Models of caring, or acting as if one cared, about the welfare of others

Julio J. Rotemberg*

November 25, 2013

Abstract

This paper surveys the theoretical literature in which people are modeled as taking other people's payoffs into account either because this affects their utility directly or because they wish to impress others with their social-mindedness. Key experimental results that bear on the relevance of these theories are discussed as well. Five types of models are considered. In the first, the utility of people is increasing in the payoffs of another. The more standard version of these preferences supposes that only consumption leads to payoffs and has trouble explaining pro-social actions such as voting and charitable contributions by poor individuals. If one lets other variables determine happiness as well, this model can explain a much wider set of observations. The second class of model that is surveyed involves people trying to demonstrate to others that they have pro-social (or altruistic) preferences. In these models, altruistic acts need not have a direct effect on utility. Third, I consider models of reciprocity where people's altruism depends on whether others act kindly or unkindly towards them. Fourth are models where inequality has a profound effect on altruism, with individuals being spiteful towards people whose resources exceed their own. Fifth and last, I discuss how specifications of altruism might have to be modified to take into account how people behave when they are able to transfer lotteries to others.

*Harvard Business School, Soldiers Field, Boston, MA 02163, jrotemberg@hbs.edu. I wish to thank Roland Bénabou, Stefano DellaVigna, Ernst Fehr, Michael Norton, Klaus Schmidt, Dmitry Taubinsky and the editor David Laibson for comments. All remaining errors are my own.

Bentham (1907, ch 5) hypothesized that all sources of utility and disutility stemmed from 14 "simple pleasures" and 12 "simple pains." Of these, the four pleasures of "benevolence," "malevolence," "amity," and "a good name" involved other people directly, and so did the corresponding "pain of benevolence," "pain of malevolence," "pain of enmity," and "pain of an ill name." According to Bentham (1907, ch 5), the pleasure (pain) of benevolence originated from seeing the pleasure (pain) of beings towards whom one was benevolent, whereas the pleasure (pain) of malevolence originated from seeing the pain (pleasure) of beings towards whom one was malevolent. In this survey, I consider efforts to provide mathematical representations of preferences and utility that incorporate these effects. Once one has such a model in hand, one can ask whether individual behavior in particular settings is consistent with these preferences, and particularly whether observed decisions maximize a utility function that captures these preferences.

When it comes to doing things for (or against) others, individual behavior is extremely heterogeneous. Some people give to charity; others do not, some explode in rage at the smallest provocation, others ignore direct insults. If one wishes to suppose that people maximize their utility, one inevitably has to suppose that these utility functions are heterogeneous. As I discuss below, people's generosity seems quite sensitive to details of the situation they find themselves in. If one does not explicitly keep track of these other determinants of preferences, individual utility functions may thus appear unstable. An obvious question, then, is whether there are common patterns suggesting that some aspects of preferences arise regularly, even if they are not universal.

The aim of this paper is to survey formal models where individuals are directly concerned with the welfare of others, as well as those in which people find it advantageous to mimic the behavior of individuals that are so concerned. I emphasize models with verifiable predictions, and try to give readers a sense of the extent to which the empirical evidence suggests that people act in accordance with these models. The most direct evidence for the relevance of such models in decision-making contexts comes from the study of situations in which individuals are allowed to incur a personal cost in exchange for affecting the welfare of others.

Much of this evidence has been gathered by experimentalists, whether they be economists or psychologists. In addition, there is naturally occurring field evidence from charitable donations and voting, two institutionalized activities that are generally thought to impose some costs on individuals while being valuable to others. One reason these activities have received disproportionate attention relative to other domains in which people act generously, including the extent to which people help strangers in the street, co-workers, or family members, is that the latter activities are harder to observe systematically.

Overall, the volume of the experimental and field evidence that bears on these models is much too vast to be surveyed in its entirety here. I thus selectively focus on two kinds of evidence. The first is evidence that a particular model seems well-suited to explain, either because it provides a particularly simple explanation of the phenomenon or because other models in the literature seem inconsistent with the observations in question. Second, I focus on observations that appear to put tight limits on the range of applicability of the model. If, for example, a model can explain a set of observations if at least y percent of individuals are of a certain type and another experiment shows that at most z percent of subjects are of this type, a large positive difference between y and z suggests that only a small subset of the original set of observations is explicable with the model in question.

Much of the literature I study can be understood as making assumptions about altruism (where malevolence is simply negative altruism), including how stable it is, what makes it vary, why people might want to pretend to have it, and how they can demonstrate having it. Section 1 thus focuses on models in which some people's preferences involve a constant degree of altruism, or caring, for certain other people. The section considers not only experimental evidence that bears on this model, but also studies the extent to which variants of this model can explain the two main "prosocial" institutional activities we observe in society at large, namely charitable contributions and mass political participation.[1] The capacity of this simple model to account for observations turns out to depend crucially on whether

[1] The direct benefits to an individual from participating in a mass political event, such as voting, are negligible. Groups, on the other hand, can gain a great deal from the participation of masses of people.

one imagines that the selfish component of utility depends only on one's access to material resources or whether one takes a more expanded perspective on human happiness. Supposing that people dislike feeling regret, or that their happiness and self-esteem depend on whether people agree with them, for example, vastly expands the explanatory power of the basic altruism model.

Section 1 of the paper, which surveys models of constant altruism, is thus divided in several subsections. After presenting the basic model in the first subsection, the second focuses on the evidence for the stability of altruism across different settings. This subsection stresses that, while altruism is triggered by environmental cues, the altruism of i for j in a new setting is correlated with i and j's past behavior. The third subsection focuses on the difficulties this model has had in explaining charitable contributions and political participation. The fourth subsection enriches this model by allowing people's direct payoffs to depend on more than individual consumption. It considers both the effect of disappointment and regret as well as the effect on self-esteem of learning that more people agree with oneself.

The second section of the paper studies models in which people gain something from demonstrating their lack of selfishness to others. This class of formal models, which was pioneered by Levine (1998), involves signaling. This is followed in section 3 by the study of models in which people are signaling their altruism to themselves as in Bénabou and Tirole (2011). Sections 4 and 5 consider two other determinants of altruism that have been the focus of theoretical work, namely the equilibrium actions taken by others (as in Rabin (1993)) and the extent to which the distribution of resources is unequal (as in Fehr and Schmidt (1999) and Bolton and Ockenfels (2000)). Section 6 focuses on models that have been built to account for transfers whose value is uncertain at the time they are made. This is still a small and growing literature that has sprung up because earlier models do not seem to account well for some aspects of these transfers. Lastly, section 7 provides some concluding remarks.

1 Exogenous (and sincere) altruism

1.1 The basic material payoffs version and some key experiments

The most established formal model in which individuals have preferences concerning the payoffs of others is one where the utility of individual i depends not only on his own consumption vector X_i, but also on the consumption vector X_j of individual j. A tractable specification that has been used often is one where there is a function $v_i(X_i)$ that gives the "material payoffs" of i and where the utility of i depends linearly on $v_i(X_i)$ and $v_j(X_j)$.[2] In particular, utility u_i is given by[3]

$$u_i(X_i, X_j) = v_i(X_i) + a_i v_j(X_j) \qquad i, j = 1, 2, i \neq j, \tag{1}$$

When, as in this section, a_i is a constant, it can be thought of as an index of i's altruism for j. Because adding or subtracting constants does not change this objective function, an agent who acts to maximize this objective with a positive a_i can equally well be interpreted as suffering a loss (due to guilt, for example) that equals a_i times the difference between the material payoffs $v_j(\bar{X}_j)$ that j would receive if i carried out a "kind" action and j's actual material payoffs. A negative a_i implies that i wants to harm j, which can be seen as involving spite. Aside from its separability, this approach is special in assuming that i agrees with j in the way that they rank bundles of goods consumed by j herself. A more paternalistic approach would allow i and j to disagree, for example, on how much they care about j's consumption of soft drinks relative to her consumption of milk. While public policy sometimes appears to include such paternalistic elements, this has not been a focus of research on other-regarding preferences.

[2] This allows one to capture Edgeworth's (1881, p. 102) idea that "between the frozen pole of egoism and the tropical expanse of utilitarianism, there has been granted to imperfectly-evolved mortals an intermediate temperate region; the position of one for whom in a calm moment his neighbor's happiness as compared with his own neither counts for nothing, nor yet 'counts for one,' but counts for a fraction."

[3] The additive separability of this function is, as in the quote by Edgeworth in footnote 2, inspired in part by utilitarian social welfare functions. See Cox et al. (2008) for a more general treatment in which there is no a_i parameter and the experimental data are used to shed light on the ratio of the marginal utility of a dollar in j's hand over the marginal utility of a dollar in i's hand, which here equals $a_i v'_j / v'_i$. Increases in this ratio can, but need not, correspond to increases in a_i.

In many theoretical and experimental applications, there is a single scalar variable such as the income that the subject will take home when the experiment is over, and I will use the scalars x_i and x_j for such payoffs. The utility function (1) can be used to interpret experiments in which one subject gives resources to another. For this interpretation to be sensible it is often important that, as noted by Levine (1998), the subjects not care about the experimenter so that they treat the funds spent on the experiment as "free." Because the outside income of experimental participants is unknown to both the experimenter and the other subjects, it is convenient to ignore it in carrying out this interpretation. This still leaves open the question of whether people's utility is defined over their lifetime consumption, in which case the derivatives of v_i and v_j should be insensitive to the amount earned by rich subjects during typical experiments, or whether these earnings have a more direct concave effect on utility.

A canonical experiment where such transfers are observed is the Dictator Game (DG) introduced by Forsythe et al. (1994), where one agent is given some cash and is told that he can give up any fraction he wishes to another agent. If the marginal utility of cash on hand for agent i exceeds a_i times the marginal utility of cash on hand for the other agent, i should not make any transfers. In Engel's (2011) meta-analysis of 131 DG papers, 36 % of subjects behave in this manner, while the rest give positive transfers. About 34 % of subjects transfer between .1 and .4 of their endowment, and the behavior of these subjects seems rationalizable with the maximization of a function such as (1) in which a is strictly less than one while v_i and v_j are smooth concave functions.

The remaining 30 % transfer either half the endowment (about 17 % do so) or more. When exactly half is transferred and subjects view their own v_i function as identical to v_j, their behavior maximizes u_i only if a_i equals one. In this case, v_i can be concave. In the nontrivial set of cases where more than half is transferred, a_j can still equal one but v_i must then be linear, so that the participants are indifferent with respect to any division of the endowment between the two parties.

One of the aims of Andreoni and Miller (2002) and Fisman et al. (2007) is to study how giving varies as one varies the exchange rate between token in the hand of the dictator and tokens in the hand of the receiver. When presented with a variety of choices of this type, most of the subjects in both sets of experiments behaved as if they were maximizing an altruistic utility function like (1). In Andreoni and Miller (2002), 6 % made choices consistent with $a_i = 1$ and linear v. An additional 14.2 % always left themselves with the same level of resources as the receiver. For this to be rational at every exchange rate, it must be the case that $a_i = 1$ and that both i and j have material payoffs given by $x^{-\gamma}$ with γ tending to infinity and x representing the amount they receive in the game. It seems likely that neither these high levels of a_i nor these unusual forms of v characterize the utility function that these subjects employ when they evaluate how much of what is in their wallet to give to beggars. Indeed, one might well argue that plausible models of altruism require that people care more for themselves than for others so that the 20% of subjects behaving in these extreme fashions must do so for other reasons. I return to this issue below.

A great many additional subjects in Andreoni and Miller (2002) behaved consistently, so that their choices were rationalizable by a utility function, but had less extreme preferences. The utility functions that Andreoni and Miller (2002) fit to these subject's behavior had an a_i strictly less than one. It is also worth noting that many of these subjects increased the percentage of their endowment that they gave away when the value to the receiver of the tokens constituting this endowment rose relative to their value to the giver.

Attempts to increase a function such as (1) can also rationalize some of the helping behavior found experimentally by psychologists. Particularly notable are experiments in which subjects are all asked to perform the same task but there is variation in the extent to which this task is depicted as being important to others. Variants that keep the reduction in v_i constant while raising the gain in v_j should increase helping in a population where the distribution of a_i is constant. Shotland and Stebbins (1983) report an experiment that involves asking strangers to look up numbers in a student directory where this occurs. In part because manipulating the perception of changes in v_j is not entirely straightforward, not all field experiments of this sort report consistently strong results.[4]

1.2 Is individual altruism stable?

Several studies suggest that the extent to which a person grants or withdraws favors from another is extremely sensitive to the context in which this potential transaction takes place. For example, Baron (1997) reports that the willingness of passerby's to give change to a stranger is higher near stores that emit pleasing food smells than in other similar locations. Varying the details of the transfer game instead, Fershtman, Gneezy and List (2012) show that transfers in dictator games are lower if the dictator "earns" his endowment by winning a contest. This still leaves open the question of whether those that behave as if they had high values of a_i in one setting also tend to do so in other settings. Volk et al. (2012) show that the contributions of individuals to a public goods game remain consistent over months. Moreover, individuals that contribute more tend to score higher in "agreeableness" when they fill out questionnaires designed to measure Big Five personality traits. In a related vein, psychologists have long asked themselves whether pro-social behavior like volunteerism was associated with questionnaire-based measures of personality traits. While the association is often not strong, several studies have found statistically significant associations of this sort. A recent example of this is Carlo et al. (2005).

On the other hand, two studies that have looked at whether generosity in an experimental public goods game translated into larger donations for actual public goods found weak and inconsistent results.[5] Still, Fehr and Leibbrandt (2011) find that contributions in this experimental game correlate strongly with the extent to which shrimpers at a lake in northeastern Brazil use traps with relatively large holes, which are more helpful to the livelihood of other shrimpers. Donations in the experiment are also correlated with the mesh size of the fishnets used by fishermen. Similarly, Benz and Meier (2008) found a fairly strong concordance between donations to charities by University of Zurich students across occasions

[4]See, Wagner and Wheeler (1969) and Baron (1978), for example.

[5]See Laury and Taylor (2008) and Voors et al. (2012)

in which these were given opportunities to donate. Lastly, Fowler and Kam (2007) show that giving in DGs is correlated with self reported measures of political participation such as voting.[6] At the individual level, then, altruism is somewhat predictable from an individual's past history but we are far from being able to draw strong conclusions from this history.

A different way in which individuals can have stable altruistic preferences is by systematically favoring one group, so that their a_i is higher for that group. The favoring of particular groups is a robust experimental finding. For example, Gino and Pierce (2010), find that their survey respondents express a greater willingness to lend their parking permit to fellow students whose car is not a luxury car. Interestingly, both the effect of car luxuriousness and of other borrower characteristics on the willingness to lend are mediated by the extent to which potential lenders would feel sorry if the potential borrower received a ticket and the extent to which they are envious of the potential borrower's car. Both these variables may, in turn, be related to the degree to which the potential lender feels connected to the potential borrower.

Consistent with this, Goette et al. (2006) find that Swiss subjects are more generous in one-shot Prisoner's Dilemma experiments when they are told that they are paired with a member of their (temporary) platoon than when they are paired with a member of a different platoon. In addition to appearing to have higher a_i's for members of their own officially constituted in-group, people seem to favor those who share their opinion. For example, Tucker et al. (1977) and Sole et al. (1975) find that, after picking up unmailed letters lying on the street, people sharing the opinion of the sender are more likely to do what is necessary for the letters to reach their destination. In an experiment using deception, Batson et al. (1981) find that subjects are more likely to wish to substitute themselves for someone receiving electric shocks after learning this individual shares their beliefs. Finally, Fowler and Kam (2007) shows that dictators transfer more resources in DG games when they are told by the experimenter that the receiver shares their political affiliation (Democrat or Republican) than when they are told that the receiver's affiliation is different from their own.

[6] These results are confirmed and extended in Dawes et al. (2011).

1.3 Altruism, Voting and Charitable Contributions

The question studied in this subsection is whether altruism defined exclusively over resources can explain institutionalized group actions such as voting and charitable contributions. The challenge is that, as more people provide resources to others, each individual's incentive to do so can be diminished. In the case of voting, in particular, it has long been understood that the presence of other voters reduces an individual's incentive to vote. If few people are voting, each voter has a reasonable chance of being pivotal so that an individual's concern for the election's outcome can be enough to induce the individual to vote. As the number of voters increases, the probability of being pivotal tends to fall exponentially, however, so that voting in large elections becomes difficult to rationalize.[7] Nonetheless, several authors have suggested that the existence of altruistic voters can resolve the puzzle of voting in large elections.

To obtain this result, it helps to suppose that, as suggested by Fowler (2006), altruists base their votes on their belief that the victory of their favorite candidate (proposition) is good for other people. This paves the way for assuming that a doubling of the number of people in an electoral jurisdiction doubles the utility that an altruist obtains from the victory of her favorite candidate. While this assumption, which is used by Jankowski (2007) and Evren (2012), raises the altruists' incentive to vote, it is not strong enough to ensure significant turnout in large elections because the probability of being decisive is so negligible when many people vote.

To obtain significant turnout, both Jankowski (2007) and Evren (2012) suppose that the number of voters is stochastic. This follows from supposing that only a random fraction of a candidate's supporters is altruistic. Jankowski (2007) then obtains significant turnout when everyone has the same voting costs but only under the condition that both candidates have the same probability of having majority support. Evren (2012), who lets voting costs be heterogeneous, allows these *ex ante* probabilities to differ. However, his result that a strictly positive fraction of individuals votes in arbitrarily large elections is based on supposing that

[7] For a recent restatement of this well known result, see Evren (2012).

the proportion of supporters of each candidate that is altruistic has positive density at zero. This creates the possibility that very few people will vote even with an enormous population, and this possibility maintains a strong incentive for altruists to vote when the population is large. In actual national elections turnout rates below 50 % appear almost nonexistent, so that observed patterns of turnout do not yet seem rationalizable by models along these lines.[8]

A substantial fraction of the help that people provide to strangers is intermediated by charities. This means that contributors to charities must realize that there are other contributors to the same cause. If altruist i's concern is with the charity's beneficiaries, the natural extension of (1) to a setting with charities is to suppose that his objective is

$$u_i(x_i, g_i, G^{-i}) = v(x_i) + a_i V(g_i + G^{-i}) \tag{2}$$

where x_i is his personal consumption, g_i his gift to charity, G^{-i} the total of other people's gifts and V is a function capturing the material payoffs of the charity's beneficiaries. If i contributes, he should thus set the marginal utility of his own consumption equal to a_i times the marginal increase in V when total contributions increase. This causes two well known difficulties. The first, which has been pointed out by Sugden (1982) and Andreoni (1989) is that any increase in the beneficiaries funding, whether due to government spending or to someone else's contributions, should lower the remaining altruist's contributions nearly one for one. As discussed in Andreoni (1989) and Rotemberg (2013), this prediction seems inconsistent with observations.

The second is that, if preferences in (1) do not differ by income, the analysis of Andreoni (1988) implies that only people with high private consumption should make contributions. The reason is that the marginal utility of their own consumption is lower for high-income individuals, so that the marginal utility of material payoffs that their contributions induce

[8] In the model of Feddersen and Sandroni (2006), it is "ethical" rather than altruistic individuals that vote, and these ethical individuals do so when it is consistent with the maximization of a group objective function which depends on the probability with which the group's favorite candidate wins and on the total society-wide costs of voting. Feddersen and Sandroni (2006) base their analysis on specific functional forms and, interestingly, also assume that the fraction of ethical voters is random with a positive density at zero. As it stands, then, their model is also incapable of explaining consistently high turnout.

among beneficiaries is too low to justify contributions by people with lower private consumption levels. However, List (2011) shows that the proportion of people with low income levels that makes charitable contributions is substantial, even if it is indeed lower than that of richer individuals. Moreover, the proportion of income donated by poorer contributors is higher on average than that of richer ones.

Andreoni's (1989) response to these problems is to suppose that, on top of whatever utility they get from the consumption of beneficiaries, people obtain direct utility from their own contributions. In Andreoni (1989) this component of utility, which he deems "selfish" and calls "warm glow", depends literally on the amount of resources that the individual gives up. The maximization of a utility function that includes this but contains no altruism ($a_i = 0$) cannot account for the fact that in experiments such as Andreoni and Miller (2002) and Goeree et al. (2002) many subjects give up more of their tokens when these tokens become more valuable to others.

Korenok et al. (2013) generalizes (1) and (2) so that individual i derives utility from his own material payoffs, the material payoffs of the person that he transfers resources to and, also, from the actual transfers he makes. In the experimental part of their paper, they change the Andreoni and Miller (2002) setup so that the recipient sometimes receives an endowment as well. They then argue that the transfers of about a third of their subjects cannot readily be explained without the third term that gives "warm glow" utility from transfers. Without this term, changes in the endowment of the recipient would have too large an effect on the amount that the first player transfers to the second while this effect is attenuated when the third term is present.

1.4 Caring about others' psychological well-being more generally

1.4.1 Avoiding disappointment and regret

While convenient, the assumption in (1) that people only care about others' material resources and not about other sources of others' psychological well-being is obviously restrictive. One psychological effect that has been studied extensively by economists is the existence of "reference points" for consumption.[9] That requires replacing the material payoff function $v_j(X_j)$ by a function $v_j(X_j, X_j^*)$ where X_j^* is the reference point. One case that has received a great deal of attention from social psychologists is where X_j^* is the consumption that the consumer could have gotten if he had made a different decision in the past. The modified "material payoff" function can then be thought of as capturing regret in a manner akin to that of Loomes and Sugden (1982).

An altruist, then, would be reluctant to cause the object of his altruism to feel regret. Rotemberg (2011) uses this idea to explain why firms that act altruistically would refrain from raising prices of necessities during hurricanes. The idea is that the people who buy these necessities at the last minute remember that they could have purchased these items earlier. Firms that keep their prices constant assuage the regret from not having done so. The notion that firms are, in fact, altruistic is controversial. It is probably most applicable in settings where rich founders have sufficient resources to indulge in their personal preferences or in situations where firms are able recruit substantially more effective employees by tapping into pools of individuals who are altruistic towards customers. In other settings, firms probably act as if they cared for their customers only in those domains in which failing to do so would lead customers to be upset. The role of developing models of altruistic firms in such settings is to understand whether these negative reactions by customers can be ascribed to their concluding that firms are not being as altruistic as customers expect. In areas where the intentions behind firm actions are opaque, one would expect firms in these settings to be ruthlessly profit-maximizing.

In economic experiments, receivers in DGs are usually told the mechanism that gives rise to the payments they receive in advance, even if they later receive nothing from the dictator with whom they are matched. One can thus expect receivers to be disappointed when they do, indeed, receive a zero transfer. Perhaps, even, they regret not having lingered before entering the room where the experiment has been carried out, on the ground this would have led to being matched with a different player. A formal model in which receivers

[9] See Koszegi and Rabin (2006) for a well-known recent model along these lines.

experience disappointment, which depends on the difference between their actual payoff and the payoff they could expect to have received by being matched to someone else, is provided by Taubinsky (2012). This effect leads altruistic dictators, who internalize this disappointment somewhat, to be more generous than they would be otherwise.

Dana et al. (2006) provide evidence that this process is operative. They show that a good fraction of dictators is willing to walk away with 9/10ths of their original endowment if they are promised that the receiver will never learn that there was a dictator game involved. Some dictators are thus willing to give up resources to prevent the receiver from "feeling bad."[10] Moreover, this "exit" option is not employed as often by dictators who are told that their contribution will unobtrusively be added to the receiver's funds without the receiver knowing that a DG was played.

As Dana et al. (2006) note, this altruistic (or guilt-based) cost from making others feel bad can explain why people avoid beggars. It also seems able to explain the fact, reported in DellaVigna et al. (2012) that some people who would have made positive contributions to a charity if a fundraiser had visited them unannounced choose not to have the fundraiser visit at all when they were given this choice. Using the Dana et al. (2006) logic, people would make this choice in part to avoid making the fundraiser feel bad and in part to avoid making contributions that are based mainly on avoiding the fundraiser's disutility.[11] Indeed, one would expect the people who would make small gifts in the presence of the fundraiser to be particularly keen to avoid the fundraiser's disappointment, and this fits with DellaVigna et al. 's (2012) observation that most of the gifts that are prevented by giving people an opt-out option are indeed small.[12]

[10]Interestingly, 3 out of the 18 subjects whose transfer in standard DG games were zero took advantage of this option. While their generosity was not strong enough to make positive transfers, they were willing to give up some resources to avoid causing disappointment.

[11]Under this interpretation, the results of DellaVigna et al. (2012) are explicable with the maximization of an "altruistic" utility, except that there are two different agents (the charity and the fundraiser) that cause i to either enjoy giving or feel guilty from not giving enough. This explanation would presumably introduce a link between the parameters of what they call the "social pressure felt by an individual" and the altruism of this individual.

[12]An alternative explanation for the DellaVigna et al. (2012) findings is that the negative signal (and resulting social stigma) associated with asking a fundraiser not to come is lower than that the negative signal

A similar form of guilt can, in some cases, lead an altruist to engage in what has been called "positive reciprocity." This refers to the tendency of second players to make moves that favor a first subject after, at an earlier stage in the experiment, the first subject made a move that favors the second. This can be rationalized by altruism and guilt in those cases where the original favorable move by the first subject leads the second to believe that, unless he acts nicely as well, the first is likely to regret his initial generosity. Another case where this guilt can be activated is when, as in Falk (2007), the first player sends a memorable gift to the second. If the second fails to reciprocate, he can expect to feel guilt every time he uses or remembers the gift. To prevent these recurrent bad memories, the second subject may thus be led to reciprocate.

When comparing the amount sent back by the second subject in a trust game (TG) to the amount that dictators who had been given comparable endowments sent to receivers, Cox (2004) did observe such reciprocity while Dufwenberg and Gneezy (2000) failed to do so. Charness and Rabin (2002) ran some simpler experiments with a similar structure and did not always find such reciprocity. In an experiment in which the second subject could choose between payoffs of $\{400, 400\}$ (where the first payoff goes to subject 1 and the second to subject 2) or $\{750, 375\}$, somewhat more subjects chose the generous allocation $\{750, 375\}$ when they were on their own than when the first subject had decided to forego $\{800, 0\}$ to given them the choice between $\{400, 400\}$ and $\{750, 375\}$. On the other hand, when the second subject had to choose between $\{400, 400\}$ and $\{750, 400\}$, so that the "generous" choice did not reduce her own material payoffs, she was much less likely to choose $\{750, 400\}$ in the DG than if she was responding to a first subject that gave up $\{750, 0\}$ to let her make the choice. Malmendier and Schmidt (2012) also observe positive reciprocity when this is "free" to the individual acting reciprocally, though in their case this reciprocity can be costly to a third party.

of making a zero contribution in the presence of the fundraiser. This explanation is not applicable to the Dana et al. (2006) findings if their dictators are truly anonymous.

1.4.2 Raising other people's self esteem

People's self-esteem is intimately linked to their happiness. Moreover, Gaillot and Baumeister (2007) say that "the desire to seek high self-esteem by maintaining positive evaluations of oneself," is "among the strongest of human motivations." Thus, true altruists can be expected to try to enhance the self-esteem of those they care for. But, how can they do this? Gaillot and Baumeister (2007) suggest that there are two key determinants of self-esteem, namely "perceptions of belongingness" and "worldview validation," where the latter reflects "the extent to which others share one's values and beliefs." Using cross sectional data from questionnaires, they show that, indeed, answers to questions such as "I think that many people agree with my values and beliefs," are strongly correlated with self-esteem.

What is more surprising, perhaps, is how easy it is to manipulate a person's reported self-esteem by telling her about the opinions of others. Johnson (1973) ask people to form judgments about (fictitious) others on the basis of booklets containing their answers to a questionnaire. They show that reported self-esteem is higher for those that evaluate individuals whose questionnaire answers are more similar to their own. Similarly, Pool et al. (1998) and Kenworthy and Miller (2001) give subjects bogus information regarding the extent to which their opinion (on issues such as the death penalty) is shared by others. Subjects told that relatively few people agree with them report feeling worse after hearing this information.

If bogus data of this sort matters, real information ought to matter even more. Moreover, there are two sorts of pro-social activities that convey hard information about the extent that others agree with oneself. The first are votes for candidates and propositions. This is the basis of Rotemberg's (2009) theory of electoral participation. This theory explains turnout in large elections on the ground that voters are sending a message of support to all those that agree with them. Sending such a message raises utility if people are more altruistic towards those that agree with them. There is, as noted above, considerable evidence for this assumption. One key benefit of this theory of voting is that it can explain votes by people who have literally zero chance of changing the election outcome, such as votes for fringe parties or U.S. presidential votes in states such as Massachusetts.

The second source of hard information about attitudes that people have access to is information about total contributions to various charities. This leads Rotemberg (2013) to suggest that these contributions should be thought of as messages to like-minded people. The benefits of sending these messages do not shrink as the number of donors rises. Rather, the theory naturally implies that increases in the number of donors to a cause tend to raise perceived benefits of donating to the cause.

2 Altruistic behavior with the potential of influencing others

We saw in the previous section that, in practice, the level of altruism of i for j depends on the commonality of beliefs and attitudes between i and j. A natural extension of this idea is to suppose that i's altruism for j depends on the altruism that i believes j has for i. This reciprocity in altruism was introduced into the literature by Levine (1998). His model, specifically is that

$$u_i = v_i + \frac{a_i + \lambda a_j}{1+\lambda} v_j \tag{3}$$

with the base levels of altruism a_i and a_j being different across people (and defining their "types") while λ is a constant between zero and one. One key implication of this is that actions by j that convey information about a_j affect how i treats j subsequently. In the domain where a_j and $a_i + \lambda a_j$ are positive, an increase in i's perceived value of a_j could lead i to sacrifice more of his resources to help j. As discussed earlier, the evidence for this sort of "positive reciprocity" is mixed. In the domain where a_j and $a_i + \lambda a_j$ are negative, an action that makes j appear to have a lower altruism leads i to be more willing to sacrifice to hurt j. The classic example of this is the ultimatum game, where proposers who offer less than an even split often find their offers rejected, something that is costly to both proposer and responder. Levine (1998) shows that the broad outline of ultimatum game (UG) outcomes can be explained with three values of a (and one of λ).

The exogenous altruism model has two related problems explaining the UG that a utility function such as (3) can solve. First, as demonstrated by Forsythe at al. (1994), a much higher fraction of proposers offers a fifty-fifty split in the UG than in a DG played under the same conditions. Second, plenty of responders turn down low offers even though every proposer in UGs offers responders a positive amount. Interpreted using the exogenous altruism model, the former requires that some a_i be negative (since people are sacrificing to hurt their opponent) while the latter requires that everyone's a_i be positive (since everyone is sacrificing to help their opponent). As long as some people have negative a_i's, the objective function (3) can easily accommodate rejections of the offers made by the least altruistic proposers to the least altruistic responders. And the reason these proposers still make positive offers is that this leads the more altruistic responders not to turn down their offers, which even somewhat spiteful proposers find worthwhile.

While Levine (1998) shows that the model can account for the main features of Roth's et al. (1991) data on the UG, it turns out to require a great deal of baseline spite to do so. In particular, 80% of individuals must have a negative a_i. This is implausible for two reasons. First, it implies that the vast majority of the population wishes to engage in petty vandalism (which can hurt the victim considerably at a trivial cost to the perpetrator) and that only fear of punishment prevents this. Second, it seems impossible to square this distribution of a_i with the fact, in a DG variant, 73% of Charness and Rabin's (2002) dictators who could have taken 700 while giving 200 to the other player preferred to sacrifice so that they would each receive 600.

One reason the Levine (1998) model seems to require so much inherent meanness among humans is that the linear specification in (3) implies that an altruist (someone with a positive a_i) only becomes willing to punish an offer that is just shy of fifty-fifty if he determines that it has been made by someone whose a_j is both negative and much larger in absolute value than a_i. Because altruists are not prone to rejecting uneven offers, the UG results do not allow them to be numerous.

If one wants a model in which an altruistic orientation is more common, one needs to suppose that the a_i of altruists becomes negative more easily. This fits with the widespread tendency of people, even ones that act altruistically most of the time, to become angry after relatively small provocations. Anger is the name of an emotion that most people understand quite well and that psychologists typically define in line with Berkowitz and Harmon-Jones (2004) as being "linked associatively with an urge to injure some target."[13] As Schieman (2010) shows, surveys that ask Americans whether they have "felt angry" in the past 7 days consistently show that only 37-45% of Americans have not done so. Moreover, Srivastava et al. (2009) demonstrate that anger plays a role in the behavior of UG responders. In their experiment, responders who previously carried out a task meant to elicit anger towards another target were much less likely to reject extremely lopsided offers in an UG. Their interpretation is that the earlier task makes it more difficult for people to attribute their anger to the low offer in the UG.

To capture the idea that people's propensity to anger need not be inconsistent with their kindness in other contexts, Rotemberg (2008) proposes a variant of Levine's (1998) model in which the utility of i is given by

$$u_i = v_i + [a_i - \xi(\hat{a}_j, \bar{a}_i)]v_j \tag{4}$$

The function ξ gives the anger of i for j. This is a function of i's information about a_j, which is encapsulated by $\hat{a}_j$, and of i's required minimum altruism $\bar{a}_i$. In Rotemberg (2008), ξ is zero unless i can reject the hypothesis that a_j equals at least $\bar{a}_i$. If he can, ξ is large and i becomes spiteful towards j. People thus give others the benefit of the doubt and react negatively only when they are relatively sure that the others' altruism is lower than they require. A low offer in the UG, on the other hand, proves to people with high $\bar{a}_i$ that altruism is insufficient and triggers rejection. As a result, offers are more generous than in the DG. By letting people's minimal altruism $\bar{a}_i$ differ from their own altruism, and thereby allowing this functional form to be quite flexible, Rotemberg (2008) can fit several aspects of UG data.

[13] As an example of this link, Cheung-Blunden and Blunden (2008) show that U.S. subjects who reported feeling angry after seeing photographs of the 9/11 attack were more supportive of having the U.S. fight in Iraq and Afghanistan.

One particular advantage of this flexibility is that it can incorporate a feature of anger that is readily observed, namely that angry individuals sometimes cause damage to their victims that is less onerous than the cost they impose on themselves. In particular, the costs of road rage and of assault can be quite substantial for the perpetrator once due account is taken of the reaction of the government. To my knowledge, this willingness to incur costs larger than the costs one imposes on another has not been tested in economic experiments. Falk et al. (2005) come closest by considering a situation in which one agent can subtract one unit from another at the cost of a unit to himself. They find that, while several individuals who cooperated in a prisoner's dilemma game meted out such punishments on defectors, no defector did so.

Like any model based on Levine (1998) in which people can increase their payoffs by signaling that they are unselfish, the equilibrium of the Rotemberg (2008) model is typically not unique if one does not impose suitable refinements. Rotemberg (2008) demonstrates that a special case that may be useful in applications outside the laboratory can have a unique equilibrium in which everyone acts as if they had the same positive altruism parameter a. In this special case, no one's a_i or $\bar{a}_i$ is greater than a and some of the individuals whose altruism does indeed equal a are naive so that they do not expect to be punished no matter what they do. Under these circumstances, selfish individuals often find it valuable to pretend to have an altruism parameter of a because the cost of appearing insufficiently altruistic can be quite high even if only a subset of the population cares about the altruism of others.

Even when there is more than one equilibrium, models of this sort are inconsistent with some outcomes. For example, in the plausible case where no one's minimum required altruism $\bar{a}$ is larger than the maximum altruism in the entire population, it is not an equilibrium in the Rotemberg (2008) model for every proposer to make a low offer and for every responder to reject it. The reason is that, in this case, low offers would not allow the responder to reject the hypothesis that the proposer has the highest possible level of altruism.

The Levine (1998) and Rotemberg (2008) models can also explain third party punishments, such as those demonstrated in Fehr and Fischbacher (2004). Fehr and Fischbacher

(2004) consider a setting where one subject can split his endowment with a second, and a third can then give up units of his own endowment in exchange for reducing the endowment of the first subject by three units. They show that 60% of subjects playing the third role give up some of their endowment if the first subjects offers less than an even split.[14]

The Rotemberg (2008) model is also consistent with the observation of Bohnet and Zeckhauser (2004) that people are more reluctant to receive a low payoff that is the result of a selfish action by another subject than they are to receive this payoff from a randomization device. The reason is that, when a person makes an ungenerous choice, this choice signals their selfishness so that many people feel spite towards this person. By contrast, people are not as spiteful towards people who have received a favorable allocation from a randomization device.

In its emphasis in the benefits of signaling one's altruism, the Bénabou and Tirole (2006) model is also related to Levine (1998). Bénabou and Tirole (2006) do not explicitly model the benefits of this signaling. Rather, they take these benefits as given and focus on the consequences of adding extrinsic reasons to carry out a generous action x. The objective they assume for these agents can be written as

$$u_i = (\theta_i y v_i + a_i v_j)x - \mu_\theta E(\theta_i) + \mu_a E(a_i). \tag{5}$$

where $\theta_i y v_i x$ is the selfish payoff from x whereas $a_i v_j x$ is the altruistic one, v_i and v_j are positive constants, θ_i and a_i are indicators of the individual's selfishness and altruism respectively and y is a feature of the environment that determines the extent to which the individual benefits directly from x. The term $\mu_a E(a_i)$ captures the benefit of being seen as altruistic while the term $\mu_\theta E(\theta_i)$ captures the cost of being seen as selfish.

Bénabou and Tirole (2006) offer two interpretations for μ_a. In one, this is a benefit that i gets from his enhanced self-esteem so that E computes his own expectation at some future date, and I return to this below. In the other, it is a benefit he gets from the favorable reactions of others so that E computes other peoples' expectation. The evidence of positive

[14] While the number of people punishing does not rise as the dictator offers become less favorable to the second subject, the size of the average punishment does rise.

reciprocity discussed above is consistent with the existence of some benefits of this type, particularly when, as in Malmendier and Schmidt (2012) it is a third party rather than j himself that is responding by benefiting i.[15]

One immediate implication of this second interpretation is that individual i can be expected to carry out more actions that benefit j if these actions are visible. The reason is that visibility is essential for raising other people's estimates of a_i. There is, in fact, considerable evidence that people act more pro-socially when their actions are public. Engel (2011) shows that the proportion of dictators offering even splits more than doubles when dictators become identified *ex post*. Similarly Ariely et al. (2009) show that people exert more effort on behalf of a charity if this effort is observable to other subjects at the end of the experiment. Lastly, in two experiments in which all potential punishers were shown the same uneven outcome from two earlier experiments, Kurzban et al. (2007) show that third-party punishment of selfish subjects is larger if the punishment is seen by the experimenter, and larger still if it is seen by other subjects. It appears that even small and misleading hints of being observed appear to be sufficient to enhance people's pro-social orientation. Nettle et al. (2013) carry out a meta-analysis of a literature that studies whether images of watching eyes make people more generous. By and large, this literature finds that they do. In their own experiment, for instance, Nettle at al. (2013) demonstrate that being surrounded by posters containing pictures of caricatured composite faces makes people more likely to transfer resources in a DG experiment.

All these experiments have been interpreted as suggesting that people's pro-social acts are based in part on an attempt to acquire a good reputation. Even without reputational effects, straight altruism may be sufficient to explain these findings as long as some individuals derive pleasure from seeing generous (or righteously furious) acts. As discussed earlier these acts can induce increases in utility among individuals who feel self-validated by their agreement

[15] They consider an experiment where a first party can either send or refrain from sending resources to a second, who then decides whether to benefit this second party or a third. When the first party sends resources to the first, he identifies himself as relatively non-selfish. By the logic of Bénabou and Tirole, he thus becomes more deserving of being favored by the second individual.

with these acts. It might be felt that this alternative explanation is more far-fetched than the idea that people are acting as altruists for personal material gain. It is important to recall, however, that a complete model of signaling one's altruism must also explain why true altruists (which is what people are pretending to be) would in fact benefit from the pro-social act in question. One reason that altruists may benefit from charitable contributions is precisely the happiness that this induces in like-minded individuals and it is then unnecessary to appeal to signaling to explain why contributions are higher in public settings.

Bénabou and Tirole (2006) emphasize a different prediction of their model. They consider the effect of increases in y and show that, even though this raises the "direct" benefit of increasing x, its overall effect on x itself can be relatively modest, and can even be negative when x is public. The reason is that a positive y mutes the extent to which a high x connotes a high a because it can now signal that the individual has a high θ. In other words, the individual may now be acting pro-socially because he is particularly sensitive to the material rewards he gets himself by doing so.

For the model they analyze, the result that there exist a nontrivial set of parameters for which an increase in y raises x by less in the public relative to the private condition is relatively general. The reason is that, for y sufficiently large, the incentive effect is so strong that x is the same under both conditions. Thus, as long as reputational concerns in the public condition lead to a higher level of x when y is zero, increases in y must typically increase x less when others see this action.

Ariely et al. (2009) provide some evidence that is consistent with this effect. They asked their subjects to carry out a task and promised they would donate more funds to a charity the more of the task their subjects completed. Their 2X2 design either made the output of the subjects visible or not (as discussed above) and either did or did not also pay the subjects for the amount of the task that they completed. Consistent with the analysis above, the increase in the subject's output when they received compensation was larger in the anonymous than in the non-anonymous condition. This suggests that, indeed, one reason that subjects carry out pro-social acts when these are public is that they want their audience to infer that they are altruistic. It still leaves open the question of whether subjects want to appear altruistic for the private benefits that they thereby earn, or because they gain vicariously as the utility of other altruists rises when they see these pro-social acts.

3 Signaling one's altruism to oneself

If a person is determined enough to act as if she maximized (1) for fixed a_i, her behavior would be hard to distinguish from that of a person whose actual preferences can be represented by the utility function (1). It might thus seem that true altruists are observationally equivalent to people who wish to convince themselves that they are altruistic even though they are selfish.[16] Under some circumstances, these two types of individuals do not behave identically.

Dal Bó and Terviö (2013) emphasize that it may be difficult to fool oneself forever. In their model, people's actions are not always under their conscious control, so that they are sometimes dictated by their true a_i, which is initially unknown to them. Dal Bó and Terviö (2013) demonstrate the existence of an intra-personal equilibrium in which individuals bolster their self-esteem by acting as if their a_i were high until the moment in which they take a subconscious (selfish) action that proves to them that their a_i is in fact low. If that moment ever comes, they give up on signaling a high a_i to themselves.

Bénabou and Tirole (2011) highlight the fact that, just as in the case where the signal is meant for others, people who wish to signal to themselves that their altruism is $\bar{a}$ will sometimes take actions which are more generous than those taken by someone who is sure that his altruism is $\bar{a}$. Such actions have the benefit of allowing people to distinguish themselves from individuals whose altruism is lower. Bénabou and Tirole (2011) demonstrate this in a setting where people receive an initial glimpse of their true altruism parameter a_i, which they then stand a good chance of forgetting. At the later stage when this forgetting is likely to have taken place, they are assumed to remember a particular action whose ultimate payoff depends on their true altruism. Utility at this stage is increasing in the probability people

[16]Notice that such people are quite different from psychopaths, who are often regarded as lacking the capacity to feel empathy and guilt, and whose signaling seems directed exclusively at others.

assign to being altruistic and this probability is computed on the basis of the remembered action as well as a prior probability. Bénabou and Tirole (2011) study how the memorable action varies both with this prior belief and with the temporary glimpse of one's true altruism. An interesting finding is that the action can be nonmonotonic in the prior belief for people whose glimpse indicates that they are relatively selfish. These individuals choose to carry out the generous action when their prior probability of being altruistic is intermediate (so that the action has a large effect on the posterior) but not when it is very low or very high. According to Bénabou and Tirole (2011), this can explain some of the instability one observes in the extent to which people act generously.

Merritt et al. (2010) survey a literature showing some of the instability motivating Bénabou and Tirole (2011). This literature displays examples where carrying out (or even thinking about carrying out) acts that are socially approved is correlated with less moral behavior in a subsequent act.[17] In Khan and Dhar (2006), for example some subjects were given money to answer a survey asking them whether they would be willing to help a foreign student for two hours while other subjects were given a meaningless task. Both groups of subjects were paid \$2 and were asked whether they wanted to donate part of this to a local charity. The students who were asked whether they would help all said they would, but their average donation was only \$1.20. By contrast, those in the control group who did not get to imagine themselves as helping, donated an average of \$1.70. Somewhat related to this Brañas-Garza et al. (2013) let subjects participate in a battery of DGs that differed by whether participants knew the gender, wealth and political orientation of either the dictator or the receiver. Consistent with the idea that generosity in one instance can lead people to be less generous in the next, they found that offers were negatively serially correlated once they controlled for the characteristics of the pairing.

Interpreting these results from the from the perspective of the Bénabou and Tirole (2011) model seems to require that people typically find themselves at a point where a slight increase

[17] This is known as "moral licensing." "Moral cleansing" refers to the ethical acts people perform after less ethical ones.

in the memory of one's good deeds is enough to reduce the incentive to carry out more of them. This seems to run counter to the results of Aknin et al. (2011), where subjects who are asked to recall their own generous acts, and whose memories of these acts is presumably enhanced as a result, become more prone to act generously once again.

4 Altruism that is driven by one's beliefs about others' actions

Rather than depending on i's views regarding j's altruism, a literature starting with Rabin (1993) has allowed i's altruism towards j to depend on i's beliefs about j's action. Suppose, in particular that agent i has an action x_i, expects j to carry out action $E_i(x_j)$, and expects j to believe that i is carrying out $E_i(E_j(x_i))$. Then, each agent is assumed to maximize a utility function of the form (1), except that the altruism parameter a_i is a function of $E_i(x_j)$ and $E_i(E_j(x_i))$. At a rational expectations equilibrium, these beliefs correspond to the actual actions x_j and x_i respectively. What is crucial, however, is that i is more altruistic towards j if he believes that the action that j is taking is beneficial to i given j's beliefs about i's action (as anticipated by i). Rabin's proposes a functional form for $a_i(E_i(x_j))$ which consists of a constant times the ratio of the difference of the payoff that i gets from $E_i(x_j)$ minus an equitable payoff divided by the difference between the highest and the lowest payoff that j can give to i. All these payoffs are computed assuming that i carries out his equilibrium action x_i and that j knows this.

In their extension of Rabin's (1993) model to sequential games, Dufwenberg and Kirchsteiger (2004) use a very similar formulation. As noted by Segal and Sobel (2007), the presence of the equitable payoff matters in this functional form and leads to counterfactual predictions in the experiment of Falk and Fischbacher (2006) that I describe below and in which the equitable payoff is not feasible. Alternative functional forms that avoid this problem are proposed by Falk and Fischbacher (2006), Segal and Sobel (2007), and Cox et al. (2007), though the latter only covers second movers in a sequential two-person game.

This class of models does not involve signaling. Even in the sequential version of Dufwenberg and Kirchsteiger (2004), no player takes into account that he can change the altruism of another by changing his own actions. Rather, players take the beliefs and expected actions of others as given and optimize accordingly. While it might be thought that the absence of signaling has the potential of preventing the existence of multiple equilibria, this turns out not to be the case. Indeed, Dufwenberg and Kirchsteiger (2004) demonstrate the existence of many equilibria, some of which are both highly implausible and avoidable in signaling models. In particular, they show that two equilibria can coexist, one in which a proposer makes an uneven offer in an ultimatum game and this offer is later rejected for sure while, in the other equilibrium, the proposer makes an even offer and this offer is always accepted. These two equilibria cannot coexist in a signaling framework because the proposer would then choose to make the even offer for sure (so that the uneven offer would never be made in equilibrium). The reason the uneven offer persists in Dufwenberg and Kirchsteiger (2004) is that, at this equilibrium, the proposer is upset at the responder as a result of his awareness that, along the equilibrium path, the responder will turn down his own unfavorable offer.

5 Altruism that becomes locally negative when another person's income exceeds one's own

Fehr and Schmidt (1999) suppose that individual i maximizes the utility function

$$u_i^f(x_i, x_j) = x_i - \alpha \max(x_j - x_i, 0) - \beta \max(x_i - x_j, 0), \tag{6}$$

where x_i is a scalar indicating i's resources. With v_i equal to x_i this leads to the same decisions as the altruistic utility function (1) as long as a_i equals $-\alpha/(1+\alpha) < 0$ when $x_i < x_j$, while it equals $\beta/(1-\beta) > 0$ when $x_i > x_j$. Equal outcomes play a key role in their theory because i is altruistic only when he receives more than j, and is otherwise (locally) spiteful. One benefit of this specification is that it is extremely tractable. In the two-player case, these preferences have similar implications to those of Bolton and Ockenfels (2000), who suppose that utility is increasing in one's own payoff and in the extent to which one's own payoff is close to the average of all the payoffs received by all players.

Both Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) motivated their specifications by alluding to the rejections of uneven offers in the UG. As I will discuss, there are reasons to suppose that the models of section 2 provide a better explanation for the UG and UG variants, particularly when results from the UG and the DG are combined. On the other hand, in part because these specifications are very tractable, there is a vast subsequent experimental literature that uses these preferences to interpret its results. While I cannot do justice to this large literature here, I will at least illustrate some experimental findings for which the utility functions in Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) provide the only known formal interpretation.

Fehr and Schmidt (1999) show that the utility function (6) can explain the results in the UG. This requires that a sufficiently large fraction of the population have an α large enough to reject uneven offers. They suppose, in particular, that 70% of individuals have an α greater than or equal to .5. What this means is that 70% of people prefer an outcome of $\{0, 0\}$ to any outcome where the other person gets three or more times what they receive themselves. This turns out to be wildly inconsistent with one of the experiments of Charness and Rabin (2002). In this experiment, which is a variant of the DG, 100% of their subjects choose to have the other subject receive four times more than they receive themselves (800 to their 200) rather than have both subjects receive nothing. Charness and Rabin (2002) also show that 70% of dictators choose to keep 400 and give 750 when they have the option of both giving and keeping 400. Thus 70% are altruistic, rather than spiteful, towards people who receive a larger allocation than themselves.[18]

Motivated by these observations, Charness and Rabin (2002) consider a utility function in which people's utility is increasing in their own payoff, in the sum of all payoffs and in the minimum of all payoffs. In this alternative specification, i experiences a gain in utility whenever j's payoff increases, so that i can be interpreted as being altruistic towards j. As in Fehr and Schmidt (1999), i's marginal benefit from an increase in j's payoff drops

[18]Using a sample that seeks to be representative of the population of Sardinia, Pelligra and Stanca (2010) find that only that 43 % of their subjects prefer to keep 400 and give 800 when they have the option of both giving and keeping 400.

discontinuously as j's actual payoff starts exceeding i's own and thereby stops being the minimum payoff. Distinguishing between these discontinuous changes and the more gradual reductions that would follow from assuming that the v's in (1) are concave should be a priority for experimental research.[19]

A subsequent literature has made it clear that the reciprocity observed in the UG is not exclusively due to a concern for equality. First of all, in an experiment alluded to earlier, Falk and Fischbacher (2006) show in a simplified version of the UG that an unfavorable offer of 8/2 is both much more likely to be accepted and more likely to be made when a symmetric 5/5 offer is unavailable than when it is. The rejection of the 8/2 offer has the same effect on inequality in both cases, so the bulk of these rejections must be due to the sort of considerations I surveyed earlier. Second, Blount (1995) and Falk et al. (2008), and Bellemare et al. (2011), show that second players who receive a disproportionately small allocation are much less likely to sacrifice to punish the first player if the allocation is made by a computer than if it is made by the first player himself.

Lastly, Rotemberg (2008) argues that these preferences cannot explain one of the central findings of Forsythe et al. (1994), namely that even splits are offered more frequently in the UG than the DG, unless α is set to an unreasonably large value.[20] The reason is that, for plausible values of α, responders should accept some offers that are less generous than even splits.[21] Thus, even splits are offered only by individuals with $\beta \geq .5$. Fehr and Schmidt (1999) suppose that 40% of people satisfy this condition, and these individuals should also offer even splits in the DG.

The aforementioned should be interpreted as saying that the combination of DG and UG

[19] López-Pérez (2008) supposes that, in addition to depending on one's own payoff, the total of all payoffs and the minimum payoff, utility also depends negatively on the maximum payoff (with the same coefficient as its positive dependence on the minimum payoff). The difference between the implications of this formulation and the specification in Charness and Rabin (2002) is likely to be most manifest when there are more than two players. Experimental work that takes advantage of these differences has yet to be carried out.

[20] Rotemberg (2008) shows that the Bolton and Ockenfels (2000) model is also inconsistent with this fact. Interestingly, Barr et al. (2009) show that there exists societies in which even splits are offered more frequently in the DG than the UG. This seems to call for rather different modeling of social preferences.

[21] Fehr and Schmidt (1999) suggest a distribution of preferences in which the highest value of α is 4. It follows that no one turns down an offer above 4/9 of the total pie.

observations is easier to interpret if one imagines that, as in the models of Section 2, people react negatively to low UG offers for reasons that go beyond their not liking the resulting allocation. At the same time, these signaling models cannot explain why any subject would give up resources of her own to reduce the resources of others when these others have not previously done anything objectionable. The models of Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), by contrast, do predict such sacrifices as long as the difference between the endowments of the two subjects is large enough. Andreoni and Miller (2002), Dawes et al. (2007), Falk et al. (2008), Leibbrandt and López-Pérez (2011), display sacrifices of this sort and, consistent with an aversion to inequality these sacrifices are more common when the subject from whom resources are taken has a higher endowment.[22] For example, Dawes et al. (2007) report that subjects are about two times more likely to take from someone whose endowment exceeds their own rather than from someone whose endowment is lower, where the earnings are randomly generated by computer.[23]

Still, Falk et al. 's (2008) results show that the effect of computer-generated inequality on taking can be modest. Even when the resources of the high earner are 3 times larger than those of the person doing the taking, the average high earner loses less than 10 % of his resources as a result of taking. At the same time, these terms lead some subjects to take even from impoverished subjects. Leibbrandt and López-Pérez (2011) show that 4% of subjects endowed with 200 units took resources from a subject who only had 60, even though this second subject had been victimized by a third subject who had chosen an allocation of $\{590, 60\}$ when he could have chosen $\{150, 150\}$.

When subjects reduce their own payments in exchange for reducing the payments made to others, the funds go to the experimenter. An alternative to the standard assumption that these funds do not affect participant's utility is to suppose that subjects have an a_i towards the experimenter which is positive as well. If v_j is concave, the value to i of additional re-

[22]Bellemare at al. (2011) demonstrate that, faced with computer generated offers that are lopsided in favor of themselves, about 40% of their respondents choose an allocation of $\{0, 0\}$ instead. Bellemare et al. (2011) includes references to other studies with similar findings.

[23]These findings can be found in the online supplement to the paper.

sources in j's pockets decreases with j's income so that endowing j with more resources ought to make redistribution from j to the experimenter more attractive. This could conceivably explain the slight increase in taking that seems to be associated with increased inequality. To my knowledge, the possibility that subjects feel altruism towards the experimenter has not been investigated in the experimental literature on social preferences. This seems like a mistake because it is well known that experimenters can induce compliance with unpleasant requests even in the absence of payments to subjects. Thus, the perceived preferences of the experimenter evidently have some effect on her subjects.

Andreoni and Miller (2002) also show that a relatively high proportion (half) of the subjects who unilaterally give up resources to reduce the payoffs of others make choices that are incompatible with convex preferences. This can be seen in the choices they display for subject 219, who seems willing both to sacrifice so that the other subject becomes rich relative to himself and to reduce the payoffs of both subjects when the other subject is richer than himself. In particular, the subject chooses $\{1, 12\}$ when he can set x to any value between 0 and 4 in $\{4 - x, 4x\}$. At the same time, he chooses the allocation $\{.8, 10.4\}$ when he can choose any allocation $\{y, 13y\}$ with $y \leq 1$. It is almost as if he treats a large sacrifice in the form of a high x as attractive when generosity is cheap (so that a positive a_i is evoked by this opportunity) while he sees the sacrifice implied by a reduction in y as attractive when this inflicts a relatively large degree of pain in the other subject (so that a negative a_i is evoked by this one). This subject may be valuing the large changes in the emotional state of the other subject that he is inducing, and this is not well captured by conventional preferences.[24]

Andreoni and Bernheim (2009) add a signaling element to the egalitarian models of Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). Individual i chooses his consumption x_i and thereby leaves individual j with $(1 - x_i)$. Rather than representing altruism, the indi-

[24]It remains an open question whether this interpretation is valid more generally for people whose behavior is inconsistent with convex preferences when they have access to both "generous giving" and "cheap mutual destruction" technologies. The behavior of this particular subject may, after all, be explicable by extremeness aversion (Simonson and Tversky 1992).

vidual's parameter a_i now captures how much weight he puts on the egalitarian allocation. And, as in Bénabou and Tirole (2006), i cares for unspecified reasons about the expectation of a_i, $E(a_i)$, held by others. This person's utility thus has some similarities with (5) and can be written as

$$u_i = v_i(x_i) + a_i w(x_i - 1/2) + \mu_a E(a_i) \tag{7}$$

where the w function has a global maximum at zero. Their model implies that there should be no transfers that are just shy of 1/2 while transfers of 1/2 should be relatively common, as is indeed the case in standard DGs. On the other hand, in the modified DG studied by Krawczyk and Le Lec (2010) in which each unit transferred by the dictator is worth one third as much to the second subject, transfers of 1/2 are neither particularly common nor non-existent. Indeed, slightly smaller transfers are somewhat more common. This evidence suggests that the equality norm has less relevance in this case.

Andreoni and Bernheim (2009) also consider a variant of the DG where there is a probability p that the allocation is chosen by a computer who then has a fifty/fifty chance of picking an allocation that is very favorable to i and one that is very favorable to j. As p rises above zero, fewer dictators choose the equitable allocation and more pick the allocation favorable to them that could have been chosen by the computer. This is consistent with their model though it would not be surprising if the signaling models discussed above would predict this as well.

Rather than supposing that people get punished for not being egalitarian, López-Pérez (2008) assumes that they get punished when they fail to maximize the objective function described in footnote 19. One strength of López-Pérez (2008) and Andreoni and Bernheim (2009) is that, relative to models where people signal their altruism, it is easier to rule out transfers of more than half, and these are indeed relatively uncommon in DG experiments. The reason these are easy to rule out is that transferring more than half does not easily convey the idea that one cares more about having a norm-complying division. In models where people signal their altruism, by contrast, the single crossing property leads more altruistic individuals to have a smaller cost of giving a bit more than less altruistic individuals, and this can lead to an incentive to give more than half. This incentive might be extinguished if altruists started to suffer losses when giving more than half. In fact, giving more than half could easily be distressing to altruistic receivers (who prefer an equal division) and thereby indirectly costly to the altruistic givers themselves. This force is currently absent from models in which people signal their altruism, and might help improve the realism of these models.

6 Altruism with transfers of uncertain value

The discussion so far has involved payoffs that are known by the time altruists have finished making their decisions. In practice, of course, the ultimate payoffs of most decisions are random at the time these decision are made. This raises the obvious question of how this uncertainty affects decisions that affect the payoffs of others. The most straightforward theoretical approach is to treat $v_i(X_i)$ and $v_j(X_j)$ in (1) as expected utility functions that depend on final outcomes, so that the maximization of u_i inherits the properties of expected utility functions.

Unfortunately, there is considerable evidence against the idea that all subjects maximize an expected utility function of this sort. Kircher at al. (2013) consider a DG variant and show that, when A and B are allocations that differ in terms of how generous they are to the decider and the receiver, about 30% of their subjects choose a lottery that gives equal probability to A and B even when the direct choices of A and B are also available. As they note, it is extremely unlikely that these subjects are strictly indifferent between A and B, so that these subjects prefer the randomization itself.

In a related experiment, Krawczyk and Le Lec (2010) and Brock et al. (2013) consider a variant in which dictators choose the probability that they win their own endowment, with the full endowment going to the receiver when dictators lose. Over 40% of Brock et al. (2013) dictators lower their own probability of winning below 1. This result is quite extreme because indifference between keeping all the available resources to oneself and giving them all to another is inconsistent with essentially every formal model of altruism that has ever been proposed. One possible explanation is extremeness aversion of the sort demonstrated in Simonson and Tversky (1992). In any event, some people evidently like to give responsibility to a randomization device over allocations that are either more favorable to the self or more favorable to another individual. As proven by Fudenberg and Levine (2012), this preference violates the independence axiom on which expected utility representations are based.

To explain results of this type, Trautmann (2009) suggested that the final outcomes x_i and x_j be replaced by their expectation $E(x_i)$ and $E(x_j)$ in the utility function proposed by Fehr and Schmidt (1999), so that i's utility is given by $u_i^f(E(x_i), E(x_j))$. The idea of taking a utility function defined over final outcomes and replacing these final outcomes with their expected value can obviously be applied also to utility functions such as (1). Aside from being somewhat arbitrary, this approach has an unrealistic prediction, namely that the variability of payoffs should not affect the expected value of transfers in DG games. As both Krawczyk and Le Lec (2010) and Brock et al. (2013) show, however, the expected value of transfers is lower when these consist of lottery tickets (with random payoffs) rather than being deterministic.

An alternative suggested by Fudenberg and Levine (2012), and for which Saito (2013) provides axiomatic foundations, is to suppose that utility is a linear combination of $u_i^f(E(x_i), E(x_j))$ and of the expected value of $u_i^f(x_i, x_j)$. Because maximizing the expected value of the Fehr and Schmidt (1999) utility function $u_i^f(x_i, x_j)$ predicts that there will be no transfers when what is transferred are lottery tickets which determine whether the dictator or the recipient receive the entire endowment, this modification could in principle explain why transfers are lower in the random than in the deterministic case. As emphasized by Fudenberg and Levine (2012), this combined utility function inherits the Fehr and Schmidt (1999) aversion to having unequal ultimate payoffs. The findings of Bolton and Ockenfels (2010) suggest, however, that the inequality of ultimate payoffs plays essentially no role in certain simple settings. They consider a DG variant in which the dictator can give up his (and the receiver's) endowment in exchange for a 50% probability of receiving x and a 50% probability of receiving zero. In some treatments the receiver obtains x whenever the dictator receives x (so that there is no ultimate inequality) while in others the receiver does so when the dictator gets nothing (so that inequality is acute). Overall, the subjects in Bolton and Ockenfels (2010) were equally likely to pick the lottery in both treatments.

While not developed to the point that they can be confronted with experimental data, two axiomatic approaches have been proposed to construct "social welfare functions" with the property that people who maximize these functions might grant others a positive probability of earning a reward large enough to induce inequity. Following Grant et al. (2010), one approach is to suppose that, when i is a dictator, he maximizes a function like (1) where v_j is a function that is concave (and increasing) in the *expected utility* that j obtains from his material payoffs. The concavity of the function then ensures that i does not want to keep j's expected utility at zero. Borah (2013) supposes instead that v consists of j's expected utility plus a "correction," which penalizes situations in which j is less likely than others to receive a high payoff.

So far, this section has considered the role of uncertainty in decisions that are taken before the uncertainty is resolved. Bolton et al. (2005) stress instead that past uncertainty about outcomes can affect decisions that are taken *ex post*. They postulate a "self-centered" utility function that, in addition to depending positively on an individual's payoff, depends negatively on both the difference between the individual's actual payoff and the actual payoff of others as well as on the difference between the payoff the individual could have expected to receive given the choices of others and the payoff that others could have expected to receive. In Bolton et al. (2005), it is the minimum of the these two differences that determines individual utility. Building on their work, Krawczyk (2011) lets both differences matter at the margin.

The key implication of this idea is that subjects are more likely to turn down offers the more likely it was, *ex ante*, that they would end up with an unfavorable offer. Bolton at al. (2005) demonstrate this effect by considering a game where a dice determines whether the allocation $A = \{200, 1800\}$, $B = \{1000, 1000\}$ or $C = \{1800, 200\}$ is available to the second subject, who can also "reject" the offered allocation and choose $\{0, 0\}$ instead. Using the strategy method, 34% of subjects say they would reject C when the *ex ante* probabilities of A, B, and C are .01, .01 and .98 respectively, so that the allocation is almost certain to be unfavorable to the deciding subject. By contrast, only 19% of subjects say they would reject C in the two specifications in which A and C were equally likely in advance. This suggests that, indeed, the *ex post* acceptability of an offer depends on the prior likelihood of other offers and not just on the properties of the offer itself.

It is not obvious, however, that one should interpret this finding as demonstrating that a key determinant of the *ex post* acceptability of an offer is the difference in the *ex ante* expected payoff of the two participants. To see this, imagine that the subject had been presented with the degenerate lottery whose *ex ante* probabilities on A, B and C are 0, 0, and 1. Since the expected value of this game is slightly lower for the second subject, this theory predicts that it should turn down C with a probability greater than 34%. The findings of Andreoni and Miller (2002) cast doubt on this conclusion. One of their treatments involves having the deciding subject endowed with 1/13 as many resources as the other subject, so that initial inequity is even larger than it is under C, where the deciding subject is endowed with 1/9 as many resources as the other subject. Andreoni and Miller (2002) also let the deciding subject choose $\{0, 0\}$ instead, though their choice set is richer because the deciding subject can multiply the endowment of both subjects by any number between zero and one. Andreoni and Miller (2002) report that only 20% of their subjects set a number strictly below 1. And, while some of the subjects who do so set this number to zero, the average multiple for subjects that sacrifice to reduce the others' resources is .5. One might conclude from this that subjects find it particularly infuriating to have a small positive probability of winning a prize and take their revenge on the subject who has a high probability of doing so, though this seems somewhat inconsistent with the willingness of dictators to transfer such small probabilities in Krawczyk and Le Lec (2010).

7 Conclusion

The study of preferences that involve the payoffs of others has a large unfinished agenda. First of all, there is the age-old question of whether these preferences are utilitarian in the sense of having individuals care about the utility of others or deontological, in terms of having them seek actions that are desirable on *a priori* grounds. To this point, the popularity of equal divisions seems easier to rationalize along the latter lines, as in Andreoni and Bernheim (2009). A richer model of altruism, in which altruists seek to benefit other altruists who themselves do not want to receive more than half the pie, might provide a utilitarian foundation for this finding.

It is also worth noting that most actions that affect others do not involve items that can be divided in equal parts, and the utilitarian approach seems easier to apply in these cases. Indeed, this survey has emphasized that the simple model where people experience vicariously some of the gains and losses made by others has the potential to explain a wide range of phenomena if one supposes that people's "direct" utility does not depend only on their consumption of material resources. If it also depends on their regret, their disappointment, their feeling that others agree with them, and so on, there is much more scope for people to take actions in order to affect the well-being of others. Thus, theoretical and empirical research on these other forces might usefully be combined with research on the importance of simple altruism.

I have also surveyed an extensive literature in which people signal their altruism, or their pro-social orientation. Such signaling models can explain a number of empirical findings. Perhaps because these signaling equilibria are not always easy to compute, the experimental literature has not devoted much effort at seeking to discover the limitations of these models. Such an effort would appear worthwhile.

The experimental literature has been much more focused on the question of whether equality of outcomes is an important determinant of altruism. As I have discussed, this force appears to be fairly muted among the university-based experimental subjects that have typically been employed in the past. As an example, Falk et al. (2008) show that very large increases in inequality lead to only small increases in spitefulness. [25] In exploring this issue, the literature has discovered that some people act spitefully without apparent provocation. This phenomenon deserves more study.

The last class of models that I have discussed in some depth are models in which people give each other gifts of uncertain value, i.e. lotteries. When it comes to getting rewards for oneself, the risk induced by randomization is usually seen as a negative. On the other hand, some people seem to like the use of randomization devices when it comes to distributing resources between oneself and others. They are, for example, quite willing to give others a small probability of receiving endowments that they could keep for themselves. One attractive aspect of this finding is that it fits with a long standing preoccupation of the social welfare literature, which has asked itself since Diamond (1967) whether using randomization devices to distribute resources could increase social welfare. The attempt to elucidate what preferences might describe individual generosity under uncertainty is thus bringing back together the study of altruistic preferences and the study of social welfare. This is an exciting development in part because traditional utilitarian social welfare functions share common roots with simple altruistic utility functions.

[25] This effect is stronger in the sample from Sardinia studied in Pelligra and Stanca (2010).

References

Aknin LB, Dunn EW, Norton MI. 2012. Happiness runs in a circular motion: Evidence for a positive feedback loop between prosocial spending and happiness. *J. Happiness Stud.* 13:347–55

Andreoni J. 1988. Privately provided public goods in a large economy: The limits of altruism. *J. Public Econ.* 35:57–73.

Andreoni J. 1989. Giving with impure altruism: Applications to charity and ricardian equivalence. *J. Polit. Econ.*, 97:1447–58

Andreoni J, Bernheim BD. 2009. Social image and the 50-50 norm: A theoretical and experimental study of audience effects. *Econometrica.* 77:1607–36

Andreoni J, Miller J. 2002. Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica.* 70:737–53

Ariely D, Bracha A, Meier S. 2009. Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *Am. Econ. Rev.* 99:544-55

Baron, RA. 1978. Invasions of personal space and helping: Mediating effects of invader's apparent need. *J. Exp. Soc. Psychol.* 14:304–12

Baron RA. 1997. The sweet smell of... helping: Effects of pleasant ambient fragrance on prosocial behavior in shopping malls. *Personal. Soc. Psychol. Bull.* 23:498–503

Barr A, Wallace C, Henrich J, Ensminger J, McElreath R, Barrett C, Bolyanatz A, Cardenas JC, Gurven M, Gwako E, Lesorogol C, Marlowe F, Tracer D, Ziker J. 2009. *Homo qualis: A Cross-Society Experimental Analysis of Three Bargaining Games.* CSAE Working Paper 2009-02 http://www.csae.ox.ac.uk/workingpapers/pdfs/2009-02text.pdf

Batson CD, Duncan BD, Ackerman P, Buckley T, Birch K. 1981. Is empathic emotion a source of altruistic motivation?" *J. Personal. Soc. Psychol.* 40:290–302

Bellemare C, Krger S, van Soest A. 2011. Preferences, intentions, and expectations violations: A large-scale experiment with a representative subject pool. *J. Econ. Behav. Organ.* 78:349–65

Bénabou R, Tirole J. 2006. Incentives and prosocial behavior. *Am. Econ. Rev.* 96:1652–78

Bénabou R, Tirole J. 2011. Identity, morals and taboos: Beliefs as assets. *Q. J. Econ* 126:805–55

Bentham J. 1907. *An Introduction to the Principles of Morals and Legislation.* Oxford:Clarendon Press

Benz M, Meier S. 2008. Do people behave in experiments as in the field? Evidence from donations. *Exp. Econ.* 11:304–12

Berkowitz L, Harmon-Jones E. 2004. Toward an understanding of the determinants of anger. *Emotion.* 4:107–30

Blount S. 1995. When social outcomes arent fair: The effect of causal attributions on preferences. *Organ. Behav. Hum. Decis. Process.* 63:131–44

Bohnet I, Zeckhauser R. 2004. Trust, risk and betrayal. *J. Econ. Behav. Organ.* 55:467–84

Bolton, GE, Brandts J, Ockenfels A. 2005. Fair procedures: Evidence from games involving lotteries. *Econ. J.* 115:1054-76

Bolton GE, Ockenfels A. 2000. ERC: A theory of equity, reciprocity, and competition. *Am. Econ. Rev.* 90:166–93

Bolton GE, Ockenfels A. 2010. Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and The United States: Comment. *Am. Econ. Rev.* 100:628–33

Borah A. 2013. *Accommodating Procedural Fairness when Harsanyi's Impartial Observer is a Utilitarian: An Axiomatization.* Unpublished manuscript, Univ. of Mainz http://www.macro.economics.uni-mainz.de/Dateien/ProcFair.pdf 2013.

Brañas-Garza P, Bucheli M, Espinosa MP, García-Muñoz T. 2013. Moral cleansing and moral licenses: Experimental evidence. *Econ. Philos.* 29:199–212

Brock JM, Lange A, Ozbay EY. 2013. Dictating the risk: Experimental evidence on giving in risky environments. *Am. Econ. Rev.* 103:415–37

Carlo G, Okun MA, Knight GP, de Guzman MRT. 2005. The interplay of traits and motives on volunteering: Agreeableness, extraversion and prosocial value motivation. *Personal. Individ. Differ.* 38:1293–305

Charness G, Rabin M. 2002. Understanding social preferences with simple tests. *Q. J. Econ.* 117:817–69

Cheung-Blunden V, Blunden B. 2008. The emotional construal of war: Anger, fear, and other negative emotions. *Peace Confl.* 14:123–49

Cox JC. 2004. How to identify trust and reciprocity. *Games Econ. Behav.* 46:260–81

Cox JC, Friedman D, Gjerstad S. 2007. A tractable model of reciprocity and fairness. *Games Econ. Behav.* 59:1745

Cox JC, Friedman D, Sadiraj V. 2008. Revealed altruism. *Econometrica.* 76:31-69

Dal Bó E, Terviö M. 2013. Self-esteem, moral capital, and wrongdoing. *J. Eur. Econ. Assoc.* 11:599–663.

Dana J, Cain D, Dawes, R. 2006. What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organ. Behav. Hum. Decis. Process.* 100:193–201

Dawes CT, Fowler JH, Johnson T, McElreath R, Smirnov O. 2007. Egalitarian motives in humans. *Science.* 446:794–6

Dawes CT, Loewen PJ, Fowler JH. 2011. Social preferences and political participation. *J. Polit.* 73:845–56

DellaVigna S, List JA, Malmendier U. 2012. Testing for altruism and social pressure in charitable giving. *Q. J. Econ.* 127:1–56

Diamond PA. 1967. Cardinal welfare, individualistic ethics, and interpersonal comparison of utility: Comment. *J. Polit. Econ.* 75, 1967, 765-66.

Dufwenberg M, Gneezy U. 2000. Measuring beliefs in an experimental lost wallet game. *Games Econ. Behav.* 30:163–82

Dufwenberg M, Kirchsteiger G. 2004. A theory of sequential reciprocity. *Games Econ. Behav.* 47:268–98

Edgeworth FY. 1881. *Mathematical Psychics.* London:Kegan

Engel C. 2011. Dictator games: A meta study. *Exp. Econ.* 14:583–610

Engelmann D, Strobel M. 2004. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *Am. Econ. Rev.* 94:857–69

Evren Ö. 2012. Altruism and voting: A large-turnout result that does not rely on civic duty or cooperative behavior. *J. Econ. Theory.* 147:2124–57

Falk A. 2007. Gift exchange in the field. *Econometrica.* 75:1501–11

Falk A, Fehr E, Fischbacher U. 2005. Driving forces behind informal sanctions. *Econometrica.* 73:2017–30

Falk A, Fehr E, Fischbacher U. 2008. Testing theories of fairness - intentions matter. *Games Econ. Behav* 62:287–303

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

Feddersen TJ, Sandroni A. 2006. A theory of participation in elections. *Am. Econ. Rev.* 96:1271–82

Fehr E, Leibbrandt A. 2011. A field study on cooperativeness and impatience in the

Tragedy of the Commons. *J. Public Econ.* 95:1144–55

Fehr E, Schmidt KM. 1999. A theory of fairness, competition and co-operation. *Q. J. Econ* 114:817–68

Fehr E, Fischbacher U. 2004. Third-party punishment and social norms. *Evol. Hum. Behav.* 25 (2004) 6387

Fershtman C, Gneezy U, List JA. 2012. Equity aversion: Social norms and the desire to be ahead. *Am. Econ. J.: Microecon.* 4:131–44

Fisman R, Kariv S, Markovits D. 2007. Individual preferences for giving. *Am. Econ. Rev.* 97:1858–76

Forsythe R, Horowitz JL, Savin NE, Sefton M. 1994. Fairness in simple bargaining games. *Games Econ. Behav.* 6:347–69

Fowler JH. 2006. Altruism and turnout. J. Politics. 68:674–83

Fowler JH, Kam CD. 2007. Beyond the self: Social identity, altruism, and political participation. *J. Politics.* 69:813–27

Fudenberg D, Levine DK. 2012. Fairness, risk preferences and independence: Impossibility theorems. *J. Econ. Behav. Organ.* 81:606–12

Gailliot MT, Baumeister RF. 2007. Self-esteem, belongingness, and worldview validation: Does belongingness exert a unique influence upon self-esteem?. *J. Res. Personal.* 41:327–45

Gino F, Pierce L. 2010. Robin hood under the hood: Wealth-based discrimination in illicit customer help. *Organ. Sci.* 21:1176–94

Goeree JK, Holt CA, Laury SK. 2002. Private costs and public benefits: Unraveling the effects of altruism and noisy behavior. *J. Public Econ.* 83:255–76

Goette L, Huffman D, Meier S. 2006.The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *Am. Econ. Rev.* 96:212–6

Grant S, Kajii A, Polak B, Safra Z. 2010. Generalized utilitarianism and Harsanyi's impartial observer theorem. *Econometrica.* 78:1939–71

Jankowski, R. 2007. Altruism and the decision to vote: Explaining and testing higher voter turnout. *Ration. Soc.* 19:5–34

Johnson CD, Gormly J, Gormly A. 1973. Disagreements and self-esteem: Support for the competence-reinforcement model of attraction. *J. Res. Personal.* 7:165–72

Kenworthy JB, Miller N. 2001. Perceptual asymmetry in consensus estimates of majority and minority members. *J. Personal. Soc. Psychol.* 80:597–612

Khan U, Dhar R. 2006. Licensing effect in consumer choice. *J. Mark. Res.* 43, May 2006, 259–66

Koszegi B, Rabin M. 2006. A model of reference-dependent preferences. *Q. J. Econ.* 121:1133–65

Krawczyk MW, Le Lec F. 2010. "Give me a chance!" An experiment in social decision under risk. *Exp. Econ.* 13:500–11

Krawczyk MW, 2011. A model of procedural and distributive fairness. *Theory Decision.* 70:111–28

Kurzban R, DeScioli P, OBrien E. 2007. Audience effects on moralistic punishment. *Evol. Hum. Behav* 28:75–84

Laury SK, Taylor LO. 2008. Altruism spillovers: Are behaviors in context-free experiments predictive of altruism toward a naturally occurring public good? *J. Econ. Behav. Organ.* 65:9–29

Leibbrandt A, López-Pérez R. 2011. The dark side of altruistic third-party punishment. *J. Confl. Resolut.* 55:761–84

Levine DK. 1998. Modeling altruism and spitefulness in experiments. *Rev Econ. Dyn.* 1:593–622

López-Pérez R. 2008. Aversion to norm-breaking: A model. *Games Econ. Behav.* 64:237–67

List JA. 2011. The market for charitable giving. *J. Econ. Perspect.* 25:157–80

Loomes G, Sugden R. 1982. Regret theory: An alternative theory of rational choice under uncertainty. *Econ. J.* 92:805–24

Malmendier U, Schmidt KM. 2012. *You Owe Me.* NBER Working paper 18543

Merritt AC, Effron DA, Monin B. 2010. Moral self-licensing: When being good frees us to be bad. *Soc. Personal. Psych. Compass.* 4/5:344–57

Nettle D, Harper Z, Kidson A,, Stone R, Penton-Voak IS, Bateson M. 2013. The watching eyes effect in the dictator game: It's not how much you give, it's being seen to give something. *Evol. Hum. Behav.* 34:3540

Pelligra V, Stanca L. 2010. *To Give or Not To Give? Equity, Efficiency and Altruistic Behavior in a Survey-Based Experiment.* University of Milano-Bicocca Working Papers 202

Pool GJ, Wood W, Leck K. 1998. The self-esteem motive in social influence: Agreement with valued majorities and disagreement with derogated minorities. *J. Personal. Soc. Psychol.* 75:967–75

Rabin M. 1993. Incorporating fairness into game theory and economics. *Am. Econ. Rev.* 83:1281–302

Rotemberg JJ. 2008. Minimally acceptable altruism and the ultimatum game. *J. Econ. Behav. Organ.* 66:457–76

Rotemberg JJ. 2009. Attitude-dependent altruism, turnout and voting. *Public Choice.* 140:223–44

Rotemberg JJ. 2011. Fair pricing. *J. Eur. Econ. Assoc* 9:952–81.

Rotemberg JJ. 2013. Charitable giving when altruism and similarity are linked. *J. Public Econ.* In press

Roth AE, Prasnikar V, Okuno-Fujiwara M, Zamir S. 1991. Bargaining and market behavior in Jerusalem, Liubljana, Pittsburgh and Tokyo: An experimental study. *Am. Econ. Rev.* 81:1068–95

Saito, K. 2013. Social preferences under risk: Equality of opportunity vs. equality of outcome. *Am. Econ. Rev.* forthcoming

Sandroni A, Ludwig S, Kircher P. 2013. On the difference between social and private goods. *B.E. J. Theor. Econ.*

Schieman S. 2010. The sociological study of anger: Basic social patterns and contexts. In *International handbook of anger: Constituent and concomitant biological, psychological, and social processes.* ed M Potegal, G Stemmler, C Spielberger, pp 329–47. New York:Springer Science

Segal U, Sobel J. 2007. Tit for tat: Foundations of preferences for reciprocity in strategic settings. *J. Econ. Theory.* 136:197–216

Simonson I, Tversky A. 1992. Choice in context: Tradeoff contrast and extremeness aversion. *J. Marketing Res.* 29:281–295

Shotland RL, Stebbins CA. 1983. Emergency and cost as determinants of helping behavior and the slow accumulation of social psychological knowledge. *Soc. Psych. Q.* 46:36–46

Sole K, Marton J, Hornstein HA. 1975. Opinion similarity and helping: Three field experiments investigating the bases of promotive tension. *J. Exp. Soc. Psych.* 11:1–13

Srivastava J, Espinoza F, Fedorikhin, A. 2009. Coupling and decoupling of unfairness and anger in ultimatum bargaining. *J. Behav. Dec. Making.* 22:475–89

Sugden, R. 1982. On the economics of philanthropy. *Econ. J.* 92:341–50

Taubinsky D. 2012. *Great Expectations and Prosocial Acts: Theory and Experiments.* Unpublished manuscript, Harvard Univ.

Tucker L, Hornstein HA, Holloway S, Sole K. 1977. The effects of temptation and information about a stranger on helping. *Personal. Soc. Psychol. Bull.* 3:416–20

Trautmann ST. 2009. A tractable model of process fairness under risk. *J. Econ. Psych.* 30:803–13

Volk S, Thöni C, Ruigrok W. 2012. Temporal stability and psychological foundations of cooperation preferences. *J. Econ. Behav. Organ.* 81:664–76

Voors M, Turley T, Kontoleon A, Bulte E, List JA. 2012, Exploring whether behavior in context-free experiments is predictive of behavior in the field: Evidence from lab and field experiments in rural Sierra Leone. *Econ. Lett.* 114:308–11

Wagner C, Wheeler L. 1969. Model, need, and cost effects in helping behavior. *J. Personal. Soc. Psychol.* 12:111–6