

A Neuroeconomic Perspective on Charitable Giving

David Yokum^{*}

dyokum@email.arizona.edu

Filippo Rossi[^]

filippor@email.arizona.edu

ABSTRACT

Psychologists and economists, particularly those assuming that people are rational egoists, have struggled to understand the causes of voluntary donation for decades. Why would a person decide to sacrifice part of his or her material payoff in order to increase the wellbeing of others? In the first part of this paper, we outline a core set of possible motivations, and then consider how those motivations can be used to construct behavioral models that can also be tested in terms of what we know about brain function. We emphasize the role of other-regarding preferences and argue that there are *moral* judgments, independent of any consideration of payoffs, that partially determine when and to whom such preferences exist. In the second part of the paper, we argue that a neuroeconomic perspective can help understand charitable giving, and then discuss recent neuroimaging studies that demonstrate this potential.

1. THE CHALLENGE OF VOLUNTARY DONATION

Homo economicus, that rational creature concerned only with its personal, material payoffs, is not an ideal candidate from whom to elicit a charitable donation. Unless there are offsetting material benefits, it will refuse to contribute anything to the public good. In this paper, we outline evidence demonstrating that people are more charitable than predicted by rational egoism alone, and then explore alternative models that use other-regarding motivations to explain the difference. Attention is given to unpacking the features determining when and to whom such motivations exist and, in particular, how neuroscientific evidence can inform the debate.

Two caveats before we get started. First, for the moment, we are restricting egoism to mean a self-centered focus on achieving *material* gains for oneself, without any regard for the interests of other people. A more inclusive definition of egoism would allow any self-centered focus, material or otherwise, to count. For example, it might be that the decision maker does not expect a material payback, but instead wants to feel the *pleasure* of giving money to the needy. This desire is still selfish in that it seeks to obtain personal benefit – the pleasurable sensation – rather than advance the welfare of the beneficiary. We will return to this possibility in detail later.

Second, egoism and altruism are defined differently by different researchers. For instance, biologists often discuss selfishness and altruism in terms of evolutionary fitness without any reference to intentionality. A different question, the one which we will focus on, is whether the proximate, psychological system motivating the decision to act altruistically is selfish, altruistic, or both. It is a question of what beliefs and desires are driving behavior. To put it

^{*} Neural Decision Science Lab Department of Psychology The University of Arizona

[^] Neural Decision Science Lab Department of Psychology The University of Arizona



loosely, do humans have a cognitively represented, ultimate desire of the form “I want others to be well-off” – a desire that exists independently of their concern for their own wellbeing? Or is the only ultimate desire instead “I want to be well-off,” perhaps coupled with auxiliary beliefs that acting charitably will somehow advance *that* goal? To say that humans act charitably because they are psychologically motivated by altruism is to say that they possess an ultimate desire to advance others’ wellbeing, irrespective of their own wellbeing. They have other-regarding preferences. Non-altruistic explanations of charity, in contrast, either deny altruistic impulses altogether or explain them away as instrumental desires: they exist only to advance ultimate desires that are selfish.

Returning to the original discussion, whatever its descriptive limitations, the assumption of rational egoism has stimulated the creation of formal models, such as in game theory,¹ that have precise, testable predictions about how persons will behave toward one another. This provides a useful benchmark from which to compare real behavior and, as we will see, helps orient research aimed at understanding underlying neural mechanisms. *Public good games* are immediately relevant for our purposes. Within a typical public good design, each individual is given a monetary endowment and must decide whether or not to contribute some portion of that endowment to a common pool from which all participants will benefit; the transferred amount is often increased by a stated factor (Ledyard, 1995; Gächter et al. 2009). For example, each subject receives ten euro, and any money donated to the common pool is multiplied by two before being evenly redistributed back to the participants. If there are four subjects, and everyone contributes his or her entire endowment, then the final common pool is $4[10 \times 2] = 80$ euro, and each individual receives back 20 euro – a two-fold increase from the initial endowment. This is obviously a better outcome for all parties than if no one contributes, for in that case each individual remains with only ten euro. Nevertheless, *Homo economicus* will contribute nothing. The reason is that the highest payoff is achieved if one defects while everyone else contributes. In that case the payoff is $10 + (3[10 \times 2])/4 = 25$ euro – a five euro increase from when personally contributing as well. The other egotistical players, of course, will also realize the benefits of unilateral deviation. The end result is that no one contributes to the common pool (Dawes & Thaler, 1988). Such an undesirable outcome is a Pareto-deficient equilibrium, since everyone would be better off if no one defected, but no one actually does so because of the even higher potential benefits of unilateral deviation.

Pareto-deficiency is a classic obstacle in many one-shot games, and it is also the basis of the free-rider problem. In the context of charitable giving, it means that each person is predicted to refuse donating time or money, relying instead on the contributions of other people – thereby getting a “free-ride” to a public good made possible by *others’* donations. However, real behavior in laboratory and naturalistic settings negates this prediction (Ledyard, 1995; Camerer, 2003; Andreoni, 1995). Rather than transferring zero euro in the one-shot public good game, for example, subjects typically donate 40-60% of their endowment (Marwell & Ames, 1981). Comparable rates are also contributed in the first round of a finitely repeated version of the game (Isaac & Walker, 1988; Kim & Walker 1984).²

Charitable behavior is not restricted to economic games either. A recent survey in the United States revealed that 89% of households donated an average of \$1,620, or 3.1% of household income, to charitable organizations, and almost half of all American adults volunteered their time to participate in such groups (Independent Sector, 2001). Such

¹ See von Neumann and Morgenstern (1944) and Osborne and Rubinstein (1998).

² For a more general discussion on the public good game, see Ledyard (1995) and Camerer (2003).



estimates fluctuate across surveys and countries (Andreoni, 2006; 2008), but in all cases charitable giving is well above the predicted level of zero contribution.

Initially it was thought that the egoism assumption could be preserved at the expense of rationality. For one reason or another, people make mistakes and donate to the public good against their own best interest. Such a hypothesis is partially consistent with data from repeated public good games. Although subjects in an iterated public good game make first-round donations that are comparable to those in a one-shot interaction, the donation rate begins to decay with each subsequent round until, in the final round, most subjects are actually behaving as egotistical players (Dawes & Thaler, 1988; Fehr & Schmidt, 1999; Kim & Walker, 1984). It was therefore hypothesized that subjects were initially confused by a rather odd experimental setting, and that behavior during later rounds, once the rules are learned, is more informative of what would actually happen in naturalistic settings.

Andreoni (1995) directly addressed the possibility of confusion. He used three different public good games: (a) a standard version; (b) a version where players were paid according to the relative rank of their earnings (it was, in other words, transformed into a zero-sum game); and (c) a modified standard version with rank information (which was inconsequential with respect to payment, and meant only to control for the effects of information about rank on behavior). Note that in condition (b) there is no incentive to cooperate, due to its zero-sum nature. Donation in this case can therefore be attributed to confusion, for it benefits no one. The results replicated the common finding for the standard condition (a): cooperation decayed with each additional round, although donation rates never dropped below 25%. Subjects in condition (b), however, quickly and drastically altered their behavior, donating less than 10% by the fourth round. Analyses across conditions revealed that confusion was indeed *one* explanatory factor for early round behavior, but that it was unable to fully explain the persistent willingness to donate. Moreover, the observed late round defection was best explained, not by learning, but rather as a reaction to free-riding: players initially try to cooperate, but after their attempts are exploited by others, they turn to defection as a form of retaliation.³

The so-called “re-starting effect” (Andreoni, 1988; Cookson, 2000) provides further evidence against the confusion hypothesis. If subjects are learning that donation is against their own interests with each subsequent round of a public good game, then it should not matter if the game is interrupted and then restarted. If players are learning that the best strategy, for a rational egoist, is not to donate then we would predict that cooperation should constantly decline across rounds of the repeated game. If, on the other hand, subjects are sensitive to the decisions of their partners (rather than merely ignorant of the game rules), then a new start may re-open possibilities for cooperation. The experimental data confirms that donation rates, contrary to the confusion hypothesis, do in fact bump back up to high rates when the game is stopped and restarted. The upshot is that confusion *per se* is unable to explain the observed rates of contribution.

³ See also Goeree et al. (2002) for an investigation of this issue within the framework of stochastic game theory. They used a series of (quasi) one-shot public good games that varied: (a) the number of players in a group; (b) the magnitude of how much a player received in return for contributing one unit of his or her endowment to the public good; and (c) how much each *other* player in the group received from the unit contributed by the player. Their data revealed that both altruism and behavioral noise are explanatory factors of behavior in a public good setting. See also Palfrey et al. (1997).



Substantial evidence therefore indicates that *Homo economicus* is an endangered species, and that for most people the assumption of raw egoism focused on material payoffs simply does not hold. So why, then, do people voluntarily donate their time and money to a public good even when there are not obvious offsetting material benefits? The terrain of candidate explanations is complex, spanning disciplines such as economics, psychology, and evolutionary biology, and there are a variety of terms used by different authors that overlap in some ways but not others. In the next section, we attempt to outline several major conceptual possibilities, but make no claim to being exhaustive. Our aim is to capture a set of essential motivations, even if necessarily oversimplified, in order to begin constructing behavioral models that can also be tested in terms of what we know about brain function.

2. COMPONENTS OF CHARITABLE DONATION

It is worth noting that the demise of *Homo economicus* in no way obliterates the role of selfishness. Most people act charitably under the right circumstances, but not everyone does. And for those with other-regarding preferences, there is no reason to think that some degree of selfishness no longer exists. The motivations outlined here should therefore be understood as expanding rather than replacing selfish models. People act charitably for selfish reasons at times, but non-selfish reasons are also necessary to explain the level of generosity that exists, which is well above that predicted by selfish models, as we saw in the previous section. Bearing this in mind, we now turn to four components of charitable decisions that are different from seeking personal material payoffs and warrant special scrutiny: warm-glow, preferences related to fairness, reciprocity, and deservedness.

2.1 WARM-GLOW

James Andreoni (1989; 1990) provides a useful distinction, often cited in the charitable giving literature, between pure and impure altruism. The dichotomy captures the possibility that a person might donate either because an increase in a public good is desirable *per se* or, alternatively, because he or she experiences a sort of selfish, personal satisfaction from the very act of giving. To consider this formally, let w_i represent person i 's wealth, and assume that there is one private and one public good. Wealth can be spent either on personal consumption of the private good, denoted x_i , or given as a charitable donation to the public good, denoted g_i . Let G represent the overall public good and be equal to the sum of all donations, that is, person i 's donation (g_i) plus everyone else's donation. People can then be assumed to have a utility function $U_i = U_i(x_i, G, g_i)$, which they seek to maximize. What this means is that people derive satisfaction from their private consumption (x_i), the level of public good (G), and their personal contribution to that public good (g_i), and thus their decision of how much to donate, if at all, will reflect a balance between these three considerations.

A person is said to be purely altruistic if he or she cares about the status of the public good but not how it is achieved, except insofar as it affects their level of private consumption. The utility function in this case reduces to $U_i = U(x_i, G)$. A person is a pure egoist, on the other hand, if he or she cares only about private consumption and personal contribution. In this case, $U_i = U(x_i, g_i)$. The g_i term, as part of the utility function, captures a phenomenon referred to as "warm-glow." It reflects a type of satisfaction that is independent of that derived from achieving a given level of G . It is instead a satisfaction derived immediately from the giving act itself – a good feeling from personally helping out, so to speak. Someone who is neither purely



altruistic nor purely egotistical, that is, a person who is sensitive to both G and g_i , is considered to be an impure altruist.

What we want to stress here is the “warm-glow” component of the model. The variable for private consumption, x_i , represents the continued presence of selfish motivations in charitable decisions, as we stressed at the onset of section two. And the sections 2.2 and 2.3 below could be interpreted as further elaborating what is represented by G , in other words, what it means to have truly other-regarding preferences. Warm-glow is unique, however, in that it does not cleanly fit into either of the traditional categories of selfishness and other-regarding preference. There is no material payoff at stake, and it seems odd to call someone who enjoys helping others selfish. Nonetheless warm-glow is, in an important sense, entirely selfish. The act of giving is merely a means to the end of personal satisfaction, and at bottom the well-being of others is not a concern. To see that this is the case, consider the following decision: should you personally donate one euro to a charitable group or allow the charitable group to receive instead one-thousand euro from someone else. The person motivated only by warm-glow will select the former, despite the substantial monetary loss to the charitable organization.

2.2 SOCIAL PREFERENCES

Theories of fairness assume that subjects have preferences regarding the way in which certain resources are allocated. An allocation that satisfies these preferences is called fair. We assume that an unfair allocation can motivate a decision-maker to act in a compensatory manner, that is, to seek the obtainment of a fair allocation. In this sense, fairness enters as a motivational force in a person’s preferences (Camerer, 2003, p. 114). A theory of charitable giving should therefore accommodate the relationship between voluntary donation and the theory of justice of the donators.

In the economical literature, four main types of social preferences are usually discussed (see Charness & Rabin, 2001, for a review): self interested, competitive, social-welfare, and differences-aversion motives. Selfishness was considered above. A competitive motive is not particularly interesting from the point of view of voluntary donation: if a subject’s preferences are such that she prefers to be better off than her opponents, it is unlikely for her to donate. Let us then focus on the last two preference types.

Consider the linear model below, adapted from Charness and Rabin (2001), which can be used as a simplified representation of different kinds of distributional preferences:

$$U_i(\pi_i, \pi_j) = \pi_j(\rho \cdot r + \sigma \cdot s) + \pi_i(1 - \rho \cdot r - \sigma \cdot s) \quad (1)$$

In a two-player interaction, equation (1) entails that the utilities derived from material payoffs for players i and j , denoted π_i and π_j , respectively, depends on: (a) the magnitude of the two payoffs and (b) the parameters ρ and σ , which provide a way of modeling different kinds of distributional preferences. The strategic interaction is described through r and s . If i ’s material payoff is bigger than j ’s (i.e., $\pi_i > \pi_j$), then $r = 1$; otherwise it is zero. If instead $\pi_i < \pi_j$, then $s = 1$; otherwise it is zero.

Andreoni and Miller (2002) and Charness and Rabin (2001) proposed a quasi-maximin model of social-welfare preferences, where subjects are concerned about their material payoff, the material payoff of the individual with the lowest payoff in the society, and finally the public good (in terms of the model in (1), this translates into the following constraints: $1 \geq$



$\rho \geq \sigma > 0$, $\sigma \leq 1/2$). The upshot of this model is that subjects prefer more for themselves, are more likely to help when they are better off, and prefer Pareto-improvements.⁴ Charness and Rabin (2001) found that the quasi-maximin model was capable of organizing the data from a battery of thirty-two dictator and response games, and thus that it may capture important features of people's distributional preferences.

Another important theory has been proposed by Fehr and Schmidt (1999), one that is particularly pertinent to the analysis of public good games.⁵ They present a two parameter model in which subjects are concerned about the differences between their payoff and the payoffs of the other players; it is essentially a type of "self-centered inequity aversion." The intuition behind Fehr and Schmidt's proposal is that when subjects have more of a given resource, they are more willing to sacrifice that resource in order to compensate for payoff differences. On the other hand, when subjects are behind, they are willing to hurt their opponents in order to shrink the distance between the material payoffs. This can be written formally as:

$$U_i(\pi) = \pi_i - \alpha_i \max\{\pi_j - \pi_i, 0\} - \beta_i \max\{\pi_i - \pi_j, 0\} \quad (2)$$

The parameter β_i captures the degree to which subjects care about the inequality of the payoff when they are ahead, while α_i represents their concern when they are behind. A reasonable constraint is that α_i should be greater than or equal to β_i .⁶ Fehr and Schmidt also assume that $0 \leq \beta_i < 1$, which distinguishes their theory from a competitive preferences model, where subjects prefer to be better off than other players. This model is also capable of explaining a wide variety of experimental data.

The above models are highly simplified, and there is no agreement on which is best for representing social preferences. Charness and Rabin therefore provide some sound advice:

Too much will be lost if experimentalists jump too quickly to calibrating highly simplified model [...] At this stage, models ought to be developed that help to interpret psychologically sound and empirically prevalent patterns of behavior common in a broad array of games. (2001, p. 821)

What is most important for our purposes is the explicit representation of some sort of fairness preference, namely, a concern about how resources are distributed that is sensitive to the payoffs of others. Whether this motivation reflects a desire to enhance the well-being of the worst off, to simply avoid inequity, both, or some other sort of other-regarding preference is, at this stage, too early to determine. However, as experimentalists tackle this issue, we think it necessary to keep in mind that people are unlikely to have social preferences that are uniformly applied across people and situations. Such preferences are not blind, so to speak, and as such models of charitable giving should be sensitive to qualifications about when and to whom charitable motivations apply. We turn now to this possibility.

⁴ See also Yaari and Bar-Hillel (1984).

⁵ See also Fehr and Fischbacher (2002) and Bolton and Ockenfelds (2000).

⁶ It would be surprising, after all, if people cared about inequality when they were ahead *more* than when they were behind, for in the latter case selfish motivations would also be predicted as a motivating factor to achieve equality. Also, as Fehr and Schmidt note, this assumption is in line with the loss aversion literature (e.g., Tversky and Kahneman, 1991), which reflects the common saying that "losses loom larger than gains."



2.3 RECIPROCITY AND DESERVEDNESS

The theories outlined above do not specify anything about the qualities of the recipients of a charitable act, but instead focus exclusively on their material payoffs. Nonetheless, issues such as whether the potential recipients *deserve* your help or satisfy some moral criteria are important candidates for inclusion in the consideration of whether or not to act charitably (see Rabin, 1993; Dufwenberg et al., 2004; Charness & Rabin, 2001). As Dufwenberg et al. (2004) note:

The assumption that individuals only care about final distributions implies that they must be indifferent concerning *how* distributions come about. This is problematic if in fact individuals regard information about their co-players' specific choices or intentions as important to their decision-making. (2004, p. 260-70)

What determines whether someone deserves help? In the economic literature, deservedness is strictly related to fairness in the form of reciprocity (Rabin, 1983; Dufwenberg et al., 2004). Charness and Rabin (2001), for example, focus their attention on *withdrawal reciprocity*. The idea is that subjects will withdraw their willingness to act charitably toward people who are unwilling to sacrifice for the sake of fairness. They observed that this kind of reciprocation is particularly important in the case of simple dictator and response games.

In the game theoretic literature, Rabin (1993, p. 298) developed the notion of “fairness equilibrium” on the basis of three intuitions: (a) people are willing to sacrifice their own material well-being to reciprocate kindness; (b) they are also willing to sacrifice well-being to punish unkindness; and (c) the first two conditions significantly affect human behavior. Dufwenberg and Kirchsteiger (2004) developed this notion for extensive form games,⁷ and their formulation provides an example of the kind of formalism that may be implemented in order to qualify when and to whom other-regarding preferences apply.

Call $A_{i \in N}$ the set of player i 's strategies, with N the set of players. Denote with A the set of strategy profiles (the Cartesian product of the sets A_i). Dufwenberg and Kirchsteiger define two components of the utility function for player i : his material payoff π_i and the *reciprocity payoff*. This latter component depends on the players' beliefs at different stages of the game (Battigalli & Dufwenberg, 2009). In particular, there are: (a) the set of possible beliefs of player i on j 's strategies, $B_{ij} = A_j$ and (b) player i 's beliefs about player j 's beliefs about player k 's strategy, $C_{ijk} = B_{jk} = A_k$. Players update these beliefs throughout the game on the basis of the previous actions of their opponents. For example, $b_{ij}(h)$ represents an updating of B_{ij} based on h , a possible history in the game.

From here we can define a kindness index, which can be used to discriminate the players with whom to be kind. Following (in part) Rabin's original formulation (1993), Dufwenberg and Kirchsteiger define a *references* payoff (π_j^e) as the average between the highest and the lowest payoffs that player i can give to player j by playing one of his efficient strategies (Dufwenberg et al., 2004, p. 276). The distance and sign of the material payoff of j from the reference payoff defines the kindness of i toward j :

⁷ This theory belongs to the more general framework of psychological game theory; see Genakopolos et al. (1989) and Battigalli and Dufwenberg (2009).



$$k_{ij}(a_i(h), (b_{ij}(h))_{j \neq i}) = \pi_i(a_i(h), (b_{ij}(h))_{j \neq i}) - \pi_j^e((b_{ij}(h))_{j \neq i}) \quad (3)$$

A further term (analogous to (3)) can be constructed that measures whether player i thinks that j was kind with i (λ_{jji}). Through the material payoff of i and the kindness functions, we can finally define the following utility function:

$$\begin{aligned} U_i & \left(a_i(h), (b_{ij}(h), (c_{ijk}(h))_{k \neq j})_{j \neq i} \right) \\ & = \pi_i(a_i(h), (b_{ij}(h))_{j \neq i}) + \sum_{j \in N \setminus \{i\}} \left(Y_{ij} \cdot k_{ij}(a_i(h), (b_{ij}(h))_{j \neq i}) \lambda_{jji}(b_{ij}(h), (c_{ijk}(h))_{k \neq j}) \right) \end{aligned} \quad (4)$$

The first term stands for i 's material payoff. The second term is of particular interest here. Notice that k_{ij} is negative when i behaved unkindly with j , while λ_{jji} is negative when j was unkind with i . When only one of the terms is negative, the value of U_i decreases; this captures the fundamental intuition underlying reciprocity: a mismatch between kind and unkind behavior triggers a violation of reciprocity. Finally, Y_{ij} is an exogenous term, which models how much player i cares about reciprocity (when it equals zero, the utility function reduces to the material payoff component).⁸

The relation between this approach to reciprocity and voluntary giving is indirect, but particularly relevant. In particular, Dufwenberg and Kirchsteiger's theory is one of the most developed ways of formally imposing constraints on other-regarding preferences. This psychological game theoretic approach can, in principle, accommodate the imposition of further constraints based on the *properties* that a recipient should have from the point of view of the donator in order to be eligible for a charitable act. This would constitute a novel and important addition. What we have in mind, specifically, is that moral considerations are relevant, considerations that are related to (but subtly different from) desert, or the idea that a person deserves something based upon his or her actions.

As an example of data motivating this hypothesis, consider the debate surrounding the system of welfare in America. This is an ideal test case, for the welfare system constitutes an egalitarian redistribution of income among total strangers, and to the extent that it is supported through votes it is a voluntary system – it is, in other words, an act of charity. Nonetheless, popular support for welfare is mixed, with many individuals adamantly opposed to its continuation. Why might people oppose welfare or, to put it differently, refuse to be charitable in this case? Surprisingly, income level, education, and a variety of demographic variables are incapable of adequately discriminating between supporters and detractors; voters are not really concerned about the cost of welfare or fraud; and even agreeing with the idea that income should be evenly redistributed is not a very good predictor. The overwhelmingly significant predictor is instead the answer to the following question: does bad luck cause poverty? Those who believe that poor people are somehow responsible for their poverty are far more likely to oppose welfare (see Fong et al., 2005, for a full discussion). This form of responsibility, or rather irresponsibility, can take several forms. Fong et al. (2005), for example, report survey results indicating that, by more than a five-to-one margin, respondents

⁸ The model as presented here is only a partial representation. For the full structure, see Dufwenberg and Kirchsteiger's original paper (2004).



believe that welfare recipients could obtain employment if they tried, 70% of respondents believe it is more financially rewarding to stay on welfare than get a job, 57% believe welfare encourages laziness, and 60% believe welfare encourage out-of-wedlock childbirth.

The welfare survey data reveal a refusal to financially support people who, it is believed, are engaging in socially unacceptable behaviors, such as refusing to contribute to the common good via employment or having children outside of marriage. To some extent the opposition to welfare can be interpreted in terms of reciprocity. What is meant by reciprocity requires specification. Most authors use reciprocity to mean reciprocal altruism in particular, or a sort of tit-for-tat in which the benefactor expects the recipient to somehow repay the favor, even if unknowingly. This is not the most parsimonious explanation, however. After all, welfare supports anonymous strangers that any given individual is unlikely to ever meet, and even many very wealthy persons – who presumably stand little to gain – support welfare. More pertinent is what is referred to as *strong reciprocity* (Gintis, 2000; Gintis et al., 2005). The theory is that people tend to behave pro-socially and punish antisocial behavior, at a cost to themselves, even when the probability of future interactions is extremely low, or even zero. This is fundamentally different from the *weak reciprocity* associated with reciprocal altruism, for it remains a motivating factor even when future compensation is unlikely or impossible.

Strong reciprocity is different from pure altruism, however, for it is still conditional. It is just that the condition is not of equivalent worth as the beneficent act. What this means is that the possible recipient must satisfy certain criteria, for instance (to continue with the welfare example) be willing to work. But it is *not* the same as saying the potential recipient has a legitimate claim to compensation in the standard sense of reciprocity. In other words, the recipient has not actually provided an offsetting benefit to the benefactor, nor does the benefactor *ever expect* the recipient to return the favor. Rather, the benefactor has a true other-regarding preference, one in which he or she is genuinely concerned with the payoff of another individual that is independent of his or her own payoff – but that concern, importantly, only exists for those who satisfy the relevant moral criteria.

Selfishness, warm-glow, preferences related to fairness, reciprocity, and deservedness are therefore all crucial components in the decision to act charitably. Other factors might be important as well, but these motivations, incorporated into economic models such as those outlined above, provide a generative starting point for the neuropsychological investigation of charitable giving. Indeed, it is difficult to even begin research into the neurobiology of charitable giving without the aid of precisely formulated models. Having outlined such models, we now turn to the very new and very small neuroscientific literature on charitable giving (see Mayr et al., 2009, for a review).

3. NEURAL EVIDENCE

Neuroeconomics is an experimental approach that couples the mathematical precision and simplifying assumptions of economic models with the methods and data from cognitive neuroscience (Glimcher et al., 2009). Neural data is useful for the study of charitable giving, as well as other economic behaviors, for several reasons. To begin with, neural evidence places a constraint on the range of viable psychological theories, since proposed cognitive mechanisms must align with possible neurophysiological function. In essence this works by enhancing the epistemic criteria of internal consistency: not only must the proposed cognitive mechanism align with behavioral data, but now it must also mesh with theories about different brain regions and the functions they perform. Second, brain activity during charitable giving is likely to overlap with brain activity observed in other contexts and for which there is already a good deal of theoretical understanding. This means that our understanding of brain function in



general can be used to help interpret the neural signals observed during charitable behavior and, potentially, stimulate hypotheses for testing at cognitive and behavioral levels. Finally, neural evidence might reveal motivations for charitable giving that are not accessible behaviorally, either because subjects are unwilling to report or unconscious of the actual impetus for their behavior.

This final point is especially relevant for uncovering the psychological motivations driving charitable behavior. As mentioned previously, egoism can be broadly construed to entail non-material selfish benefits, such as warm-glow. This is a legitimate interpretation, but the consequence is that egoism becomes seemingly impossible to refute on the basis of behavioral data alone. One problem is that self-reported reasons claiming genuine altruism can almost always be reinterpreted as an instrumental desire toward a selfish, ultimate desire, one that subjects may be unable to access. And it is difficult to bypass this obstacle with behavioral evidence other than self-report because a decision maker, whether motivated by warm-glow or a true other-regarding preference, will likely *act* the same. Helping other people for their own sake and doing it only to generate warm-glow feelings both require *actually* helping. This is *not* to say that behavioral data are irrelevant, only that it is difficult to disentangle motivational theories that make similar behavioral predictions.

But what if we could more directly observe the motivational systems at work? Sober and Wilson (1998, p. 205-208) discuss the example of marine bacterium that are obligate anaerobes (they cannot survive in the presence of oxygen) as an analogy for the problem of teasing apart psychological egoism from psychological altruism. They note that the bacterium could avoid oxygen by utilizing a device that directly detects oxygen or, alternatively, it could use a magnetosome sensitive to the gravitational field of the earth (oxygen is more abundant near the water surface, so gravity could be used to as a guide in swimming away from the surface). But bacteria with either an oxygen detector or a magnetosome would behave the same – they would swim downward. However, assuming we knew enough about bacterial anatomy, we could resolve this dilemma by dissecting the bacterium and observing what devices are inside. Does it have an oxygen detector or does it have a magnetosome? Sober and Wilson note that this is possible, in principle, for the question of how human altruism operates in the brain but, writing a decade ago, rightly concluded that “even if neurobiology will answer this question one day, it offers little guidance now” (p. 207). This, fortunately, is beginning to change.

The neural investigation of charitable giving specifically is only just beginning. Jorge Moll and colleagues (2006) conducted one of the first functional magnetic resonance imaging (fMRI) studies. Nineteen subjects were endowed with U.S. \$128. Each person was then presented with a series of real charitable organizations, including a brief mission statement for each, and given the opportunity to either donate or oppose a donation to each organization from his or her endowment. Decisions were strictly anonymous, and subjects were aware that real money was at stake, both for themselves and the stated charitable organization. The design entailed several possible payoff conditions for each decision: pure monetary reward (YOU: \$+2, ORG: \$0), non-costly donation/opposition (YOU: \$0, ORG: \$5), costly donation/opposition (accepting YOU: \$-2, ORG: \$5; refusing YOU: \$2, ORG \$5). Note that the pure monetary reward condition has no consequences for the charitable organization, and therefore the decision should reflect solely egotistical preferences. The non-costly condition, on the other hand, has no personal consequences, and therefore the decision should reflect solely other-regarding preferences for the charity. The costly conditions entail a conflict between egotistical and other-regarding motivations.



There were two main results. First, the ventral tegmental area (VTA) and striatum, components of the brain's reward system, were activated by both pure monetary rewards and decisions to donate. This finding suggests that charitable giving and personal gains share the anatomical systems underlying reward reinforcement and expectancy. Second, a direct contrast between the pure monetary reward and donation conditions revealed that neural activity was greater in the ventral striatum for donations and, most intriguingly, that the subgenual area was uniquely activated during donations. To the extent that the donation condition elicits the experience of reward above and beyond the monetary gain condition, the interpretation of warm-glow motivation is reasonable. In other words, the enhanced ventral striatum activity might represent the experience of utility from the act of giving *per se*.

But the presence of warm-glow does not rule out the existence of other motivations as well. The subgenual area has been repeatedly implicated in the expression of social attachment (e.g., Bartels & Zeki, 2004; Aron et al., 2005) and the release of neuromodulators, such as oxytocin and vasopressin, which are likewise thought to have important social functions (e.g., Zak et al., 2005). It is therefore tempting to speculate that subgenual activity reflects neural evidence of some form of conditional altruism or strong reciprocity. At the very least, the subgenual area is a neuroanatomical region outside of the putative reward system, and thus charitable decisions entail a type of processing that is perhaps meaningfully different from the utility processing of personal payoffs. Such a conclusion really is speculative though. Aside from awaiting replication, the interpretation of subgenual activity is somewhat complicated by the fact that the charitable organizations used in the study were deliberately chosen on the basis of their support or opposition to socially controversial issues, for instance euthanasia, gun control, and abortion. The observed activity might therefore reflect a reaction to the social issue that is independent of the decision to donate. This seems unlikely, however, and as we argued above social values and moral assessments are important considerations in the decision of whether or not to donate to a particular public good. To put it differently, the subgenual activity might very well reflect a reaction to stimulating socially values, but that reaction exists as a relevant factor in the decision of whether or not to behave charitably.

Harbaugh et al. (2007) conducted a second neuroimaging study, this time without the additional variable of moral assessment and with a design that more clearly tested the possibility of other-regarding preferences. Nineteen subjects were endowed with US \$100 and given a series of payoff situations which they could either accept or reject. The charity was a food bank. There were also, however, three unique conditions that did not require a decision: a tax condition, in which subject money was involuntarily transferred to the food bank; a pure mandatory payment to the subject, at no cost to the charity; and a pure mandatory payment to the charity, at no cost to the subject. Activation of the reward system was again found for both personal gains and gains to the charity. There was also greater striatum activity during voluntary rather than mandatory donations, again supporting the existence of warm-glow. However, in a contrast between the mandatory treatments, activation in the ventral striatum was found for both personal monetary gains and the mandatory tax. This latter activation cannot possibly be interpreted as a warm-glow effect. The subject, after all, did not volunteer to donate the money.

The most impressive finding, however, was the predictive model that Harbaugh et al. tested. Recall the impure altruism utility function, $U_i = U_i(x_i, G, g_i)$. They used neural activity during the pure mandatory payments to each subject as an indicator of his or her marginal utility of money (x_i), and activity during pure mandatory payments to the charity as an indicator of the marginal utility derived from increases in the public good (G); the transfers were mandatory, so g_i (personal donation) was irrelevant. Those with a larger neural response to the mandatory personal payoff were labeled "egoists," and those with a larger response to



the mandatory payoff to the charity “altruists.” As predicted, the labeled altruists were more charitable, giving nearly twice as much. The authors reasonably interpreted this as evidence of a purely altruistic motive. The larger the neural response to increases in G , no matter the source (that is, regardless of g_i), the more likely one is to give voluntarily.

4. CONCLUSIONS

We have argued in this paper that selfish motivations are only one component of decisions related to charitable behavior. People also value warm-glow feelings and have genuine other-regarding preferences, such as concerns about how resources are distributed and strong reciprocity. Such other-regarding preferences are not blind, however, and are instead constrained by a variety of criteria related to the deservedness of the potential recipient. These constraints are not related to material compensation but are rather, in an important sense, moral considerations.

The Moll et al. (2006) and Harbaugh et al. (2007) studies are merely a taste of what is to come from the neuroeconomic approach. These two studies demonstrated that it is possible to identify neuroanatomical regions uniquely activated during conditions of selfishness, pure-altruism, and warm-glow. The data at present do not conclusively resolve the psychological egoism versus altruism debate by any means, but this work is promising. Dissecting the brain, as we might to see if a bacterium has an oxygen detector or magnetosome, is not as far-fetched as it once seemed. For example, the evidence that reward areas of the brain are activated by merely witnessing a charitable act is compelling evidence that humans do, in fact, donate for other-regarding reasons, contrary to the egoism hypothesis. To appreciate the unique force of these data, consider someone who claims to enjoy the fact that another group is benefitted by charity as an end in itself. An advocate of egoism could dismiss this claim as disingenuous; perhaps the person merely *says* such things to earn the approval of the experimenter, or is somehow deluding him or herself. But to also dismiss the neuroimaging data, the critic would have to argue that the subject is somehow capable of self-generating striatal activity despite not actually experiencing a reward, or perhaps postulate a source of striatal activation other than reward. There might be such explanations, but this is a harder case to make, and now the burden of proof would lie with the proponent of egoism.

BIBLIOGRAPHY

- Andreoni J. (2008) Charitable giving. In S. N. Durlauf, & L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics* (2nd Edition ed.) Palgrave Macmillan.
- Andreoni J. (1995) Cooperation in Public-Goods Experiments: Kindness or Confusion? *The American Economics Review*, 83 (4), 891-904.
- Andreoni J. (1989) Giving with impure altruism: applications to charity and Ricardian equivalence. *Journal of Political Economy*, 97, 1447-1458.
- Andreoni J. (1990) Impure altruism and donations to public good: a theory of warm-glow giving. *Economic Journal*, 100, 464-477.
- Andreoni J. (2006) Philanthropy. In S.-C. Kolm, & J. M. Ythier (Eds.), *Handbook of Giving, Reciprocity, and Altruism* (pp. 1201-1269) Amsterdam: North Holland.
- Andreoni J. (1988) Why Free Ride - Strategies and Learning in Public-Goods Experiments. *Journal of Public Economics*, 37, 291-304.



- Andreoni J., & Miller, J. (2002) Giving According to GARP: An Experimental Test of Consistency of Preferences for Altruism. *Econometrica* , LXX, 737-53.
- Aron A., Fisher, H., Mashek, D. J., Strong, G., Li, H., & Brown, L. L. (2005) Reward, motivation, and emotion systems associated with early-stage intense romantic love. *Journal of Neurophysiology* , 94, 327-337.
- Bartels A., & Zeki S. (2004) The neural correlates of maternal and romantic love. *NeuroImage* , 21, 1155-1166.
- Battigalli P. & Dufwenberg M. (2009) Dynamic Psychological Games. *Journal of Economic Theory* , 144, 1-35.
- Bolton G. & Ockenfels A. (2000) ERC: A Theory of Equity, Reciprocity and Competition. *American Economic Review* , 90, 166-93.
- Camerer C. *Behavioral Game Theory. Experiments in Strategic Interaction*. Princeton, New Jersey: Princeton University Press.
- Charness G. & Rabin M. (2001) Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics* , 117 (3), 817-69.
- Cookson R. (2000) Framing Effects in Public Goods Experiments. *Experimental Economics* , 3, 55-79.
- Dawes R. & Thaler, R. (1988) Anomalies: Cooperation. *The Journal of Economics Perspectives* , 2 (3), 187-97.
- Dufwenberg M. & Kirchsteiger G. (2004) A theory of sequential reciprocity. *Games and Economic Behavior* , 47, 268-298.
- Fehr E. & Fischbacher U. (2002) Why Social Preferences Matter: The Impact of Non-Selfish Motives in Competition. *The Economic Journal* , 112 (478), C1-C33.
- Fehr E. & Gächter S. (2000) Cooperation and Punishment in Public Good Experiments. *The American Economic Review* , 90 (4), 980-94.
- Fehr E., & Schmidt K. M. (1999) Theories of fairness, competition, and cooperation. *Quarterly Journal of Economics* , 114, 817-868.
- Fong C. M., Bowles S. & Gintis H. (2005) Reciprocity and the Welfare State. In H. Gintis, S. Bowles, R. Boyd, & E. Fehr (Eds.), *Moral Sentiments and Material Interests* (pp. 277-302) Cambridge, MA: MIT Press.
- Gächter S. & Herrmann B. (2009) Reciprocity, Culture and Human Cooperation: Previous Insights and New Cross-Cultural Experiment. *Philosophical Transactions of the Royal Society B* (364), 791-806.
- Genakopolos J., Pearce D., & Stacchetti E. (1989) Psychological Games and Sequential Rationality. *Games and Economic Behavior* , 1, 60-79.
- Gintis H. (2000) Strong Reciprocity and Human Sociality. *Journal of Theoretical Biology* , 206, 169-79.
- Gintis H., Bowles S., Boyd R., & Fehr E. (Eds.) (2005) *Moral Sentiments and Material Interests*. Cambridge, MA: MIT Press.
- Glimcher P. W., Camerer C. F., Fehr E. & Poldrack R. A. (Eds.) (2009) *Neuroeconomics: Decision Making and the Brain*. San Diego: Elsevier Inc.



- Goeree J., Holt C. & Laury S. (2002) Private Costs and Public Benefits: Unraveling the Effects of altruism and Noisy Behavior. *Journal of Public Economics* , 83, 255-276.
- Harbaugh W. T., Mayr U., & Burghart D. R. (2007) Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science* , 316, 1622-1625.
- Herrmann B. & Gächter S. (2008) Antisocial Punishment Across Societies. *Science* , 319, 1362-7.
- Independent Sector. (2001) *Giving and volunteering in the United States 2001*. Washington, DC: Independent Sector.
- Isaac R. M. & Walker J. M. (1988) Group size effects in public goods provision: the voluntary contributions mechanism. *Quarterly Journal of Economics* 103, 179-199.
- Kim O. & Walker, M. (1984) The Free Rider Problem: Experimental Evidence. *Public Choice* , 43, 3-24.
- Leydard J. (1995) Public goods: a survey of experimental evidence. In J. H. Kagel, & A. E. Roth (Eds.), *The handbook of experimental economics* (pp. 111-194) Princeton, NJ: Princeton University Press.
- Marwell G. & Ames, R. (1981) Economists Free Ride, Does Anyone Else? *Journal of Public Economics* , 15, 295-310.
- Mayr U., Harbaugh, W. T., & Tankersley, D. (2009) Neuroeconomics of charitable giving and philanthropy. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision Making and the Brain* (pp. 303-320) San Diego, CA: Elsevier Inc.
- Moll J., Krueger, F. Zahn, R. Pardini, M., de Oliveira-Souza, R., & Grafman, J. (2006) Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences* , 103 (42), 15623-15628.
- Osborne M., & Rubinstein, A. *A Course in Game Theory*. Cambridge, MA: MIT Press.
- Ostrom E., Walker, J., & Gardner, R. (1992) Covenants With and Without a Sword: Self-Governance is Possible. *American Political Science Review* , 86 (2), 404-17.
- Palfrey T., & Prisbrey, J. (1997) Anomalous Behavior in Public Goods Experiment: How Much and Why? *American Economic Review* , 87, 829-46.
- Rabin M. (1993) Incorporating Fairness into Game Theory and Economics. *American Economic Review* , 83, 1281-1302.
- Sober E., & Wilson, D. S. (1998) *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.
- Tversky A., & Kahneman, D. (1991) Loss Aversion in Riskless Choices: A Reference-Dependent Model. *The Quarterly Journal of Economics* , CVI, 1039-62.
- Von Neumann, J., & Morgenstern, O. (1944) *Theories of Games and Economic Behavior*. Princeton: Princeton University Press.
- Yaari M., & Bar-Hillel, M. (1984) On Dividing Justly. *Social Choice and Welfare* , 1, 1-24.
- Zak P. J., Kurzban R. & Matzner W. T. (2005) Oxytocin is associated with human trustworthiness. *Hormones and Behavior* , 48, 522-527.