A FUNCTIONAL IMAGING STUDY OF WORKING FOR SELF AND OTHER

by

Stephen J. Saletta
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Economics

Committee:

Director

Department Chairperson

Richard E Wagner — Program Director

Dean, College of Humanities and Social Sciences

Date: October 16, 2007

Fall Semester 2007
George Mason University
Fairfax, VA

A Functional Imaging Study of Working for Self and Other

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Stephen J. Saletta
Bachelor of Science
Eastern Michigan University, 2001

Director: Kevin A. McCabe, Professor
Department of Economics

Fall Semester 2007
George Mason University
Fairfax, VA

DEDICATION

To my family and friends who are a constant source of both support and inspiration, and to my scout leaders, Dwight Benner and Bob Wolf who taught me about altruism and working for the benefit of others.

ACKNOWLEDGEMENTS

The intellectual support of family and friends have been of inestimable value in this effort, additionally, I would like to thank: Prof. Bruce Allen, Prof. Terry Chorvat, Prof. Abdullah Dewan, Dr. Arthur Gulick, Prof. Daniel Houser, James Hudler, Prof. Constance Hunt, Dr. Frank Krueger, Prof. Kevin McCabe, Prof. Ryan Oprea, Prof. Russell Roberts. I would also like to thank Hank Reynolds for helping me analyze the imaging data, Sharjeel Jarvaid for managing the behavioral lab, Marcianne Moe for her patience with me and expertise with the magnet, and Christopher Kaighen who came to the desert and supported me on my first sabbatical. For engineering assistance, a soldering iron, and a workbench to build the interface between the MRI and the stimulus response box, I would like to thank Dr. Andy Mitz at the National Institutes of Mental Health as well as the Research Services Branch. I am especially grateful for the support I have received from Genny Flattery, Kate Thompson and Dana Vogel of the Economics Department, Dawn Moore and Rusty O'Conner of the Law School, Christine Donati at the National Institute of Neurological Disorders and Stroke, Jennifer Sturgis and Jane Wendelin at the Krasnow Institute, Bess Dieffenbach at the Office of Research Subject Protections, Ray Loomis, Michelle Morelli and Tina Sims at the Office of Sponsored Programs, and Jenn Arce, formerly of the Office of the Provost. Financial support for my research was provided by the National Science Foundation (SES -0129744) and the following organizations at George Mason University: the James Buchanan Center for Political Economy, the Mercatus Center, and the Office of the Provost.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

A FUNCTIONAL IMAGING STUDY OF WORKING FOR SELF AND OTHER

Stephen J. Saletta, Ph.D.

George Mason University, 2007

Dissertation Director: Dr. Kevin A. McCabe

Altruistic behaviors can be defined as those actions which are costly to self, beneficial to another, and do not convey a benefit from reputation or reciprocity on the part of the recipient. Behaviors which meet these criteria are widely observed in behavioral experiments utilizing the dictator game. It has been suggested altruists may receive direct utility in the form of "warm glow" which offsets the cost of their behavior. Alternatively, it has been suggested that social norms exist which supporting reciprocity and reputation, the salient features of those norms are reproduced in the experimental setting, and altruism will decrease over time as subjects gain experience in the experimental environment. We explore other-regarding behavior while subjects undergo functional magnetic resonance imaging in the context of a modified dictator game where money cost is either replaced or augmented with effort cost. We find behaviorally that subjects are willing to exert effort to benefit their counterpart, but will not expend money, even when the cost to the subject is trivial compared to the gains available to the counterpart.

Neurologically, we find evidence that superior-temporal regions and temporo-parietal junction is active when subjects observe reward accruing to the counterpart but not to self. These regions are frequently implicated in theory of mind tasks where subjects must imagine the mental state of another individual, and in social contextual knowledge tasks, where subjects must access and utilize norms proscribing appropriate conduct in social settings. Our results suggest that regions of the brain associated with social knowledge and interaction are required to interpret those outcomes associated with other-regarding behavior, even in the context of a one-player game where social interaction is absent. This activation pattern is more consistent with a theory that other-regarding behavior is modulated by social norms than the "warm glow" of utility directly experienced from increased payments to the counterpart.

1. Introduction

1.1 Overview

Other regarding behavior has been problematic for economists because it is observed in a robust set of circumstances, and contradicts the standard theory of *homo economics* - the rational and self-interested agent. Economists have responded by incorporating uncertainty, incomplete information, longer time horizons, or incomplete rationality into standard utility theory. (Sugden 1984; Selten 1987) These efforts attempt to explain altruistic behavior as a result of individuals who simply do not understand the payoffs in a particular game, or within the context of an agent who engages in altruistic behavior as a form of social insurance against uncertain future payoffs, i.e. where the cooperative strategy is also the payoff dominant strategy. (Coate and Ravallion 1993) The field of experimental economics has thus far demonstrated that changing the institutional framework or context in which participants interact will significantly reduce but not eliminate other-regarding behavior.

In this paper we review the existing literature from experimental economics on other-regarding behavior in chapter 2 as well as the related studies from cognitive neuroscience utilizing functional neuroimaging methods in chapter 3. In chapter 4 we propose our own model of other-regarding behavior which uses social knowledge to interpret payoff-salient stimuli in the context of a social norm which influences the decision making process. We present a new task similar to the dictator game and present behavioral results in chapter 5. The imaging design is presented in chapter 6, and chapter 7 presents the imaging results and concludes with a discussion of what implications these results may have on existing economic models.

1.2 Perspectives on Other-Regarding Behavior

The fundamental issues behind other-regarding behavior are related to the question of why someone would choose to perform a favor for another individual, why someone would enter into an employment agreement with another party, or why an individual would choose to interact with another individual at all. Given that other-regarding behavior is so frequently observed, why engage in it at all?

Economists have been drawn to this issue because the story about altruism and weaker forms of other-regarding behavior is the story of surplus from voluntary exchange. By exchange, we mean an interaction where is each agent trades away something which is less valuable than whatever that agent hopes to gain from the transaction. The difference between cost and value represents surplus, frequently described in money value or the more abstract term “utility”, which brought into existence *ex nihilo* as a result of the trade.

Because the principle that voluntary exchange is good and benefits all parties of a transaction is so fundamental to the economist's discipline as a positive science, the fields of experimental economics and economic system design have evolved to study means by which to increase the number of transactions in society and hence, the welfare of mankind. The notion of humans organizing into societies for their mutual benefit is not particularly new or controversial. (Hobbes 1946) Such "mutual benefit" can easily be interpreted under the surplus rubric of economists, and under such a framework, extend the study of altruism to the study of how agents maximize and allocate surplus under particular game forms lending themselves to cooperation or, alternatively, exploitation.

The formal modeling of economic behavior examines how individuals interact with economic institutions to affect the volume of exchange in an economy. Other-regarding behavior, loosely interpreted, is of particular importance, because it represents a human strategy for coordinating on cooperative outcomes, which, if we take Hobbes at his word, is what the entire enterprise of human society is about.

What *is* controversial is the notion of *altruism* as opposed to the broader term 'other-regarding behavior.' We will take the common definition of altruism as behavior which is costly, benefits another individual, and where there is no possibility for reciprocity or reputation effects. As reviewed by Burnham and Johnson (2005), four *self-interested* explanations for other-regarding behavior have been proposed: kin selection (Hamilton 1964a; Hamilton 1964b), reciprocal altruism (Trivers 1971), indirect reciprocity

(Alexander 1987), and signaling (Gintis, Smith et al. 2001). Furthermore, Burnham and Johnson assert behaviors arising from individual selection as opposed to group selection are not “genuinely altrusistic”. (ibid., p124) Because of the ongoing debate addressed most recently by Trivers (2005) and Fehr and Henrich (2003) about the definition and characteristics of altruism, we will take a moment to place our inquiry into context.

1.3 The False Dichotomy of Reciprocal Altruism and Strong Reciprocity

From the perspective of pure behavioral observation, the division between recirpocal altruism (Trivers 1971) and strong reciprocity (Bowles and Gintis 2004) is primarily one of the evolutionary origins of altruistic behavior rather than whether or not altruistic behavior actually exists. As we will explain, this debate is primarily about whether group selection or individual selection drove the development of altruistic behavior, and whether or not group selection presents a valid framework for the evolution of any trait whatsoever. Proponents of strong reciprocity have linked the theory of group selection and have almost gone as far as to claim exclusive dominion over the term ‘altruism’ as something which *can not* have as a component any motivation stemming from concern over the agent’s own reproductive success. (Fehr, Fischbacher et al. 2002) (Fehr and Henrich 2003).

Because this debate is primarily an attack on the theory of group selection rather than human behavior, it has paradoxically led some supporters of reciprocal altruism to claim that because group selection theory is false, it is a logical fallacy to describe any observed motivations as altruistic. (Burnham and Johnson 2005) The alternative explanation offered is that behaviors which appear to be altruistic are maladaptive and driven by the same emotional mechanisms which enable the self-interested pro-social behaviors (direct altruism, kin selection, reputation, etc…) (Burnham and Johnson 2005; Trivers 2005).

There are, of course, counter-arguments, and counter-counter-arguments, which we will not delve into here. Our immediate purpose, however, is to highlight the common ground between both groups; namely, behaviors which appear to be other regarding are driven by emotional states. In the case of reciprocal altruism, this emotional state is self-serving in because those agents who do not satisfy their own emotional reaction by punishing a defector will suffer the harm associated with defectors who obtain a fitness advantage as a result of their unjust deeds. (Table 1)

Table 1. Quotes from proponents of strong reciprocity and reciprocal altruism
Both forms of reciprocity are products of a fundamentally emotional process.

“We are expected to react negatively to unfair offers by others, not out of envy of their extra portion, but because they chose to inflict this unfair offer on us and the unchallenged repetition of such behavior is expected in the future to inflict further costs on our inclusive fitness.” p79 … “Anger is not a mere emotion, it is (costly) physiological arousal for immediate aggressive action.” p80 (Trivers 2005)	“… Thus, although the existence of emotions affects our tastes, humans seem to cognitively weigh the costs and benefits of different courses of action, irrespective of whether they are emotion-driven or not. If this argument is correct and if emotions like guilt, shame and anger are driving forces of strong reciprocity strongly reciprocal behavior patterns should quickly respond to changes in the costs and “benefits”. Experiments strongly confirm this argument…” (Fehr and Henrich 2003) p10

We believe that arguments put forth from the reciprocal altruism camp can be reduced to: (1) the conditions necessary for group selection did not exist in the evolutionary environment, (2) individual selection is the only possible alternative for trait selection, (3) individually selected traits are by definition, self-interested, (4) all behavior that appears to be altruistic must be the product of self-interested traits. We are left with the question of whether or not altruism is defined by the evolutionary selection mechanism (group vs. individual) or the three observable characteristics of a decision’s consequence: (1) costly to self, (2) beneficial to another, and (3) absence of reputation or reciprocity effects (Fehr and Henrich 2003; Burnham and Johnson 2005). Because neither functional MRI nor behavioral experiments can demonstrate whether or not the appropriate migration patterns or group size were present in the evolutionary environment (Bowles and Gintis 2003), studies such as the one we present in section 4.2 will not be useful in resolving this question.

The reciprocal altruism vs. strong reciprocity dispute can be useful in our own discussion and definition of the term “preferences”. The notion of revealed preference is that when we observe someone exhibiting a behavior, and label that behavior altruism, that person has exhibited a preference for altruism. fMRI *does* have the capability to observe which neural networks are active during a particular decision making task. If there is agreement that some form of altruism is consistently observed in experimental settings, the next step (and perhaps underlying dispute) will be to determine the strength or “innateness” of that preference.

It is important to note that not all preferences are created equal. While a preference for fairness is like a preference for money in that individuals are capable of trading off these goods against each other as well as all attainable consumption bundles, they are different in that the ability to appropriately attach value to money does not require the on-line access of social constructs. Although humans are capable of trading off a wide variety of goods like food, reproductive opportunities, and world peace, we believe that the ‘innateness’ of the preference should be inversely related to the neural activity which is concomitant with reward system activation, e.g. in the case of altruism, more complex cognitive processes when processing reward accruing to another should support the more complex utility functions attempting to explain how the utility of others is factored along with other kinds of knowledge about the other individual, social norms, or other kinds of variables that are not required to evaluate utility for self. (See section 7.6) As the study of social preferences progresses, we believe it will become more important to qualify observed preferences along this axis.

The present challenge is to develop a notion of other-regarding behavior which accounts for the observations about anonymity and property rights can reducing altruism, (Kahneman, Slovic et al. 1982) and the classification and accounting of residual, other-regarding behavior which persists under the three conditions for altruism mentioned earlier. (Andreoni and Miller 2002)

Cognitive neuropsychology will inform these models with an established literature correlating specific regions and systems in the brain with the functions and activities of daily living. (Hendelman 2006) Relating other-regarding preferences to psychology provides the additional benefit of including established information on the biology of the brain and how it influences decisions in other contexts, which can be compared to the economic models at hand.

Although neuroeconomics is a relatively new, the methodology has proved to be relatively useful in implementing paradigms from experimental economics and imaging methods from cognitive psychology to explain those theories in both disciplines where observed behavior deviates from expected predictions. (Kenning and Plassmann 2005) The goal of economics, cognitive neuroscience, and efforts all along the spectrum is to explain the relationship between stimuli and observed behavior through the unobservable processes of human thought. In the case of behavioral economics, the utility function represents this internal process, and we suggest that economists adjust the model utility function on the basis of how experimental manipulations produce changes in observed behavior. In cognitive neuroscience, the goal is to learn what regions of the brain are involved in a particular task; instead of a utility function, psychologists seek a consistent network of brain regions associated with a particular type of task.

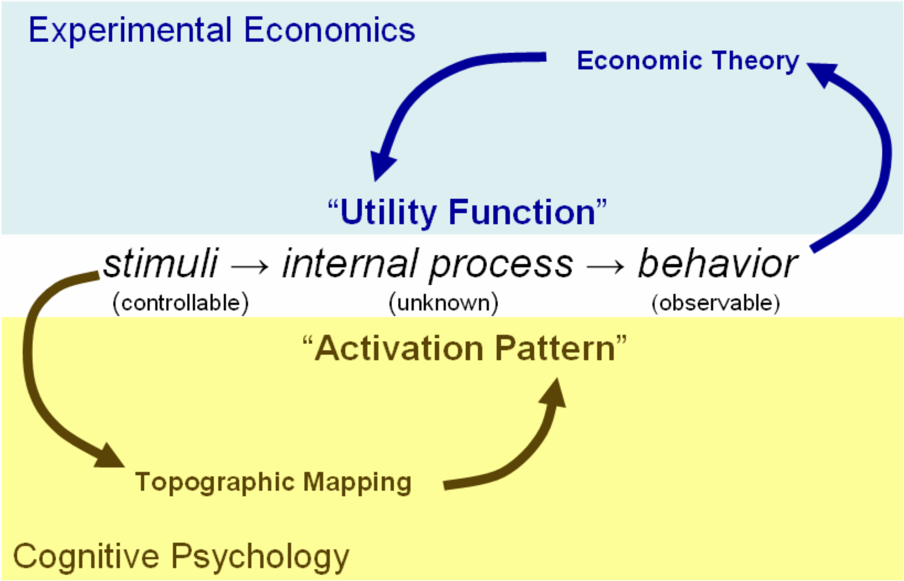


Figure 1. Emotional state model
Economists and psychologists use different names to describe the same internal process between stimuli and behavioral observations. Describing the unseen with the seen is a challenging enterprise and requires vigilance in design and interpretation.

Although fMRI allows the observation of brain activations, on it's own, fMRI cannot explain why the subject is pursuing a specific behavior. Integrating both economic theory and the brain topography of cognitive neuroscience allows the researcher to develop a hypothesis about the regions of the brain that might be implicated in a particular utility function suggested by economic theory on the basis of the domains of information associated with the functional form being tested.

1.4 The Challenge to Homo Economicus

By demonstrating the existence of other-regarding behavior which appears to be altruistic in nature, experimental and behavioral economics has issued a serious challenge to *homo economicus*. One approach is that certain situations invoke behavior which is not rational, similar to a notion of classical conditioning where certain game forms or contexts evoke altruistic behavior. (Fehr and Henrich 2003) In this perspective, parents and other authority figures punish or reward the appropriate behavior in childhood, which is reinforced later in life with weaker, but significant social cues. Ultimately, this leads to the notion that rationality is a variable dependent on the context of exchange.

Other studies incorporate equity as a *preference* which is optimized over. (Bolton and Ockenfels 2000; Engelmann and Strobel 2007) Preference models are different because rather than a response which is evoked by context, altruism is something which is weighed along with other variables such as own money income. The implication is that the economic decision making process of optimization is expanded to incorporate variables that are inherently non-economic, i.e. virtue is its own reward.

The conditioned model and preference model both share the notion that altruism may be maladaptive for the individual, and rely on group selection as a significant contributor to the evolution of these traits. (Burnham and Johnson 2005) The *second* main theme is that of individuals optimizing over a longer time horizon, with all activity described in terms of its effect on long-term income, a proxy for reproductive success. (Barkow, Cosmides et al. 1992) Under this model, the altruistic behavior in the lab is spurious, and the function of mechanisms which are actually adaptive in a more ecologically rational environment. (Trivers 2005) This does not necessarily exclude notions of "interdependent preferences" where individuals value the preferences of others. (Sobel 2005) A long run equilibrium model only suggests that those preferences, when properly factored into a utility function, should produce accurate predictions of behavior that confers a long run benefit on those who express that preference. It also leaves open the possibility for those preferences to generate error in an internal prediction function that an individual might use to predict the likelihood of cooperation or reciprocation from a trading partner, and hence make a decision as to whether or not to exchange with that person. (Willinger and Ziegelmeyer 2001)

Prediction error in this context can result in suboptimal outcomes arising from exploitation; a potentially friendly trading partner turns out not to be the case. The long run equilibrium model also leaves room for behaviors that were adaptive in the evolutionary environment but now prove to be maladaptive. ‘Cleaning your plate’, i.e. consuming all of the food available in the environment, may have been adaptive in the evolutionary environment of relative nutritional scarcity, but it has been argued that evolutionary strategies for dealing with uncertain nutritional availability are implicated in the widespread obesity observed in developed nations. (Ulijaszek 2007) Proponents of reciprocal altruism argue that if traits which were adaptive in the evolutionary environment can make us fat, they can also induce maladaptive behavior, like giving in the ultimatum game, in the laboratory. (Burnham and Johnson 2005; Trivers 2005)

Early attempts to incorporate observations about altruism into economic theory tended toward the notion that because the standard economic model does not include non-economic values like fairness, and people do in fact behave in a way that appears to pursue these values, economic models are incorrect. (Rabin 1993; Conlisk 1996; Yee 1997) While values like fairness have now been factored in to decision making models, this addition now requires a mechanism to quantify the impact of non-economic preferences on behavior. Adding non-economic and non-quantifiable preferences as independent variables only serves to point out the current limits of economists’ understanding of human behavior, and economists will always be wary of models incapable of yielding predictions based on observable variables.

There now seems to be general agreement that non-economic factors do, in fact, influence human behavior. What fMRI analysis can bring to bear on this problem, is the correlation of particular classes of games (such as the investment game or the ultimatum game) with brain regions to identify how those decisions rely on social, emotional, and reward mechanisms to generate an output. In order for the argument agreed on by proponents of both strong reciprocity and reciprocal altruism that emotions play a significant role in other-regarding behavior (see section 1.3), we must demonstrate the involvement of emotional brain regions in those decision making tasks. If economists have become more open to notions of non-economic preferences, is because fMRI data have provided some support for those behavioral studies positing the involvement of social and emotional systems across the widest variety of themes: reward (Knutson, Adams et al. 2001; Knutson and Cooper 2005; Kuhnen and Knutson 2005; Oya, Adolphs et al. 2005; Delgado, Labouliere et al. 2006; Padoa-Schioppa and Assad 2006), trust (Benjamini and Hochberg 1995; McCabe, Houser et al. 2001; Decety, Jackson et al. 2004; Rilling, Sanfey et al. 2004; Damasio 2005; King-Casas, Tomlin et al. 2005; Kosfeld, Heinrichs et al. 2005; Delgado, Labouliere et al. 2006), and altruism (Hoffman, McCabe et al. 1996; Sanfey, Rilling et al. 2003; de Quervain, Fischbacher et al. 2004; Moll, Krueger et al. 2006; Tankersley, Stowe et al. 2007).

In section 4.2 we propose a model similar to one presented by Cox, Friedman and Gjerstad (2007) (see section 7.5) where learned mental constructs like social values and social norms emerge to reinforce non-physical assets that have long-term economic value. Assuming that agents live in a society with significant possibilities for coordination gains, instead of having a preference for reciprocal altruism, or a preference for punishing non-cooperators, we posit that agents value non-physical assets like reputation and security. People do not value justice, per-se, but rather the long term value of living in a just society. These non-physical assets are, in many ways, the factors of production that facilitate the attainment of coordination gains. The value of "justice" itself, is a secondary reinforcer which, invokes an emotional state that binds short term significance to a secure environment or positive reputation as a trading partner – non-physical assets that convey significant long-term value to the organism.

1.5 Definitions

Before proceeding further, it will be necessary to define both 'rationality' and 'altruism', or other-regarding behavior. Rational has traditionally referred to the use of thought or reason in arriving at a conclusion. Reason itself involves proceeding from facts and principles to conclusions and, under this definition, would explicitly exclude concepts of "evolutionary rationality" where the organism itself has no knowledge or is not intuitively or consciously implementing the principles which are being used to deem a particular action rational. This still allows to an appeal to kin-selection (Hamilton 1964a; Hamilton 1964b) or similar arguments in humans, but only on the grounds that actions which benefit kin provide a positive change in utility (from warm glow, or otherwise) in the present. It also excludes from rationality actions which yield unexpected returns in future periods, i.e. if the action yields a payoff that is greater than its cost, but the individual is surprised by that outcome, the behavior is not rational on the individual level. This provides, perhaps, the largest domain for social norms for which we do not currently have an explanation. For a norm to be rational to the individual, his expected value must exceed the expected cost. For a norm to be rational for the group, however, this is not the case. Norms can be rational to the group if they, in fact, confer a competitive advantage to that group, even if the individual enforcing or practicing that norm is not made better off for it. Not all social norms need be rational either. There are certainly some activities and values that are enforced as social norms but convey a disadvantage to groups who adopt them.

As we discuss in section 1.3, altruism is defined by behavior producing an outcome with three characteristics: (1) it must be costly to the decision maker (2) it must benefit another, and (3) it must occur in the absence of reciprocity or reputation effects. Altruism also includes the case where individuals have a preference for their estimation of benefits generated on behalf of the commonweal. If an individual contributes to a charitable cause which, in and of itself, does not experience utility, however he may reasonably expect that organization to confer a positive benefit on others in which case he can be considered to have altruistic preferences towards the beneficiaries of that organization.

Based on the three criterion enumerated above and elsewhere, we exclude from altruism actions taken to avoid sanction or improve a reputation that is expected to yield future economic benefit. Here too, we imagine that there may be multiple components to a behavior: a man may buy his wife a present for Valentine’s Day partially because his utility increases directly as a function of his wife’s, partially because his expectation is that a gift will confer positive reputational effect which his wife will reciprocate in the near future with favor, and partially because he expects not doing so will almost certainly result in negative consequences. The first may be considered altruistic, while the latter two are self interested. It is also assumed that all actions have a cost; even opening the door for a stranger can be considered a form of altruism.

2. Economic Arguments for Other-Regarding Behavior

2.1 Introduction

Researchers operating from the perspective of economics have first and foremost sought to introduce other-regarding behavior into the utility function. This project has been undertaken through means of theory, experiment, and computational modeling in an attempt to explain why such behavior exists in conjunction with a more descriptive inquiry into the rules which might govern altruism in individuals. A brief selection of studies which have clear implications for cognitive neuroscience are reviewed here.

2.2 Warm Glow

Andreoni formalized the notion of altruism with the term "Warm Glow" within the framework of public goods. (1990) To put the public goods issue in context, briefly stated, economic theory holds that individuals should free-ride on public goods which are voluntarily provided by the community. In its strictest form, no individual should ever make a contribution to such an enterprise which is clearly at odds with the existence of the myriad donor-supported non-profit organizations which exist in the modern environment.

The warm glow model provides three sources of utility: private consumption, consumption of the public good, and utility arising from the level of one's contribution to the public good. Because the model describes a public goods environment, it presents us with a curious vocabulary. In Andreoni's model, individuals can allocate their endowment, *w,* between individual income, x_i, and their gift to the public good g_i*.* The total public good is represented as *G*, the sum of individual gifts. Utility, is derived from three sources: own income, one's enjoyment from the total public good, and warm glow, which is one's own level of contribution to the public good.

$$w_i = x_i + g_u \qquad \textit{(2.1)}$$

$$G = \sum_{i=1}^{n} g_i \qquad \textit{(2.2)}$$

$$U_i = f(x_i, G, g_i) \qquad \textit{(2.2)}$$

For Andreoni, the pure altruist is someone who enjoys use of the public good but derives no utility from his own contribution to the public good ($g_i=0$) i.e. the altruist gives only so that others may enjoy the public good. An impure altruist enjoys both the public good and the contribution to the public good. We frame this distinction as one where the utility of the pure altruist results from the public good, whereas the utility of the impure altruist also introduces a utility component which is based on the gift. A pure egoist does, in fact, contribute to the public good, but only receives a benefit from the contribution and does not enjoy the public good. ($g_i>0$, $G=0$)

The warm glow model provides us with a formalization of how those internal neural processes (like emotional states) that are correlated with the act of giving (g_i) might be incorporated into a utility function. In our own model of social norms influencing the decision of whether or not to work for another, warm glow from following the social norm can also be extended to disutility arising from negative emotional states associated with defection, on the part of the agent or the agent's trading partner, from those norms which govern charity, favors, or work for hire.

Andreoni's distinction among motivations for other-regarding behavior is important because the phrases "impure altruism" and "egoist" convey a pejorative context, but it is precisely the egoist component of Andreoni's model which is described as altruism, or other regarding behavior, in most current literature. Stated differently, current studies would not describe an agent who contributes to a public good only for the benefit that they derive from that public good as a pure altruist. We think this problem can be partially avoided by describing behavior, rather than intentions. Andreoni's warm glow is very specific to the case of *why* someone is engaging in a particular behavior.

We are interested primarily in describing *what* results from the behavior. Those results which appear to benefit another are altruistic or other-regarding, while those that benefit the individual are self-interested. We believe it is better to label the outcome as altruistic, beneficent, evil, or otherwise, and to describe the neural activations as being associated with behavior described by non-economic concerns. We are aware that intentionality is an important component of social decisions (Hoffman, McCabe et al. 1994), however our view on the goal of neuroeconomics is not one that aims to change intentions, but rather to modify institutions in such a way that realized outcomes provide more surplus from exchange.

While initially, warm glow (specifically, impure altruism) was used to describe the majority of observed other regarding behaviors in both experimental settings and real-world behaviors, experimental evidence from Andreoni as well as Houser and Kurzban (1995; 2002) conclude that approximately half of the altruism which is observed experimental settings can be reduced to confusion, i.e. experimental subjects don't understanding how incentives are structured in games traditionally associated with other regarding behavior such as the voluntary contribution mechanism. (Isaac and Walker 1988)

One interpretation of the results on confusion, is that there is a widespread misunderstanding among people in the real world about how the incentives in our day-to-day interactions are structured, and confusion in the lab is an accurate observation of that confusion. Another interpretation is that individuals are not confused about the incentives in their daily life, but rather the confusion data represent an implicit critique of experimental economics in the failure to accurately represent those incentives. Although both Andreoni, and Houser and Kurzban (1995; 2002) report that confusion describes a significant portion of other-regarding behavior on the lab, confusion falls over time and there is a significant level of other-regarding behavior that confusion cannot account for.

The significance of “warm glow” is the notion of alternative forms of behavioral reinforcement that can be derived in addition to the usual utility resulting from income to self. Even authors like Burnham and Johnson (2005) who are highly critical of the strong reciprocity model suggested by Fehr, Fischbacher, et. al. (2002) concede that a form other regarding behavior exists; while the notion of non-economic values factoring into the utility function has been hotly debated, as we discuss in section 1.3, the current debate concerns whether or not something like warm glow could have evolved as the result of individual versus group selection.

In addition to the theoretical developments resulting from this literature, the methodological contribution associated with the confusion research makes a strong case for the inclusion of a “confusion control” in any experimental design. (Andreoni 1995; Houser and Kurzban 2002) The likelihood of a subject misinterpreting the incentives will increase with task complexity. As economists and neuroscientists look towards the more subtle components of human behavior, this issue will likely remain of concern. Control conditions for motor response are common, and although confusion is difficult to “control” in the traditional sense of the word, the design presented here in section 5.1 includes a condition to detect confusion by presenting an option where decisions are costly to make with no benefit to any party. We will consider responses in this condition as evidence that the incentives have not properly been communicated to the subject.

2.3 Other Regarding Behavior Emerges in the Evolutionary Environment to Obtain Superior Outcomes from Cooperation

We suggest that individuals may have a preference for something beside minimizing one's own cost and maximizing monetary payoffs (section 4.2), and are therefore sensitive to the charge that ours is an *ad hoc* solution which may not apply in most situations. However, it seems that most arguments citing evolutionary pressures on mankind have a similar deficit, and to this end cognitive psychology has, for some time, relied on computational models to buttress their own claims about under what conditions particular traits may or may not have evolved. Economists, particularly macroeconomists, are not strangers to the practice of developing theoretically sound models and fitting human behavior to those models after the fact, and we believe our own model can be informed by looking at how different forms of non-economic preferences influence fitness in a computational environment incorporating various problems in game theory.

Danielson (2002) provides evidence from computational models in which one agent has not only a preference for its own altruism, but also a preference for the altruism of that agent's trading partner. These agents engage in a variety of sequential games, and the model states that in environments with a possibility for gains from coordination and a risk from deception: (1) reciprocal altruism is necessary to achieving those cooperative outcomes, (2) cooperation among reciprocal altruists is not necessarily an optimal or stable outcome, and (3) altruists, that is agents that value the utility of their trading partner who defect on partners with lower levels of altruism, will be evolutionarily successful but in particular games will display behavior which appears irrational.

More generally, the evolutionary explanation argues that other-regarding behavior exists because in environments where there are gains from cooperation, organisms that are capable of other-regarding behavior will out-compete those that are not. What is particularly interesting about Danielson's model, is that in an environment with a wide variety of social problems where agents prefer other agents with a similar level of altruistic preference, the level of that preference will change over time.

Under this model, agent A will cooperate if A "likes" the trading partner with which it is currently matched (B). How much A likes that partner depends on A's own inherent preference for the welfare of any other agent-his altruistic tendencies- and the absolute value of the difference between A's altruistic tendencies and those of B. If this difference is large enough, the absolute value of that will swamp A's altruistic tendencies as well as A's preference for own payoff. In this case A may "spite" B, that is, engage in costly punishment. (A's reproductive success however depends solely on income earned by A.) Agents are motivated by their own payoff, the payoff of their partner, or spite.

In pure mutual advantage games, defection will only occur as a result of "spite". In constant sum games, the pie will be fixed but will never allow an opportunity for spite: if the first mover is behaving in accordance with rational self interest, that agent will give the second mover the option of taking none or a small portion of the surplus, if the first mover is behaving altruistically, it will give the second mover the option of taking all or the majority of the surplus. If one's own payoff is valued more than one's counterpart, than other-regarding behavior will never be rational in this context. In the prisoner's dilemma, agents at the edges of the altruism continuum will spite each other, those in the middle will always attempt cooperation, and those at the lower bound will always defect.

What emerges from this model is a sort of "scissors, rock, paper" equilibrium. The simplification used in this model is that agents are of high, medium, or low altruism. High altruism agents will spite low altruism agents in coordination games, and high agents can cooperate with each other in the prisoner's dilemma, whereas low agents will defect and wind up with lower total surplus amongst them. Medium agents, however, will exploit high agents in the constant sum game, even if only on the margin.

In a repeated interaction of the constant sum game, high altruists will take turns providing their counterpart with a larger share of the surplus, while medium and low altruists will take turns taking their own larger share. The worst that the high altruist motivated by spite towards a low altruist can do in this situation, is act in accordance with self-interest. Furthermore, although we will make the small assumption here that a medium altruist by definition is altruistic enough to take most, but not all of the surplus if offered the choice, when a high and medium altruist interact the medium altruist will cheat; while high altruists reciprocate by first giving the second mover the opportunity to defect and then not doing so, medium altruists will always interact with the high altruist in such a way that medium always gets the larger share of the constant sum.

This computational model argues for reciprocal altruism on the basis that it confers a benefit on its participants, that is, behaving in a manner consistent with this pattern provides the benefits of cooperative outcome while protecting against cheating by those who are not likely to reciprocate. Danielson explains that in the case where spite on the basis of the absolute value between altruism coefficients is relaxed to a condition where high altruists spite low altruists, but not visa-versa, evolutionary pressure drives the coefficient of altruism to its upper bound. (In the continuous model, medium altruists experience spite towards high altruists that limits cooperation between the two groups. Removing the absolute value component of spite allows high altruists to asymmetrically capture gains from coordination.)

The model is limited insofar as it depends wholly on the games that are selected to represent the environment, and while it allows for agents to adjust their preferences in the short term, long-term memory is conveyed only through evolutionary pressures on those parameters. Agents make decisions about how to interact with their counterpart without considering the particular game. That is, preferences about altruism, reciprocity, and parameter similarity are constant for each game. One could imagine an extension that would multiply these preferences by an array of dummy variables representing each type of game, allowing agents to adjust their sensitivity toward these parameters independently in each situation.

In line with our earlier categorization of explanations that rely on contextual modifiers on the rationality behavior vs. optimization of rational behavior over a longer time period, evolutionary agent based models implicitly invoke extended time horizons as those agents compete and adjust their preferences. The repeated interactions with similar games define the Folk theorem environment on which long-run equilibrium models are based. The coefficients on altruism and reciprocity change over time through evolutionary pressure, but remain fixed during each round. A contextual extension that might produce a more successful agent would ask how should context (i.e. the type of game) influence this agent's preference for altruism, where preferences for altruism in a particular game are variable and are a function of the type of game presently being played. This modification would allow agents to be especially sensitive to exploitation in the constant sum game, while still allowing for the attainment of all surplus in cooperative settings.

In the model of brain as a scarce resource, should the brain expend as much effort estimating relative differences of altruistic preference in the constant payoff game where a factor like the absolute altruism of one's counterpart might prove more useful?

One criticism is that such flexibility might reduce the advantages that a pre-commitment strategy provides such as the strategy for high altruists to spite low ones. Indeed, this describes an interesting feature of societies to hire individuals who implement punishment strategies as well as Ernst Fehr’s notion that individuals’ utility functions might include reward for punishing norm violators to offset the cost of imposing such punishment. (Fehr, Fischbacher et al. 2002) However, even in this particular class of computational modeling, it would be possible to impose lagged or sticky preferences which provide an optimal rate of change to those preferences or if it is the case that, for example, a strict strategy of spite towards low altruists will result in a future competitive advantage through evolutionary pressure that will offset one’s own gains in the coordination game, such a preference could still emerge in an evolutionary context. Alternatively, it may be that a small population of agents with a strict spite rule might provide a positive externality to other altruists with a more relaxed one.

This suggestion of a more complicated computational model is not meant to criticize the study at hand, which in its simplicity has surely provided interesting results that are plausible to a wide range of social scientists, but merely to suggest the possibility that an agent based model considering the impact of context on preferences on a game-by-game basis should not be excluded from similar treatment.

If it is true that agents engage in altruism as a signal, and that ability conveys an advantage, it follows that evolutionary pressures would necessitate the creation of a neurological mechanism to reward altruism. Let us briefly return to the constant sum game where high altruists allow each other to take the larger portion of the surplus as the second mover. This strategy is dangerous under a rational self interest model insofar as it allows for the possibility of exploitation. Agents with altruism functions that are sensitive to game type, can cooperate where it is required, but display strict self interest in the constant sum game.

But what if the altruistic strategy in the constant sum game is actually strategy of gift exchange? If there is uncertainty about the possibility of a defection, then that action can be said to have a true cost associated with the expected value imposed by such a defection. This cost might have as an offsetting benefit the value of a signal that a particular individual is, in fact, a cooperator by opening themselves up to this risk. In fact, if and where it is cheap to demonstrate that one is a cooperator, it would make sense for not only cooperators but defectors and other nefarious types to behave like cooperators in order to gain the trust of potential trading partners upon whom they can later defect.

One response to this kind of cheating behavior would be to raise the net cost gift giving to the point where its marginal cost approached the marginal value in detecting true cooperators from defectors. Furthermore, if the large gifts in and of themselves convey significant monetary benefit to the recipients, the true marginal cost of reciprocal gift giving is lowered and would lower the value of such exchanges as a signal of who is and is not a cooperator.

The computational modeling literature has a significant impact on the existence of reputations or rules as an unconscious or implicit strategy or good. Computational agents are obviously unconscious, but their behavior demonstrates a situation where the cooperative nature of the group, i.e. the percentage of group members which cooperate rather than exploit, possesses value insofar as it allows agents to obtain cooperative surplus. The return an individual agent is likely to obtain depends on the characteristics of other agents, and the concept of value associated with this ‘group characteristic’ is a significant component to the models of strong reciprocity discussed in section 1.3.

2.4 Efficiency Wage Hypothesis

The study we present is primarily about why individuals exert effort on behalf of another, particularly in the case where work is costly and they receive no financial benefit for doing so. The efficiency wage hypothesis relates to this problem as an explanation of why observed *wages* exceed those which an otherwise competitive market equilibrium might suggest (Akerlof 1982). Fehr, et. al. studies a scenario where *effort* exceeds equilibrium, a scenario similar to our own, in the efficiency wage context. (1998)

The efficiency wage hypothesis explores the principal agent problem namely, the optimum level of wage, effort & monitoring given their marginal effects on utility to both principal (profit earned by employer) and agent (total compensation earned by employee including rents from shirking, pilferage, etc…) as well as the costs and benefits of engaging in monitoring.

Fehr (1998) studies this problem experimentally in labor markets where employers and employees interact with either incomplete or complete labor contracts. In the incomplete case, both parties negotiate a wage in the first stage, followed by the employee making a decision as to how much effort to exert on behalf of the employer. The labor contract is incomplete because the employee is free to shirk in the second stage of the game, and anonymous matching among participants provides protection from the effects of individual reputation effects. In the complete contract case, the effort level is exogenously defined and automatically enforced. The institution also gives market power to employees by assigning more employees to employee roles than in employer roles. This design has the nice effect of demonstrating that any change in wage from the complete to incomplete environment is a wage above the competitive level revealed in the complete case.

Indeed, Fehr finds that in the Bilateral Gift Exchange (BGE) wages increase under the incomplete labor market –that surplus described as a gift– and that employees reciprocate by providing a higher level of effort than predicted in the case of a one-shot game where employees should defect and provide the lowest amount of effort according to game theory. This result does seem to provide evidence for the case of the efficiency wage as the increased wages induce increased effort on behalf of the workers which creates a greater level of overall surplus for both parties to share. This result, however, extends beyond the efficiency wage hypothesis insofar as the anonymous, one-shot nature of each worker's interaction with employers would provide effort for defection (shirking). This extends beyond the efficiency wage hypothesis which holds that workers are motivated to increase their effort level to retain the efficiency wage provided by the employer. (Yellen 1984)

The authors suggest two possible motivating forces:

> *"(i) Workers may have felt an obligation to share the additional income from higher wages at least partly with firms. (ii) Workers may have had reciprocal motives; that is, they were willing to reward good intentions and punish bad intentions." p334, Fehr, Kirchler, et. al. (1998)*

Both of these explanations appeal to the argument that some aspect of the environment of the bilateral gift exchange environment which induces individuals to behave in a fashion contrary to self interest. In the case of a one-shot interaction with an employer, this would appear to be the case. Taking into account a framework of rational individuals optimizing over a longer time horizon, an alternative explanation might be that subjects are not operating under the notion that they are engaging in multiple, discrete labor contracts but rather that they are engaged in continuous employment with the experimenter or the experimental environment. In this environment, the game begins to look more like a

Voluntary Contribution Mechanism because defection in the present round reduces the potential for future earnings. The presence of a surplus of labor (i.e. more workers than employers) might at first blush appear to reduce the incentives for effort because workers should defect in the present round to earn income now, to insure against the possibility that they would never work again. However, in the parameters reported by the authors, the ratio of firms to workers is either 2:3 or 7:11. So, even in the worst case, a worker has a .63 chance of striking a bargain in any given round.

In a 2 round game with an employer who pays an efficiency wage the first round and, if the employee defects, a competitive wage in the second, a strong case for defection can be made because its only effect on income is in the second round. Only a true single shot game with the same employer would a stronger incentive for defection be present. It would only be rational defect if the marginal benefit of defection in round 1 exceeds 63% (7:11 – the lowest ratio of employers to employees) of the cost in terms of the lost wages in moving from the efficiency to the competitive wage.

Without the full data set it is hard to perform a complete analysis. We reproduce results from (Fehr, Kirchler et al. 1998) below in figure 2. We note the largest group of workers reciprocates with an effort of .4 in exchange for a wage interval of 51-60. Assuming workers earn the lowest value in this range, the cooperative outcome yields a payoff of 27, and defection a payoff of 31. In the complete contract case, the equilibrium wage appears to be 20, yielding a payoff to the worker of 20. Using the estimate of .63 as the probability that a particular worker expects to work in the second round, defecting in the first yields an expected loss of 4.41 in future earnings at a gain of 4. If time preference is controlled for by the fact that subjects are paid the earnings for the entire experiment at the end, risk aversion would be required to support an explanation for defection in the first round of this game.

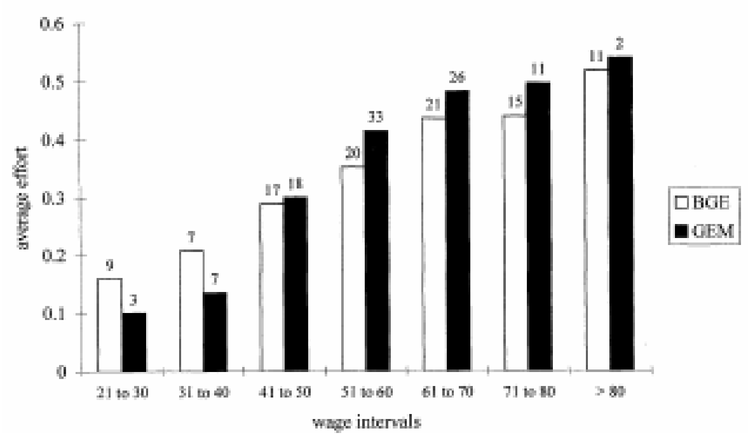


Figure 2. Payoffs in the bilateral gift exchange vs. gift exchange market
Results reproduced from p334 of (Fehr, Kirchler et al. 1998)

We suggest two alternative designs which might be considered in BGE market that might provide a stronger case for the existence of other-regarding behavior. One would be the implementation of a rank payoff scheme such as the one used by Andreoni (1995) where subjects are paid according to their relative earnings rather than their absolute earnings, the other would be a strict one shot game where subjects engage in only one game in the entire experiment. Both of these treatments would eliminate any long term cooperative strategy among employees. Cooperation in a strict one-shot game would provide the cleanest evidence in either direction, as it would reduce the likelihood of error introduced by subjects who may be confused as a result of erroneous multi-round strategizing as seen in both Andreoni and Houser and Kurzban. (2002)

In the multi-round model with multiple workers, the strategy of a single worker always defecting strictly dominates if that single worker is the only defector among a population of cooperators, exploiting each employer in turn. However, if a worker believes that his defection will induce his employer to offer sub-efficiency wages in future periods to other workers, and that other workers hold similar beliefs, the game appears to provide similar incentives to the VCM.

In the VCM, every player has an incentive to free ride off of the contributions of others, however that free riding has the result of influencing the amount that other players make available to the public good. In the labor market case the common pool resource is the goodwill of employers who expect their gift will be reciprocated with higher effort levels. In both BGE and VCM, cooperation provides a positive externality where a surplus can be attained through cooperation. However, the BGE seems to support outcomes that provide cooperative surplus, where the VCM fails to do so.

If it is true that BGE incentives are similar to VCM and the BGE can sustain cooperation where the VCM cannot, it may prove to be a concrete example of how context does not induce irrational behavior, but rather enables the attainment of a long term cooperative outcome by allowing the brain to relate to a similar set of incentives in a way which adds salience to the cooperative rather than the competitive outcome.

Returning to our inquiry into gift exchange in the computational model, it is rational to ask: what is different about a gift exchange like the one observed in a constant game vs. a system predicted by a strict rational self-interest model? It may be that the gift adds value as a signal of one's willingness or ability to provide resources in a cooperative endeavor. The value of the BGE is that it allows employers to communicate intention, i.e. signal, in a sequential setting where the VCM does not.

The efficiency wage description of Fehr, Kirchler, et. al. (1998) argues that because *employers* earn more when workers exert a super-optimal level of effort, *employers* have an incentive to pay those workers a super-competitive wage. An alternative explanation is that the *employees* have an incentive to work harder because it represents a contribution to a kind of public reputational good associated with all workers for which the private future benefit associated with the super-competitive wage exceeds the present value of defection. This notion is similar to later work by Fehr, Fischbacher, et. al. (2002) who argue for the group selection condition of strong reciprocity, which we discuss in section 1.3 in greater detail. We believe this is closely related to the computational results from

Danielson (2002) reviewed in section 2.3 where agents are likely to earn more from interactions when trading partners are cooperators.

2.5 The Pure Social Norm Model

Hoffman, McCabe, and Smith (1996) formulate a social norm model for reciprocity that a social norm exists which prescribes that one should provide a benefit to another when the cost for doing so is low. The social norm is a rule; compliance or non-compliance with the social norm has an effect on reputation which provides value in terms of influencing the behavior of future trading partners. The social norm explanation of other-regarding behavior in the dictator game is, simply stated, that expectations regarding reputation effects are carried by subjects into the experimental environment.

The authors are able to reduce but not eliminate giving in the dictator game by increasing what they term as “social distance” through manipulating the experimental environment to allow for varying degrees of privacy in the choices of the decision maker. In the weakest setting of privacy, or the shortest length of social distance, subjects write their decision on a piece of paper which they hand to an experimenter who immediately and in-person pays that subject in cash according to that decision. In the treatment with the greatest amount of privacy, subjects are given an envelope with both dollar bills and blanks slips of paper, allowed to make their decision in private, and deposit the envelope with the remaining bills or blank sheets of paper in a box that is not disturbed until all decisions have been made and all envelopes deposited in the box. The decision makers are allowed to leave before the envelopes are opened.

If these subjects believe that deception is involved during the period of time when they are making their decision and their behaviors are being monitored, a self interested explanation involving reciprocity might be invoked, but this critique is nonspecific to this experiment. The implementation of subjects walking out of the experiment with bills in hand also seems to rule out an explanation of confusion. The fact that even in the treatment with the strongest enforcement of privacy, subjects still offer the second player a nonzero sum of money provides evidence for something that can be called other-regarding behavior.

In the model of purely rational, self-interested agents, other-regarding behavior in the dictator is the point at which the greatest friction between behavioral and traditional economics occurs. We have only irrationality, confusion, or preferences for non-economic factors as descriptors for this behavior Traditional economics cannot incorporate those non-economic factors because a preference for other, equality, and the like, represents an unobservable internal phenomenon. While traditional economics can offer the observation that the dictator in the double blind case appears to be displaying an irrational preference for the welfare of the other player, it cannot incorporate other factors. However, even a perfect imaging study that can make a parametric prediction of how much money will be given on the basis of some activation pattern in the brain will not improve the standard economic model of rational self interested actors.

At this point we respect the preference for observable behavior in economic models, however it still seems cruel to abandon our subjects to the rough seas of irrational preferences. In a later paper, Hoffman, McCabe and Smith highlight a possible solution:

> "…what is it that is being consumed when someone rejects an offer in the ultimatum game, or when someone gives money away in either the ultimatum or dictator experiments.[sic] From the perspective of this experiment the answer, which we will call reputation (or image) , is largely explained as self-regarding, that is, people act as if they are other regarding because they are better off with the resulting reputation. Only under conditions of social isolation are these reputational concerns of little force."p659, Hoffman, McCabe and Smith (1996)

Another way of putting this is that under conditions of low social distance, the marginal benefit of other regarding behavior is high. However, other regarding behavior in the high social distance case where its marginal benefit is low still needs to be explained. The solution we suggest that extends this explanation is that reputation itself which is not valued, but rather a more generalized environment of social reciprocity which allows individuals to reach cooperative outcomes that produce economic surplus in other games. We think a possible description of the anonymous donor's actions, is that giving in the dictator game is an attempt to coordinate with the recipient on a long run strategy of a social norm (reciprocity) which allows for the attainment of cooperative surplus in other settings.

Returning to Danielson's (Danielson 2002) most cooperative computational agents, the high altruists are outcompeted by the medium altruists because they are too generous in the constant sum game. But, this behavior is an artifact of the parameters which allow for the attainment of all cooperative surplus which is available in other circumstances.

Let us suppose that these computational agents are mobile and move from population to population and that they are unable to observe the parameters of their counterparts except through revealed preference. Suppose also that they interact with agents whom they meet only once, but participate in the entire block of games which occur in random order, and in those games, gains from coordination in the prisoner's dilemma outweigh the losses of being exploited in the common sum game. Furthermore, those agents cannot employ a strategy which can shift altruistic preferences from game to game, but they can utilize information learned from interactions in early games as a parameter of their spite function (Sethi and Somanathan 2001) which governs behavior in towards the same agent in later games of the same block. It is possible to conceive of a situation where altruism in the constant sum game would have value as a signal of whether or not cooperation is prudent in the prisoner's dilemma.

As a thought experiment, this notion can be extended to the case where an agent remains in the same population of agents who have staggered, multi-round life-spans and interact each round with a random agent only once in a random game. Each agent must ask "How am I to know the distribution of parameters employed by other agents at any given time and therefore this particular agent will defect if the cost for defection is high?" In this situation, cooperation in the constant sum game or giving money away in the dictator game is used as a low cost device to sample the preferences of other agents; it is the method by which altruists communicate to each other that gains from coordination are available.

These behavioral results clearly identify an additional motivational factor beyond self-interest as defined by self payoff. The notion of a motivating social norm along with the value of a cooperative pool of agents to trade with has led us to the model we present in the next section.

3. Imaging Studies of Other-Regarding Behavior

3.1 The Role of Neuroimaging

The imaging data represent information that is difficult to integrate into traditional economic theory insofar as it describes internal unobservable phenomenon which are correlated by observed behavior. Economists are interested in how the behavior of individuals influence the behavior of other individuals, when mediated through economic institutions or otherwise. Economists are interested in variables which can be observed or deduced without appeal to these unobservable states.

Neuroeconomists are interested in how these unobservable factors might be influenced by the environment, and in turn influence the observable economic behavior of interest. The traditional model of economics is that modifying the incentives that influence human behavior can generate higher overall surplus. The relative question for economists is: *what* changes in incentives can modify behavior? (Roth 2002; Saaristo 2006) Behavioral economics as a subset experimental economics, asks the question of *how* those incentives are perceived by the brain to generate different behavioral responses depending on how those incentives are perceived by the mind, *how* institutions can effect the outcome, and *how* those institutions can be modified to increase available surplus.

The distinct area of overlap to both groups is the institution, that governs both the *what* and *how* of the incentives faced by individuals. We say that the imaging literature invokes the *how* question because it identifies specific regions of the brain and, based on studies from other disciplines linking region to function, posits a story of how psychological concepts like intentionality, agency, and emotion factor into an individual's decisions.

Briefly reviewed here is a subset of the imaging literature as it relates to economic exchange and a discussion of how the literature relates to economic phenomenon as well as how the authors couch their economic explanation in terms of context, equilibrium, or both.

3.2 Intentionality Detection and Theory of Mind

The trust game (Hoffman, McCabe et al. 1994; Benjamini and Hochberg 1995) is one of the most well studied problems in experimental economics as, especially in the multi-round case, it directly deals with the principal agent problem of one agent managing resources on another's behalf. Behavioral experiments have demonstrated that a trust game with simultaneous decisions yields lower levels of surplus compared to the sequential, extensive form game. (Deck 2001) Trust is intimately linked with intention, and cooperation arises on the part of the second mover who chooses to reciprocate the intention of the other player. In the simultaneous choice situation, the first mover is incapable of communicating an intention and in the absence of the social context associated with that intention, subjects are more likely to defect.

Described another way, the second mover is effected by the first mover's intention communicated by a decision to trust. Simultaneity or interactions with a computer eliminates the intention to trust signal responsible for evoking reciprocal behavior.

Intentionality detection describes how the brain would interpret the exact same piece of information (the decision by the first mover in the trust game which allows the second mover to make a decision) in a different way depending on whether the information was the result of a computer's decision or the decision of another individual. It is the change in how the message is perceived from non intentional in the computer case, to intentional in the personal case which induces a change in behavior on behalf of the second mover.

McCabe, Houser, et. al. (2001) study a trust game in pairs of individuals with one player participating from within the MRI. For half of the trials subjects interact with a computer counterpart that plays a fixed random strategy, and for the remainder subjects interact with their human counterpart. Players are classified into two groups based on the number of cooperative moves they make with human counterparts. Activations are similar among both groups in the condition where players are interacting with the computer. Activation patterns for non-cooperators is similar to the computer case, and activation among cooperators shows significant differences in occipital, parietal, thalamus, medial frontal gyrus and frontal pole.

The authors suggest that prefrontal activation is required to achieve cooperative outcomes because it involves binding information about long run mutual gains and the inhibition of immediate reward to achieve a greater long-run behavioral outcome. Cooperators attend significantly more to the information on the screen to form beliefs about their own payoffs and the payoffs of their counterparts. Because that information is presented in the spatial map of a game tree, greater demands are placed on the occipital and parietal regions which process visual information (occipital) and information about which points of space on the screen are payoff salient (parietal).

This experiment is listed under intentionality detection not because the authors claim that it is the presence and detection of intentionality that is enabling the attainment of cooperative outcomes; non cooperators playing against a human are also interacting with subjects who possess intentionality, but that intentionality does not induce a specific neurological response. In the case of the cooperators, however, there is a shift in behavior that results from moving from the computer to the human case. In addition to intentionality, the human case also introduces the host of social concerns representing non-economic factors of decision making.

This experiment suggests an approach of separating the effect of intentionality *per se* with a design using a silent counterpart that cannot invoke reciprocity or intentionality, but can invoke preferences for equality, altruism, and similar social norms.

3.3 Counterpart's Behavior Invokes Values That Effect Own Behavior

King-Cassas, Tomlin, et. al. (2005) find that the head of the caudate nucleus is active when the second mover intends to repay the first mover in the trust game. Furthermore, early in the experiment this activation occurs at the point in time when the second mover sees the result of the first mover's decision, this activation response moves earlier in the game. By the end of the experiment, this activation occurs during the time period immediately preceding the moment when the second mover learns of the first mover's decision.

In other studies, the head of caudate nucleus is associated with the processing of feedback, hence one explanation of this result is that individuals process the feedback of the first mover's decision in a different way when they intend to reciprocate than when they intend to defect. The head of the caudate nucleus also receives afferent projections from the ventral tegmental area, (VTA) which is the network of dopaminergic structures that receive and process reward in the brain. (Rolls, Thorpe et al. 1983) Furthermore, Schultz, Dayan, et. al. (1997) demonstrate that a reward signal tends to shift forward in time when that reward is predictable. They utilize single-cell recordings in non-human primates of dopamine neurons to show that a reward associated with a conditioned stimulus initially provokes a dopamine response when the reward is received, but as the subject gains experience with the task, that response shifts to the earlier point in time between the stimulus and the expected response.

King-Cassas, Tomlin, et. al. (2005) find a similar timing of activation in the head of the caudate, with that activation shifting earlier in time as the experiment progresses. Noting that VTA and caudate are closely related in the brain possible, they suggest that the decision as to whether or not to reciprocate occur earlier in time as subjects gain more experience with their counterparts (p82, King-Cassas, Tomlin, et. al, 2005). The question remains as to whether or not associated reward signal shifts earlier in time as well, or this pre-decision activity represents some other reward state in the brain, related to uncertainty, expected reward, or otherwise.

The alternative explanation drawn from the ‘liking vs. wanting’ literature (Berridge and Robinson 2003) is that dopamanergic activation associated with revelation of the first mover’s decision draws attention to the decision of the counterpart as a payoff-salient feature of the environment. It is a monitoring state in which the individual is trying to interpret and value the information displayed on the screen during the time period when the first decision maker’s response is revealed. Over repeated games, learning associates the counterpart’s decision with monetary reward, and the dopamine signal shifts to an earlier point in time because the second player begins to “search” for this payoff-relevant information when waiting for the first mover’s decision.

The problem with the authors' explanation of sensitivity to fairness, is that it seems unsupported by the design. It is not clear whether the subjects are effected by fairness at all because it is the authors and not the subjects who have labeled the two possible choices of the first mover as either "benevolent" or "malevolent". "Benevolent reciprocity" is explicitly defined by the authors as the case where DM1 attempts to cooperate subsequent to DM2's defection. If DM2 expects that defection will result in DM1 choosing exit in the next round, it makes sense for DM2 to behave "benevolently" not out of altruism, but to coordinate on the long-term strategy of reciprocal trust which is predicted by folk theorems. (Rubinstein 1979) We assert that this design lacks a reasonable control condition where benevolence or malevolence can be abstracted from the other relevant information about uncertainty and that what is needed in order to identify altruism is a task allowing for contributing to the welfare of another in the absence of any reputation effects or possibility of coordinating which we address in our own design in section 5.1.

3.4 Social vs. Nonsocial Decision Making

Rilling, Gutman, et. al. (2002) report on an experiment in which subjects interact in a prisoner's dilemma with a unconstrained human opponent, a constrained confederate, a computer playing a fixed strategy, or a control task where subjects press buttons to earn money. Sometimes, subjects are told that they are interacting with an unconstrained human opponent when in actuality they are sometimes playing against the unconstrained human, the confederate, or a computer strategy. When subjects are told that they are interacting with the computer, they always interact with the computer, and the control task is non-strategic.

Cooperative social outcomes in this study consist of those games in which both players are cooperating in the prisoner's dilemma. This includes the case where subjects have knowledge of and in fact do play against a computer partner, as well as the deception and non-deception cases. This case finds activation in anterioventral striatum, rostral anterior cingulate cortex, and orbitalfrontal cortex. In the control task, subjects pick one of 4 boxes with randomly assigned values that are similar those earned in the prisoner's dilemma. These activation patterns are contrasted with the case where the subject is, in fact, playing with a computer implementing a strategy which makes a decision based on those of the subject rather than a purely random choice. The cingulate cortex has recently been implicated in binding information presented in decision making games to the particular agent responsible for that information (Tomlin, Kayali et al. 2006), and theories regarding the striatum suggest a function that binds reward to motor response. (Apicella, Ljungberg et al. 1991)

The description presented to subjects in the computer condition is a "preprogrammed computer strategy that does not play a fixed sequence of choices. Instead, it responds to your choices from earlier rounds with specified probabilities." The authors generate these strategies from the behavioral data in earlier experiments with an unconstrained player outside of the scanner, and calculate the probability that such an individual would ‘defect’ or ‘cooperate’ as a function of the immediately two previous decisions of the MRI subject. The authors do not report the parameters implemented in the computer strategy, but it would be interesting to note what, if any, strategy on the part of the human would result in the highest payoff and if that strategy could be used to describe the behavior of any of the subjects.

This seems to have the interesting feature of removing the social component concerning norms regarding fairness, reciprocity, etc… while retaining the intentionality detection involved in ‘figuring out’ the strategy implemented by the computer or searching for the equilibrium given the decisions of the computer. The authors note the role of OFC in “reward processing”. Indeed the OFC has been implemented in reward processing, (O'Doherty, Dayan et al. 2004) but why a differential response in the computer partner task than in the control task? It may be that OFC activation in this specific context is a result of the subject’s attempt to ‘figure out’ the strategy of the computer opponent. We note that OFC is consistently active in the Iowa Gambling Task (Bechara, Damasio et al. 2000) in which subjects attempt to learn the variance and expected value of 4 decks. The ideal parallel for this notion of attempting to discover the rule would be found in the Wisconsin Card Sorting Task in which the task is explicitly one of rule discovery. A significant difference between the IGT and the WCST however can be found in the reward schedule. The IGT provides monetary feedback after every decision made by the subject while the WCST does not involve varying monetary feedback as in the IGT. A clarifying study in this area might compare results from a subject interacting during the prisoner's dilemma which subjects are fully informed that the computer player is either utilizing a purely random strategy or a strategy incorporating the previous decisions of the subject.

3.5 Agency as a Component of Economic Decision Making

An interesting feature of most economic tasks in the MRI is that individuals interact with a computer program rather than individuals. In order for the simplest reputation system to function, information about a particular transaction must be bound to the specific agents involved in that transaction. As discussed earlier, the anterior cingulate has been implicated in a two-person trust game with imaging data from both decision makers: medial anterior cingulate activation correlates with the time period when one's own decision is made, and anterior/posterior activation with observing information about the decision of another party. (Tomlin, Kayali et al. 2006)

Tankersley, Stowe, et. al. (2007) explore the relationship between agency and altruism with a work task which yields either reward for themselves, or reward for a charity which they have selected. In addition to playing the game themselves, they observe the computer playing the game as well. When comparing the conditions where subjects are watching the task, to those where they are engaging in the task themselves, they find greater activation during the watching condition in the right posterior superior temporal cortex than in the condition where the subjects are engaging in the task themselves, to the benefit of both self and charity. Furthermore the magnitude of this difference predicts self-reported altruism on the survey instrument.

The authors hypothesize that the pSTC attributes agency to the actions of the computer feeding into, perhaps, a generalized sense of empathy which is related to altruistic behavior. This describes the case where the opportunity to give provokes an internal state which is correlated jointly with both pSTC activation and the measure of altruism reported by the personality instrument.

Like the playing a prisoner's dilemma against the computer, this design has the property of removing the strategic component from the social one in terms of decisions made in the experiment when subjects have the opportunity to do so. The period of time when the subject interacts with the computer strategy by observing its behavior provokes activation in regions that is different from the OFC activation observed in the strategic case of the prisoner's dilemma. (Rilling, Gutman et al. 2002)

The other important feature of this experiment is that there are no gains from exchange. Additional surplus is being generated when the subject works for the charity, but this is not a surplus which will benefit the subject later. That is, unless the subject exercises substantial control over or draws significant benefits from the specific charity, there can be no long run coordinated strategy of a greater level of effort on behalf of the charity being reciprocated by a greater level of benefit derived from that charity. However, especially in the MRI environment where there is usually only one subject in the MRI being observed by at least one and most frequently two or more experimenters, social distance is extremely small. Insofar as the subject is seeking to earn reputation effects from the experimenter, the chances of the experimenter attending to the decisions of the subject seem especially good. Furthermore, if we extend this possibility to the instrument itself, if subjects are similarly motivated and the questions are fairly obvious as to what they are after, those who are particularly motivated to appear altruistic may represent themselves as being exceedingly so on the instrument. It is also interesting to note that after the initial data had been collected and analyzed Tankersley, Stowe, et. al. revised their instrument "specifically for our young adult population." Supplementary Methods (Tankersley, Stowe et al. 2007)

These observations are not meant to seriously suggest that the activation reported is associated with some strategic behavior attempting to deceive the experimenter. However, the only variable ascribing economic/social relevance to this activity are those responses to a particular questionnaire that has been given the description of altruism by the experimenter. As an experiment attempting to describe economic behavior, it lacks the purely objective value of economic surplus or a significant variation in behavioral response associated with the institution. The decision making tasks contributes to the results only as a task which influences behavior, the actual decisions themselves do not influence the results. This task affects activation in the same way that the Wisconsin Card

Sorting or working memory task influences dorsolateral activation, i.e. as a rule implementation task rather than a decision making task.

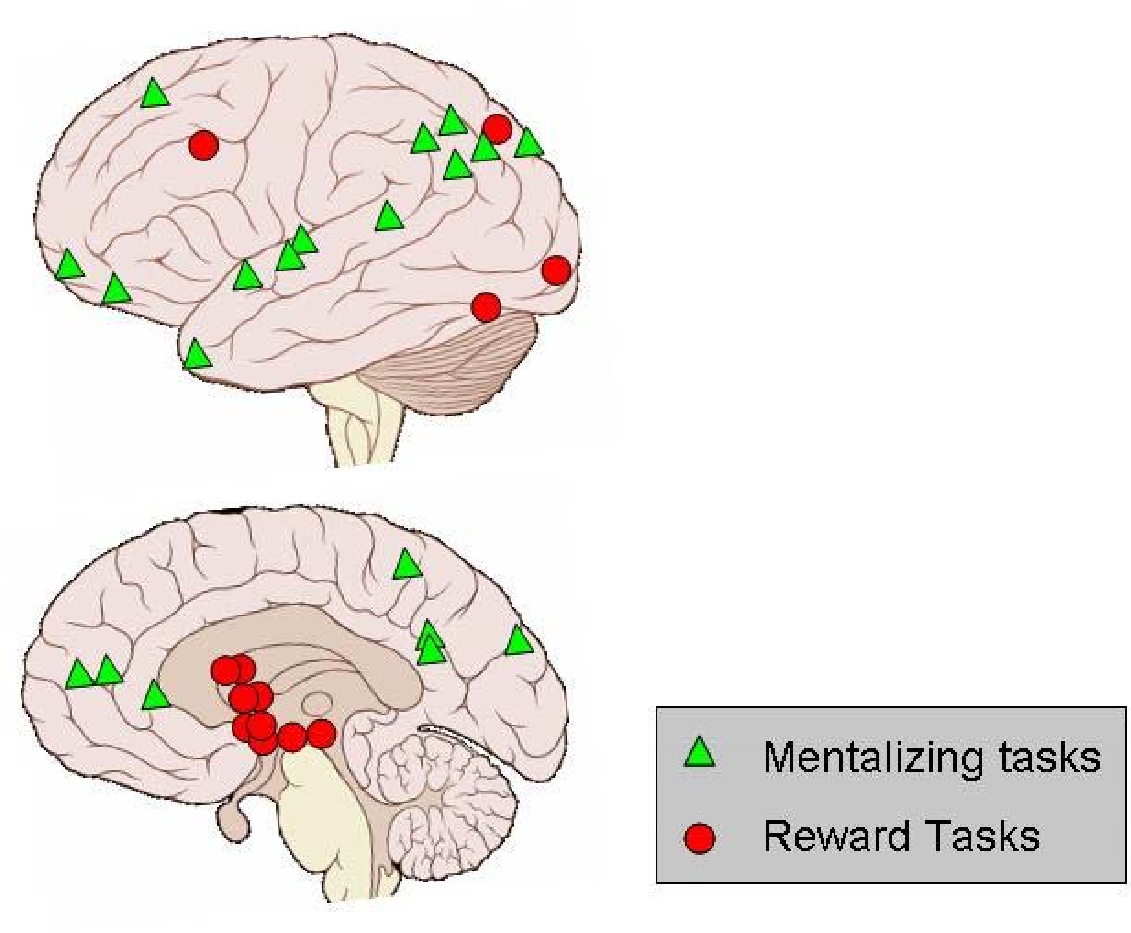


Figure 3. Cartoon showing location of activations associated with mentalizing and reward tasks

References in Appendix A. Brain image by Patrick J. Lynch (2006), used with permission.

3.6 Implications

A stylized fact from the behavioral literature is that cooperation increases when moving from the case of multiple partners interacting in a *simultaneous*, randomly matched, anonymous environment to the case of personal, one-on-one exchange involving a *sequential* decision making task. It may be the case that the institution in this case (i.e. the experimental software) is suppressing enough of the social context to thwart cooperative behavior in the case of the non-cooperators. It is not that they are behaving in a self interested fashion, or are somehow 'more rational' than the cooperators. Indeed, the property of cooperation is that higher payoffs are available than in non-cooperative cases. Starting from the supposition that a better institution is one that merely yields a higher amount of surplus, it may be the case that a simultaneous two-person trust game mediated by the experimental software is an inferior institution from the sequential game, insofar as it reduces that total surplus available to both parties. (Deck 2001)

If we accept the notion that these particular situations achieve improved economic performance, the question raised is: what specific features of the socially intimate setting enables the attainment of cooperative outcomes? The context description of psychology and the economic framework of equilibrium over preferences may not be adequate to pursue this research in the interdisciplinary setting.

We suggest the following three categories for explanations of other regarding behavior: Internal, Social, and External:

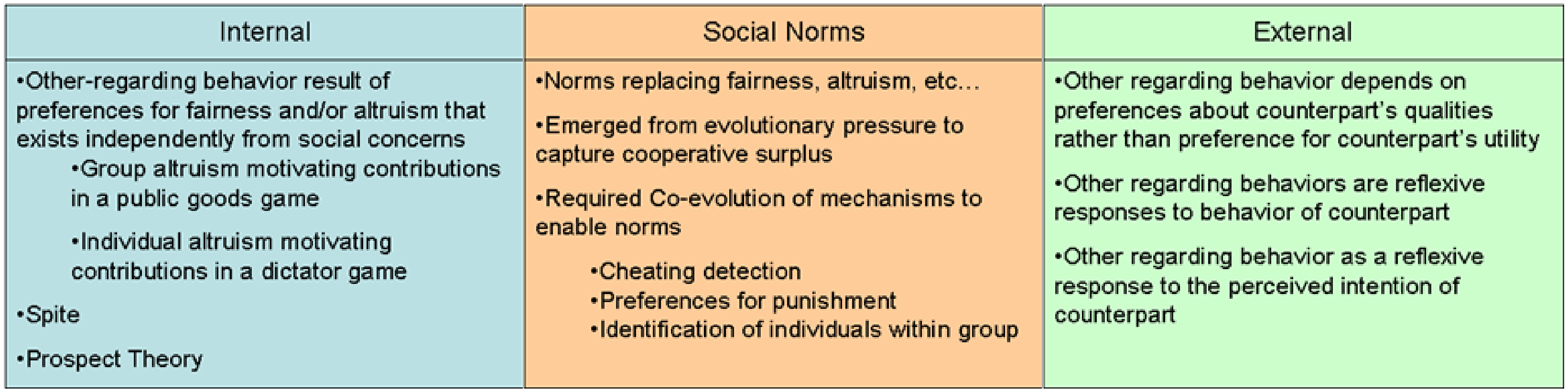
Internal explanations are those which are innate to the individual. This includes Andreoni's pure and impure altruism, (1990) and Fehr and Henrich's notions of preferences for equity and fairness where those preferences are assumed to exist a priori. (1999) Although not reviewed here, these also include normative Rawlsian arguments proceeding from the *Original Position*, (Rawls 1971) other social contract arguments, and natural rights. They are preferences that appeal the properties of how man is, was, ought to be, might be, or should be.

Social Norm explanations are those which relate to preferences that individuals have over abstract social norms which are proscribed by society. Preferences over Social Norms are distinct from internal preferences because they function like an institution. They are dynamic in that the preference is for the social norm, and those preferences can change depending on theoretical changes in society that might be observed by that individual. Social norms allow for those preferences to in practice remain fixed in the short run, but have flexibility in the long run due to evolutionary pressures. Social norms evolve in response to competitive pressures and convey a benefit to groups (but not necessarily every individual in every interaction) that can implement and enforce them. The defining feature of a social norm is that it is learned. (Axelrod 1986; Kandori 1992)

External explanations are those that describe a cause an effect explanation similar to operant conditioning where the stimulus arising outside of an individual's mind evokes an almost automatic response. In the King-Cassas study reviewed here (2005), trust induces reciprocity (they also describe possible “complex goal states” which are not a part of their hypothesis or framework), and in the Fehr, Kirchler, et. al. study (1998), gifts of employers induce reciprocity on behalf of the employee. Unlike the internal category, external explanations do not appeal to any intermediate value which an individual has a preference over, only that one behavior induces another. These explanations describe a cause and effect. Deviations occur not because the subject preferred some alternative to the predicted effect, but because the prediction is not 100% accurate. In external explanations, when the predicted effect does not follow the observed cause, it is a result of an error in the model, or an error in behalf of the subject rather than incomplete explanations of internal and unobservable preferences.

The *internal, Social Norm,* and *external* categories are not a rank order of these explanations, only an attempt to describe the types of claims that these explanations intend to make. It is easy to describe behavior in the trust game under all three: internally, the subject has a preference for equity, there is a social norm for reciprocity, and in the external case a signal to trust sometimes provokes a reciprocal response. Some authors provide explanations from more than one category, but the literature reviewed above tends to favor one over the others.

The same game can be used in two different studies to make different claims over different domains of information. Thinking about these issues can inform the design of new experiments. If the goal of a proposed study is to make a claim about how social norms influence behavior, employing methods from a study designed to examine stratgic behavior may not is making a claim about how social norms influence behavior, importing design elements designed to control for strategic behavior may not be useful unless the strategic behavior from the prospective design is a component of the social norm under consideration.



Internal	Social Norms	External
•Other-regarding behavior result of preferences for fairness and/or altruism that exists independently from social concerns •Group altruism motivating contributions in a public goods game •Individual altruism motivating contributions in a dictator game •Spite •Prospect Theory	•Norms replacing fairness, altruism, etc… •Emerged from evolutionary pressure to capture cooperative surplus •Required Co-evolution of mechanisms to enable norms •Cheating detection •Preferences for punishment •Identification of individuals within group	•Other regarding behavior depends on preferences about counterpart's qualities rather than preference for counterpart's utility •Other regarding behaviors are reflexive responses to behavior of counterpart •Other regarding behavior as a reflexive response to the perceived intention of counterpart

Figure 4. Three models of other-regarding behavior

4. Social Norms Influence Behavior

4.1 Summary

Evidence from the imaging literature seems to support the following stylized facts:

1. Cooperation/Other regarding behavior is supported by neural mechanisms that activate differentially from those which support self-interested behavior
2. Both watching and participating in economic exchange yields changes in neural activation
3. The utility function incorporates other-regarding preferences or preferences for abstract values, but institutions can modify the valence of those values for example:
    a. suppressing information about intentionality that is conveyed in a sequential but eliminated when the decisions occur simultaneously
    b. introducing complexity
    c. Eliminating information about the identity of one's training partner that might otherwise inform reputational mechanisms

What the imaging data do not support is a judgment as to whether or not a decision is economically and individually rational in a given situation. We think a more useful approach is the notion that the utility function is itself a function of the context in which the decision is presented. We hope to include (1) the value of reputation or sanction avoidance associated with other-regarding behavior while, (2) still retaining the ability to describe some behaviors as irrational. Ideally this model should produce predictions which describe behavior of healthy individuals as rational, while allowing for the behavior of individuals with a cognitive or neurological deficit as irrational. (Cox, Friedman et al. 2007)

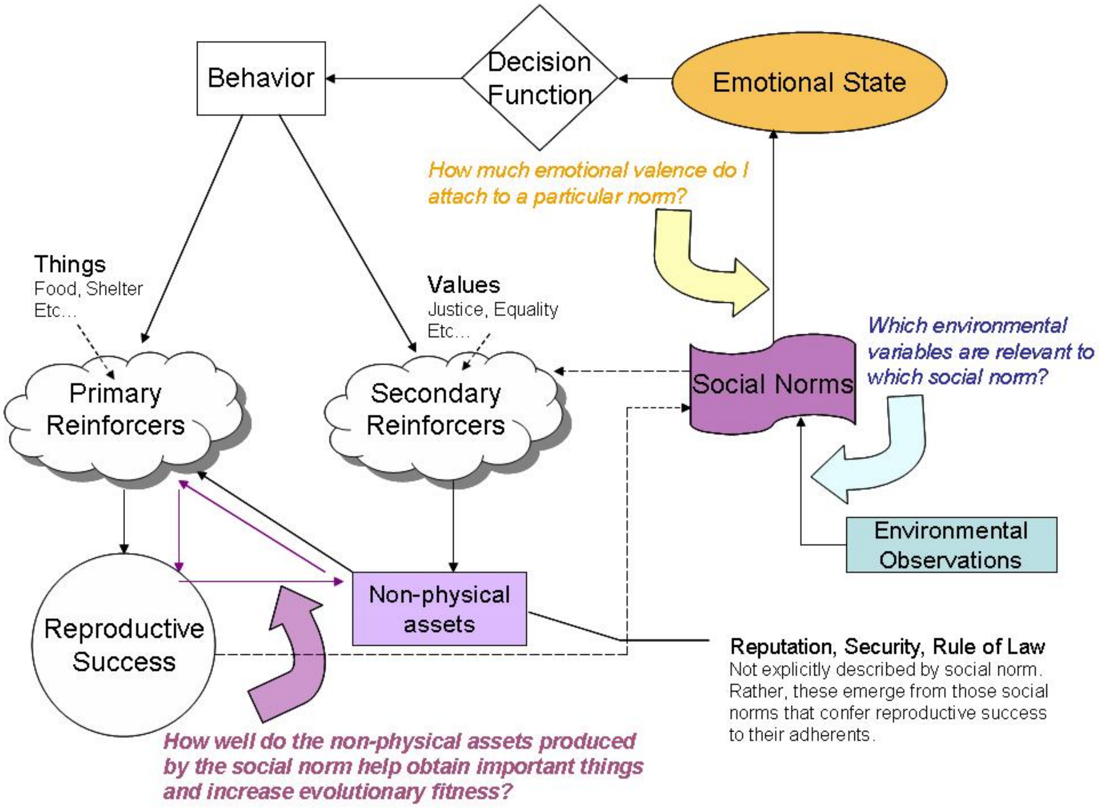

Behavior
Decision Function
Emotional State
How much emotional valence do I attach to a particular norm?
Things
Food, Shelter
Etc…
Values
Justice, Equality
Etc…
Which environmental variables are relevant to which social norm?
Primary Reinforcers
Secondary Reinforcers
Social Norms
Environmental Observations
Reproductive Success
Non-physical assets
Reputation, Security, Rule of Law
Not explicitly described by social norm. Rather, these emerge from those social norms that confer reproductive success to their adherents.
How well do the non-physical assets produced by the social norm help obtain important things and increase evolutionary fitness?

Figure 5. A model of social norms influencing emotional states

4.2 The Socal Norm-Emotional State Model

In this model, individuals directly value the environmental variables that allow them to prosper or maximize their reproductive success. These variables can be as specific as one's personal reputation as a trading partner, or as abstract as a governmental system in which that individual can prosper. These variables do not directly motivate an individual's behavior, but through associative learning, individuals learn that these variables have an effect on nutrition, survival, etc… We do not include money in this category, because it is so readily convertible to these primary reinforcers. However, it is important to note that primary reinforcers do not strictly dominate secondary ones, insofar as we have the ability to share food, shelter and the like in order to build goodwill with others which can be beneficial in the future.

We call these abstract variables like a stable government, world peace, etc…"non-physical assets", and because they have a generalized effect over a longer time period, humans have a developed value system that allows attention and motivation to be attached to emotional states which are evoked by these assets. Social norms both describe the asset, as well as the emotional effect that prospective or observed changes in those assets should evoke. In addition, the norm defines the context in which it should apply. By invoking a particular social norm, the context of the decision making problem colors how observations about one's own possible decisions, the decisions of others, and other environmental information will influence the emotional state of the individual.

The emotional state of the individual, not any preference for a value or even non-physical assets, is the variable which is optimized by the decision-making function. Decisions produce behaviors that effect either the primary goals of survival or these non-physical assets which can later be used to achieve those primary goals, and the reproductive success of individuals adhering to those norms, influences their future development and refinement. In this fashion, emotions bind short-term value to these abstract variables which have a long run significant impact on the individual that is not immediately observable or salient.

Trading institutions and behavior of one's counterpart invokes the context that defines how evaluations will be made about one's own prospective decisions, as well as the decisions of one's counterpart., are judged. Context and option space set the stage for the weighting of preferences The social norm propagates under evolutionary pressure and is learned by that particular individual. Available information about the possible decisions and outcomes associated with all parties is evaluated against the social norm assigning emotional valence to outcomes and decisions according to the context. In cases where multiple social norms are relevant in a particular context they may compete with each other.

Two variables describe the impact of an individual social norm on the emotional state, imagining the emotional state as a traditional utility function, the relationship can be described as $U_x = P_xX(C)$ where U_x is the utility derived from social norm x, P_x is the preference for that norm, and X(C) represents the social norm as a function of C which is the context presented by the decision making problem. This is analogous to the notion of the intention-dependent models we explore in section 7.5 where utility is a function of the perceived intention of one’s counterpart. Negative outcomes may not be as strongly identified with defection if one’s trading partner faces a constrained decision process or is dependent on random processes, compared to situations in which one’s counterpart has complete control over the transaction or environment.

Similarly, in our model, if an agent places a value on benevolence, in the context of a charitable donation that value may be less binding on our emotions if the prospective recipient is the willful cause of his own fortunes, as opposed to the case of a recipient who is the victim of a natural disaster or similar event. Insofar as a social norm like benevolence can function as insurance against natural calamity, it is reasonable to expect a capacity for sensitivity to these kinds of context to avoid exploitation of donors.

If the emotional state is dependent on a variety of norms as well as primary reinforcers and monetary income, it is possible to conceive of our model as a utility function where agents optimize over a set of social norms as well as additional factors that directly contribute to reproductive success. However, without a comprehensive description of these norms, this model is not particularly useful in predicting behavior. Where it may prove to be of some use is over a longer time horizon of neuroeconomic research aimed at identifying the particular social norms that may be involved in a particular game form which produces behavioral outcomes that contradict economic theory. If at some point in the future it becomes possible to identify these norms it may, at least within an individual, be possible to identify that person’s relative preference for those norms that trade off against money income, and identify what could be considered a “price” for altruism, fairness, and the like.

4.3 Model Predictions

In a simple reputation model of the dictator game, a positive reputation as someone who shares, benefits the individual by allowing him to receive benefits in the future in exchange for present income. Individuals give in the dictator game because they believe it provides them with a reputation that has a positive value. By increasing social distance, giving is reduced, but not completely extinguished. (Hoffman, McCabe et al. 1996)

Under the model proposed here, individuals might value equity in addition to reputation if equally distributing resources confers a benefit even in the absence of reputation. (Rawls 1971) In this case, than individuals pursuing emotional states resulting from equitable behavior would convey an evolutionary advantage on the group, but would not necessarily be rational. Individuals who contribute in a double-blind dictator game because they do not believe the double blind and think that subsequent outcomes will depend on their present decision are behaving rationally. Individuals who believe the double blind and make a contribution solely because they are pursuing the emotional state which results from that equitable action are behaving irrationally according to the definition we have provided here.

Under a strict definition of rationality, the explanation would stop here. However, by adding the assumptions made by our model, we can make the conjecture that the emotional state resulting from the altruistic behavior is preferable to the state which results from rational self interest.

If it is true that an environment which results in an equitable distribution of payoffs is associated with a social value and emotional state associated with fairness, we can target our manipulations to those factors which influence the competing emotional states at decision-making time. Behavioral experiments have demonstrated that earned income is much less likely to be shared equitably than a random endowment. Because there is a well established literature surrounding emotions, social norms, and norm violations, an economic decision making model incorporating values and emotional states might provide additional insight into how factors like property rights and social distance influence exchange.

A more generalized prediction is that when social norms compete with one's own monetary payoff, we would expect to see a network of activations associated with social knowledge such as the superior temporal sulcus (Zahn, Moll et al. 2007), and areas of agency detection/attribution, such as the tempo-parietal junction (Castelli, Happe et al. 2000). Behaviorally, in the context of altruism, we would expect contributions to decrease as the game form drifts away from the social norm by adding layers of anonymity or multiple recipients. Hoffman, McCabe and Smith (1996) suggest an identical explanation of the predicted behavior, but suggest that lower levels of giving result from the reduction of social cues that “automatically” prompt other-regarding behavior that would be consistent with a long-run cooperative strategy of reciprocal altruism. This model suggests that altruism is rather like hitting a tennis ball as a learned response to an expected environmental cue and suggests a minimal requirement for effort associated with the decision making process. The model we have presented contrasts with that explanation only insofar as we believe the evaluation of environment against social norm is a more active and conscious process arising from the requirement to compare task and outcome to a specific domain of social contextual knowledge.

In light of this model as well as the existing literature, we next present the design of our task to control for interpersonal social factors in the context of other-regarding behavior.

5. Behavioral Design

5.1 Task

The root of the problem explored by the neuroeconomcs literature is the principle agent problem. The overarching theme is one of disparate incentives faced by individuals in the situation of available cooperative surplus. Society itself is a long run coordination game where on the whole, gains are available from cooperation, but in individual cases an on the margin, each individual is better off defecting. Concepts like altruism, fairness, and kindness enter the picture because they can be used to describe short run observations in this long run game. They are potential strategies in the individual rounds of our extended coordination game.

One conclusion of our hypothesis is that individuals value the emotional state resulting from warm glow. Actions which generate the most warm glow and are therefore the most motivating, are situations where warm glow can be obtained at the lowest cost. Reputation effects are basically an expected value good. It should be pursued where there is a high probability and magnitude of return. We assume that this reputational effect is encoded by some abstract ‘warm glow’ which secures reputational improvements by motivating other regarding behavior through a positive emotional state.

If the brain is pursuing reputation efficiently, it must be able to encode the value of that reputation as such. We want to explore specifically the pursuit or reputation as its own motivational goal as it differs from the goal of monetary reward, and we want to control for the value of that reputation by controlling the probability and magnitude of potential reciprocation.

The initial behavioral study is a search for the price of the other regarding behavior which is at the root of the ability to achieve cooperative surplus. The experiment manipulates the value of other-regarding behavior relative to self-regarding behavior while attempting to control for confusion effects, or at least provide the ability to identify confused subjects.

The game is a single person decision making, delayed match-to-sample task. The second subject interacts only as the recipient of the payoff. Our goal is to maximize the motivating factors associated with other regarding behavior while abstracting from strategic concerns; if reputation is what is being pursued, working for a single individual will have a greater benefit that working for a charity or large group. The subject is displayed one screen displaying 5 letters in a horizontal row in the center of the screen for three seconds, immediately followed by a screen displaying 3 letters. The task is to identify whether all three letters on the second screen were displayed on the first. The subject can press a button corresponding with “yes”, “no” or can choose not to press a button at all. We classify the case where the subject presses the appropriate button as

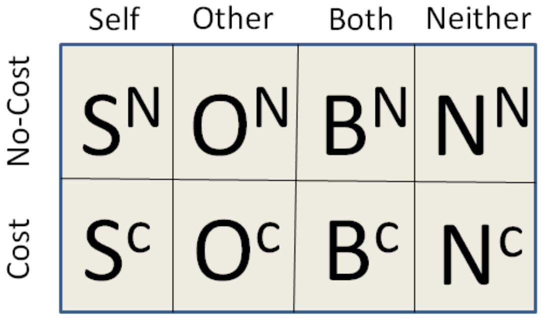
"Correct"; pressing the wrong button is "Incorrect"; not pressing a button is "No Decision".

One subject will be randomly assigned as the decision maker and the other, the counterpart. The decision maker engages in the working memory task and correct responses yield a payoff in points to either the decision maker or the counterpart. Points are converted into dollars at a fixed rate and paid to the subject in cash at the conclusion of the experiment. The sole role of the counterpart is to collect the money earned for him or her by the decision maker at the end of the experiment.

Three payoff conditions (Self, Other, Both, Neither) and two cost conditions (Cost, No Cost) were presented which describe the payoff for a given presentation: Self-Cost (SC), Self-No Cost (SN), Other-Cost (OC), Other-No Cost (ON), Both-Cost (BC), Both-No Cost (BN), Neither-Cost (NC), and Neither-No Cost (NN). In the Self payoff conditions, the worker (and only the worker) earns a variable money reward for a correct answer, in the Other conditions only the counterpart earns. The Neither conditions do not generate a payoff to either party. In our experiment cost is only borne by the worker; the counterpart never pays a cost for the response of the worker. Decision makers incur the cost for either a Correct or Incorrect response. Therefore, a Correct response in the Both-Cost condition will result in the counterpart earning more points than the subject for that game.

These conditions allow us to identify confused subjects (Andreoni 1995) (those who respond in the NC condition) and the inherent utility for completing the task (response rates in the NN condition).

Table 2. Payoff conditions from behavioral experiment.
S_N denotes the condition where only the working subject receives a payoff for answering correctly and there is no cost for responding (Self, No-Cost), O_C denotes the condition where only the counterpart receives a payoff for the correct answer, and a cost is imposed (Other, Cost)



	Self	Other	Both	Neither
No-Cost	S^N	O^N	B^N	N^N
Cost	S^C	O^C	B^C	N^C

The number of points given for a correct response varies from 10 to 90 points, in 10 point increments. The distribution of payoff values is shown in Table 3. In Cost conditions, 15 points are deducted from the subject's total. At the end of the experiment subjects are paid in cash based on the number of points earned.

Table 3. Distribution of payoff values from behavioral experiment

Payoff Level	Percent
10	10.44
20	13.94
30	10.38
40	10.13
50	10.25
60	11.19
70	11.13
80	12.88
90	9.69

Incentives are communicated to the subject through a potential payoff screen (Fig. 2) before the task which consists of three numbers representing the net payoff possible to self, payoff possible to their counterpart, and the decision cost for the current block of tasks (which is either 0 or 15). Results are displayed following the decision making task on the results screen which is similar to the potential payoff screen. The design in the behavioral study consisted of 16 blocks which contain 20 decisions per block. The payoff and cost condition remained the same within each block. The payoff condition changed after every block 20 decisions, and the cost condition changed after every 4 blocks as shown in Figure 2. All eight conditions are run twice during the experiment. A thirty second rest period was provided after each block of 20 decisions. At the end of the experiment subjects were shown the net number of points they earned for themselves as well as for their counterpart.

5.2 Methods

80 undergraduate students at George Mason University were enrolled in the behavioral pre-study. Subjects were recruited from introductory-level university classes. For every 200 points earned, subjects were paid $1. In addition to the money earned from the decisions, subjects were paid $10 for arriving at the session on time. The average subject payout was $35.57, inclusive of the payment for showing up on time. Experimental procedures were approved by the Human Subjects Review Board at George Mason University.

Multiple subject pairs participated in each experimental session. Upon arrival subjects were randomly paired together and assigned roles within that pair by drawing cards. Each pair was split and subjects were placed in separate rooms. Once segregated, the subjects were administered identical written instructions detailing both Decision Maker and Counterpart roles, followed by a quiz testing for comprehension. Subjects were offered two chances to successfully complete the quiz. To reduce the number of confused subjects in our dataset, roles were assigned so that the individual filling the worker role during the subjects was someone who had successfully completed a quiz; if the subject initially selected to be the decision maker (by drawing the decision maker card at the beginning of the experiment) failed the quiz twice, they were assigned to a counterpart role and the counterpart was assigned to a worker role. As we intended to eliminate confused subjects from our analysis, this procedure was implemented to increase usable data. A case of both subjects in a randomly matched pair failing the quiz did not occur.

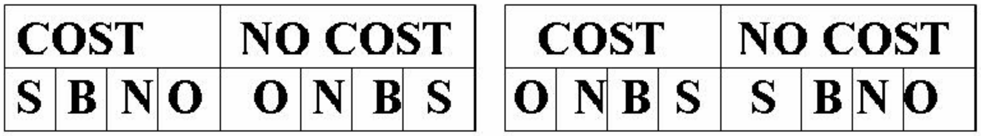
Next, subjects were told which role they would be participating in during the experiment. Decision makers were seated at computer terminals and counterparts were brought into the room and instructed to shake hands and greet the worker with the same color card. This is the only interaction between worker and counterpart that occurred during the experiment. While both groups were in the same room, the counterparts were told they were free to leave the lab until the end of the experiment.

Decision makers completed the letter matching task for approximately 65 minutes preceded by 15 minutes of training on the experimental where no money was earned. At the end of the experiment, both workers and counterparts were paid their earnings privately in cash based on the decisions made by the workers.

5.3 Behavioral Results

The results of the behavioral study indicate that subjects always work for themselves where it is rational to do so. That is, they will always attempt the task in the self payoff conditions, except the case of Self-Cost when the payoff level is at 10 points. Because our cost is fixed at 15 points, responding in this case would result in a net loss. Response rates in the Both payoff conditions are indistinguishable from Self Conditions. (α=.05)

Table 4. Block ordering from behavioral experiment.



COST				NO COST				COST				NO COST			
S	B	N	O	O	N	B	S	O	N	B	S	S	B	N	O

We define four behaviors that could be exhibited by subjects: Strict Self-Interest, Effort Altruism, Strong Altruism, and Cost-Benefit Altruism. Under Strict Self-Interest, individuals respond constantly in conditions where they are earning a payoff at no cost to themselves. In conditions where there is a cost, they respond at a rate which maximizes their payoff. That is, they will not respond when the payoff level is less than the cost (i.e. in cost conditions where the payoff level is 10. Under rational self interest, work is costly, and subjects will never respond in conditions where only the counterpart or neither party is receiving a payoff. The Strict Self-Interest is described by *5.1-5.3* Where *A(x)* is the attempt rate for condition *x*.

$$A(SN) = 1 \tag{5.1}$$

$$A(SC) < 1 \tag{5.2}$$

$$A(ON) = A(OC) = A(NN) = A(NC) = 0 \tag{5.3}$$

Under Effort Altruism, effort is costly but subjects are willing to expend as much effort on behalf the counterpart as they are on themselves as long as no cost is involved. In the case of strong altruism, subjects respond on behalf of another individual under cost at the same rate they respond for themselves in the cost condition and this response rate is greater than in the Neither-Cost condition. In Cost-Benefit Altruism, subjects are somewhat willing to exert effort in the Other-Cost condition, but only where the benefit to the counterpart is large.

Effort Altruism is defined in (*5.4)*, Strong Altruism in (*5.5)* and Cost-Benefit altruism in *5.6*:

$$A(SN) = A(ON) > A(NN) \qquad (5.4)$$

$$A(SC) = A(OC) > A(NC) \qquad (5.5)$$

$$A(SC) > A(OC) > A(NC) \qquad (5.6)$$

Effort Altruism is not exclusive with Strong Altruism or Cost-Benefit Altruism. Indeed, one would expect that if a subject responds in OC where a cost is imposed on the worker to benefit only the counterpart, they would also respond in ON where they can earn money for their counterpart at no cost to themselves. Strong Altruism, however, requires that subjects treat payoffs to their counterpart exactly the same as they treat payoffs to themselves, and is therefore exclusive of Cost-Benefit Altruism.

All three of these altruism conditions require that within the Cost and No Cost conditions, effort exerted for an individual must be greater than effort that does not benefit either one of the subjects. Subjects who respond in this condition get enough utility out of answering the task that they are willing to pay a cost to do so. Because these individuals demonstrate that they receive significant utility from responding which is above the loss imposed by the monetary cost, when these individuals respond in conditions where the counterpart is receiving a payoff, it can not be said that this is due to other-regarding behavior. Five different subjects responded 45% 60% 90% 90% and 100% during the second presentation of the Neither-Cost condition. Data from these subjects has been excluded.

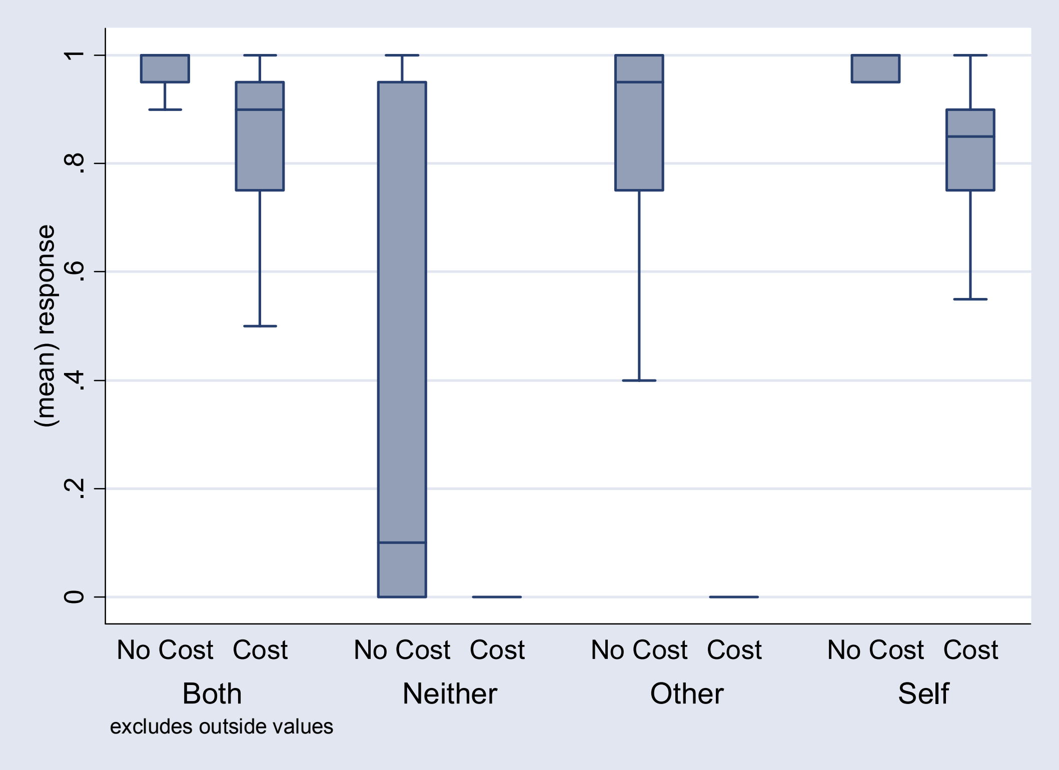


Figure 6. Attempt rates by condition from the behavioral experiment

Attempt ratio is defined as the percentage of comparisons that a subject attempted to answer, and is used as a measurement of willingness to work. Response rates by payoff levels (Figure 6) show that when there is a decision cost, subjects will make fewer attempts to respond.

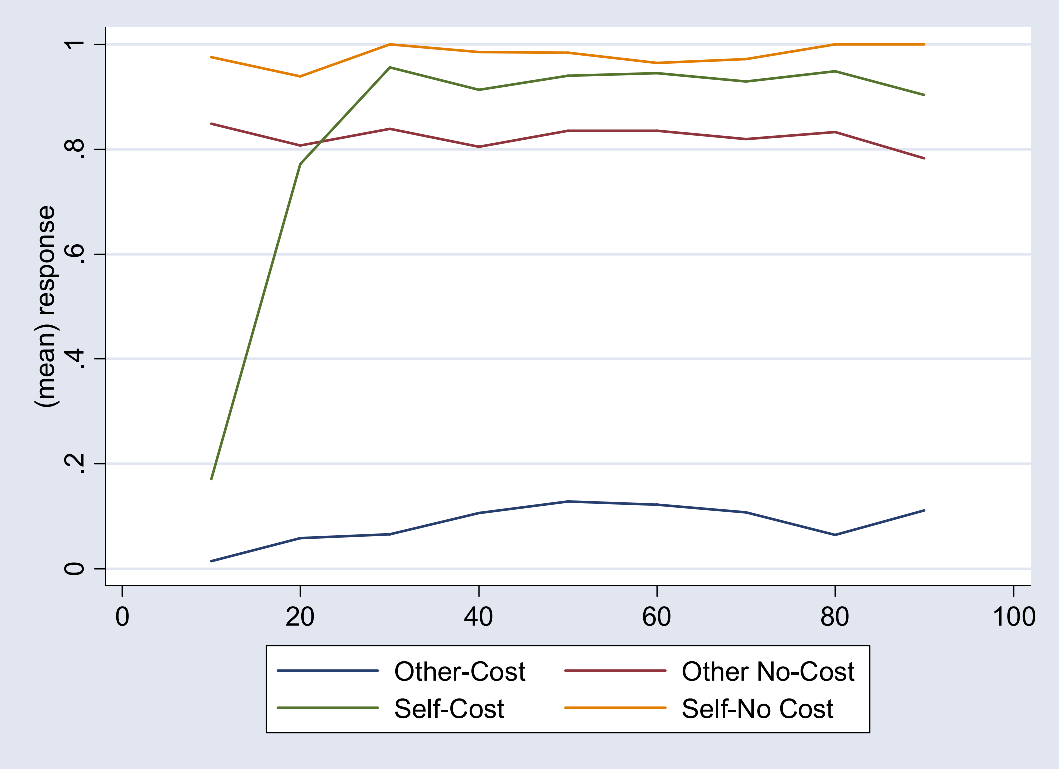


Figure 7. By payoff level, attempt rates in Self-No Cost, Self-Cost, Other-Cost, and Other-No Cost

Subjects respond at a rate of 98% in the Self-No Cost condition, and 82.7% in the Self-Cost condition, which are statistically different, suggesting that subjects are not completely sure of their ability to answer correctly. Response rates in Other-No Cost condition are 82.3%, which is also significantly less than the Self-No Cost condition rate of 98%.

Four subjects had response rates in the Other-Cost condition of 30% 80% 90% and 90%. The remaining 29 subjects did not respond in the Other-Cost condition. The responses of the four who did respond in the Other-Cost condition correlate positively with payoff level. This indicates that although they are willing to sacrifice their own earnings, they do so when it achieves the greatest benefit to their counterpart, which is not consistent with a notion of strong altruism.

For the remaining subjects, the effort seen in the Other-No Cost condition is completely extinguished by the presence of a cost of 15 points which equates to 7.5 cents. In each block of 20 presentations within the Other-Cost condition subjects could have earned an average of $5 for their counterpart at a cost of $1.50 to themselves.

6. Imaging Experiment

6.1 Rationale

We have identified several elements from the imaging literature discussed in section 3.6 that we think will allow for the identification of other-regarding activation abstracted from social and strategic interaction. After Houser and Kurzban (2002) we identify the need for a control condition to identify those subjects who may be confused about the incentives of the experiment. This is accomplished primarily through our Neither-Cost condition. Additionally, the Self-Cost and Self-No cost have the feature of identifying payoff salient activations which are specific to the subject, allowing them to control for visio-spatial activations and motor responses involved.

The most significant feature is the elimination of an interacting counterpart. Although a counterpart is necessary by definition for studying altruism, a counterpart which interacts with the subject is problematic for three reasons. First is the issue of those strategic interactions that are self-interested, but can be mischaracterized as altruism. This issue is highlighted by King-Cassas, Tomlin, et. al. (2005) who paint as altruistic those behaviors which under a folk theorem equilibrium can be considered self-interested. (Rubinstein 1979) Our study is vulnerable to this effect insofar as subjects may actually know one another, and may choose to maximize surplus extracted from the experimenter by working in the Other-Cost condition in the hopes that the counterpart will share some portion of that surplus with the decision maker. We believe this possibility is minimized through our procedure to both ask subjects if they do, in fact know each other, and through recruiting from a central database of undergraduates, rather than recruiting subjects for a specific experimental session from a single classroom recruiting presentation.

Another reason for removing the counterpart is to reduce the intentionality confound that may be associated with strategically thinking about the other person's decisions. We believe that it is possible and reasonable that an decision maker would reflect on how the counterpart might appreciate or disapprove of one's own actions, but this is a process we would like to disentangle from mentalization about what move one's counterpart might make as part of an interactive game. In a zero-sum game like chess, players spend a great deal of time attempting to deduce the possible moves of one's counterpart, this strategic mentalizing is contrasted from that which occurs when one contemplates the purchase of a gift. In the first case one is mentalizing about the counterpart's actions, in the other, one is contemplating about the counterpart's utility. This kind activation is the subject of research by McCabe, Houser, et. al. (2001), but is more particularly highlighted by the results of Rilling, Gutman, et. al. (2002) who demonstrate a similar strategic interaction pattern when subjects are interacting with a computer player as well as with another human. By removing the opportunity for subjects to observe decisions made by any process, human or random, we hope to identify utility mentalization by controlling for intentionality detection associated with strategic predictions.

Because the dependent variable in the behavioral experiment was the response rate, that experiment was designed to maximize the number of games during the experiment. The imaging study will focus on the neural activity of the subject during the potential payoff and results screen, and therefore the design of the imaging experiment will increase the time spent by the subject looking at those two screens. The limitations inherent to fMRI research will necessitate fewer observed decisions, however we have a large set of behavioral data with which to compare these results. Our prior data should be sufficient to support a claim that the choices of subjects in the imaging experiment are similar to those in the behavioral experiment.

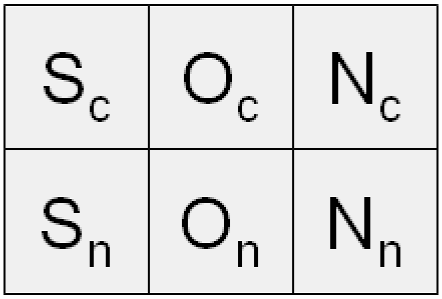
Because the behavioral task was unconcerned with the thought process of the decision to work, and only the result of that decision as reflected by the attempt rate, the behavioral task blocked games so that adjacent games had differing payoff levels, but not differing cost or payoff conditions, i.e. there would be a string of games in a particular condition (e.g. Self-Cost) with the payoff level in each of those games changing. Our purpose in this was to economize on the time spent in the potential payoff screen to increase the number of decisions, by making it easy for the subject to adopt a mental rule ("Always respond in the Self-No Cost condition") to implement the desired strategy. In the imaging experiment, the opposite is true. By changing payoff levels and conditions in each round, we hope to increase the cognitive load associated with the potential payoff screen. We hypothesize this modification will increases the probability that a subject will actually be making a decision while observing potential payoffs. Additionally, more time observing potential payoffs allows for more time to make and implement a decision, hopefully reducing mistakes where subjects ‘automatically’ respond when they did not intend to.

We also allow for additional time during the reward stimulus in an attempt to observe the outcome because we believe it is this period of time in which the differential activation motivating self-directed behavior vs. other-directed behavior will occur. If changing payoff conditions each game increases cognitive load in the potential payoff screen, we believe that it also increases attention during the results screen. Increased attention on results should enhance the self vs. other signal we hope to observe.

6.2 Design

The design implemented in the imaging study is identical to the behavioral study reported in section 5.1 except for the elimination of the “Both” payoff condition. This change was made because we wanted to clearly identify differences in activation between Self and Other and, given that response rates between Both and Self did not differ significantly in the behavioral pilot, we decided to increase the number of observations per cell in the hopes of clearly identifying a difference in activation between Self and Other. Changes in stimulus presentation were made to eliminate the letter labels “S”, “O”, and “C” used to identify the three payoff-salient parameters of our task. Instead we consistently placed the number associated with self in the upper left-hand corner of the screen, the number associated with Other in the upper right hand corner of the screen and the Cost variable at the bottom of the screen. (Figure 8) Pilot-testing of the new software design (n=16) produced behavioral results identical to our earlier behavioral experiment. Because the behavioral results of the imaging experiment closely matched both the results of the previous study and pilot data, we believe the change in how these values were presented did not significantly change the underlying nature of the task.

Table 5. Imaging experiment design.
Letter indicates payoff recipient condition (Self, Other or Nether)
and subscript represents cost condition (Cost or No-Cost)



S_c	O_c	N_c
S_n	O_n	N_n

Table 6. Games per condition in imaging experiment.

Each Quarter	*Entire Experiment*	*Condition*
6	*24*	*Other No-Cost*
6	*24*	*Other Cost*
6	*24*	*Self No-Cost*
6	*24*	*Self Cost*
6	*24*	*Neither No-Cost*
6	*24*	*Neither Cost*

The experiment consisted of a 2 x 3 design (Table 5), in each condition subjects played the 24 games for a total of 144 games in each session. Payoffs followed the same distribution as in the behavioral experiment, (Table 3**),** and were balanced so that the same distribution of payoff levels were presented in each condition. Each observation will consist of an entire game from the prospective payoff screen through the results screen, and will be followed by a jitter screen of approximately 2s in length (random length will be drawn from a geometric distribution per Burock, Buckner, et. al. (1998) The gradient field in the MRI degrades as a function of how long the scanner is on which results in signal loss if the MRI remains on for the duration of the entire experiment. This field can be reinitialized by pausing the experiment briefly - we therefore broke the experiment into 4 blocks of 36 games each. Payoff conditions and levels were assigned in such a way to give each block an equal number of games from each condition, which distributes any effect from learning or fatigue evenly across conditions. Additionally, the order of presentation was randomized between subjects subject to this balancing rule.

6.3 Methods: Subjects

36 right handed males participated in our study. Approval was obtained from the Human Subjects Review Board at George Mason University. Subjects were recruited from the experimental economics subject pool at George Mason University. Subjects were pre-screened in a telephone interview the day prior to the experiment. Upon arrival, subjects gave informed consent to participate and one subject was selected at random and evaluated for MRI contraindications by the MR technician. If the first subject could not be scanned for medical reasons, the other subject was screened and scanned if suitable. Subjects were paid $20 for showing up on time plus 1.7¢ for each point earned in the experiment.

The experimenter read the instructions (Appendix B**)** aloud with the subjects, after the instructions subjects completed a quiz to assess for comprehension of instructions. Two subjects were not able to complete the quiz with a pass rate of 80%. For these two sessions, the counterpart was screened by the technician and participated from within the scanner. After the quiz, we informed the subjects who would be the decision maker and who would be the counterpart. Subjects then engaged in a practice round of the experiment lasting for 10 minutes. During the practice experiment, the MR interacted with the experimental software using the button box utilized in the MRI, and the counterpart used the keyboard. After the practice session, while both subjects were in the room, we informed the counterpart that he would be paid based on the decisions made by the MR subject - this feature of the experiment was also outlined in the instructions. We asked the counterpart to shake hands with the MR subject and return in 90 minutes to collect his earnings.

The presentation sequence of each game will differ from the behavioral condition in that the payoff possibilities screen and the results screen will each persist for 6 seconds. In addition, at the beginning of each quarter, we will display a black and white photo of either the stranger or acquaintance for 6 seconds, depending on which of them will be earning money in that particular quarter.

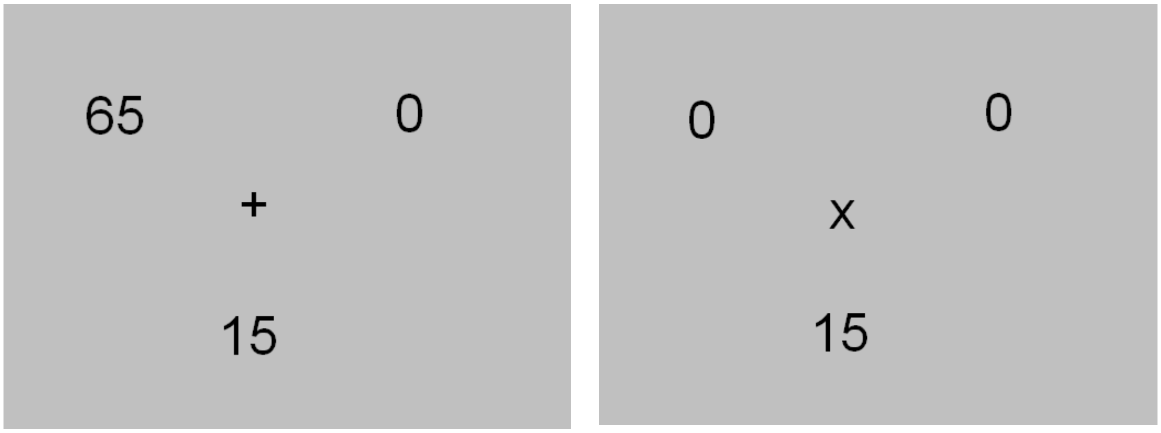


Figure 8. Potential payoff (left) and results screen (right) from a Self-Cost condition where the subject answered incorrectly

In the payoff presentation screen, the fixation point in the middle of the screen will appear as a cross, while in the results screen, the cross will be rotated 45 degrees to appear as an X. The jitter screen will appear as the plus symbol alone in the center of the screen. The jitter screen will display for a random period of time with a mean of 2 seconds to deconvolve a linear combination effect associated with the same region of the brain activating and deactivating at constant interval.

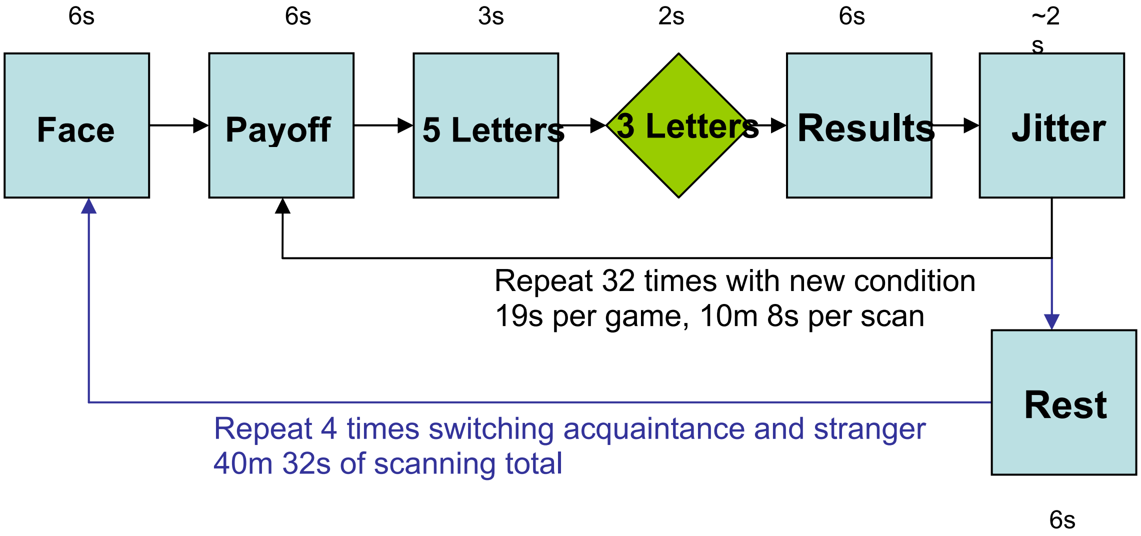

6s
6s
3s
2s
6s
~2 s
Face
Payoff
5 Letters
3 Letters
Results
Jitter
Repeat 32 times with new condition
19s per game, 10m 8s per scan
Rest
Repeat 4 times switching acquaintance and stranger
40m 32s of scanning total
6s

Figure 9. Order and timing of stimuli within experiment

6.4 Methods: Imaging Contrasts

Our aim is to first, verifying that the observed decisions (response rates) are consistent with those observed in the behavioral control, and then identify the effect of reputation as both a motivator and reward in the brain. The main effect we will attempt to establish a theory of the pursuit of reputation, is a reward signal that varies with the expected value of reputation according to a hierarchy of Self, Acquaintance, Stranger-No Cost, Stranger

Cost. We originally anticipated the dummy variables assigned to these conditions would significantly predict magnitude of striatal activation.

In the potential payoff screen, we also aimed to replicate the Tankersly, Stowe, et. al. (2007) result of superior temporal sulcus associated with agency. Ideally, that activation of this kind in the acquaintance conditions would exceed activation in stranger conditions. We predicted greater STC activation in the acquaintance case than in the stranger case during the potential payoff phase of the experiment.

We have proposed that giving in the dictator and ultimatum games may be a long run strategy for coordination within society. Work for the other individual in the cost condition is a similar situation to that of the dictator game, but with more social distance as the payoff to one's partner depends on multiple decisions. Consistent with literature that prefrontal control is required for delay of gratification and coordination, we expect more superior/anterior prefrontal activation during the potential payoff screen in cost conditions involving the acquaintances than in cost conditions for the stranger, as well as activation significantly greater than self conditions where no coordination problem exists.

The cingulate model of agency predicts medial activation associated with own agency, and anterior/rostral activation associated with outcomes attributable to another individual. It remains to be seen whether this effect is specific to agency, i.e. stimulus associated with the cause of that stimulus, or whether that effect is generalizable to the case of observing an outcome that affects self as opposed to another agent. We predicted that results associated with self will generate activation in the cingulate that is more medial than in results associated with other individuals, against the null hypothesis that there will be no observable effect.

7. Results

7.1 Data Acquisition and Analysis

fMRI scans were conducted on a Siemens 3.0T Allegra head-only scanner with a T_2 weighted EPI sequence (TR=2s, TE=23ms, FOV=192x192, 90° flip angle) resulting in 3x3x3mm functional voxels. 348 volumes were collected during each functional run. Data was collected over 4 functional runs each lasing 11m 36s, followed by a high-resolution MPRAGE T_1 weighted structural image.

Data were analyzed using BrainVoyager QX 1.9.9. (Goebel, Esposito et al. 2006) Functional images were co-registered onto the high-resolution structural image from each subject. Structural images were mapped into and coordinates reported from Talairach space. (Talairach and Tournoux 1988; Fox and Uecker 2005) Functional images in Talairach space were smoothed at 8mm. For all contrasts, uncorrected activation maps were calculated at p=.005 (*df=17*) and corrected at the α=.01 level of significance with the cluster-level statistical threshold estimator method implemented in BrainVoyager. (Worsley, Evans et al. 1992; Forman, Cohen et al. 1995; Hagler, Saygin et al. 2006) Significance of interaction effects is reported at the *q*=.05 level of significance.

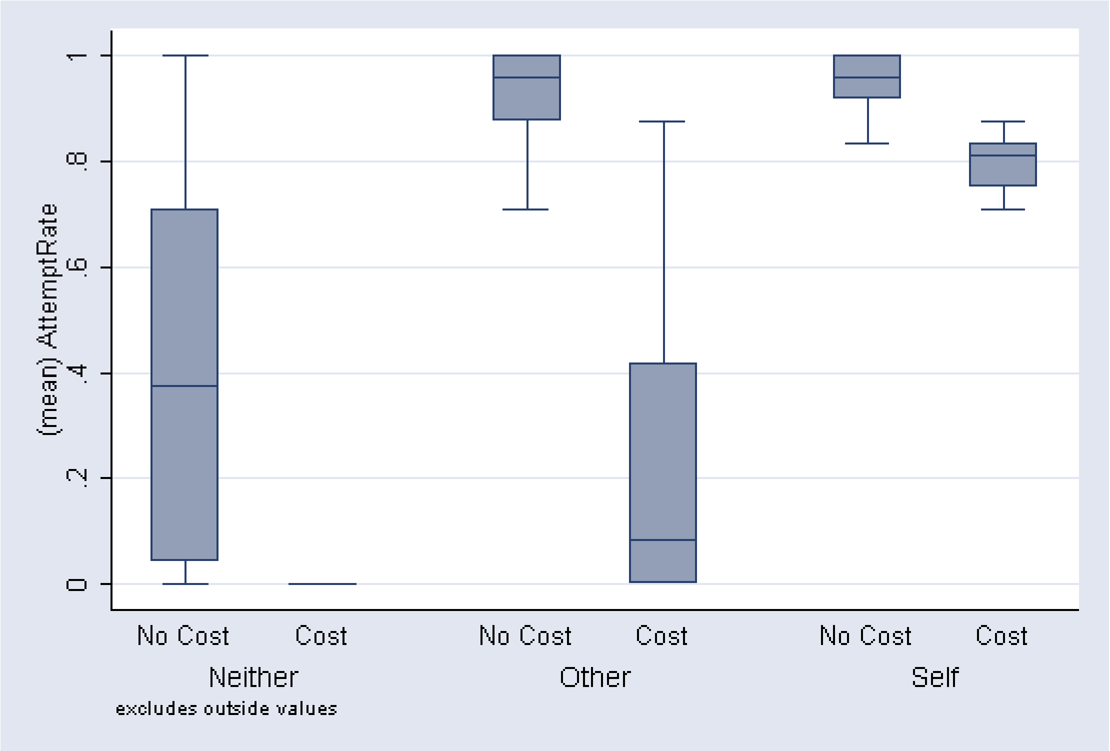


Figure 10. Behavioral response rates by condition from subjects participating in the MRI experiment

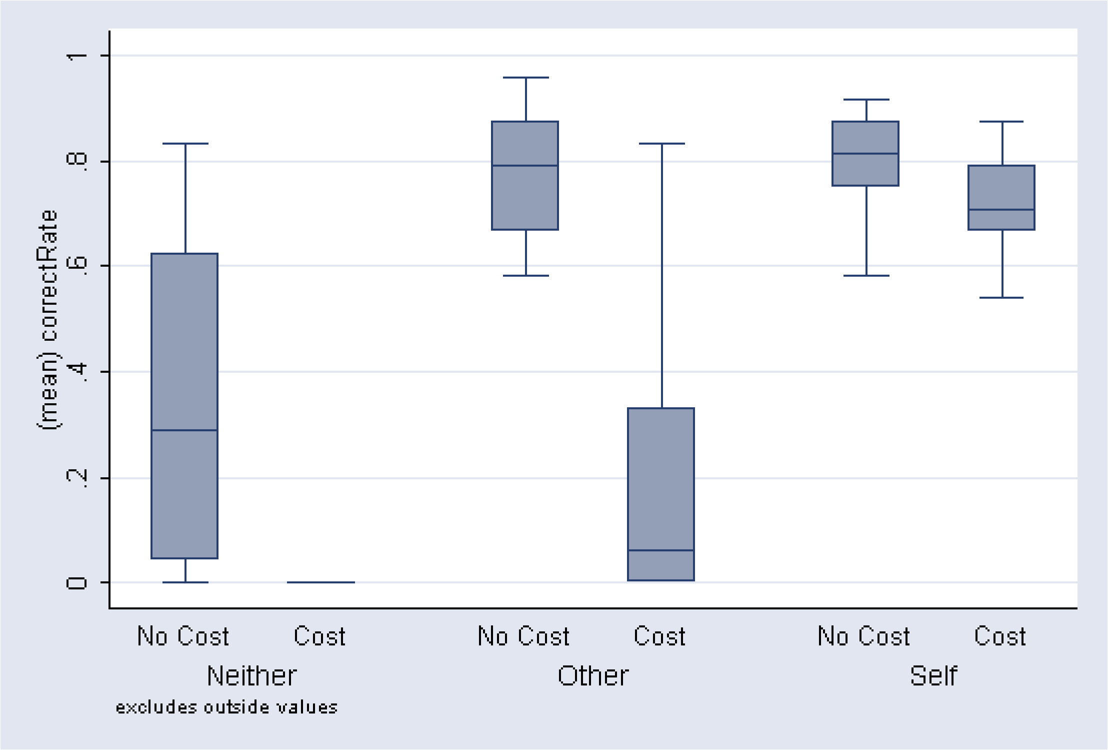


Figure 11. Ratio of correct responses (correct responses/total number of questions asked) by condition from subjects participating in the MRI experiment

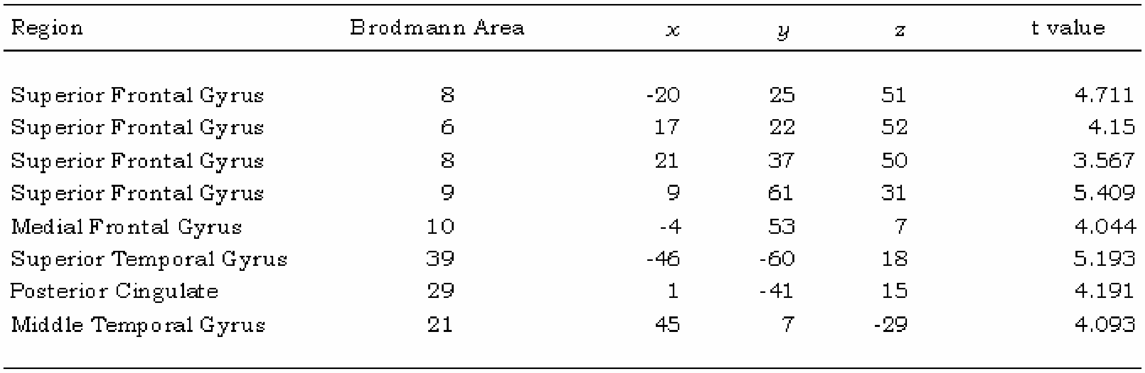
7.2 Behavioral Data

Behavioral responses were similar to those in the behavioral pre-study. Figure 6 in comparison with Figure 11 shows the similarity in behavioral responses in the imaging experiment when compared to the behavioral pilot study. In the other-cost condition, the response rate in the imaging experiment is greater than in the behavioral experiment at the α =.066 level of significance.

7.3 Contrasts

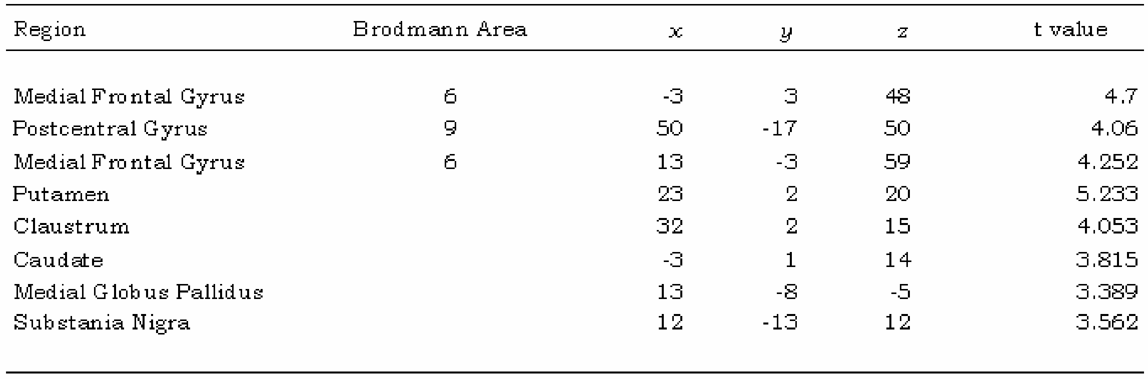
The primary aim of this study was to examine the difference in activation associated with earning reward for self and reward earned for another in addition to theory of mind and payoff expectancy activations that may exist in the potential payoff screen when subjects observe how many points could be earned for a correct answer.

Table 7. Regions of increased brain activity when observing earned results in [OC+ON-NN-NC] > [SC+SN-NN-NC]
All activations reported with p<.05, corrected for multiple comparisons.



Region	Brodmann Area	*x*	*y*	*z*	t value
Superior Frontal Gyrus	8	-20	25	51	4.711
Superior Frontal Gyrus	6	17	22	52	4.15
Superior Frontal Gyrus	8	21	37	50	3.567
Superior Frontal Gyrus	9	9	61	31	5.409
Medial Frontal Gyrus	10	-4	53	7	4.044
Superior Temporal Gyrus	39	-46	-60	18	5.193
Posterior Cingulate	29	1	-41	15	4.191
Middle Temporal Gyrus	21	45	7	-29	4.093

Table 8. Regions of increased brain activity when observing earned results in [SC+SN-NN-NC] > [OC+ON-NN-NC]
All activations reported with p<.05, corrected for multiple comparisons.



Region	Brodmann Area	x	y	z	t value
Medial Frontal Gyrus	6	-3	3	48	4.7
Postcentral Gyrus	9	50	-17	50	4.06
Medial Frontal Gyrus	6	13	-3	59	4.252
Putamen		23	2	20	5.233
Claustrum		32	2	15	4.053
Caudate		-3	1	14	3.815
Medial Globus Pallidus		13	-8	-5	3.389
Substania Nigra		12	-13	12	3.562

7.4 Main Effect: Reward Perception, Self v. Other

We examine the main effect of payoff recipient with the contrast Result_OC + Result_ON – Result_NC – Result_NN > Result_SC + Result_SN – Result_NC – Result_NN which reveals significant differences in the medial frontal gyrus, anterior cingulate, posterior cingulate, right mid-temporal gyrus, right temporal pole, and bilateral superior temporal gyrus. (Table 7) Areas that were significantly more active in self conditions were L pulvinar, bilateral thalamus, L caudate, right insula, and mid-cingulate. (Table 8)

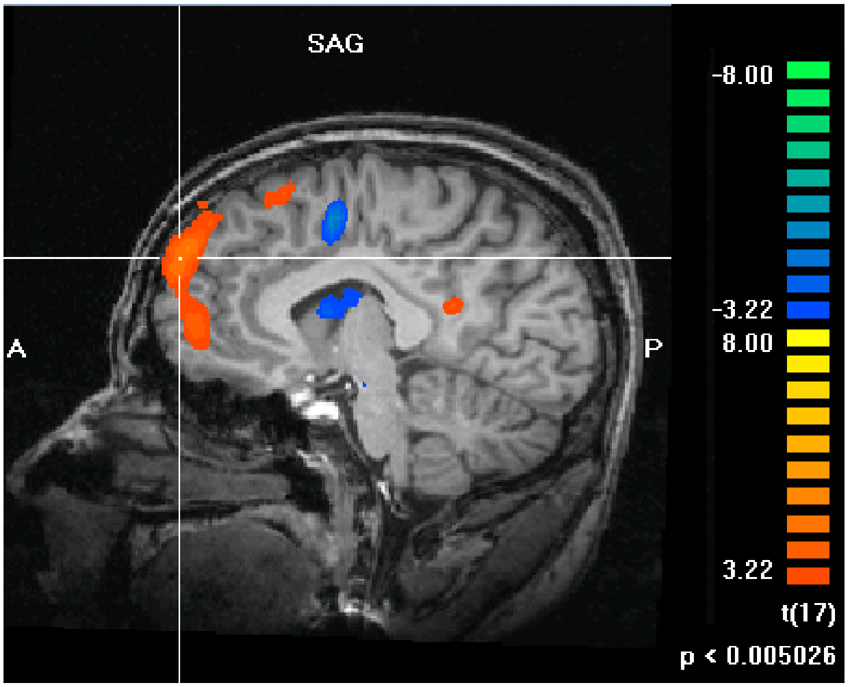


Figure 12. [Other>Neither] – [Self>Neither] medial frontal activation in Other>Self conditions when viewing results. Corrected at p<.05

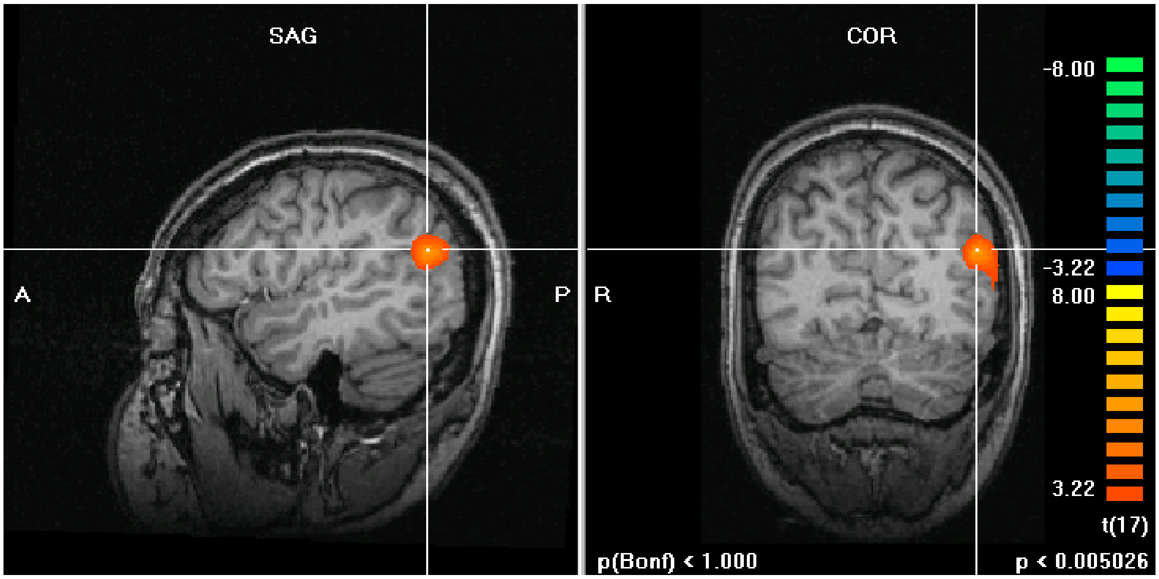


Figure 13 [Other>Neither] – [Self>Neither] Left Superior Temporal activation in Other>Self conditions when viewing results. Corrected at p<.05

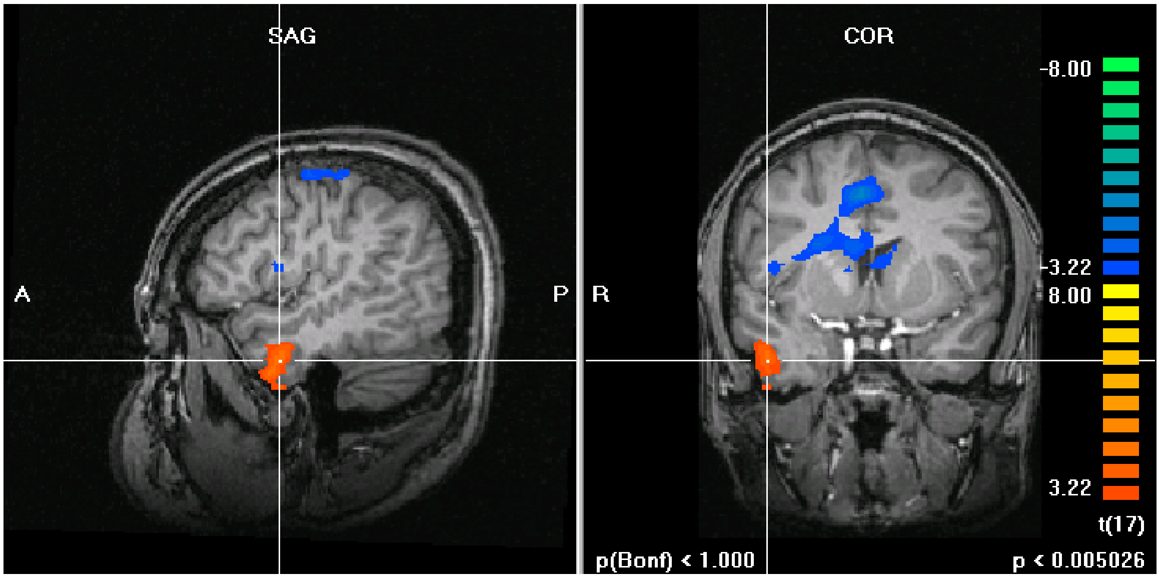


Figure 14 [Other>Neither] – [Self>Neither] Right temporal pole activation in Other>Self conditions when viewing results. Corrected at p<.05

7.5 Discussion

The objective of this study was to identify brain regions associated with reward earned for self in a context that removes the opportunity for reputation, reciprocity, and signals of intention to pass between individuals. Subjects participated in a delayed match to sample task that yielded either reward to self or to a socially distant counterpart. The inability of the counterpart to effect the subject's payoff in a meaningful fashion suggest the comparison of our task to a dictator game.

Our results reveal that regions of the brain commonly associated with a two player interactive game are involved with the processing of reward earned for another individual, even in the absence of strategic interaction between those two individuals.

Specifically, our results show activity in the Medial Frontal. Temporal pole, and superior temporal regions that are frequently implicated in theory of mind (Olson 1965; Gallagher, Jack et al. 2002; Saxe and Kanwisher 2003; Rilling, Sanfey et al. 2004), social and emotional processing (Farrow, Zheng et al. 2001; Carr, Iacoboni et al. 2003; Vollm, Taylor et al. 2006; Zahn, Moll et al. 2007), and agency detection (Castelli, Happe et al. 2000; Blair 2005; Tomlin, Kayali et al. 2006; Tankersley, Stowe et al. 2007).

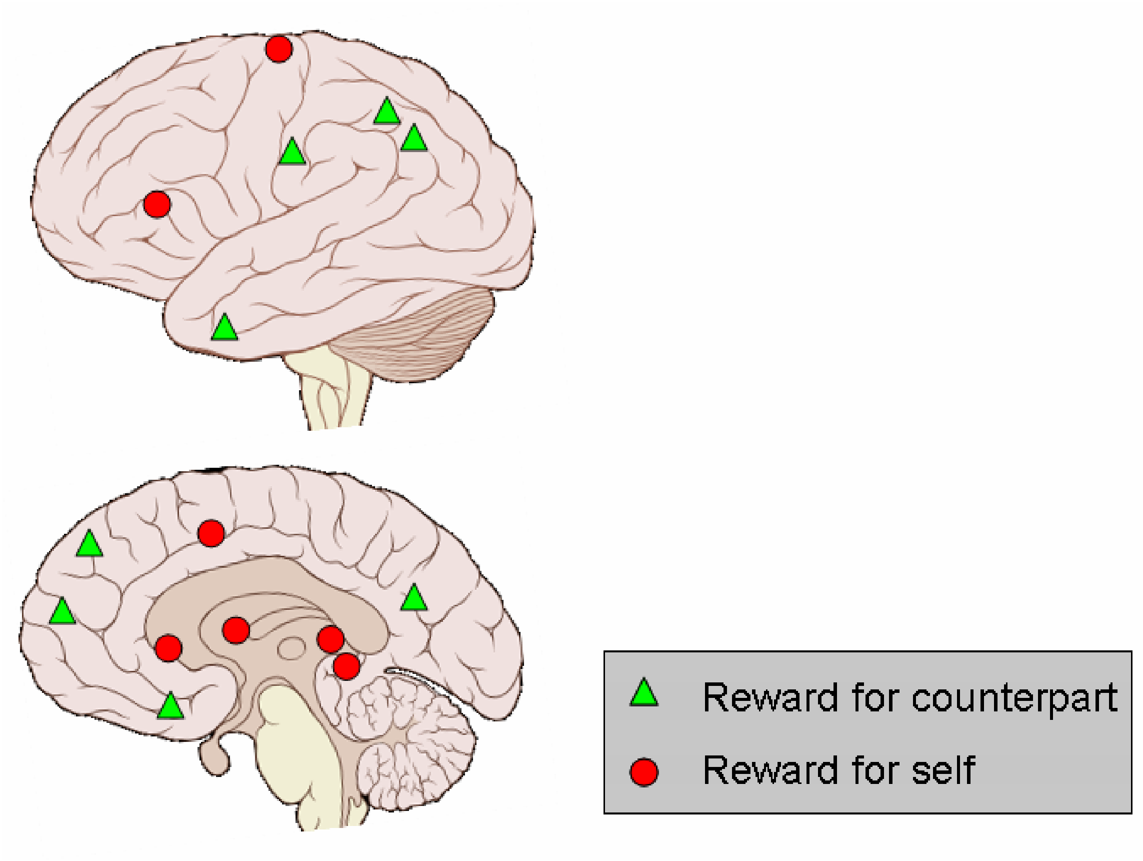


Figure 15. Cartoon representing results from present study
Green triangles represent activations when viewing results earned for counterpart, red circles identify activations observed when viewing results for self. Brain image by Patrick J. Lynch (2006), used with permission.

Our results also directly support work identifying the cingulate cortex in agency processing (McCabe, Houser et al. 2001; Gallagher, Jack et al. 2002), as well as the specific result of anterior and posterior activations in the cingulate cortex associated with other, and medial cingulate activations associated with self. (Tomlin, Kayali et al. 2006) (Figure 15)

We find both posterior and anterior regions of the cingulate cortex respond to outcomes associated with the counterpart, and mid-cingulate activations associated with outcomes associated with self. Tomlin and colleagues find mid-cingulate activation associated with decisions submitted by subject, and anterior/posterior cingulate when viewing the decision made by the counterpart. Of note is the fact that these activations in this study are present when observing outcomes for either self or counterpart, suggesting that this phenomenon may be associated more generally with the issue of “who” is associated with a stimuli, rather than the more specific notion of agency describing “who caused” the stimuli.

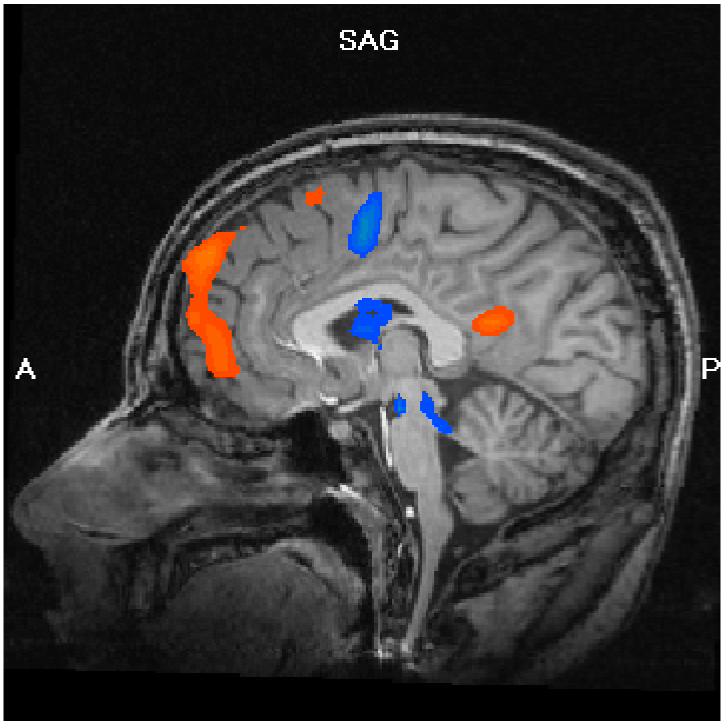


Figure 16. Cingulate activation when observing results in Other>Self
Orange identifies regions active when viewing results for the counterpart, blue identifies regions associated with reward for self.

Based on the difference in activation associated with reward earned for self and a counterpart, we believe that it is unlikely that the motivation to work for another individual, or give in a dictator game, is the result of a specific preference for the payoff of another individual represented in the utility function. Rather, our results seem to be more consistent with the active processing of the observed payoff results through a social process integrating the observed stimuli with stored concepts in the socio-emotional domain.

We suggest this process associated the observed stimulus about a social norm and that this associational process requires mentalizing about either the subject's own hypothetical behavior in the context of the norm, or the subject’s attempt to estimate how the counterpart might act in a similar situation. The integration of this information creates a mental representation of what one “ought” to do in the context of the social norm, and it is the difference between this mental representation of what “should” happen, and the homeostatic state of passive response that provides the motivational impulse to respond.

In the context of observing the results associated with the subject’s action, we suspect that the social constellation of temporal pole, temporal junction, superior temporal cortex, and dorsal medial prefrontal cortex are always active during the results phase as a result of the maintenance or updating of the heuristic used when deciding whether or not to undertake effort which occurs during the potential payoff screen. The development and tuning of this heuristic during the results phase would explain the lack of such an activation pattern during the potential payoff screen.

Such a process would not be inconsistent with an ex-post Bayesian updating rule associated with those heuristics that might be employed during a time-limited task such as ours. (Kahneman, Slovic et al. 1982) This would imply an ‘unconscious’ decision is briefly made during the potential payoff stimulus; an alternative explanation would be that the decision is made during the letter matching task. Regardless of when the decision itself occurs, an ex-post reward learning mechanism that activates regions associated with the decision-making task is well supported in the literature. (Berridge and Robinson

2003; McClure, Berns et al. 2003; Tobler, O'Doherty et al. 2006; Bray and O'Doherty 2007)

7.6 Relationship to Existing Models of Other-Regarding Preferences

Cox, Friedman, and Gjerstad (2007) propose their own model of other-regarding behavior and classify the existing models into two categories: relative-payoff models, and intention free models. In this section we will review one of the functional forms representing each type as presented by Cox et. al. and review the relationship of our data to these models, and to the model we have presented in section 4.2.

Using the utility functions developed by economists to predict a neural activation pattern is not a reasonable critique of economic theory because it does not represent the more accurate justification of a utility function emerging from activation associated with the stimulus. We caution that this enterprise is highly speculative, at best, and, at least unsporting, as we will attempt to use these models to generate the answers to questions which the original authors had no intention to answer. We intend not to criticize their work, but use it as a tool to critique our own.

$$u(m,\, y) = \begin{cases} m - \alpha\,(y - m), & \text{if } m < y, \\ m - \beta\,(m - y), & \text{if } m \geq y, \end{cases}$$

Figure 17. Representation of Fehr-Schmidt model
m represents "my" payoff and "*y*" represents "your" payoff

The Fehr-Schmidt model is represented as a utility function which depends on one's own payoff as well as the difference in payoffs between both parties. (Fehr and Schmidt 1999) In the context of our experiment, this would require the subject maintain a running total of the payoffs associated with both parties, however the explicit decision function would require a simple comparison between these two registers, or at least the maintenance of one running-total "difference" register that would identify the inequity between the two parties. During the outcome phase of our task, we would expect the updating of this register in both cases of payoff to self and to counterpart.

This seems inconsistent with our data showing a distinct social pattern of activation which is inconsistent with the simpler type of comparison required in evaluating the magnitude of this difference. An alternative explanation might be the involvement of such social systems in the evaluation of α and β against this value. However, we think the more serious inconsistency is revealed in the behavioral data demonstrating an almost negligible response rate in the Other Cost condition, where a small decision cost can remediate a difference in relative payoff, in addition to the effect of that payoff on such a difference through reducing a subject's overall earnings. It may also be the case that the subject hopes to satisfy this inequity aversion in the long run, which is addressed in the intention-sensitive models.

Intentionality models are similar to the Fehr-Schmidt model, with the exception that instead of preferring equity of payoffs, preferences are attached to the perceived intention of the other party as measured by the effect that the decision of that other party has on one's own payoff options. (Rabin 1993) In this model, the payoff of one's counterpart is replaced by a perception of how the counterpart is behaving towards the decision maker. In the Fehr-Schmidt model, the fact that one's counterpart would be earning 0 during the experiment would result in disutility to the decision-maker. In Rabin's model, the subject is concerned about the counterpart's decision only inasmuch as that decision effects the subject's payoff.

Because our study was specifically designed to abstract from any notions of intentionality, it does not provide much explanatory power to this model. In our design, the only possible perception about intention or type would occur at the beginning of the experiment when the subject is introduced to the counterpart, and we find it unlikely that the subject is actively recalling that impression or judgment independently of other cognitive processes associated with social behavior when that subject observes payoff information about the counterpart.

$$u(m, y) = m + \frac{a_m + \lambda\, a_y}{1 + \lambda}\, y,$$

Figure 18. Levine’s intention-free model

In Levine’s intention-free model (1998), subjects have an altruism coefficient α_m, an estimate for the altruism coefficient of one’s trading partner, α_y, and a weighted preference for altruism λ. In this model, preference for equity as a function of relative payoffs is replaced by a preference for the payoff of another, weighted by an altruism parameter for both parties, as well as a preference for the altruism of the counterpart. This is different from the previous models insofar as the payoff level of the counterpart directly factors into the utility function of the decision maker.

Behaviorally, this model seems plausible given our results: if a subject estimates that his counterpart would behave in an altruistic fashion if the roles were reversed, the subject should behave altruistically towards the counterpart even in the absence of an expectation of reciprocity. However, it appears that our imaging data describe the case where social processing occurs throughout the length of the experiment in each round when reward is perceived. Once the subject has estimated the altruistic preference of his counterpart, under what conditions would he change his behavior towards that individual? Because social thinking occurs throughout our experiment, our data seem inconsistent with this model.

$$u(m,\, y) = \begin{cases} \frac{1}{\alpha}\,(m^{\alpha} + \theta\, y^{\alpha}), & \alpha \in (-\infty, 0) \cup (0, 1]; \\ m\, y^{\theta}, & \alpha = 0. \end{cases}$$

Figure 19. Cox, et. al. (2007) emotional state model

In addition to reviewing the existing models, Cox, et. al. (2007) propose their own model (Figure 19) where α represents the elasticity of substitution between both parties and *θ* represents the *emotional state* of the decision maker which depends on perceived changes in reciprocity from the counterpart, and the relative social standing of both individuals. This model seems closest to our own model proposed in (section 4.2) because the emotional state depends on a *social* process**.** While their model explicitly excludes relative payoffs as a factor in that social standing, if we remove that restriction and allow for relative payoff changes to factor into the utility function both as a coefficient of the emotional state and as a dynamic component of the emotional state this would necessitate the kind of social computation each and every round that seems consistent with our imaging results.

As in the Fehr-Schmidt model (1999), this social standing will change every round based on the payoffs, however the crucial difference is that it is no longer a simple mathematical calcualation which takes place, but an evaluation on how those payoffs affect the relative social standing of each party. Although we believe it unlikely that subjects explicitly worry about the trivial changes in social standing arising from a 40¢

change in wealth, we do think it likely that the subject's emotional state continuously varies and that a more complex process involving social reasoning and/or mentalizing about how that payoff fits in with social norms associated with our task are a constant input into the emotional state driving the decision-making process.

What is common among all of these models is the preference for fairness, reputation, or a preference for altruism in a trading partner. In light of our model in section 4.2, we believe these can be generally defined under the social norm rubric. Regardless of what it is individuals prefer, it is clear that those preferences depend on some notion which is distinct from the monetary value of one's own payoff. We believe that the activation patterns observed during the results phase of our experiment are consistent with the hypothesis that subjects evaluate stimuli associated with the counterpart's payoff along with social concepts that are relevant to the task at hand. Subjects who behave altruistically are not being motivated because their counterpart has earned money, but because the subject has recognized that he has behaved altruistically. It is the altruism that is valued, not the payoff. The payoff is evaluated against the social norm, but it is the social norm which is reinforced by the reward signal and provides motivation for similar behavior in future rounds.

7.7 Conclusion

In the case of our study, we hypothesize that an other-regarding preference as suggested by Fehr-Schmidt would invoke the reward system in the same way demonstrated by studies of altruistic contributions toward a charity. (Fehr and Schmidt 1999) Instead, we find a network of social activation that is similar to those observed in two-player, interactive games, (McCabe, Houser et al. 2001; Gallagher, Jack et al. 2002; Decety, Jackson et al. 2004; Tomlin, Kayali et al. 2006) which suggests a process that is more complex than comparing relative payoffs over the long run.

We have reviewed the existing literature of imaging and economic experiments on the topic of other-regarding behavior, and have proposed our own model of social norms that motivate behavior by creating a mental representation of the appropriate action in a particular circumstance. We present a novel task which is similar to a dictator game where both effort and money cost on the part of the subject are traded off against the payoff of the counterpart and new evidence demonstrating that brain activations associated with social processes are involved in evaluating payoff to a counterpart in this context.

Our design is significant because it abstracts from the social interaction games, explicit theory of mind tasks, or agency detections tasks which have, until now, been primarily associated with these kinds of activations. These results support the hypothesis of a reward system capable of motivating abstract social norms which have evolved to capture opportunities for economic surplus associated with cooperation. More importantly, they suggest the ongoing and active integration of reward-salient stimuli with the social contextual knowledge associated with social norms describing altruism, rather than a utility calculation that relies on non-social comparisons or evaluations of the counterpart's payoff.

The issue of cooperative surplus is important because although the evolution of social groups is likely driven by the evolutionary advantages afforded by cooperation. Although nature has provided the human brain with cognitive strategies to capture cooperative surplus, it may be the case that modern society presents those opportunities in manner different from those present in the small social groups in which the human brain evolved. Hopefully studies like ours about how humans coordinate on these gains in small groups will lead to new ideas about how existing social institutions can be modified to promote coordination on the kinds of long-run cooperative equilibria available from the more complex social groups in which we currently live.

Appendix A
References from Figure 3

Bray, S. and J. O'Doherty (2007). "Neural Coding of Reward-Prediction Error Signals During Classical Conditioning With Attractive Faces." Journal of Neurophysiology **97**(4): 3036-3045.

Elliott, R., K. J. Friston, et al. (2000). "Dissociable Neural Responses in Human Reward Systems." Journal of Neuroscience **20**(16): 6159-6165.

Gallagher, H. L., F. Happe, et al. (2000). "Reading the mind in cartoons and stories: an fMRI study of `theory of mind' in verbal and nonverbal tasks." Neuropsychologia **38**(1): 11-21.

Harbaugh, W. T., U. Mayr, et al. (2007). "Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations." Science **316**(5831): 1622-1625.

Moll, J., F. Krueger, et al. (2006). "Human fronto-mesolimbic networks guide decisions about charitable donation." Proceedings of the National Academy of Sciences **103**(42): 15623-15628.

Rilling, J. K., A. G. Sanfey, et al. (2004). "The neural correlates of theory of mind within interpersonal interactions." NeuroImage **22**(4): 1694-1703.

Saxe, R. and N. Kanwisher (2003). "People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind"." NeuroImage **19**(4): 1835-1842.

Schultz, W., P. Dayan, et al. (1997). "A Neural Substrate of Prediction and Reward." Science **275**(5306): 1593-1599.

Tankersley, D., C. J. Stowe, et al. (2007). "Altruism is associated with an increased neural response to agency." Nature Neuroscience **10**(2): 150-1.

Tomlin, D., M. A. Kayali, et al. (2006). "Agent-Specific Responses in the Cingulate Cortex During Economic Exchanges." Science **312**(5776): 1047-1050.

Vollm, B. A., A. N. W. Taylor, et al. (2006). "Neuronal correlates of theory of mind and empathy: A functional magnetic resonance imaging study in a nonverbal task." NeuroImage **29**(1): 90-98.

Zahn, R., J. Moll, et al. (2007). "Social concepts are represented in the superior anterior temporal cortex." Proceedings of the National Academy of Sciences **104**(15): 6430-6435.

Appendix B
Experiment Instructions

Welcome to today's experiment. For the remainder of today's session, please keep your eyes on your own computer monitor, and refrain from speaking with other participants.

Your earnings today will be determined by the choices made by you and other participants in the experiment.

You will be paid your earnings in cash at the end of the experiment. You will be paid 1.7¢ for each point earned in the experiment, between 4-5 experimental points will earn you one dollar.

In this experiment, you will be matched with a person selected at random who you will be introduced to before the experiment begins.

If you are selected to be the decision maker, you will be the only person engaging in a task during the experiment. The other individual will be asked to leave the lab and come back at the end of the experiment to collect their earnings.

In this task you will be shown a screen with 5 letters. For example:

FJTSB

This screen will be followed by a screen with 3 letters:

BTS

Your task is to determine whether or not every letter on the second screen was present in on the first screen. In the case where all three letters in the second screen are present on the first, you should answer *YES*. In the case where one or more letters on the second screen were not present on the first, you should answer *NO*.

You will answer by pressing either the "Q" or "P" keys on your keyboard:

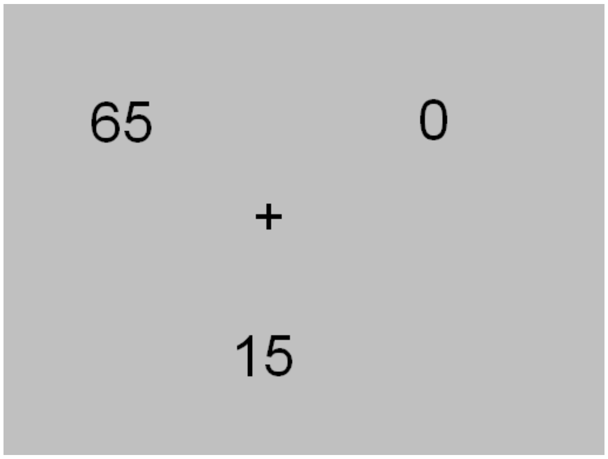
Q = *YES*
P = *NO*

Sometimes, a correct answer will earn you points, and sometimes it will earn the person you are paired up with points; sometimes, nobody will earn points for a correct answer.

At any given time, you will only be earning points for <u>one</u> person, yourself, or your counterpart.

Sometimes it will cost you money to press a button. In this case, points will be deducted from your total if you press a button, regardless of whether or not you answer correctly. Sometimes the points earned for getting a correct answer will be larger than the cost, and sometimes the cost will be larger than the points for getting a correct answer. However, if there is a cost, you will be charged points each and every time you press the button.

Before each series of letters you will see a screen which indicates how many points pressing the button will cost, and how many points will be earned for a correct answer:

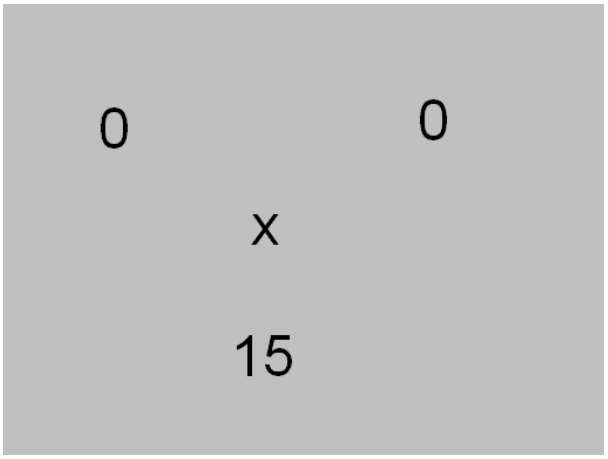


The points you will earn for a correct answer are always displayed in the above left corner of the screen. In the upper right hand corner, are the points earned for your counterpart.

The number on the bottom of the screen represents the number of points you will be charged for pressing a button. It will either be 15 or 0. You are the only person who is charged for pressing the button, the other person never pays that cost.

In the above example, you will earn 65 points for a correct answer, the other person will earn 0 points, and it will cost 15 points to press a button. If you answer correctly, you will have 65 – 15, or 50 points added to your total. If you answer incorrectly, you will have 15 points deducted.

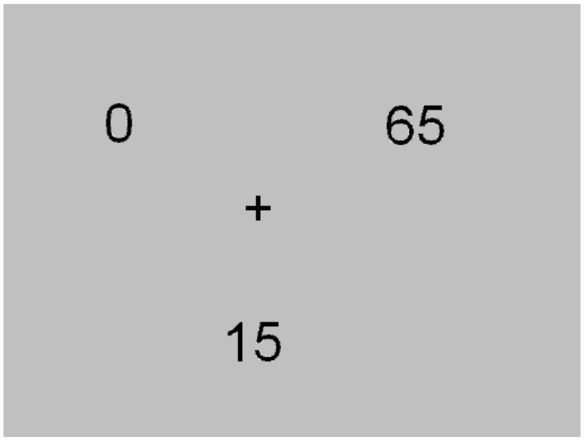
After you answer *YES* or *NO* you will see a screen representing the number of points earned based on your response. For example:



Note the “X” in the middle of the screen which is different from the “+” on the screen before the decision. When you see “+”, you are looking at the number of points you could earn and when you see an “X” you are looking at the number of points you have earned.

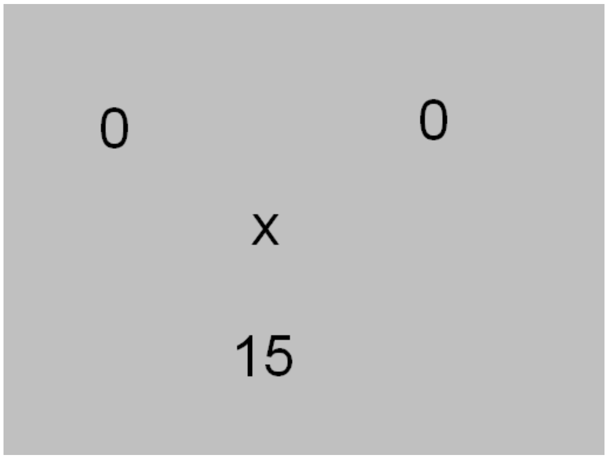
The screen above displays the case where you have answered incorrectly and there was a 15 point cost for pressing a button.

Let’s look at a few more examples.



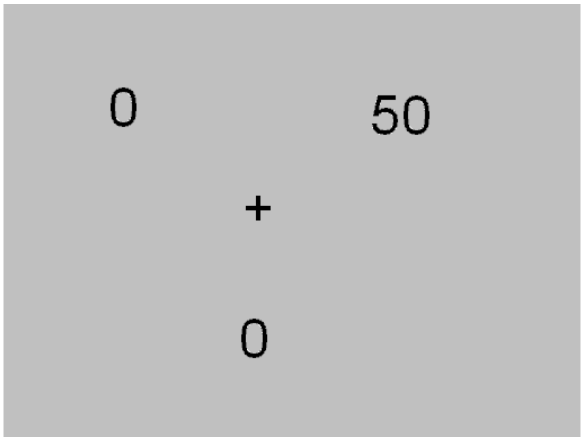
In the above example, your counterpart will earn 65 points if you answer correctly, and 15 points will be deducted from your total. If you answer incorrectly, your counterpart will earn 0 points and 15 points will be deducted from your total. If you do nothing, your counterpart will earn 0 points, and 0 points will be deducted from your total.

Supposing you answer incorrectly, you would see the following screen:



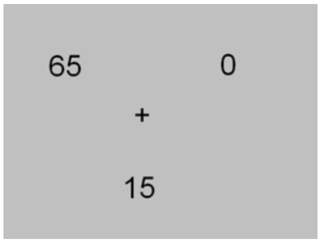
Again, 15 points would have been deducted from your score, and your counterpart would have earned 0 points.

Suppose that the next screen you see looks like:



If you answer correctly, your counterpart will earn 50 points, and 0 points will be deducted from your total. If you answer incorrectly, your counterpart will earn 0 points, and 0 points will be deducted from your total.

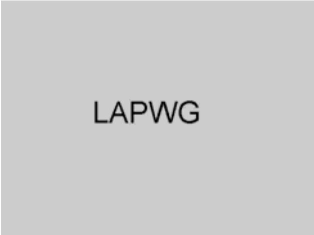
Let’s continue with a few examples. Please fill in the answers to the questions following each example. For your own payoff, please write down the total number of points added to your total (points earned – cost)



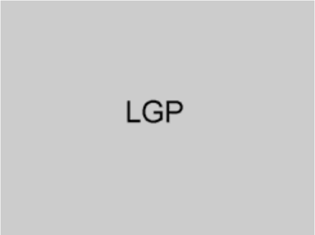
65 0

+

15



LAPWG



LGP

If you press the “P” key:

Points added to or
removed from your total: __________________

Points earned by
your counterpart: __________________

References

References

Akerlof, G. A. (1982). "Labor Contracts as Partial Gift Exchange." The Quarterly Journal of Economics **97**(4): 543-569.

Alexander, R. D. (1987). The biology of moral systems. New York, Aldine de Gruyter.

Andreoni, J. (1990). "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." The Economic Journal **100**(401): 464-477.

Andreoni, J. (1995). "Cooperation in Public-Goods Experiments: Kindness or Confusion?" The American Economic Review **85**(4): 891-904.

Andreoni, J. and J. Miller (2002). "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." Econometrica **70**(2): 737-753.

Apicella, P., T. Ljungberg, et al. (1991). "Responses to reward in monkey dorsal and ventral striatum." Exp Brain Res **85**(3): 491-500.

Axelrod, R. (1986). "An Evolutionary Approach to Norms." The American Political Science Review **80**(4): 1095-1111.

Barkow, J. H., L. Cosmides, et al. (1992). The Adapted mind : evolutionary psychology and the generation of culture. New York, Oxford University Press.

Bechara, A., H. Damasio, et al. (2000). "Emotion, Decision Making and the Orbitofrontal Cortex." Cereb. Cortex **10**(3): 295-307.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society. Series B (Methodological) **57**(1): 289-300.

Berridge, K. C. and T. E. Robinson (2003). "Parsing reward." Trends in Neurosciences **26**(9): 507-513.

Blair, R. J. R. (2005). "Responding to the emotions of others: Dissociating forms of empathy through the study of typical and psychiatric populations." Consciousness and Cognition **14**(4): 698-718.

Bolton, G. E. and A. Ockenfels (2000). "ERC: A Theory of Equity, Reciprocity, and Competition." The American Economic Review **90**(1): 166-193.

Bowles, S. and H. Gintis (2003). The Origins of Human Cooperation. Genetic and Cultural Evolution of Cooperation. P. Hammerstein. Cambridge, MIT Press.

Bowles, S. and H. Gintis (2004). "The evolution of strong reciprocity: cooperation in heterogeneous populations." Theoretical Population Biology **65**(1): 17-28.

Bray, S. and J. O'Doherty (2007). "Neural Coding of Reward-Prediction Error Signals During Classical Conditioning With Attractive Faces." Journal of Neurophysiology **97**(4): 3036-3045.

Burnham, T. C. and D. D. P. Johnson (2005). "The Biological and Evolutionary Logic of Human Cooperation." Analyze & Kritik **27**: 113-135.

Burock, M. A., R. L. Buckner, et al. (1998). "Randomized event-related experimental designs allow for extremely rapid presentation rates using functional MRI." Neuroreport **9**(16): 3735-9.

Carr, L., M. Iacoboni, et al. (2003). Neural mechanisms of empathy in humans: A relay from neural systems for imitation to limbic areas. **100:** 5497-5502.

Castelli, F., F. Happe, et al. (2000). "Movement and Mind: A Functional Imaging Study of Perception and Interpretation of Complex Intentional Movement Patterns." NeuroImage **12**(3): 314-325.

Coate, S. and M. Ravallion (1993). "Reciprocity without commitment : Characterization and performance of informal insurance arrangements." Journal of Development Economics **40**(1): 1-24.

Conlisk, J. (1996). "Why Bounded Rationality?" Journal of Economic Literature **34**(2): 669-700.

Cox, J., D. Friedman, et al. (2007). "A Tractable Model of Reciprocity and Fairness." Games and Economic Behavior **59**(1): 17-45.

Damasio, A. (2005). "Human behaviour Brain trust." Nature **435**(7042): 571-572.

Danielson, P. (2002). "Competition among cooperators: Altruism and reciprocity." Proceedings of the National Academy of Sciences **99**(90003): 7237-7242.

de Quervain, D. J. F., U. Fischbacher, et al. (2004). "The Neural Basis of Altruistic Punishment." Science **305**(5688): 1254-1258.

Decety, J., P. L. Jackson, et al. (2004). "The neural bases of cooperation and competition: an fMRI investigation." Neuroimage **23**(2): 744-51.

Deck, C. A. (2001). "A Test of Game-Theoretic and Behavioral Models of Play in Exchange and Insurance Environments." The American Economic Review **91**(5): 1546-1555.

Delgado, M. R., C. D. Labouliere, et al. (2006). "Fear of losing money? Aversive conditioning with secondary reinforcers." Social Cognitive and Affective Neuroscience **1**(3): 250-259.

Engelmann, D. and M. Strobel (2007). "Preferences over Income Distributions: Experimental Evidence." Public Finance Review **35**(2): 285-310.

Farrow, T. F. D. C. A., Y. Zheng, et al. (2001). "Investigating the functional anatomy of empathy and forgiveness." Neuroreport **12**(11): 2433-2438.

Fehr, E., U. Fischbacher, et al. (2002). "Strong reciprocity, human cooperation, and the enforcement of social norms." Human Nature **13**(1): 1-25.

Fehr, E. and J. Henrich (2003). Is Strong Reciprocity a Maladaptation? On the Evolutionary Foundations of Human Altruism. Genetic and Cultural Evolution of Cooperation. P. Hammerstein. Cambridge, MIT Press.

Fehr, E., E. Kirchler, et al. (1998). "When Social Norms Overpower Competition: Gift Exchange in Experimental Labor Markets." Journal of Labor Economics **16**(2): 324-351.

Fehr, E. and K. M. Schmidt (1999). "A Theory of Fairness, Competition, and Cooperation." The Quarterly Journal of Economics **114**(3): 817-868.

Forman, S. D., J. D. Cohen, et al. (1995). "Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold." Magnetic Resonance in Medicine **33**(5): 636-647.

Fox, M. and A. Uecker (2005). Talairach Daemon database. San Antonio, Texas, University of Texas Health Science Center.

Gallagher, H. L., A. I. Jack, et al. (2002). "Imaging the Intentional Stance in a Competitive Game." NeuroImage **16**(3, Part 1): 814-821.

Gintis, H., E. A. Smith, et al. (2001). "Costly Signaling and Cooperation." Journal of Theoretical Biology **213**(1): 103-119.

Goebel, R., F. Esposito, et al. (2006). "Analysis of functional image analysis contest (FIAC) data with Brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis." Human Brain Mapping **27**: 392-401.

Hagler, J. D. J., A. P. Saygin, et al. (2006). "Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data." NeuroImage **33**(4): 1093-1103.

Hamilton, W. D. (1964a). "The genetical evolution of social behaviour. I." Journal of Theoretical Biology **7**(1): 1-16.

Hamilton, W. D. (1964b). "The genetical evolution of social behaviour. II." Journal of Theoretical Biology **7**(1): 17-52.

Hendelman, W. (2006). Atlas of functional neuroanatomy. Boca Raton, FL, CRC Taylor & Francis.

Hobbes, T. (1946). Leviathan. Oxford, B. Blackwell.

Hoffman, E., K. McCabe, et al. (1994). "Preferences, Property Rights, and Anonymity in Bargaining Games." Games and Economic Behavior **7**(3): 346-380.

Hoffman, E., K. McCabe, et al. (1996). "Social Distance and Other-Regarding Behavior in Dictator Games." The American Economic Review **86**(3): 653-660.

Houser, D. and R. Kurzban (2002). "Revisiting Kindness and Confusion in Public Goods Experiments." The American Economic Review **92**(4): 1062-1069.

Isaac, R. M. and J. M. Walker (1988). "Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism." The Quarterly Journal of Economics **103**(1): 179-199.

Kahneman, D., P. Slovic, et al. (1982). Judgment under uncertainty : heuristics and biases. Cambridge ; New York, Cambridge University Press.

Kandori, M. (1992). "Social Norms and Community Enforcement." The Review of Economic Studies **59**(1): 63-80.

Kenning, P. and H. Plassmann (2005). "NeuroEconomics: An overview from an economic perspective." Brain Research Bulletin **67**(5): 343-354.

King-Casas, B., D. Tomlin, et al. (2005). "Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange." Science **308**(5718): 78-83.

Knutson, B., C. M. Adams, et al. (2001). "Anticipation of Increasing Monetary Reward Selectively Recruits Nucleus Accumbens." Journal of Neuroscience **21**(16): 159RC-.

Knutson, B. and J. C. Cooper (2005). "Functional magnetic resonance imaging of reward prediction." Current Opinion in Neurology **18**(4): 411-7.

Kosfeld, M., M. Heinrichs, et al. (2005). "Oxytocin increases trust in humans." Nature **435**(7042): 673-676.

Kuhnen, C. M. and B. Knutson (2005). "The Neural Basis of Financial Risk Taking." Neuron **47**(5): 763-770.

Levine, D. K. (1998). "Modeling Altruism and Spitefulness in Experiments." Review of Economic Dynamics **1**: 593-622.

Lynch, P. J. (2006). Brain human sagittal section, Wikimedia Commons.

McCabe, K., D. Houser, et al. (2001). "A functional imaging study of cooperation in two-person reciprocal exchange." Proceedings of the National Academy of Sciences **98**(20): 11832-11835.

McClure, S. M., G. S. Berns, et al. (2003). "Temporal Prediction Errors in a Passive Learning Task Activate Human Striatum." Neuron **38**(2): 339-346.

Moll, J., F. Krueger, et al. (2006). "Human fronto-mesolimbic networks guide decisions about charitable donation." Proceedings of the National Academy of Sciences **103**(42): 15623-15628.

O'Doherty, J., P. Dayan, et al. (2004). "Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning." Science **304**(5669): 452-454.

Olson, M. (1965). The logic of collective action; public goods and the theory of groups. Cambridge, Mass.,, Harvard University Press.

Oya, H., R. Adolphs, et al. (2005). "Electrophysiological correlates of reward prediction error recorded in the human prefrontal cortex." Proceedings of the National Academy of Sciences **102**(23): 8351-8356.

Padoa-Schioppa, C. and J. A. Assad (2006). "Neurons in the orbitofrontal cortex encode economic value." Nature **441**(7090): 223-226.

Rabin, M. (1993). "Incorporating Fairness into Game Theory and Economics." The American Economic Review **83**(5): 1281-1302.

Rawls, J. (1971). A theory of justice. Cambridge, Mass., Belknap Press of Harvard University Press.

Rilling, J. K., D. A. Gutman, et al. (2002). "A Neural Basis for Social Cooperation." Neuron **35**(2): 395-405.

Rilling, J. K., A. G. Sanfey, et al. (2004). "The neural correlates of theory of mind within interpersonal interactions." NeuroImage **22**(4): 1694-1703.

Rolls, E. T., S. J. Thorpe, et al. (1983). "Responses of striatal neurons in the behaving monkey. 1. Head of the caudate nucleus." Behavioural Brain Research **7**(2): 179-210.

Roth, A. E. (2002). "The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics." Econometrica **70**(4): 1341-1378.

Rubinstein, A. (1979). "Equilibrium in supergames with the overtaking criterion." Journal of Economic Theory **21**(1): 1-9.

Saaristo, A. (2006). "There Is No Escape from Philosophy: Collective Intentionality and Empirical Social Science." Philosophy of the Social Sciences **36**(1): 40-66.

Sanfey, A. G., J. K. Rilling, et al. (2003). "The Neural Basis of Economic Decision-Making in the Ultimatum Game." Science **300**(5626): 1755-1758.

Saxe, R. and N. Kanwisher (2003). "People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind"." NeuroImage **19**(4): 1835-1842.

Schultz, W., P. Dayan, et al. (1997). "A Neural Substrate of Prediction and Reward." Science **275**(5306): 1593-1599.

Selten, R. (1987). Equity and coalition bargaining in experimental three-person games. Laboratory Experimentation in Economics: Six Points of View. A. E. Roth. Cambridge, U.K., Cambridge University Press**:** 42-98.

Sethi, R. and E. Somanathan (2001). "Preference Evolution and Reciprocity." Journal of Economic Theory **97**(2): 273-297.

Sobel, J. (2005). "Interdependent Preferences and Reciprocity." Journal of Economic Literature **43**: 392-436.

Sugden, R. (1984). "Reciprocity: The Supply of Public Goods Through Voluntary Contributions." The Economic Journal **94**(376): 772-787.

Talairach, J. and P. Tournoux (1988). Co-planar stereotaxic atlas of the human brain : 3-dimensional proportional system : an approach to cerebral imaging. Stuttgart ; New York, G. Thieme ; New York : Thieme Medical Publishers.

Tankersley, D., C. J. Stowe, et al. (2007). "Altruism is associated with an increased neural response to agency." Nature Neuroscience **10**(2): 150-1.

Tobler, P. N., J. P. O'Doherty, et al. (2006). "Human Neural Learning Depends on Reward Prediction Errors in the Blocking Paradigm." Journal of Neurophysiology **95**(1): 301-310.

Tomlin, D., M. A. Kayali, et al. (2006). "Agent-Specific Responses in the Cingulate Cortex During Economic Exchanges." Science **312**(5776): 1047-1050.

Trivers, R. L. (1971). "The Evolution of Reciprocal Altruism." The Quarterly Review of Biology **46**(1): 35-57.

Trivers, R. L. (2005). Reciprocal altruism: 30 years later. Cooperation in Primates and Humans: Mechanisms and Evolution. P. M. Kappeler, Springer**:** 67-83.

Ulijaszek, S. J. (2007). "Obesity: a disorder of convenience." Obesity Reviews **8**(s1): 183-187.

Vollm, B. A., A. N. W. Taylor, et al. (2006). "Neuronal correlates of theory of mind and empathy: A functional magnetic resonance imaging study in a nonverbal task." NeuroImage **29**(1): 90-98.

Willinger, M. and A. Ziegelmeyer (2001). "Strength of the Social Dilemma in a Public Goods Experiment: An Exploration of the Error Hypothesis." Experimental Economics **4**(2): 131-144.

Worsley, K., A. Evans, et al. (1992). "A three-dimensional statistical analysis for CBF activation studies in human brain." Journal of Cerebral Blood Flow and Metabolism.

Yee, A. S. (1997). "Thick Rationality and the Missing "Brute Fact": The Limits of Rationalist Incorporations of Norms and Ideas." The Journal of Politics **59**(4): 1001-1039.

Yellen, J. L. (1984). "Efficiency Wage Models of Unemployment." The American Economic Review **74**(2): 200-205.

Zahn, R., J. Moll, et al. (2007). "Social concepts are represented in the superior anterior temporal cortex." Proceedings of the National Academy of Sciences **104**(15): 6430-6435.

CURRICULUM VITAE

Stephen J. Saletta attended classes at Oakland Community College, Washtenaw Community College, the University of Michigan, Michigan State University, and Wayne State University before graduating in 2001 with a Bachelor of Science in Economics from Eastern Michigan University, Ypsilanti, Michigan. He enlisted in the U.S. Air Force Reserve in 1998, earned a commission as Second Lieutenant in 2004 and the Joint Service Commendation Medal in 2006.