untrustworthy partners that reciprocated at the same frequency (mean = 3.38, se = 0.14), t(60) = 3.35, p < 0.01. Finally, there was no significant difference in trustworthiness ratings between untrustworthy partners that reciprocated frequently (mean = 3.38, se = 0.14) as compared to trustworthy control pictures (mean = 3.43, se = 0.08), t(60) = −0.35, p > 0.05, suggesting that positive experiences can override initial negative impressions.

Participants were quite accurate in their estimation of their partners’ behavior, as gauged by the percentage of the time they believed the partners sent money back. Participants estimated that partners in the high probability condition (mean = 65.19, se = 1.55) reciprocated more often than partners in the low probability condition (mean = 28.08, se = 1.74) F(1, 60) = 186.09, p < 0.001, $\eta^2 = 0.76$. Participants did not differ in their probability estimation as a function of their partner’s trustworthiness F(2, 120) = 1.0, p > 0.05. However, there was a significant partner by probability interaction F(2, 120) = 3.47, p = 0.03, $\eta^2 = 0.06$, where high trustworthy looking partners (mean = 70.82, se = 2.3) were estimated to reciprocate more frequently than both low trustworthy partners (mean = 61.80, se = 3.03) t(60) = −2.57, p = 0.01 and computer controls (mean = 61.80, se = 3.03) t(60) = −2.57, p = 0.01. This indicates that participants explicitly judged the probability of reciprocation to be higher for partners that looked more trustworthy in comparison to other partners.
3.3. RL models

Our results demonstrate the notable effect that initial perceptions of trustworthiness interact with experience to influence both the amount of trust actually placed in a partner and the perceived judgments of trustworthiness revealed via participants’ post-experiment subjective ratings. However, these analyses cannot speak to how these two variables might be interacting. There are several plausible explanations for this effect. First, initial trustworthiness judgments might merely influence the starting trust value, but not how participants update their beliefs after feedback. Second, it is possible that the initial expectations influence how feedback is interpreted, where feedback that is consistent with the expectation (e.g., cooperation by partners that are perceived to be trustworthy) is weighted more heavily than feedback that is inconsistent with the expectation (e.g., cooperation by partners that are perceived to be untrustworthy). Third, it is possible that initial expectations influence how
feedback is interpreted, but the feedback can also in turn influence the expectations. For example, partners that repeatedly violate trust are eventually perceived to be untrustworthy, which means that feedback suggesting otherwise (i.e. cooperation) will be less likely to change participants’ behavior. In an effort to further characterize our behavioral results, we employed a class of RL models referred to as value learning (Sutton & Barto, 1998) to test these different computational accounts. These models attempt to calculate how an individual learns the value \( V \) of a stimulus \( S \) for a given trial \( t \) using a minimal number of parameters. In our approach, value refers not to the probability of making a discrete decision, but to the actual value of a given partner’s trustworthiness (in dollars). We utilized a cross-validation procedure that is robust to differences in model complexity to compare three different processing hypotheses (initialization, confirmation bias, and dynamic belief) against a baseline model.

3.4. Gain Loss model (baseline)

Considerable evidence supporting Prospect Theory (Kahneman & Tversky, 1979) has demonstrated that people prefer avoiding losses as compared to acquiring gains of the same magnitude. Given these findings, we use a model that differentially updates gains and losses via separate learning rates as a baseline model (Doll et al., 2009; Frank, Moustafa, Haughey, Curran, & Hutchison, 2007; Yechiam, Busemeyer, Stout, & Bechara, 2005). This model computes a predicted value for the next trial for each stimulus based on the experienced outcome:

\[
V_s(t + 1) = V_s(t) + \alpha_G \delta^+ + \alpha_L \delta^-
\]

where \( \alpha_G \) is the amount that a positive outcome (notated by the + superscript) is weighted and \( \alpha_L \) is the amount that a negative outcome (notated by the − superscript) is weighted in the update \((0 < \alpha < 1)\). This allows people to learn from losses differently than gains, which we think is particularly important due to the social nature of the task. For example, if a participant believes their goodwill has been violated, they will likely adapt their behavior quickly (Bohnet & Zeckhauser, 2004). We chose to set the initial value \( V_s(1) \) for all conditions to the average amount sent by the participants on the first trial of the game (mean = 14.52). This value is calculated by inputting an initial investment of $4.52 into Eq. (9).

3.5. Gain Loss Initialization model

We used two different models to test the initialization hypothesis (see Trust Decay model below for alternative approach). This model initialized the starting values of the baseline GL model based on the
trustworthiness ratings from an independent sample. The values for high trustworthiness ($T_{HT} = 3.6$) and low trustworthiness ($T_{LT} = 2.5$) were scaled by a free parameter $\lambda$ when $t = 1$, where $0 < \lambda < 20$.

$$V_S(t) = \begin{cases} 14.52 + \lambda \cdot T_{HT} \\ 14.52 - \lambda \cdot T_{LT} \end{cases}$$

This model predicts that the perceived facial trustworthiness will only influence the initial expectations and will have no bearing on the update rule.

### 3.6. Confirmation Bias model

Previous research has examined the effect of explicit information on decision-making (Biele et al., 2009; Doll et al., 2009). These studies have proposed models that give a higher weight to feedback that is consistent with the advice, and lower to feedback inconsistent with the advice. To examine this hypothesis we will test a model that is similar to Doll et al.'s (2009) instructed learning model and Biele et al.'s (2009) outcome bonus model, formally defined as

$$V_S(t + 1) = V_S(t) + \alpha_1 \delta^+ + \alpha_2 \delta^- + \phi[T_S]^{\text{replay}} - \phi[1 - T_S]^{\text{abuse}}$$

where $0 < \phi < 10$. In this model, participants receive a fixed bonus that is proportional to their partner’s level of facial trustworthiness $T_S$ scaled by a free parameter $\phi$, when their partner reciprocates their investment (denoted by the ‘replay’ superscript). If their partner fails to reciprocate, this amount serves as a deduction (denoted by the ‘abuse’ superscript). This means that higher levels of facial trustworthiness will promote quicker learning from gains, while lower facial trustworthiness will facilitate greater learning from losses. Facial trustworthiness was determined by ratings from an independent sample that were transformed to range between 0 and 1 ($T_{HT} = 0.72$ for trustworthy faces; $T_{LT} = 0.5$ for untrustworthy faces). This model tests the confirmation bias hypothesis, and predicts that initial facial trustworthiness judgments will influence how feedback is interpreted consistently over the course of the experiment.

### 3.7. Trust Decay model (initialization)

An alternative hypothesis is that facial trustworthiness judgments influence initial behavior, but then become less important with increased experience with a partner. This is a different test of the initialization hypothesis from the Gain Loss Initialization model because it predicts that facial trustworthiness will influence the update early on and not just start at a higher value. This model decreases the influence of the trustworthiness bonus as a function of time by $\rho$ and is formally defined as

$$V_S(t + 1) = V_S(t) + \alpha_1 \delta^+ - \alpha_2 \delta^- + \frac{e^{-\rho t}[T_S]^{\text{replay}} - e^{-\rho t}[1 - T_S]^{\text{abuse}}}{\rho}$$

where $0 < \rho < 1$. This model is similar to the Confirmation Bias model initially, but exponentially decays the influence of the trustworthiness bonuses and deductions over time. This model tests the initialization hypothesis (also see Gain Loss Initialization model) and predicts that facial trustworthiness judgments will provide a preliminary estimate of an individual’s level of trustworthiness, but will eventually be overcome by experience – a prediction that has previously been framed from a multiple systems perspective (Frank et al., 2007).

### 3.8. Dynamic Belief model

The final hypothesis that we will test also treats trustworthiness as a bonus in the update function, but rather than being a fixed bonus based on the initial trustworthiness judgment like the Confirmation Bias and Trust Decay models, it adapts over time based on the perceived level of trustworthiness.

---

1. This model uses an additive bonus, as was employed by Biele et al. (2009). We were unable to get a multiplicative implementation of Doll et al.’s (2009) Instructed Learning Model to converge.
This means that the model can learn the level of trustworthiness $T$ for each partner $S$ and will use this information as a bonus or deduction in the update. This model is formally defined as

$$T_S(t+1) = T_S(t) + \phi \text{Outcome}(t) - T_S(t)$$

$$V_S(t+1) = V_S(t) + 2 \delta_t^+ + 2 \delta_t^- + \theta [T_S(t+1)^{\text{repay}} - T_S(t+1)^{\text{abuse}}]$$

where $\phi$ represents the trustworthiness learning rate, and $\theta$ is a free parameter used to scale the amount of influence of the trustworthiness bonus in the value update. Outcome = 1 if partner reciprocates trust or 0 if the partner abuses trust. This model can dynamically learn the trustworthiness of each partner and will add a bonus that is proportional to the level of perceived trustworthiness if the partner reciprocates, or alternatively will deduct a value proportional to the level of trustworthiness if the partner defects. This model differs from the Confirmation Bias model because defections by a partner with lower perceived trustworthiness result in smaller deductions. Another important conceptual distinction is that the model allows facial trustworthiness to influence the update like the other two bonus models, but it also allows feedback to influence the trustworthiness beliefs. Because this model learns the perceived trustworthiness of each partner, it can be used to predict each individual participants' subjective trustworthiness ratings that were measured at the end of the experiment. This model tests a processing hypothesis and predicts that perceived trustworthiness changes over time but is continually used to update the value associated with a given partner.

3.9. Model evaluation

To evaluate the models, we employed a cross-validation procedure in which the models were initially trained on half of the data and then subsequently tested on the other half. This approach substantially reduces over-fitting and also provides a useful way to compare models of differing complexity. During the training phase, models were fit to the participant’s actual behavioral data by minimizing the sum of the squared error (SSE) and parameters were estimated for the entire group. Models were compared using a metric that rewards the most parsimonious model. We then compared the models in their ability to predict the behavioral data out-of-sample using the parameters estimated during the training phase. This approach controls for models having different numbers of free parameters when fitting to behavioral data (Hampton & O’Doherty, 2007).

We calculated the reward for a given stimulus $s$ for trial $t$ using the following equation

$$r_s(t) = \left(10 - \frac{\text{Investment} + 4 \times \text{Investment}}{2}\right)$$

Investments were multiplied by a factor of 4 and were divided by 2 because they were split equally between both parties (if they were reciprocated). The participant’s reward takes into account the amount of money they kept plus the amount that was reciprocated, which allows the possibility for participants to still receive reward when there is negative prediction error (i.e. the partner did not reciprocate).

Parameters were estimated during the training phase by minimizing the SSE between the behavioral data and the predictions from the various models on every odd trial using fmincon (Coleman & Li, 1996), a multivariate constrained nonlinear optimization algorithm implemented in Matlab (Mathworks, Cambridge, MA).

$$\sum (r_s(t) - V_s(t))^2$$

The value $V$ for a given state $s$ at time $t+1$ is updated using the functions specified above. Because of the small number of trials for each condition ($n = 15$), the parameters were estimated for the entire group. This means that while the models were fit to each participant’s individual behavioral data, the error in the parameter estimation was pooled across subjects. This procedure has been previously utilized when individual parameter estimates are not stable as a result of a small number of trials and collinearity between parameters (Daw, O’Doherty, Dayan, Seymour, & Dolan, 2006). While we observe
a similar hierarchy of model fits when an individual parameter is fit for each participant, we report the more stable group fits. Multiple start locations were used to minimize the risk of the optimization routine getting stuck in local minima.

All models were compared to a baseline RL model (Eq. (3)), using the Akaike Information Criteria (AIC), a value that provides a metric of model fit by taking into account the complexity of the model (i.e. the number of estimated parameters). AIC rewards the most parsimonious model by penalizing for additional free parameters and is formally defined as

\[ \text{AIC} = -2 \ln(L) + 2k \]

where \( L \) is the maximum likelihood of the observed data and \( k \) is the number of estimated parameters. AIC performs a trade-off between model fit and model complexity, penalizing models with more parameters to avoid overfitting.

**Fig. 4.** RL model cross validation. Note: (A) this graph depicts the AIC fits from the training phase, in which each of the models were fit to the odd trials and parameters were estimated for the entire group. LG = baseline Gain Loss model (Eq. (3)) with two parameters, LG_Init = Gain Loss Initialization model (Eq. (4)) with three free parameters, CBias = confirmation bias model (Eq. (5)) with three parameters. TrustDecay = Trust Decay model (Eq. (6)) with three parameters. DynB = Dynamic Belief model (Eqs. (7) and (8)) has four parameters. (B) This graph depicts the AIC fits of the test phase in which the parameters estimated from the odd trials were used to predict the even trials. This out-of-sample test controls for the number of free parameters when comparing model fits because no parameters are actually estimated during this procedure.

It is important to note that our models are accounting for each individual participant’s trial-by-trial sequence of choices and outcomes despite estimating the model parameters for the group. As a demonstration, randomly shuffling the trial sequence for each participant using the Gain Loss Baseline model results in a dramatic decrease in model fit (Normal Sequence AIC = 25326.27; Random Sequence AIC = 26600.49).

We also calculated the Bayesian Information Criterion (Schwarz, 1978), which provides a slightly larger penalty for the number of free parameters and observed results consistent with the AIC (LG = 6859.94; LG_Init = 6871.57; Confirmation Bias = 6770.03; Trust Decay = 6779.08; Dynamic Belief = 6725.86).
AIC = 2k + n \left[ \ln \left( \frac{2nRSS}{n} \right) + 1 \right] \quad (11)

where $k$ is the number of free parameters, $n$ is the number of observations, and RSS is the residual sum of squares (Akaike, 1974).

To provide a more stringent model comparison procedure, we then tested the models out-of-sample by using the parameters estimated from the training phase and minimizing the SSE between the behavioral data and the model predictions on the even trials. Because no parameters are actually being estimated during this process, the model fits cannot be artificially inflated as a result of additional free parameters. While this type of procedure is often used in financial forecasting to predict the last 20% of the data, we chose to evenly sample throughout the trials to avoid unfairly penalizing our models that predict changes in the update function as a function of experience (e.g., trust decay and dynamic belief).

3.10. Modeling results

Overall, most of our models provided a better explanation of the data than the baseline GL model ($AIC = 21620.29; \text{RSS} = 18895.58, N = 5248$; see Fig. 4, Panel A). Both the Confirmation Bias model ($AIC = 21496.84; \text{RSS} = 18449.23, N = 5248$) and the Trust Decay model ($AIC = 21485.47; \text{RSS} = 18409.31, N = 5248$) provided a better fit of the data than the baseline GL model and the Dynamic Belief model ($AIC = 21460.57; \text{RSS} = 18315.19, N = 5248$) provided the best fit of all models tested. The one exception was the GL Initialization model, which did not appear to fit the data any better than the baseline model ($AIC = 21615.64; \text{RSS} = 18871.64, N = 5248$). These results suggest that facial trustworthiness judgments do not just merely influence the initial value, but rather seem to affect how feedback is interpreted. The Dynamic Belief model, which allows the initial expectation to influence the update function and also allows for the feedback to update the expectation, appeared to be the best account of the behavioral data. However, it is important to note that all of these models have a different number of free parameters, with the Dynamic Belief model having the most. While the AIC and BIC metrics are standard ways of penalizing for additional free parameters, a stronger test is to look out-of-sample.

We find a similar hierarchy of results in our out-of-sample prediction procedure (see Fig. 4, Panel B). The GL Initialization model ($AIC = 21764.75; \text{RSS} = 19437.74, N = 5248$) fits about the same, in fact slightly worse, than the baseline GL model ($AIC = 21753.11; \text{RSS} = 19394.69, N = 5248$). The Confirmation model ($AIC = 21663.21; \text{RSS} = 19065.27, N = 5248$) and the Trust Decay model ($AIC = 21672.25; \text{RSS} = 19098.15, N = 5248$) both fit better than the baseline model and the Dynamic Belief model ($AIC = 21619.04; \text{RSS} = 18905.48, N = 5248$) exhibits the best fit. These results provide additional evidence that initial beliefs about trustworthiness influence the update function, and that feedback in turn can update the trustworthiness beliefs. A simulation of the Dynamic Belief model using the parameters estimated from the training phase can be seen in Fig. 5.

Consistent with our prediction, participants appeared to adapt their behavior more radically when their partner defected (mean $\alpha_L$ across models = 0.35) as compared to when their partner cooperated (mean $\alpha_G$ across models = 0.02). In addition, there seems to be little evidence that high initial values can explain our behavioral results. The Gain Loss Initialization model which specifically manipulated the starting value based on the trustworthiness ratings did not explain the data any better than the baseline model. Also, the Trust Decay model, which only uses the trustworthiness bonuses early on, did not explain the data any better than the Confirmation Bias model. In fact, the parameter that was estimated for the decay, $\phi$, was essentially 0 indicating that the model reduced to the Confirmation Bias model with a $\phi$ of 1. A summary of the estimated parameters for each model can be seen in Table 1.

3.11. Predicted trustworthiness ratings

While the Dynamic Belief model appeared to provide the best explanation of the behavioral data, it is also possible to examine how well it can capture subjective perceptions of trustworthiness. The
model makes specific predictions about how trustworthiness beliefs should adapt based on experiences in the game. In this analysis we averaged the last five trials of the Trustworthiness Beliefs \( T_S \) derived from the learning model and converted them back into ratings (multiplied them by 5) to predict the participants' post-experiment subjective trustworthiness ratings (see Fig. 6). This analysis is useful because it provides a method to test the predictive validity of the Dynamic Belief aspect of the model using participants’ actual trustworthiness ratings. To evaluate whether the ratings produced by the model were better than the initial normed trustworthiness ratings at predicting the actual post-task participant ratings, we used a Williams’s T2 statistic (Steiger, 1980) to compare
the magnitude of the standardized beta values derived from a mixed effects regression (Baayen, Davidson, & Bates, 2008). The regression was performed on data that was z-transformed and allowed participants' intercepts to vary randomly. Consistent with our hypothesis, this analysis found that the trust ratings predicted from the model $\beta = 0.62$ (se = 0.05) were a better account of the participant's actual post-experiment trust ratings than were the normative trust ratings $\beta = 0.13$ (se = 0.05), $t(57) = 3.15$, $p < 0.05$. This effect appeared to be specific to trustworthiness, as the model derived ratings did not predict other ratings better than the normative trust ratings such as attractiveness $t(57) = 0.47$, ns, aggressiveness $t(57) = 0.94$, ns, competence $t(57) = 0.66$, ns, and likeability $t(57) = 1.85$, ns. In addition, the model predicted the participants' reported probability of reciprocation better than the normed ratings, $t(57) = 3.94$, $p < 0.05$, indicating that the model-predicted ratings captured subjective perceptions of the experienced probabilities. Thus, the Dynamic Belief model appears to not only predict participant’s behavior in the game, but it can also predict how their perceptions of trustworthiness change after interacting with a partner.

4. Discussion

This study investigated the processes underlying the decision to trust (or not trust) a partner in a consequential interaction. Previous research has reported that both initial impressions (Delgado et al., 2005; van ‘t Wout & Sanfey, 2008) and direct experience (King-Casas et al., 2005; Singer et al., 2004) play important roles in influencing judgments of trustworthiness. This experiment provides the first account of how these variables interact in a social interactive financial investment game that has been explicitly designed to study trust (Berg et al., 1995). Consistent with our hypotheses, both the initial trustworthiness judgment of a partner as well as subsequent experience with that partner synergistically influence behavior in this game, in terms of how much money players were willing to entrust to their partner.

4.1. Behavioral measures of trust

Consistent with our group’s previous finding, we found that facial trustworthiness influenced participant’s initial investment amount (van ’t Wout & Sanfey, 2008). On the first round, participants invested more money if their partner looked trustworthy than if the partner looked untrustworthy. This provides further support to the notion that social signals can be conveyed through facial expressions (Oosterhof & Todorov, 2008), and that participants are sensitive to differences in perceived
trustworthiness. In our study, it appears that participants believe trustworthy faces predict a higher probability of reciprocation, and therefore facial trustworthiness may serve as a risk signal which influences the amount of money an individual expects to be sent back. However, there is likely something unique to the social nature of this signal as it is able to be selectively manipulated (compared to pure risk) using a hormone induction (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005). This hormone, known as oxytocin, acts as a neurotransmitter in the brain and is likely mediating the effect on trust via the amygdala (Baumgartner, Heinrichs, Vonlanthen, Fischbacher, & Fehr, 2008). We also replicated findings which indicate that people use experience as a basis for their trustworthiness judgments (King-Casas et al., 2005). Over time, participants in our study learned to invest more money in partners that reciprocated frequently, and less money in partners that reciprocated infrequently. Together these results suggest that investment behavior in the Trust Game is influenced by both implicit social signals, revealed here through facial trustworthiness, and also by direct social signals conveyed via experience in the game. Finally, these variables appear to act synergistically to influence behavior. Partners that were initially viewed as more trustworthy, and actually turned out to be more trustworthy, were entrusted with the most money in the game, indicating the twin influences of both first impressions and experience.

One potential limitation to our study is that we did not ask whether participants believed they were playing a “real person” at the conclusion of the study. It is unlikely that many participants approached this task believing the person they saw was an active participant due to the rather dated nature of the photographs. However, people do appear to easily anthropomorphize when stimuli behave like real people (Blakemore & Decety, 2001), and more pointedly, the pattern of behavior that we observed is similar to all other reported versions of the repeated Trust Game (Delgado et al., 2005; King-Casas et al., 2005; Krueger et al., 2007). We interpret this to mean that people approached the task at least “as if” they were playing with a real life partner.

4.2. Subjective ratings of trust

Trustworthiness judgments were assessed by participants’ behavior towards their partners in the game, and also by their subjective ratings of these partners at the conclusion of the game. Consistent with previous research, we found that post-task subjective trustworthiness ratings were influenced not only by the facial appearance of the partner (van’t Wout & Sanfey, 2008; Winston, Strange, O’Doherty, & Dolan, 2002), but also by the partner’s behavior in the game (Delgado et al., 2005; Singer et al., 2004). Partners that reciprocated more frequently were rated as more trustworthy than partners that reciprocated infrequently. We also observed a significant appearance by behavior interaction, where partners that both looked and behaved trustworthy were rated as the most trustworthy of all. These subjective ratings perfectly mirror our behavioral measures of trust.

Several interesting phenomena emerged from examining the post-game ratings of partner trustworthiness. Firstly, partners that looked untrustworthy, but behaved trustworthy, were rated at the same level of trustworthiness as were the control trustworthy faces (with whom participants did not play). Similarly, those partners that were initially viewed as trustworthy, but who behaved in an untrustworthy fashion (i.e. reciprocated trust infrequently), were rated as similar in trustworthiness as the control untrustworthy faces in the post-game subjective ratings (see Fig. 3). This suggests that these fast automatic judgments of trustworthiness can be overridden by experience, even relatively minimal experience. This finding may have important implications for social psychologists interested in stereotype and prejudice. For example, Cunningham et al. (2004) has demonstrated that black faces presented quickly are associated with increased amygdala activation compared to white faces and has suggested that this effect is related to implicit levels of racial bias. However, this effect seems to disappear when the faces are presented for longer lengths of time, presumably when controlled processing can override this automatic evaluation. Our results suggest that repeated positive interactions may also be able to reshape these automatic evaluations. It is important to note, however, that we still observed a significant difference between high trustworthy high reciprocators and low trustworthy high reciprocators, indicating that not all pre-existing judgments were erased by experience. In addition to perceived trustworthiness, we also found that participants believed that the high trustworthy high reciprocators reciprocated at a higher probability compared to the low trustworthy...
high reciprocators and the high probability computer controls. Together, these findings indicate that the initial impressions interact with experience when they are congruent to influence both cognition and behavior.

4.3. Modeling trust

To better understand the individual learning processes underlying our behavioral findings, we modeled three possible learning processes. While there are of course many models that could have been tested, we chose to focus on four that have a strong conceptual grounding. First, we tested a pure initialization hypothesis by manipulating the initial starting value based on the normative trust ratings in the GL Initialization model. Next, we tested models that allow the trustworthiness beliefs to influence how feedback is interpreted. The Confirmation Bias model proposes that information consistent with the initial trustworthiness judgment will receive a learning bonus, while inconsistent information (i.e. a high trustworthy face defecting) will be ignored (Biele et al., 2009; Delgado et al., 2005; Doll et al., 2009). The Trust Decay model proposes that initial trustworthiness judgments will influence behavior early in the interactions, but will eventually be overridden by experience, a prediction akin to a dual process model (Frank, O’Reilly, & Curran, 2006; Frank et al., 2007; Poldrack et al., 2001). Finally, the Dynamic Belief model proposes that trustworthiness beliefs consistently serve as a learning bonus in the update function, but are themselves updated based on their partners’ behavior after each interaction.

Our cross-validation procedure revealed that a pure initialization account could not explain the data as well as models that allowed the initial beliefs to influence the actual update function. Supporting this finding, we find that our behavioral interaction is not driven by early trials, but rather that it persists until the last trial of the learning sequence (see Fig. 2, Panel B). In addition, our model which allows a learning bonus to only influence the update function during early trials essentially reduced to a constant bonus model, providing additional evidence that this process is not something that is overridden by experience.

Further, we find that a model that allows trustworthiness beliefs to both influence how feedback is interpreted and allow the beliefs to be updated based on this feedback provides the best explanation of the data. Importantly, this model makes two specific predictions that are supported by our data. First, as illustrated in our simulation (Fig. 5), the Dynamic Belief model predicts that by the end of the experiment participants should be investing the most money in trustworthy looking partners that reciprocate frequently and the least amount of money in trustworthy looking partners that reciprocate infrequently. We observed support for this prediction in the last trial of our behavioral data (Fig. 2, Panel B) and moreover observe the exact pattern of behavior for all of the conditions predicted by our model. Second, our Dynamic Belief model predicts that perceptions of trustworthiness will change over time based on actual experiences. This prediction was also supported, as the model-predicted trustworthiness ratings were able to better predict the post-experiment subjective trustworthiness ratings than the initial perceptions of trustworthiness (i.e. the normative trust ratings). Because the model-predicted ratings update based on experience, they were also better able to predict the post-experiment probability ratings than the initial trustworthiness ratings. These findings support and extend previous research, which has found that experience overwhelms description in risky choice (Barron & Erev, 2003; Hertwig, Barron, Weber, & Erev, 2004; Jessup, Bishara, & Busemeyer, 2008). While this notion of dynamically updating beliefs is certainly not new, this study provides, to our knowledge, support for the first computational model of this effect in an iterative social exchange.

5. Conclusion

Our study integrates theories and methods from psychology, economics, and reinforcement learning to gain a greater understanding of how high-level social cues such as trustworthiness are acquired and utilized in a consequential social decision. The findings suggest that trustworthiness judgments may serve as a risk belief (i.e. probability of reciprocation). This belief is based on initial judgments...
of perceived trustworthiness and is dynamically updated based on experiences through repeated interactions. This study illustrates the conceptual and methodological advantages of an interdisciplinary approach and provides a novel quantitative framework to conceptualize the notion of trustworthiness as well as an approach to bridge the division between descriptive information and experienced information in the judgment and decision-making literature (Jessup et al., 2008). More broadly, our study provides an important and timely contribution to a growing literature interested in the neural computations underlying social learning (Behrens et al., 2008; Biele et al., 2009; Delgado et al., 2005; Hampton, Bossaerts, & O’Doherty, 2008; King-Casas et al., 2005; Olsson & Phelps, 2007) and trustworthiness (Krueger et al., 2007; Osterhof & Todorov, 2008; van ‘t Wout & Sanfey, 2008; Winston, Strange, O’Doherty, & Dolan, 2002) and illustrates the importance of social beliefs in decision-making behavior.

Acknowledgments

The authors thank Niko Warner and Carly Furgersen for their help in the collection of the data and Dr. Mike X. Cohen and three anonymous reviewers for their helpful comments.

References


APPENDIX B: TRIANGULATING THE NEURAL, PSYCHOLOGICAL, AND ECONOMIC BASES OF GUILT AVERSION (STUDY 2)

Triangulating the neural, psychological, and economic bases of guilt aversion

Luke J. Chang\textsuperscript{1}  
Alec Smith\textsuperscript{2}  
Martin Dufwenberg\textsuperscript{2}  
Alan G. Sanfey\textsuperscript{1,3,4,*}

\textsuperscript{1}Department of Psychology  
\textsuperscript{2}Department of Economics  
University of Arizona  
1503 E. University Blvd  
Tucson, AZ 85721  
\textsuperscript{3}Donders Institute for Brain, Mind \& Behavior  
\textsuperscript{4}Behavioral Science Institute  
Radboud University Nijmegen  
6525EN Nijmegen  
The Netherlands

*Corresponding Author:  
Phone: 520.621.1477  
Fax: 520.621.9306  
Email: 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
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 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 \((E_2E_1;S_2)\),
can be used as a proxy.

\[
U_2 = M_2 - \Theta_1(E_2E_1;S_2 - S_2)^+ 
\]

(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 \((\Theta_{12})\), 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 \((E_1 S_2)\). 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 \((E_2 E_1 S_2)\). 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

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 than 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.
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 dIPFC 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 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.
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 fsloroi 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 \( \theta_{12} \) to vary for every Investor/Trustee interaction. There are two possible maxima of the utility function depending on \( \theta_{12} \). 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 \( \Theta_{12} \).

**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


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 acknowledge 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 to 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 1\textsuperscript{st} order belief ($E_1S_2$) by the Trustee’s 2\textsuperscript{nd} order belief ($E_2E_1S_2$). B) The amount returned by the Trustee ($S_2$) by their 2\textsuperscript{nd} 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.
Figure 3.

A. 

B. 

C. 

Investor’s 1st Order Belief (E1S2) vs. Trustee’s 2nd Order Belief (E2E1S2)

Amount Returned (S2) vs. Trustee’s 2nd Order Belief (E2E1S2)

Counterfactual Guilt vs. Difference from Behavior
Figure 5.
APPENDIX C: TRIANGULATING THE NEURAL, PSYCHOLOGICAL, AND ECONOMIC BASES OF GUILT AVersion – SUPPLEMENTARY INFORMATION (STUDY 2)
Supplemental Information

Figures

Figure S1
Figure S2
Figure S3
Figure Legends

**Figure S1 related to Equation 1.** Simulation of Guilt-Aversion Model
The guilt-aversion model makes two behavioral predictions depending on $\Theta_{12}$. 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 $\Theta_{12}$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 2\textsuperscript{nd} 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.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Parameter Estimate</th>
<th>SE</th>
<th>t - Value</th>
<th>p - Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.09</td>
<td>0.03</td>
<td>3.54</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Offer Amount</td>
<td>0.17</td>
<td>0.03</td>
<td>6.82</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>E2E1S2</td>
<td>0.46</td>
<td>0.04</td>
<td>10.83</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>
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.

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

<table>
<thead>
<tr>
<th>Hemisphere</th>
<th>Region</th>
<th>BA</th>
<th>Z Value</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less &gt; Equal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Caudate</td>
<td>25</td>
<td>3.08</td>
<td>-10</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>L</td>
<td>Lateral OFC Cortex</td>
<td>47</td>
<td>4.06</td>
<td>-36</td>
<td>32</td>
<td>-18</td>
</tr>
<tr>
<td>L</td>
<td>Middle Frontal Gyrus</td>
<td>9</td>
<td>3.52</td>
<td>-28</td>
<td>24</td>
<td>42</td>
</tr>
<tr>
<td>L</td>
<td>Nucleus Accumbens</td>
<td>25</td>
<td>3.3</td>
<td>-6</td>
<td>10</td>
<td>-8</td>
</tr>
<tr>
<td>L</td>
<td>Paracingulate Gyrus</td>
<td>9</td>
<td>3.7</td>
<td>-6</td>
<td>54</td>
<td>6</td>
</tr>
<tr>
<td>L</td>
<td>Rostral ACC</td>
<td>32</td>
<td>3.18</td>
<td>-10</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>L</td>
<td>Sub Genual ACC</td>
<td>25</td>
<td>3.54</td>
<td>-4</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>L</td>
<td>Superior Frontal Gyrus (DMPFC)</td>
<td>9</td>
<td>4.11</td>
<td>-12</td>
<td>46</td>
<td>50</td>
</tr>
<tr>
<td>L</td>
<td>Superior Frontal Gyrus (DMPFC)</td>
<td>8</td>
<td>3.69</td>
<td>-20</td>
<td>26</td>
<td>58</td>
</tr>
<tr>
<td>L</td>
<td>Superior Frontal Gyrus (DMPFC)</td>
<td>10</td>
<td>3.92</td>
<td>-12</td>
<td>64</td>
<td>24</td>
</tr>
<tr>
<td>L</td>
<td>Temporal Pole</td>
<td>38</td>
<td>3.51</td>
<td>-42</td>
<td>20</td>
<td>-30</td>
</tr>
<tr>
<td>R</td>
<td>ACC</td>
<td>11</td>
<td>3.38</td>
<td>0</td>
<td>30</td>
<td>-6</td>
</tr>
<tr>
<td>R</td>
<td>Caudate</td>
<td>25</td>
<td>3.19</td>
<td>12</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>R</td>
<td>Lateral OFC Cortex</td>
<td>11</td>
<td>3.25</td>
<td>22</td>
<td>36</td>
<td>-16</td>
</tr>
<tr>
<td>R</td>
<td>Medial OFC Cortex</td>
<td>11</td>
<td>3.51</td>
<td>4</td>
<td>44</td>
<td>4</td>
</tr>
<tr>
<td>R</td>
<td>Nucleus Accumbens</td>
<td>25</td>
<td>3.08</td>
<td>8</td>
<td>16</td>
<td>-4</td>
</tr>
<tr>
<td>R</td>
<td>Paracingulate Gyrus</td>
<td>10</td>
<td>4.05</td>
<td>14</td>
<td>46</td>
<td>0</td>
</tr>
<tr>
<td>R</td>
<td>Posterior Insula</td>
<td>48</td>
<td>3.60</td>
<td>40</td>
<td>0</td>
<td>-14</td>
</tr>
<tr>
<td>R</td>
<td>Rostral ACC</td>
<td>11</td>
<td>3.21</td>
<td>8</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td>R</td>
<td>Superior Frontal Gyrus (DMPFC)</td>
<td>10</td>
<td>3.94</td>
<td>2</td>
<td>56</td>
<td>32</td>
</tr>
<tr>
<td>R</td>
<td>Temporal Pole</td>
<td>28</td>
<td>3.35</td>
<td>26</td>
<td>10</td>
<td>-26</td>
</tr>
<tr>
<td>R</td>
<td>Temporal Pole</td>
<td>38</td>
<td>3.24</td>
<td>42</td>
<td>20</td>
<td>-20</td>
</tr>
<tr>
<td>Equal &gt; Less</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Angular Gyrus</td>
<td>19</td>
<td>3.45</td>
<td>-44</td>
<td>-56</td>
<td>44</td>
</tr>
<tr>
<td>L</td>
<td>Cerebellum (Crus I)</td>
<td>NA</td>
<td>3.61</td>
<td>-40</td>
<td>-74</td>
<td>-26</td>
</tr>
<tr>
<td>L</td>
<td>Dorsal ACC</td>
<td>32</td>
<td>4.02</td>
<td>-8</td>
<td>14</td>
<td>38</td>
</tr>
<tr>
<td>L</td>
<td>Fusiform</td>
<td>19</td>
<td>3.57</td>
<td>-44</td>
<td>-70</td>
<td>-20</td>
</tr>
<tr>
<td>L</td>
<td>Lateral Occipital Cortex, Inferior Division</td>
<td>37</td>
<td>3.91</td>
<td>-58</td>
<td>-64</td>
<td>8</td>
</tr>
<tr>
<td>L</td>
<td>Lateral Occipital Cortex, Inferior Division</td>
<td>19</td>
<td>3.81</td>
<td>-52</td>
<td>-74</td>
<td>-14</td>
</tr>
<tr>
<td>L</td>
<td>Middle Frontal Gyrus (DLPFC)</td>
<td>45</td>
<td>3.88</td>
<td>-44</td>
<td>28</td>
<td>24</td>
</tr>
<tr>
<td>L</td>
<td>Middle Frontal Gyrus (DLPFC)</td>
<td>46</td>
<td>3.08</td>
<td>-32</td>
<td>40</td>
<td>24</td>
</tr>
<tr>
<td>Side</td>
<td>Region</td>
<td>MNI Coordinates</td>
<td>T-value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------------------------------------</td>
<td>-----------------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Postcentral Gyrus</td>
<td>40 -38 -34 40</td>
<td>4.74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Precentral Gyrus</td>
<td>6 -34 -6 48</td>
<td>4.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Precentral Gyrus</td>
<td>4 -58 0 30</td>
<td>4.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>SMA</td>
<td>NA -4 -2 48</td>
<td>3.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Superior Parietal Lobule</td>
<td>40 -34 -52 58</td>
<td>4.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Supramarginal Gyrus (TPJ)</td>
<td>2 -58 -28 44</td>
<td>3.88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Central Opercular Cortex</td>
<td>48 48 -2 10</td>
<td>3.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Cerebellum (I-IV)</td>
<td>NA 4 -52 -16</td>
<td>3.71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Cerebellum (Right Crus I)</td>
<td>NA 36 -58 -30</td>
<td>4.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Cerebellum (Vermis VI)</td>
<td>NA 6 -64 -22</td>
<td>4.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Cerebellum (VI)</td>
<td>NA 30 -46 -38</td>
<td>4.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Cerebellum (X)</td>
<td>NA 30 -36 -44</td>
<td>3.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>DLPFC</td>
<td>46 36 42 28</td>
<td>3.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Dorsal ACC</td>
<td>24 6 16 36</td>
<td>3.53</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Fusiform</td>
<td>37 42 -48 -20</td>
<td>3.68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Inferior Temporal Gyrus</td>
<td>37 48 -60 -12</td>
<td>4.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Insula</td>
<td>48 46 14 -2</td>
<td>3.73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Lateral Occipital Cortex, Inferior Division</td>
<td>19 50 -72 -8</td>
<td>4.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Lateral Occipital Cortex, Superior Division</td>
<td>7 26 -62 42</td>
<td>4.12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Occipital Pole</td>
<td>18 30 -92 -8</td>
<td>3.97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Precentral Gyrus</td>
<td>6 46 -4 56</td>
<td>4.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Precuneous</td>
<td>7 6 -70 48</td>
<td>3.81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>SMA</td>
<td>6 6 -6 64</td>
<td>3.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Supramarginal gyrus</td>
<td>40 44 -40 48</td>
<td>4.19</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

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


<table>
<thead>
<tr>
<th>Region</th>
<th>MNI Coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parietal Operculum Cortex (TPJ)</td>
<td>48, 3.23, 56, -22, 16</td>
</tr>
<tr>
<td>Precentral Gyrus</td>
<td>6, 4.27, 60, 10, 32</td>
</tr>
<tr>
<td>Precentral Gyrus</td>
<td>6, 3.92, 28, -8, 48</td>
</tr>
<tr>
<td>Supramarginal Gyrus, Posterior Division</td>
<td>40, 4.09, 42, -44, 46</td>
</tr>
<tr>
<td>Supramarginal Gyrus, Posterior Division (TPJ)</td>
<td>22, 3.93, 66, -44, 18</td>
</tr>
<tr>
<td>Temporal Occipital Fusiform Cortex</td>
<td>37, 3.86, 38, -48, -24</td>
</tr>
</tbody>
</table>
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 than 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.
References

APPENDIX D: GREAT EXPECTATIONS: NEURAL COMPUTATIONS UNDERLYING THE USE OF SOCIAL NORMS IN DECISION-MAKING (STUDY 3)

Great expectations: Neural computations underlying the use of social norms in decision-making

Luke J. Chang\textsuperscript{1}
Alan G. Sanfey\textsuperscript{1,2,3,*}

\textsuperscript{1}Department of Psychology
University of Arizona
1503 E. University Blvd
Tucson AZ 85721

\textsuperscript{2}Donders Institute for Brain, Mind & Behavior

\textsuperscript{3}Behavioral Science Institute
Radboud University Nijmegen
6525EN Nijmegen
The Netherlands

*Corresponding Author:
Phone: 520.621.1477
Fax: 520.621.9306
Email: asanfey@u.arizona.edu
Abstract

Social expectations play a critical role in everyday decision-making. Here we outline a novel, expectation-based, neurocomputational model of social preferences, and find that it outperforms the standard inequity-aversion model in explaining decision behavior in a social interactive bargaining task. This is supported by fMRI findings showing that the tracking of social expectation violations is processed by anterior cingulate cortex, extending previous computational conceptualizations of this region to the social domain.
Introduction

Navigating our daily environment requires us to routinely engage in social interactive decision-making. These types of decisions necessitate taking into consideration the beliefs and preferences of others, and range from mundane decisions such as where to eat for dinner to those of more consequence, such as deciding whether or not to take a new job offer. An important, though understudied, question is how we integrate our own desires with the expectations of others in order to arrive at a satisfactory outcome.

To date, there have been several theoretical accounts that endeavor to explain how we make choices in these social contexts. These models typically posit that we care about both the payoffs and intentions of others, and have been influential in advancing economic theory beyond considerations of pure self-interest.

For example, these models can broadly describe performance in key tasks of social decision-making, such as the Ultimatum Game (UG). In this game a proposer is charged with splitting a sum of money with a partner. The responder then decides whether to accept the proposed offer, or to reject it in which case both players receive nothing. While the latter course of action is never predicted by a model of self-interest, empirically about half of ‘unfair’ offers are rejected, indicating that motivations other than financial ones are factored into social decision-making.

One prominent model of social preferences argues that people value fairness and prefer to minimize inequity between players’ payoffs, hence the decision to ensure both players get nothing after an unequal proposal. Indeed, a recent study has suggested that brain areas associated with valuation and reward processing could underlie this preference.

However, despite its popularity and intuitive appeal, there is increasing evidence that inequity aversion is not a complete account of social decision-making. For example, unfair offers will be accepted if there is evidence that the proposer’s intentions were ‘noble’. In addition, participants will sometimes reject unfair offers even when knowing that their partner will still receive a share, which of course increases the inequity between each player’s payoff.

An alternative approach to understanding social decision behavior is to focus on the expectations people have regarding a social interaction. These expectations may reflect a social norm about what a majority of people would do in a given interaction. Using this framework we propose that people develop context specific expectations of social scenarios, which are subsequently used as behavioral reference points. For example, rejection rates in the UG increase
when participants are provided with a prior belief about responder behavior\textsuperscript{7} and decrease when they believe that a ‘typical’ offer is unfair\textsuperscript{8}. These results suggest that expectations about context appropriate behavior, rather than pure payoff equity, may provide a better account of motivation in bargaining behavior.

Neurally, unfair offers have been associated with increased activation in anterior insula, anterior cingulate cortex (ACC), and DLPFC, with offer rejection selectively associated with the insula\textsuperscript{9}. Computing violation of expectations is likely instantiated in ACC, as suggested by its involvement in tracking low-level sensory expectation violations\textsuperscript{10}, weighting social prediction errors\textsuperscript{11}, and also adapting behavior following prediction errors\textsuperscript{12}.

Therefore, we compared this novel model of expectations (Smith, Working Paper) to a prominent inequity aversion model in an effort to understand the motivations underlying social interactive decision-making. Further, we used our model in conjunction with fMRI to identify brain networks that track expectation violations and may ultimately contribute to social decision-making.

As expected\textsuperscript{9}, participants were more likely to reject offers as they became less equitable (parameter estimate = -6.90, se = 1.24, odds ratio = 987.62, z = -5.56, p < 0.001). Importantly however, after controlling for offer amount, participants were more likely to reject offers when they had higher initial expectations, parameter estimate = 2.98, se = 0.39, odds ratio = 19.60, z = 7.71, p < 0.001. These results (Figure 2 panel B), suggest that participants’ prior beliefs about the social norm are important in determining whether or not an offer will be accepted.

To better understand the computational mechanism underlying this result, we compared the ability of the two competing models to explain the behavioral data. The modeling results, summarized in Table 1, revealed that the expectation model was a better account of the data than the inequity-aversion model. Simulations of the models (Figure 1, Panel A) illustrate that they both make identical predictions when participants believe that the social norm is $5 (50\% \text{ of a } $10 \text{ pot}). However, the models make divergent predictions for behavior as expectations decrease, with the largest difference being for the intermediate $3 offers.

We then used the model predictions to highlight neural processes specifically associated with the tracking of expectation violations, which are computationally distinct from tracking inequity. A whole brain analysis for regions that linearly track with the model’s predictions for the $3 offers reveals that only the ACC underlies this process (Figure 1, Panel C). Participants with higher expectations demonstrated increased activity in ACC when deciding about the intermediate $3 offers. This suggests that a neural signal, perhaps akin to a prediction error signal in basic reinforcement learning, may be leveraged to calculate social norm
violations and thereby bias decision-making. In addition, using a parametric contrast of deviation from expectations on a trial-by-trial level, we find increasing activity in left insula, ACC and pre-supplementary motor areas as offers increasingly violated individual participants’ expectations (Figure 1 Panel D). These results overlap with previous UG findings, suggesting that expectations can at least partially account for neural mechanisms underlying decisions to reject unfair offers. Moreover, these results are consistent with a network that signals conflict between individual preferences and social norms, which in turn promotes conformity to the norm.

In summary, these results provide compelling evidence for the role of expectations in social decision-making behavior. Our model provides not only a better account of the behavioral data than an inequity aversion model, but reveals that the computational process of detecting expectation violations underlying decisions to reject is associated with the same neural network that has previously been demonstrated to underlie other expectation based effects such as anticipating aversive events, detecting novel events, placebo effects, and conforming to others’ expectations. Overall, these results demonstrate that people do not use simple heuristics such as equal splits in considering their responses to financial proposals, but rather rely on their context-specific beliefs about the social norm to make their decisions.

Acknowledgments
The authors wish to thank Mascha van’t Wout and Katia Harle for their assistance in data collection and Alec Smith and Michael Frank for their helpful comments. This work was supported by awards from the NIA to AGS (R21AG030768) and the NIMH to LJC (F31MH085465).

Author Contributions
LJC conducted the experiment and analyzed the data and LJC and AGS wrote the paper.
Legends

Figure 1. Ultimatum Game Trial Timeline. (I) Fixation cross, (II) Picture of their partner for the round, (III) Offer revealed and participant decision (accept or reject offer), (IV) Summary of earnings for the round.

Figure 2. Ultimatum Game Results. Panel A: results of the model simulations for theta=0.3. The Expectation model predicts the same probability of accepting offers as the Inequity Aversion model when expectations are $5. However, the Expectation model makes divergent predictions for lower expectations at the intermediate $3 offers. Panel B: average acceptance rates for each offer amount for varying expectations. Importantly, the acceptance rates closely follow the pattern predicted by the Expectation model. Panel C: neuroimaging results for the linear contrast of expectation across subjects for $3 offers. Activity in the ACC/preSMA increases as a function of expectations. Panel D: neuroimaging results of a linear contrast of deviations from expectation, revealing linearly increasing activity in left anterior insula, anterior cingulate cortex, and supplementary motor area. The colorbar reflects z-statistics. Imaging analyses are corrected for multiple comparisons using cluster correction, Z>2.3, p<0.05.

Table 1. Modeling results. Values given for the subject average log likelihood estimate (LLE), Bayesian information criteria (BIC), pseudo-$r^2$, and the estimated theta parameter for each model.
Figures

Figure 1. Ultimatum Game Trial Timeline.

I. Fixation

II. Face

III. Offer

101’s proposal is $3:
Accept or Reject this Proposal

You have accepted this offer
101 gets $7
You get $3

IV. Summary
Figure 2. Ultimatum Game Expectation Results

A. Model Predictions Theta=0.3

B. Average Behavior for Each Level of Expectation

C.

D.
### Tables

Table 1. Modeling results.

<table>
<thead>
<tr>
<th>Model</th>
<th>LLE (SD)</th>
<th>BIC (SD)</th>
<th>Pseudo-R² (SD)</th>
<th>Theta (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>-3.84 (2.36)</td>
<td>10.86 (4.71)</td>
<td>0.77 (0.14)</td>
<td>0.28 (0.26)</td>
</tr>
<tr>
<td>Fehr-Schmidt</td>
<td>-4.45 (1.75)</td>
<td>12.07 (3.49)</td>
<td>0.73 (0.11)</td>
<td>0.26 (0.23)</td>
</tr>
</tbody>
</table>
Methods

Participants

Eighteen participants (mean age=19.9, female=56%) were recruited via advertisements posted on the campus of the University of Arizona. All participants were screened for any significant health-related or neuropsychiatric disorders and none were currently taking psychoactive medication. One participant was excluded from the analysis for technical reasons (missing data). All participants gave informed consent according to procedures approved by the University of Arizona’s Institutional Review Board.

Procedure

This study consisted of two separate sessions. In the first session, participants met in small groups and were screened for eligibility and instructed on the Ultimatum Game. To increase plausibility that they would be playing with real people, participants first played in the role of the proposer and made offers to 20 people. Participants then had their picture taken and were told that other participants in the experiment would later propose offers to them, and that they would decide whether to accept or reject these offers while being scanned in the second session.

Expectations: Prior to being scanned, we elicited participants’ beliefs about the kinds of offers they expected to encounter, with participants being asked the number of people out of 100 that they believed would make a $0, $1, $2, $3, $4, $5, $6, or $7 offer. These elicited expectations were used to create a distribution of the frequency of offers that they expected to encounter. The weighted mean of this distribution was used to represent each participant’s initial expectation.

Ultimatum Game: Participants played a standard single-shot Ultimatum Game in the role of Responder with 48 different Proposer partners while undergoing fMRI. Twenty-four of these partners were human, 12 were computers, and 12 were non-intentional humans (i.e. humans whose responses were randomly generated). Each offer was preceded by a picture of their partner for that round. Though participants were told that the human-intentional offers would be made by other players, in actual fact all offers were controlled by the experimenter, and all participants saw the same set of offers. This set consisted of equal numbers (12 each) of $1, $2, $3 and $5 offers, all of which were made from a $10 pot. For each participant, partner pictures were randomly paired to an offer amount, ensuring that there was no potential picture by offer amount interaction. Only the human intentional offers were included in the modeling analyses. Participants were paid $20 for participating with an additional $5 bonus, which they believed was based on their performance in the game. While participants were not
directly queried after the experiment about whether or not they believed that they were interacting with real partners, no participant expressed doubt towards the experimenter at any time during the experiment and our behavioral results mirror previous work \(^9, 16\).

**Data Analyses**

All behavioral statistics were computed using the R statistical package \(^{17}\). A mixed logit model \(^{18}\) with the amount of money offered and participant’s initial expectations was used to predict participant’s decisions to accept or reject. We allowed the slopes for the offer parameter to randomly vary by subject, but did not estimate a parameter for the intercept because of the linear dependence on participant’s expectations.

**Modeling**

To demonstrate the importance of considering expectations in the UG, we compared a novel model \(^{19}\), which incorporates expectations to the inequity-aversion model \(^{1, 20}\), which posits that people have distributional preferences and are motivated to minimize the difference between their payoffs.

**Expectation Model** \(^{19}\): This model was developed in the context of Psychological Game Theory \(^{21, 22}\), which describes a mathematical framework in which beliefs can be modeled in the utility function. Similar to other models of emotion \(^{23-25}\), this model operationalizes anger as a belief dependent emotion and predicts that people experience anger when their beliefs about behavioral norms in a given context are violated. In the UG, offers that are lower than participants’ expectations, that is the weighted mean of their beliefs about the distribution of offers they will likely encounter, should bias participants to reject the offer. Formally, Player 2’s utility \(U\) of a given action \(i\) can be defined as

\[
U_i = \begin{cases} 
M_2 - \theta (E_2 S_1 - S_1) \cdot (10 - S_1) & \text{where } i = \text{Accept} \\
0 & \text{where } i = \text{Reject} \end{cases} 
\]

(1)

where \(M_2\) is the amount of money Player 2 will receive and \(E_2 S_1\) is Player 2’s beliefs about Player 1’s strategy \((S_1)\). The anger term is scaled by a free parameter which is constrained \(0 < \theta < 1\).

**Inequity-Aversion Model** \(^{1, 20}\): This model was one of the early Behavioral Economic models which attempted to explain the apparent contradiction of classical economic theory’s predictions in the Ultimatum Game \(^3\) and is a popular model of social preferences \(^4, 26, 27\). This model predicts that people value
fairness and will be biased to reject offers as inequity increases. Formally, Player 2’s utility for a given action i can be defined as

$$ U_i = \begin{cases} M_2 - \theta(M_1 - M_2)^+ & \text{where } i = \text{Accept} \\ 0 + \theta(M_1 - M_2)^+ & \text{where } i = \text{Reject} \end{cases} $$

(2)

where $$ M_1 $$ is the amount of money that Player 1 will receive and $$ M_2 $$ is the amount of money that Player 2 will receive. The inequality term is scaled by a free parameter which is constrained $$ 0 < \theta < 1 $$.

Choice Rule: The probability $$ P $$ of taking an action $$ i $$ (i.e. accept or reject) was computed by placing the utility values for each decision into a softmax function.

$$ P_i = \frac{e^{U_i}}{e^{U_{\text{accept}}} + e^{U_{\text{reject}}}} $$

(3)

Parameter Estimation: Best fitting parameters were derived using the MATLAB fmincon function by maximizing the log likelihood of the data under each model on a trial-to-trial basis. Multiple start locations were used to reduce the likelihood of the optimization algorithm getting stuck in local minima. LLEs were calculated separately for each participant as

$$ \text{LLE} = \sum_{t} \ln(P_{i,t}) $$

(4)

where $$ i $$ denotes the participant’s choice for a given trial $$ t $$.

To evaluate the model fits we calculated the Bayesian Information Criteria (BIC), which is a metric of model fit that rewards the most parsimonious model by adding a penalty for additional free parameters.

$$ \text{BIC} = -2 \cdot \ln L + k \ln(n) $$

(5)

where $$ L $$ is the maximized value of the likelihood function for the model, $$ k $$ is the number of free parameters estimated, and $$ n $$ is the number of observations.

We also computed a Pseudo R$^2$ measure, which compares the improvement in LLE gained by the model compared to a model that chose randomly (i.e. probability=0.5 for each trial).
Pseudo $R^2 = \frac{LLE - r}{r}$

where $r$ is the LLE for the random model.

Model Simulation:
The behavioral predictions for the models were computed by using a theta value of 0.3 (the approximate value estimated in the model fitting procedure) and then calculating the probability of accepting each offer from the set $[0,5]$ using equations 1, 2, & 3. For the expectation model, we varied expectations between $[3,5]$, which was the range we encountered in our behavioral sample (See Figure 1, Panel A).

Neuroimaging Methods

Data Acquisition: Each scanning session included a T1-weighted MPRAGE structural scan (TR=11ms, TE=4 ms, matrix=256X256, slice thickness=1mm, gap=0mm), followed by five functional runs. The first 3 functional runs contained the Ultimatum Game trials and the last two contained the memory trials (see for more details about the memory study). Functional scans used a 3-shot multiple echo planar imaging (MEPI) GRAPPA sequence using parameters selected to maximize signal in regions associated with high susceptibility artifact, such as orbitofrontal cortex and medial temporal lobe (TR=2000ms, TE=256ms, matrix=96X96, FOV=192mm, slice thickness=3.0mm, 42 axial slices, voxel size 2 X 2 X 3).

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. Scanner artifacts, physiological artifacts (i.e., cardiac and respiration), and head movement related artifacts were removed from the data using independent components analysis. 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. 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 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 $p < 0.05$.

Analysis 1: Model Prediction: To examine neural responses that follow from the expectation model predictions, we fit a model with the following regressors: face phase, Human $1$ offers, Human $2$ offers, Human $3$ offers, Human $5$ offers, Non-intentional human control trials, Computer trials, Summary phase trials, Missing trials, and head motion parameters. Together with the temporal derivatives, this produced a model with a total of 24 regressors. For this analysis we used a linear contrast of expectations (i.e. expectations that were between $2.5-3.4$ (n=3), $3.5-4.4$ (n=11), & $4.5-5.4$ (n=3)) to examine neural responses that linearly tracked with expectations for the intermediate $3$ offers across participants. We chose to focus only on the $3$ offers for this analysis to avoid any confounds introduced by inequity and also because the model simulation revealed that these offers should be most susceptible to expectation effects (see Figure 2, Panel A).

Analysis 2: Parametric Contrast: To examine the neural responses to linear deviations in expectation violation, we subtracted the amount of money offered at each trial from the participants’ initial expectations and coded them as 0, 1, 2, or 3+. Thus, this model consisted of the four expectation violation predictors, two additional predictors for the other phases of the task (i.e. face phase, summary phase, and computer and non-intentional human control trials), a regressor indicating missed trials, the temporal derivatives of these 9 regressors, and the six motion parameters, which resulted in a GLM with a total of 24 predictors. We then used a within subject linear contrast of the expectation deviations (i.e. $-2 -1 1 2$) to examine the neural signals that parametrically tracked these deviations.
References

17. R_Development_Core_Team. (Vienna, Austria, 2008).
APPENDIX E: UNFORGETTABLE ULTIMATUMS? EXPECTATION VIOLATIONS PROMOTE ENHANCED MEMORY FOLLOWING ECONOMIC BARGAINING (STUDY 5)

Originally published in *Frontiers in Behavioral Neuroscience*. Reprinted with permission from the authors.

Unforgettable ultimatum games? Expectation violations promote enhanced social memory following economic bargaining

Luke J. Chang and Alan G. Sanfey*
Department of Psychology, University of Arizona, Tucson, AZ, USA

Recent work in the field of neuroeconomics has examined how people make decisions in interactive settings. However, less is currently known about how these social decisions influence subsequent memory for these interactions. We investigated this question by using functional magnetic resonance imaging to scan participants as they viewed photographs of people they had either recently played an Ultimatum Game with in the role of Responders, or that they had never seen before. Based on previous work that has investigated “cheater detection,” we were interested in whether participants demonstrated a relative enhanced memory for partners that made either fair or unfair proposals. We find no evidence, either behaviorally or neurally, supporting enhanced memory based on the amount of money offered by the Proposer. However, we did find that participants’ initial expectations about the offers they would experience in the game influenced their memory. Participants demonstrated relatively enhanced subjective memory for partners who made proposals that were contrary to their initial expectations.

In addition, we observed two distinct brain systems that were associated with partners that either offered more or less than the participants’ expectations. Viewing pictures of partners that offered less than initial expectations was associated with bilateral anterior insula, anterior cingulate cortex/premotor area, striatum, and bilateral posterior hippocampi, while viewing partners that offered less than initial expectations was associated with bilateral temporal-parietal junction, right STS, bilateral posterior insula, and precuneus. These results suggest that memory for social interaction may not be guided by a specific cheater detection system, but rather a more general expectation violation system.

Keywords: neuroeconomics, memory, social, expectation, ultimatum game, cheater detection, neural, decision-making

INTRODUCTION

Despite its relative youth, neuroeconomics as a field has made significant progress in describing the neural mechanisms that underlie decision-making (Glimcher et al., 2009). One approach within this domain has focused on circumstances in which the participant must consider the desires and intentions of another agent in reaching his or her eventual decision (Sanfey, 2007). These interactive situations have examined decisions made in a social environment, such as whether to trust or not trust another player or how to negotiate the division of a sum of money with another. The simplicity of these tasks and their ease of quantification provide not only a useful framework for developing mathematical models of optimal behavior within a social interaction (von Neumann and Morgenstern, 1944; Rabin, 1993; Fehr and Schmidt, 1999; Battigalli and Dufwenberg, 2007), but also a controlled environment within which to understand how social interaction interacts with more general cognitive processes such as memory.

One commonly used task in this domain is the Ultimatum Game (Guth et al., 1982). In this simplified bargaining scenario, one player known as the Proposer is endowed with a sum of money and told that their task is to make a proposal to the other player, the Responder, as to how this money should be divided between the two. The Proposer can make any offer he or she wants, from keeping all of the money for themselves to giving all of it away, and any division in between. Once the offer is made, the Responder must decide to either accept or reject the proposal. If the offer is accepted, then the money is simply divided as suggested. However, if the offer is rejected, then neither player receives any money. Both players are fully aware of the rules of the game, and once the Responder makes a decision the game is over.

Many studies across a multitude of disciplines and utilizing a variety of methods have examined social decision-making using the Ultimatum Game, and the behavioral results are generally strikingly similar (Camerer, 2003). Contrary to classical predictions, which suggest that Responders should accept any non-zero offer and as a consequence Proposers should make the lowest offer possible, the modal offer to Responders is typically a little less than half of the total pot, and this amount is almost always accepted. Offers of around 30% of the pot are accepted only about half of the time, and acceptance rates diminish as offers get lower.

One suggested mechanism as to why respondents turn down what is in effect “free” money when rejecting low offers is that people severely dislike inequality (Fehr and Schmidt, 1999), and consequently feel anger in response to unfair offers (Páliata and Murnighan, 1996; Xiao and Houser, 2005). There is compelling physiological evidence supporting this argument. Unfair offers
are associated with increased autonomic tone (van 't Wout et al., 2006) and increased activity in the anterior insula (Sanfey et al., 2003). In fact, greater insula activity in response to an unfair offer results in an increased likelihood of rejection of that offer (Sanfey et al., 2003). Other studies have found that when neural systems involved in emotion regulation are disrupted in various ways, from using tryptophan depletion (Crockett et al., 2008) to lesions of the ventromedial prefrontal cortex (Koenigs and Tranel, 2007), the result is increased rejection rates of unfair offers.

Though the response to fair and unfair offers provides an interesting window into how the competing motivations of maintaining one’s reputation and maximizing one’s financial gain interact in decision-making, other relevant questions can be answered using these type of tasks. Of perhaps equal importance to examining the processes that underlie performance in this task is to ask what happens, both behaviorally and neurally, when we re-encounter a player who has made a fair or unfair offer to us in the past. How do our perceptions of others shift when these people have previously treated us either well or poorly? In this initial attempt to investigate this question, we focus on memory for players with whom we have recently interacted, and specifically examine whether the way in which another player has treated us has an impact on how we in turn remember them.

Several theoretical proposals have been made as to whether we are more attuned to remembering those who have treated us either fairly or unfairly in the past. In their highly influential theory of social exchange, Cosmides and Tooby (1992) argue that humans have evolved specific cognitive abilities to promote reciprocal altruism, a construct that has been associated with positive evolutionary fitness (Trivers, 1971). Of particular importance to their theory is the ability to detect, remember, and punish “cheaters” – individuals who benefit themselves by violating a social contract (Cosmides and Tooby, 1992). However, despite the intuitive appeal of this theory, the primary evidence presented in favor of the selective detection of cheaters is that experimental participants demonstrate improved conditional reasoning when asked to detect violations of a social contract, when compared to non-social contract violations (Cosmides, 1989; Gigerenzer and Hug, 1992). Evidence supporting these theoretical claims in the domain of memory is more mixed.

Several studies have directly examined explicit memory for cheaters. There is some evidence that after 1 week participants had better memories for pictures of people with behaviors associated with cheating (e.g., “A is a bishop who was caught embezzling money from his own church.”) as compared to pictures of those that were associated with trustworthy behaviors (e.g., “H. is a vendor at baseball games who, after finding a wallet containing $250, located the owner using the driver’s license.”) (Meyl et al., 1996; Chappe et al., 2004). However, more recent studies that have attempted to address some of the methodological limitations of these experiments have failed to replicate this finding, with no differences between cheaters and trustworthy pictures emerging (Barclay and Lahmien, 2006; Mehl and Buchner, 2008). There is even some preliminary evidence for increased confidence, though not accuracy, in remembering altruists (Barclay and Lahmien, 2006), and also that people may have better source memory than recognition memory for cheaters, meaning that people were better at remembering that an individual was a cheater than actually correctly identifying that they had seen the person before (Buchner et al., 2009).

One explanation of these mixed findings is that the memory manipulations used were not particularly socially relevant for the participants. As outlined above, these paradigms typically involve participants reading a vignette describing either a cheating or trustworthy act by a pictured person, and then subsequently performing a recognition memory test on the set of photographs. There are surprisingly few studies that have attempted to have participants first actually engage in meaningful social interactions with other people, and then test their memory for these partners. In one study with a variant of this methodology, participants were asked to imagine playing a constant strategy (i.e. cooperate or defect) in a Prisoner’s Dilemma game, and were then shown pictures and the strategies of their partners (Oda, 1997). After being tested 1 week later, the experimenters found that participants remembered defectors better than cooperators and that this effect interacted with gender. However, there was no clear explanation of the interaction with gender, nor was it clear that participants were actually engaged in the game as they were forced to stick with the same strategy.

Within neuroeconomics, there is clear evidence that people use information about a partner’s history to inform decisions in future social interactions, such as to avoid trusting a cheater in a subsequent interaction or to punish them if given the opportunity. People are more likely to invest in partners perceived to be initially trustworthy as opposed to untrustworthy (Delgado et al., 2005; van’t Wout and Sanfey, 2008), and also seem able to then disregard this prior information when these partners actually abuse their trust. There is also evidence supporting the notion that people are willing to punish cheaters, even at the risk of incurring a financial cost to themselves (Guth et al., 1982) and Public Goods Games (Fehr and Gächter, 2002; de Quervain et al., 2004). While these findings suggest that people can learn both who to trust and who not to trust and will punish cheaters given the opportunity, there is no conclusive evidence directly supporting better explicit memory for either cheaters or cooperators. One study (Singer et al., 2004) attempted to investigate this question both behaviorally and neurally by scanning participants using functional magnetic resonance imaging (fMRI) as they viewed faces which had previously behaved in either cooperative or non-cooperative ways in a modified repeated Prisoner’s Dilemma game. Behaviorally, the authors report that cooperators were rated as more likeable and deflectors as less likeable than control faces. In addition, participants were more accurate in recalling the behavior of both cooperators and defectors as compared to the null games. However, because there was neither money at stake for the null games nor an equal distribution of trials for each condition, the results of this forced choice memory task should be interpreted cautiously. In terms of neural findings, the authors reported that when asked to make a gender assessment of pictures of cooperators as compared to those who played null games, participants had increased activity in the left ventral putamen and left amygdala. In contrast, when participants viewed pictures of defectors compared to null trials, they showed increased activity in the vmPFC. These preliminary findings suggest that viewing faces of defectors and
cooperators from a socially relevant task may be associated with distinct neural systems. However, it remains an open question as to whether or not there may be selectively better explicit memory for cheaters and what processes might underlie this.

A possible mechanism that could explain the aforementioned pattern of results is the notion of deviation from expectation, that is, when partners play in a way differently than we predict. While it is known that deviations of expectation can affect decision-making in the Ultimatum Game (Sanfey, 2009), up to now there has been relatively little investigation of how expectations, and specifically deviation from expectations, can alter patterns of memory in social decision-making.

Some limited evidence comes from a recent study using the Trust Game, which found that participants did not have selective memory for either cooperators or defectors per se, but rather demonstrated enhanced memory for both types of opponents in certain circumstances, these circumstances being that the better-remembered opponent played a relatively infrequent strategy. That is, at different times they remembered both cheaters and defectors better, but only when they comprised merely 20% of the total number of interactions (Borchert, 2005). It is important to note that participants in this experiment knew a priori that they were playing with computer partners, so it is not clear if these results could be generalized to games played with real opponents. Nonetheless, this study provides compelling evidence that people may have enhanced memory for partners that behave contrary to social conventions, regardless of their behavior.

This suggests therefore that people may not rely on a specific cheater detection system, but rather a more general expectation violation system — a notion within the field of memory that has been known for some time (Eron et al., 1973; Ranganath and Rainer, 2003), often discussed in this literature as a “novelty detection” mechanism. It is therefore possible that a more general novelty detection system can potentially be employed as a cheater detection system. Because interactions with cheaters in the real world are likely to be relatively infrequent, the expectation of cheating behavior should be low and as a result incidences of cheating should be particularly memorable. However, importantly, if we do expect substantial cheating behavior in our environment, this account would predict that partners who treat us well should be preferentially encoded and remembered.

We sought to investigate this question by using fMRI to scan the brains of participants immediately after they played a series of Ultimatum Games with a variety of partners. Firstly, we examined if our participants demonstrated more accurate memories for partners that had treated them either fairly (an equal offer) or unfairly (an unequal offer in the partner’s favor). Secondly, we were particularly interested in the neural response to viewing a photograph of a previous partner as compared to a photograph of a previously unseen person, and whether the offer that had been made to the participant mediated this neural activity in our participants. Contrary to most prior behavioral studies of memory for cheaters, players in this study engaged in an actual social decision interaction and we directly assessed their social memory while they were being scanned using a standard recognition task. This study therefore can potentially inform an ongoing debate about whether people actually have enhanced memory for cheaters, and if so, whether the brain is equipped with a system to complete this task.

MATERIALS AND METHODS

Participants

Eighteen participants (mean age = 19.9, female = 56%) were recruited via advertisements posted on the campus of the University of Arizona to participate in this study. All participants were screened for any significant health-related or neuropsychiatric disorders and none were currently taking psychotropic medication. Two participants were excluded from the analysis for technical reasons (corrupted data). All participants gave informed consent according to procedures approved by the University of Arizona’s Institutional Review Board.

Procedure

Expectations

Prior to being scanned, we elicited participants’ beliefs about the kinds of offers they expected to encounter, with participants being asked the number of people out of 100 that they believed would make a $0, $1, $2, $3, $4, $5, $6, or $7 offer. Participants’ elicited expectations prior to playing the game were used to create a distribution of the frequency of offers that they expected to encounter. The mode of this distribution was used to represent each participant’s initial expectation.

Ultimatum Game

Participants then played a standard single-shot Ultimatum Game in the role of Responder with 48 different partners while undergoing fMRI. Twenty-four of these partners were human, 12 were computers, and 12 were non-intentional humans (i.e., humans whose responses were randomly generated). Each offer was preceded by a picture of their partner for that round. Though participants were told that the human-offer would be made by another player, in actual fact all offers were controlled by the experimenter, and all participants saw the same set of offers. This set consisted of equal numbers (12 each) of $1, $2, $3 and $5 offers, all of which were made from a $10 pot. For each participant, all pictures were randomly paired to an offer amount, ensuring that there was no potential picture by offer amount interaction. Participants were paid $20 for participating and an additional $5, which they believed was based on their performance in the game. Further details of the Ultimatum Game portion of the experiment will be described in greater detail in a separate paper. While participants were not directly queried after the experiment about whether or not they believed that they were interacting with real partners, no participant expressed doubt towards the experimenter at any time during the experiment.

Memory Experiment

After completing the Ultimatum Game trials, participants were given an incidental memory test while undergoing fMRI (see Figure 1 for a trial timeline). Participants had not been forewarned about this task. Participants viewed randomly presented pictures of the 24 human-intentional partners they had previously encountered, as well as 24 new faces they had not seen in the Ultimatum Game task. Participants did not view pictures of the computers or the non-intentional human partners. There was an equal number of male and female faces for both the new and old sets of faces.

To begin each trial, a jittered fixation was seen for 6 s on average
and then a face was shown for 4 s. After being presented with this photograph, participants were asked to rate their confidence that they had played the Ultimatum Game with this person on a scale of 1 to 5 (1 = “I’ve definitely never seen this person before”; 2 = “I don’t think I’ve seen them before”; 3 = “I’m not sure if I’ve seen this person before”; 4 = “I think I may have seen them before”; 5 = “I’ve definitely seen this person before”). Participants were given 8 s to make this judgment, and ratings were entered by scrolling through the possible options with one response button and then selecting the desired rating with the other button. While the rating system was always the same (e.g., 1 = never seen; 5 = definitely seen), the ratings randomly scrolled up or down on each trial to eliminate any possible motor confounds. Therefore, the number of button presses were orthogonal to the actual ratings provided. Stimuli were presented via EPrime software (Psychology Software Tools, Inc., Pittsburgh, PA, USA) using MRI-compatible goggles and responses were recorded using a fiber optic button box (Resonance Technologies, Van Nuys, CA, USA).

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 general linear model package (Bates et al., 2008). Participants were treated as a random effect with varying intercepts and slopes. We report the parameter estimates (b), standard error, 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 observations minus the number of fixed effects (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 results were identical unless otherwise noted. For robust regressions we used the rlm function from the MASS package using an MM-estimator (Venables and Ripley, 2002).

D’
To measure participants’ ability to discriminate old from new faces, we used D’, a signal detection metric (Wickens, 2002). D’ controls for individual participants’ response bias (i.e., their propensity to say yes) and was calculated as the difference between the standardized z-score for hits (indicated by a 4 or 5 on the confidence rating for an old face) and the standardized z-score for false positives (indicated by a 4 or 5 on the confidence rating for a new face). Because this analysis emphasizes hits and false positives, it ignores differences in levels of subjective confidence, that is, the difference between a 1 and 2, or a 4 or 5 rating. D’ scores were calculated separately for every level of offer amount.

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), followed by five functional runs. The first 3 functional runs contained the Ultimatum Game trials and the last two contained the memory trials (240 volumes per run). Functional scans used a 3-shot multiple echo planar imaging (MEPI) GRAPPA sequence using parameters selected to maximize signal in regions associated with high susceptibility artifact, such as orbitofrontal cortex and medial temporal lobe (Stockel 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, voxel size 2 × 2 × 3). The MEPI sequence employs parallel imaging and allows for increases in
signal intensity, image resolution, the number of slices that can be acquired in a 2000 ms TR, as well as substantial decreases in geometric distortion (Newbould et al., 2007).

**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 the functional runs 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 5 mm full width at half maximum Gaussian kernel. A high-pass filter was used to cut off temporal periodicities longer than 66 s. All images were initially co-registered to the participant’s high-resolution structural scan and were then co-registered 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 imaging 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. face phase), convolved with a canonical double-gamma hemodynamic response function. 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 GLM was corrected for temporal autocorrelations using FLM prewhitening. A second-level fixed effects model was fit for each participant to account for intra-run variability. For each participant, contrasts were calculated between predictors for different regressors of interest at every voxel in the brain. A one-sample t-test was used at the third level for each contrast using a Bayesian implementation of mixed effects inference (Forstmann et al., 2008). We corrected for multiple comparisons with cluster correction utilizing Gaussian random field theory with an initial cutoff of Z > 2.3 and a FWE p < 0.05.

We report the results of three analyses. The offer amount analysis included individual regressors during the face phase for players who had previously made offers of $1, $2, $3, or $5, a regressor indicating the duration of the response time during the memory phase, a regressor indicating a distractor face, a regressor for missed trials, the temporal derivatives of each of these predictors and 6 motion parameters (20 predictors total). We report the results for the Unfair (i.e. $1 and $2 offers) vs Fair (i.e. $5) contrast. For the expectation violation analysis we included regressors at the face phase for offers below expectation (i.e. standardized expectation error (SEE) > 0), offers above expectation (i.e. SEE < 0), and offers at expectation (i.e. SEE = 0). SEE is the within-subject z-score of the numerical deviation of an offer amount from a participant’s initial expectations. In addition, we included a regressor modeling the memory phase for the duration of the response, a regressor indicating a distractor face, a regressor modeling missed trials, and their temporal derivatives and 6 motion parameters (18 predictors total). We report a linear contrast of prediction error (i.e. +1 0 – 1) for Positive, Zero, and Negative SEE regressors. Finally, the third analysis was identical to the second analysis except the third level linear contrast was weighted by each participant’s standardized initial expectation, effectively utilizing a correlation analysis rather than a one sample t-test. This analysis tests the interaction between participant’s initial expectation and their SEE. All trials in which the participants indicated that they were unsure (i.e. a rating of 3) were excluded from all analyses (78 trials total for all subjects, or 10.2% of observations).

**RESULTS**

**Ultimatum game** Consistent with previous research (Sanfey et al., 2003; van ’t Wout et al., 2006; Harlé and Sanfey, 2007), acceptance rates decreased as offers got lower, and participants were significantly more likely to accept fair ($5) as opposed to unfair ($1, $2) offers, illustrated using a mixed effects logit model (Jorg, 2008), b = 4.24, se = 0.64, odds ratio = 69.21, Wald Z = 5.07, p < 0.05. The average acceptance rate for all intentional offers was 62.2%, with three participants accepting all offers. Consistent with previous research, participants expected most participants to make fair offers (mean = 4.5, sd = 0.63) (Sanfey, 2009).

**Memory** A one-sample t-test revealed that participants were on average accurate in their ability to discriminate between old and new faces (mean = 0.79, t(15) = 14.68, p < 0.001. Participants were able to correctly identify both old faces (mean correct = 70%, se = 0.04) and new faces (mean correct = 76%, se = 0.04). We used a mixed effects linear model to test whether or not the amount of money offered by the proposer would influence participant’s memory for that person, but did not observe a significant effect, b = 0.03, se = 0.05, t = 0.62, ns. These results indicate that participants were sensitive in their ability to discriminate between new and old faces, but that, on average, this discriminability was not influenced by the amount of money offered by the partner.

However, closer examination of these results indicate that participants demonstrated considerable variability in their ability to remember proposers that made either fair or unfair offers (Figure 2). Some participants appeared to demonstrate improved memory for proposers that made unfair offers, while other participants remembered proposers that made fair offers indicated by the random effect slope coefficient). Using robust regression we found that participants’ initial expectations predicted the random effects parameter estimates from the previous offer amount analysis, parameter estimate = 0.16, se = 0.04, r = 4.32, p < 0.05. This analysis indicates that as initial expectations increased, the slope of offer amount on D* decreased. In other words, participants with low initial expectations had positive memory slopes, meaning that they demonstrated augmented memory for proposers that made fair offers, while participants with high initial expectations had negative memory slopes, indicating increased memory for proposers that made unfair offers (Figure 2A).

To test this expectation violation hypothesis more explicitly, we used a mixed effects linear model treating subjects as a random intercept. Specifically, we attempted to predict participant’s
subjective confidence ratings using their centered initial expectation, the centered deviation of the offer amount from their initial expectation, and the interaction between these two variables. We observed an initial expectation by expectation deviation interaction, $b = 0.17$, $se = 0.08$, $t = 2.21$, $p < 0.05$, with no significant main effects. Participants demonstrated enhanced confidence ratings for faces that violated their initial expectations (see Figure 2).

Imaging Results

Offer amount

As noted above, we observed no significant effect of offer amount in predicting participant’s ability to discriminate between old and new faces. Similarly, for the corresponding imaging analysis, we did not observe any significant voxels above threshold for this previously fair vs. previously unfair contrast, even at a more liberal $p < 0.001$ (uncorrected) threshold. Therefore, at least on average across participants, there is no particular neural signature for either previously fair or previously unfair partners.

Expectation violation

Our more detailed behavioral analysis indicated that expectation violation, and not offer amount, was associated with enhanced subsequent memory for partners. To explore the neural systems underlying this effect we ran two separate imaging analyses. The first analysis examined the effect of expectation deviation. This contrast was associated with bilateral anterior insula, pre-supplementary motor area (pre-SMA), anterior cingulate cortex (ACC), the striatum (including the caudate and nucleus accumbens), and bilateral posterior hippocampi/parahippocampi. Negative expectation deviations were associated with bilateral temporal parietal junction (TPJ), right superior temporal sulcus (STS), posterior insula, and precuneus (see Figure 3). The second analysis examined the interaction between the initial expectation and the expectation error, by weighting the first analysis by each participant’s standardized initial expectation at the third level. No voxels survived our threshold for this analysis. Thus, this set of analyses reveals a network previously associated with expectation violation (i.e. insula, pre-SMA, and NAcc), and memory retrieval (i.e. hippocampi/parahippocampi) when participants view faces of partners who offered more than the participants initially expected, and a network associated with theory of mind processing (i.e. STS/TPJ) and memory (i.e. precuneus) when viewing partners that offered less than the participant initially expected (see Table 1 for a complete list of regions).

Discussion

This study investigated how economic exchange impacts subsequent memories for social partners. Following a standard Ultimatum Game paradigm, participants were shown photographs of both previously seen and unseen people, and asked to rate their confidence that they had viewed these pictures before. This question is important in understanding the behavioral and neural effects of reappraising a partner with whom one has previously been engaged in social economic interaction. In addition, this research was also interested in investigating the notion of “cheater detection”, that is, the idea of relatively enhanced memories for social partners who have treated us badly in the past (Cosmides and Tooby, 1992; Mealy et al., 1996; Singer et al., 2004; Barclay, 2008).

We were primarily interested in whether participants exhibited a relative memory enhancement for partners that made either fair or unfair proposals. A demonstration of the latter (i.e. enhanced memory for unfair proposers) would provide evidence supporting the existence of behavioral cheater-detection effects. However, we
did not observe a significant effect, either behaviorally or neurally, for an influence of offer fairness on memory. Instead, we found that participants demonstrated considerable variability in their ability to discriminate between partners associated with different levels of fairness (e.g., some participants demonstrated relative enhanced memory for fair partners and some for unfair partners). Importantly, this variability was predicted by their initial expectations about the range of offers they would see in the game. Those who had low initial expectations (operationalized by their modal reported expected offer) were more likely to demonstrate augmented memory for partners that exceeded their expectations, while those who had high initial expectations demonstrated better memory for partners who made offers lower than their initial expectations. We also observed a significant interaction between participants’ initial expectations and their expectation error (i.e., the proposal’s deviation from initial expectation) in predicting subjective confidence ratings. Participants who had low initial expectations were more likely to remember partners that made offers that were greater than their initial expectations, while participants that had high initial expectations were more likely to remember partners that made offers that were lower than their initial expectations.

This finding is consistent with the results of a recent cheater detection study (Barclay, 2008), in which participants demonstrated enhanced memory in a recognition test for whichever behavior (e.g., cooperate or defect) was more infrequent in a previously-played Trust Game. When the majority of partners cooperated, participants remembered defectors better, whereas when most partners defected, participants remembered cooperators better. Our study employed a different approach than that of (Barclay, 2008), namely use of an Ultimatum as opposed to the Trust Game, and additionally we used the participants own expectations as the “baseline”, as opposed to examining violations from experienced probabilities, but the two set of results converge on the same interpretation – that deviations from prior expectations result in greater salience and thus better memory encoding.

This conjecture has been posited for a long time in the memory literature dating to von Restorff (1933). The Von Restorff effect refers to memory enhancement occurring when an item is isolated either by manipulating the context (e.g. item is printed in red in a list of items printed in black) or content (e.g. inserting a nonsense syllable into a list of meaningful words) (Wallace, 1965). This effect is thought to be associated with unexpected change rather than
Table 1 | Brain regions associated with expectation error.

<table>
<thead>
<tr>
<th>Hemisphere</th>
<th>Region</th>
<th>BA</th>
<th>Z-value</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS &gt; NEG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Angular gyrus</td>
<td>39</td>
<td>3.56</td>
<td>−46</td>
<td>−60</td>
<td>36</td>
</tr>
<tr>
<td>L</td>
<td>Anterior insula</td>
<td>48</td>
<td>3.15</td>
<td>−42</td>
<td>12</td>
<td>−4</td>
</tr>
<tr>
<td>L</td>
<td>Frontal operculum</td>
<td>40</td>
<td>3.04</td>
<td>−42</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>L</td>
<td>Lateral OFC</td>
<td>36</td>
<td>3.33</td>
<td>−38</td>
<td>18</td>
<td>−14</td>
</tr>
<tr>
<td>L</td>
<td>Medial substantia nigra</td>
<td>NA</td>
<td>3.76</td>
<td>−19</td>
<td>−14</td>
<td>−10</td>
</tr>
<tr>
<td>L</td>
<td>Occipital cortex</td>
<td>18</td>
<td>4.08</td>
<td>−34</td>
<td>−84</td>
<td>10</td>
</tr>
<tr>
<td>L</td>
<td>Occipital cortex (primary visual)</td>
<td>17</td>
<td>4.99</td>
<td>−6</td>
<td>−96</td>
<td>14</td>
</tr>
<tr>
<td>L</td>
<td>Posterior hippocampus</td>
<td>27</td>
<td>4.14</td>
<td>−29</td>
<td>−32</td>
<td>−4</td>
</tr>
<tr>
<td>L</td>
<td>Superior parietal lobule</td>
<td>7</td>
<td>3.6</td>
<td>−30</td>
<td>−56</td>
<td>42</td>
</tr>
<tr>
<td>L</td>
<td>Temporal pole</td>
<td>56</td>
<td>3.96</td>
<td>−56</td>
<td>14</td>
<td>−6</td>
</tr>
<tr>
<td>R</td>
<td>ACC</td>
<td>24</td>
<td>3.73</td>
<td>2</td>
<td>20</td>
<td>96</td>
</tr>
<tr>
<td>R</td>
<td>Anterior insula</td>
<td>47</td>
<td>3.44</td>
<td>42</td>
<td>16</td>
<td>−8</td>
</tr>
<tr>
<td>R</td>
<td>Fusiform gyrus</td>
<td>37</td>
<td>3.03</td>
<td>38</td>
<td>−46</td>
<td>−24</td>
</tr>
<tr>
<td>R</td>
<td>Inferior frontal gyrus</td>
<td>48</td>
<td>5.05</td>
<td>54</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>R</td>
<td>Middle frontal gyrus</td>
<td>44</td>
<td>3.04</td>
<td>52</td>
<td>14</td>
<td>36</td>
</tr>
<tr>
<td>R</td>
<td>Occipital cortex</td>
<td>18</td>
<td>5.11</td>
<td>8</td>
<td>−96</td>
<td>20</td>
</tr>
<tr>
<td>R</td>
<td>Parahippocampus</td>
<td>27</td>
<td>4.19</td>
<td>18</td>
<td>−34</td>
<td>−8</td>
</tr>
<tr>
<td>R</td>
<td>Posterior hippocampus</td>
<td>20</td>
<td>4.04</td>
<td>24</td>
<td>−26</td>
<td>−10</td>
</tr>
<tr>
<td>R</td>
<td>Pre-SMA</td>
<td>32</td>
<td>3.94</td>
<td>4</td>
<td>20</td>
<td>44</td>
</tr>
<tr>
<td>R</td>
<td>Cerebellum (right VI)</td>
<td>19</td>
<td>5.08</td>
<td>28</td>
<td>−69</td>
<td>−20</td>
</tr>
<tr>
<td>R</td>
<td>Superior frontal gyrus</td>
<td>8</td>
<td>3.31</td>
<td>0</td>
<td>36</td>
<td>50</td>
</tr>
<tr>
<td>R</td>
<td>Temporal pole</td>
<td>56</td>
<td>3.43</td>
<td>56</td>
<td>20</td>
<td>−20</td>
</tr>
<tr>
<td>NEG &gt; POS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Angular gyrus</td>
<td>39</td>
<td>3.56</td>
<td>−46</td>
<td>−60</td>
<td>36</td>
</tr>
<tr>
<td>L</td>
<td>Posterior insula</td>
<td>48</td>
<td>3.01</td>
<td>−38</td>
<td>−12</td>
<td>−2</td>
</tr>
<tr>
<td>L</td>
<td>Precuneus</td>
<td>7</td>
<td>3.22</td>
<td>−4</td>
<td>−60</td>
<td>52</td>
</tr>
<tr>
<td>L</td>
<td>Superior parietal lobule</td>
<td>5</td>
<td>3.17</td>
<td>−18</td>
<td>−60</td>
<td>66</td>
</tr>
<tr>
<td>L</td>
<td>Superior temporal gyrus</td>
<td>22</td>
<td>3.5</td>
<td>−62</td>
<td>−30</td>
<td>12</td>
</tr>
<tr>
<td>L</td>
<td>STS</td>
<td>48</td>
<td>3.73</td>
<td>−48</td>
<td>−12</td>
<td>−8</td>
</tr>
<tr>
<td>L</td>
<td>Supramarginal gyrus anterior division (TPJ)</td>
<td>40</td>
<td>3.68</td>
<td>−42</td>
<td>−30</td>
<td>40</td>
</tr>
<tr>
<td>L</td>
<td>TPJ (part of operculum cortex)</td>
<td>48</td>
<td>3.41</td>
<td>−58</td>
<td>−38</td>
<td>26</td>
</tr>
<tr>
<td>R</td>
<td>Posterior insula</td>
<td>20</td>
<td>3.79</td>
<td>49</td>
<td>−12</td>
<td>−10</td>
</tr>
<tr>
<td>R</td>
<td>Precuneus</td>
<td>5</td>
<td>3.08</td>
<td>4</td>
<td>−40</td>
<td>60</td>
</tr>
<tr>
<td>R</td>
<td>Superior temporal gyrus</td>
<td>42</td>
<td>3.47</td>
<td>54</td>
<td>−34</td>
<td>16</td>
</tr>
<tr>
<td>R</td>
<td>STS</td>
<td>22</td>
<td>3.5</td>
<td>62</td>
<td>−14</td>
<td>−6</td>
</tr>
<tr>
<td>R</td>
<td>Supramarginal gyrus posterior division (TPJ)</td>
<td>48</td>
<td>3.57</td>
<td>54</td>
<td>−36</td>
<td>26</td>
</tr>
<tr>
<td>R</td>
<td>TPJ (part of operculum cortex)</td>
<td>48</td>
<td>3.79</td>
<td>62</td>
<td>−28</td>
<td>22</td>
</tr>
</tbody>
</table>

This table reflects the contrast positive expectation error compared to negative expectation error and shows the local maxima of clusters surviving cluster correction Z > 2.3, p < 0.05 in MINI space. Cortical and subcortical regions were identified using the Harvard-Oxford Probabilistic Anatomical Atlas and Allin et al., 2007, while the cerebellar regions were identified using a probabilistic cerebellar atlas (Collins et al., 2008). Abbreviations: TPJ = temporal-polar junction; SMA = supplementary motor area; STS = superior temporal sulcus; OFC = orbitofrontal cortex; ACC = anterior cingulate cortex.

actual isolation (Green, 1956). The source of this mechanism has been the focus of considerable research in the memory, attention, and cognitive control literatures and has even served as one of the primary paradigms in studying cognition in preverbal infants (Finnie, 1964). Detecting novel stimuli embedded within more frequent background stimuli has been extensively studied using a paradigm known as the “oddball task”. The ability to detect novel stimuli or expectation violations is associated with a distinct event related potential that occurs about 300 ms after the novel stimulus onset (Sutton et al., 1965). While the precise neural origins of this signal are still being worked out (Bangardt and Reiner, 2003), the hippocampus (Knight, 1996; Tuning et al., 1986), ACC (Baudela et al., 1995; Berns et al., 1997), and insula (Linden et al., 1999; Kielh et al., 2001) have been shown to be reliably involved. Our findings support the existence of this more general system that detects violations of expectations and, importantly, extends these ideas into the domain of social interactive decision-making and neuroeconomics.
In terms of our imaging results, we also found distinct networks consistent with systems previously identified with expectation violation and memory. Viewing faces of partners whose offers exceeded expectations was associated with bilateral posterior hippocampal/parahippocampal, bilateral anterior insula, pre-SMA/ACC, and striatum. These regions have previously been associated with expectation violation, social cognition, and memory. Considerable research has demonstrated that posterior hippocampal regions are critical in successful recognition memory (Eldridge et al., 2006; Yonelinas et al., 2005). In addition, patients with hippocampal damage have been demonstrated to have a selective impairment in generating the characteristic P300 and autonomic skin response following unexpected events while the processing of expected events remained preserved (Knight, 1993). Our observed activation in the striatum, which included the caudate and nucleus accumbens is consistent with the literature on reward prediction error (Schultz et al., 1997) and repeated play with cooperators (Billing et al., 2002). Participants that exceed initial expectations are associated with a positive prediction error, which is likely to promote further cooperation with these partners in the future (Billing et al., 2002; Delgado et al., 2005; King-Casas et al., 2005). We also observed activity in the pre-SMA area/ACC and bilateral anterior insula. This network appears to be functionally coupled (Fox et al., 2005; Margulies et al., 2007; Craig, 2009), and has consistently been associated with detecting violations of expectation in a multitude of contexts including stimulus frequency (Bever et al., 2001), changes in sequences (Berns et al., 1997; Huet et al., 2002) and multi-modal sensory changes (Downar et al., 2000). Thus, viewing pictures of partners who exceeded participants’ expectations resulted in increased activity in regions of the brain that have been consistently associated with detecting violations of expectations in paradigms investigating more basic aspects of novelty detection and also in successful memory retrieval.

In contrast, when viewing partners that had made lower offers than the player had expected, we found activation in bilateral TPJ, right STS, bilateral posterior insula, and precuneus. These regions have been implicated in a variety of processes including memory, expectation violation, social cognition, and pain processing. The TPJ has been shown to be involved in expectation violation (Downar et al., 2005) and plays a key role in generating the brain’s P300 novelty response (Knight, 1989) and in orienting attention (Corbetta et al., 2000). In addition, the TPJ has received attention for its role in thinking about others’ mental states (i.e., theory of mind) (Saxe and Kanwisher, 2003), but it is currently unclear if these two processes can be explained by a more general cognitive process (Mitchell, 2008). Thus, viewing pictures of partners who offered less money than was expected was associated with a region of the brain that has been implicated in both social cognition and novelty detection. We also observed increased activity in the right STS, a region which has been hypothesized to detect and evaluate intentions and actions of other’s behavior (Frith and Frith, 1999; Saxe et al., 2004). This region has been associated with updating expectations about an opponent’s strategy based on their behavior in a repeated Inspection game (Hampton et al., 2008). In addition, the STS and TPJ have been demonstrated to be involved in social prediction error – specifically in both making a prediction about the value of a social partner’s advice and updating this prediction after feedback (Beltens et al., 2008). We also observed activity in the bilateral posterior insula, which has been primarily associated with interoceptive processing, that is processing of the physiological condition of the body (Craig, 2002). This region is reliably associated with processing pain from external stimuli (Koyama et al., 2005; Singer et al., 2004) and also direct cortical stimulation (Ostrowsky et al., 2003) and suggests, at least tentatively, that viewing pictures of participants who offered less than expectations is perhaps associated with processing a negative somato-visceral state. Finally, the precuneus has been demonstrated to be involved in memory, and social cognition (Cavanna and Trimble, 2006). The range of these memory processes is diverse and includes episodic memory retrieval (Shallice et al., 1984; Tulving, 1994), recognition memory (Henson et al., 1999; Yonelinas et al., 2005), source memory (Landstrom et al., 2005), and autobiographical memory retrieval (Aldis et al., 2004). The precuneus has also been involved in mentalizing perceived intentionality (den Ouden et al., 2005), and in reasoning about another’s beliefs (Saxe and Kanwisher, 2003). These results suggest that viewing a picture of a partner that offered less money than was initially expected is associated with brain regions that have been thought to be involved in processing negative somatic states, mentalizing about another person’s beliefs, updating expectations about behavior, and memory. Interestingly, despite the methodological differences between the present study and that of Singer et al. (2004), both studies yield somewhat similar results. Singer et al. (2004) had participants repeatedly make a gender discrimination on photographs of partners with whom they had previously encountered in a repeated Prisoner’s Dilemma Game. In contrast, our study employed a single shot design using the Ultimatum Game and a recognition task that included an equal number of old and new faces. Our imaging analyses focus on partners that made offers that were either higher or lower than the participant’s initial expectation, while Singer et al. (2004) independently compared partners that were cooperators or defectors to partner’s associated with null games. Despite these methodological discrepancies both studies identify the anterior insula and different components of the striatum as being linked to partners with positive associations (i.e., cooperators or positive expectation error). While Singer et al. (2004) only found vmPFC associated with defectors, we observed activity in the posterior insula, STS, TPJ. Because our study was explicitly designed to study social memory, we were also able to observe activity in regions that have previously been associated with memory retrieval – most notably the hippocampal/parahippocampal regions and precuneus. Thus, our results extend those of Singer et al. (2004), by providing a different perspective on cheat detection (i.e., expectation violation) as well as methods that are more conducive to studying social memory.

In contrast to our behavioral results, we did not observe activation in the brain for the interaction between initial expectations and expectation error. One possible reason why we failed to observe a significant finding for the imaging interaction is the combination of a stringent statistical threshold and a lack of
statistical power. While our mixed effects procedure can account for unequal variances in the interaction analysis, there is an under representation of cases in which there are low initial expecta-
tions in this sample. Thus, while our behavioral analyses utilized a p-value of p < 0.05, our imaging analyses were restricted to a more stringent criteria to account for multiple comparisons. Indeed, when the statistical threshold is dropped to a more liberal p < 0.005 uncorrected level, we find almost identical results to our expectation error analysis including bilateral anterior insula, SMA, bilateral posterior hippocampus, bilateral caudate, left vent-
tral putamen, and bilateral amygdala.

As noted, we did not observe any evidence for a significant behavioral or imaging finding for enhanced memory for part-
ers based on the amount of money they offered. Nor did we observe evidence of a salience detection system, in which part-
ners who made either extremely fair or unfair offers were bet-
ter remembered. The literature on cheater detection is rife with conflicting results, with some studies finding enhanced memory for cheaters (Mealy et al., 1996; Oda, 1997; Chiappe, et al., 2004), others finding enhanced memory for altruists (Barclay and Lafamme, 2006), and others, like the present study, finding no significant differences (Barclay and Lafamme, 2006; Mehl and Buchner, 2008). The more general expectation deviation system outlined here is a potential mechanism that could account for the inconsistent results in this domain. However, at present it is not immediately clear why we identified two distinct expectation violation systems that track with the valence of the deviation. In addition, it is important to note that despite the attractiveness of the expectation violation hypothesis, our imaging results do not necessarily rule out the possibility that some of the regions associated with negative expectation violations may be involved in cheater detection. Addressing these issues could be fruitfully explored further in future research.

Like all studies, there are a number of limitations that should be considered before drawing firm conclusions from the results. First, it is always difficult to interpret null findings. Our lack of significant results for offer amount on memory cannot necessarily be interpreted as an absence of an effect. Neuroimaging studies are inherently underpowered (Mumford and Nicholas, 2003) and as such are greatly at risk for making Type II errors. Second, it is unclear if participants actually believed they were engaged in a real social interaction. While participants were not explicitly probed about whether they believed that they were playing with a real partner, no partner expressed any doubt, nor did their behavior deviate remarkably from other published studies that utilized actual human partners (Camerer, 2003). In addition, consistent with previous research (Sanfey, 2009), most participants indicated that they expected their partners to make fair offers. Finally, it is somewhat of an open question as to whether a single UC interaction is sufficient to label a partner as a “cheater.” It is possible that making such a judgment would require multiple interactions. However, a single interaction would be enough to develop an initial impression and there is considerable evidence demonstrating that participants generate negative emotional responses in response to a single unfair offer (Sanfey et al., 2003; van ’t Wout et al., 2006).

In summary, our results support a more general system that detects violations of expectations as opposed to a more specialized system engineered to detect cheaters. We found that participants on average were no better or worse at remembering partners who made either fair or unfair proposals, but that individual participants exhibited selectively better memory for partners who made offers which violated their initial expectations. Two dissociable neural systems were found to be underlying this effect. While both systems have been previously associated with expectation violation, social cognition, and memory, these regions tentatively suggest that there is distinct processing for positive and negative expectation violations. Positive expectation violations are associated with a system that may incorporate error detection, conscious awareness of the error, reward processing, and enhanced recognition memory, while negative expectation violations are associated with expectation violation, evaluating intentions, pain processing, and autobiographical episodic memory. By incorporating the strengths of several fields—the tasks of behavioral economics, the methodologies of psychology and the sophisticated techniques of neuroscience—we can uniquely investigate how social exchange operates, not just in terms of the immediate decisions but also how these interactions can reverberate over time.

ACKNOWLEDGMENTS

We would like to thank Mascha van’t Wout and Katia Harle for their help with the collection of this data and the two anonymous reviewers for their thoughtful suggestions. This research was sup-
ported by NIMH R01MH077058 and NIA R21AG030768.

REFERENCES


Bandura, A., Helger, J., Bost, J., and Cloe, J. (1995). Intrasubjec-
tial potentials to rare target and distrac-


Campana, C. F. (2003). Dictator, ultima-

tional anatomy and behavioral correla-

Chiappe, D., Brown, D., Dow, R., and Jones, I., Rodriguez, M., and McCalloch, K. (2004). Cheaters are looked at longer and remembered better than...
cooperates in social exchange situa-
tions. Dev. Psychol. 2, 186-129.
Corbett, M., Kinnel, J.M., Oliff, G.,
McAwen, P., and Braithwaite, G.L.
(2009). Voluntary spinning is dis-
sociated from target detection in
human posterior parietal cortex. Nat
Neurosci. 12, 137-140.
exchange: the natural selection of
shape and how humans react to it with
the Wason selection task. Cognition 31,
187-278.
tive adaptations for social exchange. In
The Adopted Mind: Developmental Psy-
chology and the Generation of Cultures, J.
Barlow, L. Cosimides, and J. Tooby, eds
(New York, Oxford University Press),
pp. 163-228.
intercept the sense of the physi-
ological condition of the body. Nat
Craig, A. D. (2009). How do you feel-
now? The anterior insula and human aware-
Crockett, M. J., Clark, L., Tabbata, G.,
Lieberman, M. D., and Robbins, T. W.
(2007). Serotonin modulates behavioral
reactions to unfairness. Science 317,
1705-1708.
de Quervain, D., Fleischer, U., Troyer,
V., Scheithammer, M., Schedel,
substrates of altruistic social punish-
Dolcos, M. R., Frankl, R. H., and
character modulate the neural systems of reward during the trust game. Nat.
Neurosci. 9, 1618-1619.
don Ouden, H. E., Frith, U., Frith, C.,
and somatosensory processing in the
D’Hondt, L., Balsdrum, I., Fioleau, I.,
A probabilistic model of the human cerebral
Downar, L., Cranley, A. P., Milikowsk,
cortical network for the detection of
changes in the sensory environment.
Nat. Neurosci. 12, 177-182.
Erdkamp, L. L., Kinnel, J. M., Oliff, G.,
185
152
148
A Theory of Fairness, Competition, and
Foa, M. L., Snyder, A. Z., Vincent, J. L.,
Corbett, M., van Essen, D. C., and
Rakic, P. M. E. (2005). The human brain is
intrinsically organized into dynamic, auto-regulated functional
102, 8675-8678.
Friston, K. J., and Thoir, U. (1999). Inter-
acting minds—a Biological theory. Science
286, 1490-1495.
specific reasoning: social contracts,
coaching, and perspective change. Cognition 45, 137-171.
Glachman, P. W., Carretero, C.,
2009). Neuroeconomics: Decision-
Green, R. T. (1956). Surprise as a factor in
the von Mises effect. J. Exp. Psychol.
32, 216-234.
Grath, W., Schmitter-Schwarz, D., and
analysis of ultimatum bargaining. J. Econ.
Behav. Organ. 3, 167.
Hampton, A. N., Bosworth, P., and
O’Heeny, E. P. (2004). Neural corre-
lates of mentalizing: related computa-
101, 6745-6748.
Bilateral subdural biosignals of social
economic decision making in the Ultimatum Game. Emotion 7, 876-881.
Henson, R. N., Rugg, M. D., Shahly,
The confrontation and familiarity in recogni-
tion memory: an event-related fMRI
Houstek, A. M., Shroyer, P. R., and
McCarthy, G. (2002). Detecting patterns in random
dynamic processing of sequence in prefrontal cortex. Nat. Neurosci. 5,
495-496.
ysis from ANOVA to mixed factorial
Kim, B., Lumer, K. R., Ditty, T. L.,
sensors involved in anterior insula
target detection and novelty process-
ing: an event-related fMRI study.
Psychophysiology 43, 134-142.
King, Ursula, C., Bielby, W., and
Montague, P. R. (2005). Getting to
know you: reputation and trust in a two-
player economic exchange. Science
308, 78-83.
Kishig, R., Nose, S., and Lambrecht, L.
(2007). Preview: a numerical decision-
related effect from word + 2. J. Exp.
Psychol. Hum. Percept. Perform. 33,
1250-1255.
Knight, B. T. (1994). Contribution of
human hippocampal regions to novelty
Kluger, B., T., Skaheib, D., Woods, D. L.,
Brain Res. 502, 149-166.
Lateral-temporal decision making after retrotemporal preferential damage:
evidence from the Ultimatum Game. J.
Neurosci. 27, 994-998.
Kohyama, T., McNaughton, L. J., and
Godill, R. (2007). Neural correlates of
experiential-experience of pain when expect-
U.S.A. 104, 12190-12195.
Koln, D. E., Prokhorov, D., Forrest, D.,
Volling, L., Zemella, Y., Gobel, B.,
necromuscle of target detection: an fMRI
study of visual and auditory odd-
Kohyama, T., McNaughton, L. J., and
Godill, R. (2007). Neural correlates of
experiential-experience of pain when expect-
U.S.A. 104, 12190-12195.
Kohyama, T., McNaughton, L. J., and
Godill, R. (2007). Neural correlates of
experiential-experience of pain when expect-
U.S.A. 104, 12190-12195.
Kohyama, T., McNaughton, L. J., and
Godill, R. (2007). Neural correlates of
experiential-experience of pain when expect-
U.S.A. 104, 12190-12195.
Kohyama, T., McNaughton, L. J., and
Godill, R. (2007). Neural correlates of
experiential-experience of pain when expect-
U.S.A. 104, 12190-12195.
Kohyama, T., McNaughton, L. J., and
Godill, R. (2007). Neural correlates of
experiential-experience of pain when expect-
U.S.A. 104, 12190-12195.
Kohyama, T., McNaughton, L. J., and
Godill, R. (2007). Neural correlates of
experiential-experience of pain when expect-
U.S.A. 104, 12190-12195.
on echo time; results from olfactory BOLD experiments. Neuroimage 30, 151–159.
REFERENCES


