Eager for Fairness or for Revenge?

Psychological Altruism in Economics

CHRISTINE CLAVIEN
University of Lausanne
REBEKKA A. KLEIN
University of Heidelberg

Abstract

To understand the human capacity for psychological altruism, one requires a proper understanding of how people actually think and feel. This paper addresses the possible relevance of recent findings in experimental economics and neuroeconomics to the philosophical controversy over altruism and egoism. After briefly sketching and contextualising the controversy, we survey and discuss the results of various studies on behaviourally altruistic helping and punishing behaviour, which provide stimulating clues for the debate over psychological altruism. On closer analysis, these studies prove less relevant than originally expected because the data obtained admit competing interpretations—such as people seeking fairness versus people seeking revenge. However, this mitigated conclusion does not preclude the possibility of more fruitful research in the area in the future. Throughout our analysis, we provide hints for the direction of future research on the question.

1 Three types of debate over altruism

One can distinguish three types of debate that make use of the notion of altruism, each of them employing different senses of the term.[1] The first type of debate centres on the notion of '*biological* altruism'; it asks how behaviours that increase other organisms' *Darwinian fitness*—measured in terms of expected number of offspring—at a cost to the actor's own fitness come to be selected in evolution. Since the publication of Darwin's *Origin of species*, this question has been one of the main challenges to the theory of evolution. Thanks to the efforts of William Hamilton (1964) and others, this difficulty has been resolved.[2]

The second type of debate makes use of the notion of '*behavioural* altruism'; it asks why ordinary people often fail to behave in the way predicted by the neoclassical model of human agency often used in economics. From the behavioural point of view, neoclassical models con-

[1] We would like to thank Philip Kitcher for helping us frame the question in this way.

ceive of human beings as rational maximizers of expected utility. This view is often more or less explicitly combined with a hedonistic assumption of human psychology and with the idea that utility must be understood in monetary terms: according to the 'false consciousness' assumption, "we all like to think of ourselves as nice, caring, altruistic beings, but then when put in the appropriate circumstances—when money is at stake—, we just cannot help but act as the cynical agents postulated by economic models" (Guala 2005: 241).

This way of conceiving human beings as rationally self-interested and always aiming to maximize profits has been powerfully challenged by work in the field of experimental economics. In a large number of studies, it was shown that ordinary people violate this expected utility paradigm; they are ready to contribute to others' welfare and to the common good at their own expense, even where a monetary incentive is at stake (Fehr & Fischbacher 2004a). In the context of this economic debate over human altruism, fitness measures do not enter the picture; behavioural altruism is defined in terms of individual costs and benefits at the end of one or a series of social actions, whereas biological altruism is measured in terms of expected number of offspring at the end of a completed life cycle. More formally, economists label a behaviour 'altruistic' if it benefits other individuals or the common good at some cost for the agent *and* if the cost cannot be compensated in the future. A paradigmatic study case of behavioural altruism is called 'altruistic punishment'. It consists of a disposition to punish unfair agents even if costly and when the punishment provides neither present nor future material rewards (Fehr & Fischbacher 2005b, Fehr & Fischbacher 2003, Fehr & Gächter 2002).

Experimental economists' research has immensely improved our understanding of human social behaviour. Much of the research is also confirmed by a growing body of data stemming from psychology and economics (Camerer*, et al.* 2003, Gigerenzer 2008, Kahneman 2003). This has had a major impact on economics and more generally on the future of social sciences.

The biological and behavioural notions of altruism should not be confused with the psychological or philosophical notion of it. The former two focus strictly on a cost-benefit analysis of behavioural outcomes. In contrast, common 'everyday' use of the term does not refer to *outcomes* but to the subjective *motivation* of the agents. This leads us to the psychological and philosophical debate over altruism (Batson 1991, Sober & Wilson 1998).[3] The traditional debate over the possibility of '*psychological* altruism' centres on the nature of primary motives. The notion of *motive* is a broad category that includes different things, such as desires, intentions, or

[2] More precisely, an expanded version of 'kin selection' theory, which refers to the broader notion of inclusive fitness can explain all cases of biological altruism (Hamilton 1970, Hamilton 1975, Lehmann & Keller 2006).

[3] "Altruism is a motivational state with the ultimate goal of increasing another's welfare. (…) Motivation is energy, a force within the individual (…) [which] is directed toward some goal (…) and [draws] the person toward this goal." (Batson 1991: 6)

judgments. There are two sorts of motives: the first are *primary* motives,[4] usually conceived as the starting points of causal chains that lead to action—they are also the driving force present until actions have come about; the second are *instrumental* motives whose function is to help achieve the aims of the primary motives. Here is an example:

> *Raymond seeks pleasure [primary motive] → Raymond knows that if he does x, he will obtain pleasure [instrumental practical reasoning] → Raymond desires to do x [instrumental motive in order to achieve pleasure] → Raymond does x*

If a primary motive is directed towards the needs and well-being of other individuals, it earns the label 'altruistic'. If a primary motive is aimed at some personal benefit for oneself—as in the example of Raymond—it is considered 'self-interested'.

'Psychological altruism' is the view according to which *at least some* actions are motivated by altruistic primary motives. On the contrary, 'psychological egoism' denies the possibility of primary altruistic motives. According to this latter view, human actions are *always* motivated by the expectation of some personal benefit, usually conceived of in terms of pleasure and avoidance of pain—the hedonistic version—or such things as power, resources, or reputation.

Psychologists (such as Batson 1991, Cialdini, *et al.* 1987) and philosophers (such as Butler 1991, Hutcheson 2004, Nagel 1970) are the traditional actors of this debate. There is a surge of theoretical arguments and experimental data designed to support one or the other thesis. Excellent studies (Batson 1991, Smith, *et al.* 1989, Stocks, *et al.* forthcoming) provide strong evidences in favour of psychological altruism by way of systematically disqualifying some of the most popular 'egoistic' interpretations. These important achievements however are not sufficient to settle the dispute. Not all 'egoistic' explanatory schemes have yet been rejected (Doris, *et al.* forthcoming, Sober & Wilson 1998, Stich 2007) and the findings need to be replicated by independent research groups. Since there are not yet decisive arguments for or against the possibility that *any* primary unconscious motive is necessarily self-directed, the debate still remains open.

Recently, some economists concerned with the study of social preferences have become interested in this debate, although they do not frame their questioning in the precise form preferred by philosophers and psychologists. Fehr and Fischbacher for example, claim that "the self-interest hypothesis assumes that all people are exclusively motivated by their economic

[4] In the philosophical literature, 'primary motives' are usually called 'ultimate motives' (Sober & Wilson 1998: 217-22). However, in order to avoid confusion with the notion of "ultimate cause" as described in biology, we prefer to use the former term.

self-interest. (…) However, the evidence presented in this paper[5] also shows that fundamental issues in economics and the social sciences in general cannot be understood solely on the basis of the self-interest model" (Fehr & Fischbacher 2005a: 183).

In fact, the debate over psychological altruism is particularly relevant for experimental economists; a detailed analysis of human motivation can help to describe how people actually make decisions and why and when they tend to violate the narrow model of rational behaviour usually associated with neoclassical economists. Indeed, as Mayr and colleagues note, typical neo-classical "economic assumptions of stable preferences and rational choices" imply "the assumption of selfish preferences". One indirect way to challenge such a model is to show that its predictions are at odds with the behaviour observed in laboratory experiments. Another more straightforward and decisive argumentative path is to investigate directly the "hidden motives behind altruistic behaviour" (Mayr*, et al.* 2009: 303-04). As we shall see, the use by some economists of fMRI technology provides new hope for a better understanding of our unconscious motives, which is a crucial aspect of the altruism debate.

In must be noted at the outset that, despite appearances, this literature is extremely difficult to read because most authors have not carried out the necessary conceptual work and there is no consensus on the various notions of altruism used (West*, et al.* 2007). Consequently, this literature is filled with conceptual confusions and unwarranted switches from behavioural to psychological altruism.

At times, it is difficult to decide whether the authors are talking about b-altruism or motivation to act b-altruistically. Here is one example: "People are often neither self-regarding nor altruistic. Strong reciprocators are conditional cooperators (who behave altruistically as long as others are doing so as well) and altruistic punishers (who apply sanctions to those who behave unfairly according to the prevalent norms of cooperation)" (Gintis*, et al.* 2005: 8).

Such an imprecise use of terminology facilitates switches from one domain to the other without further argument. For example, after having explicitly said that "throughout the paper we rely on a behavioural – in contrast to a psychological – definition of altruism as being costly acts that confer economic benefits on other individuals", Fehr and Fischbacher add on the very same page that "a combination of altruistic and selfish concerns motivates them [the strong reciprocators]. Their altruistic motives induce them to cooperate and punish in one-shot interactions and their selfish motives induce them to increase rewards and punishment in repeated interactions or when reputation-building is possible" (Fehr & Fischbacher 2003: 785). Similarly, in a context in which the authors are clearly talking about b-altruism, one finds the statement

[5] The evidence they refer to are results from experimental studies based on various games such as the public good or the ultimatum game.

that "altruistic cooperators are willing to cooperate, that is, to abide by the implicit agreement, although cheating would be economically beneficial for them" (Fehr & Rockenbach 2003: 137). In neither case are sufficient arguments provided by the authors in support of their claims about human *motivation* to act b-altruistically.

Other authors (Mayr, *et al.* 2009) appear to distinguish between motives and b-altruism, but since they do not explicitly define their concepts and use the term "altruism" in both domains of discourse, the reader can easily be confused.

This paper uses simple conceptual distinctions as a tool for sorting out and assessing the specific contribution of experimental and neuroeconomics to the debate over psychological altruism. The paper is organised as follows: section 2 surveys and discusses some studies in experimental economics that seem to be relevant to the controversy; as it will turn out, with respect to the debate over psychological altruism, the results of these studies are less satisfying than those obtained in social psychology; section 3 surveys new developments in neuroeconomics, a young research field which makes use of a technology that might help to overcome the limitations of previous studies; we pay particular attention to a study on altruistic punishment (de Quervain, *et al.* 2004) that presents promising results; section 4 presents and discusses various possible interpretations for altruistic punishment; section 5 contains critical conclusions concerning the relevance of the studies discussed as well as constructive suggestions for possible future research.

2 Some interesting data from experimental economics

Many experimental economists are concerned with pro-social behaviour, which is defined as behaviour that benefits group interaction and increases cooperation in human societies.[6] In particular, these economists investigate the necessary conditions and the outcomes of *behavioural altruism*—to avoid confusion in what follows we will use the term 'b-altruism' when referring to this notion. More recently, some researchers have also become interested in the *psychological* aspect of this behaviour. The question we address in this paper is the possible contribution of their studies to the philosophical debate over psychological altruism. Hence, we will not discuss the most well-known and important aspects of experimental economic research, namely its results relating to a better understanding of the modalities and *consequences* of b-altruism, but their relevance for a better understanding of the *causes*—the psychological motives—of this behaviour. The hope is that interesting and concurring results can be obtained in a context other than that of social psychology, where the altruism debate has its roots.

Part of the literature in experimental economics concentrates on *helping and rewarding behaviours*. More precisely, the type of behaviour investigated is helping—or rewarding—other players at some personal cost, even in circumstances where no reputation or future gain can be expected—which is a typical instance of b-altruistic behaviour. Repeated studies have shown that people are ready to share their income with others in a b-altruistic manner (Croson & Konow 2009, Dawes, *et al.* 2007). One experimental situation often used to test this caretaking behaviour is the dictator game. It is a two-person game in which the first player receives an amount of money and is asked to divide it between himself and the second player. The other player cannot reject the split proposed.[7] Contrarily to the prediction of classical economics, experiments show that more than half of the dictators are willing to leave on average roughly 30% of their endowment to the other player (Croson & Konow 2009, Forsythe, *et al.* 1994, Kahneman, *et al.* 1986). Similar results are obtained even in settings where the decision maker's absolute privacy is guaranteed; neither the second players nor the experimenters will be able to trace the identity of the first players (Charness & Gneezy 2008, Hoffman, *et al.* 1996). This result is remarkable because the dictators could keep the whole sum for themselves without expecting any unfavourable consequences, such as punishment or a bad reputation. Moreover, as the game is conducted under anonymity conditions, generous dictators receive neither material benefit nor increase their reputation in return.

One way to interpret these results is to attribute non self-directed fairness motives[8] to the dictators (Bolton & Ockenfels 2000, Fehr & Schmidt 1999). However, such an interpretation seems too simplistic; it could be shown that experimental settings that more closely resemble everyday life positively affect the dictator's generosity (Charness & Gneezy 2008, Hoffman, *et al.* 1996). Therefore, there is reason to suspect the interference of other internal self-directed motives, such as guilt aversion, the unpleasant feeling resulting from acting too selfishly (Charness & Dufwenberg 2006), or aversion to disappointing the second player (Dana, *et al.* 2006, Dufwenberg & Gneezy 2000, Koch & Normann 2008), or an expectation of feeling a warm glow, the pleasant feeling associated with the thought of oneself as a caretaker (Andreoni 1990, Eckel, *et al.* 2005). The latter hypothesis seems even stronger in the light of Haselhuhn and Mellers' study (2005); these researchers report a correlation between fair or cooperative behaviour and a self-reported degree of pleasure derived from acting fairly.[9] It might even be argued that one never knows how realistic the anonymous conditions really are from the sub-

---

[6] For a definition of pro-sociality, see Henrich & Henrich (2006).
[7] In some versions of the game, the second player is replaced by a charity (Eckel, *et al.* 2005).
[8] As we shall see below, a fairness motive can be self-directed.
[9] Recent neuroeconomic data seem to back up this hypothesis (Harbaugh, *et al.* 2007, Mayr, *et al.* 2009).

jects' point of view. Despite experimenters' efforts, the experimental settings might fail to avoid subtle reputation cues (Haley & Fessler 2005). After all, God's eyes bypass the strongest anonymity conditions. Or, as Hoffman and colleagues suggest, there is the possibility that "people have unconscious, preprogrammed rules of social exchange behavior that suit them well in the repeated game of life's interaction with other people. These patterns are imported into the laboratory" (Hoffman, *et al.* 1996: 659).

In light of this review of the literature, one has to admit that current results are not yet fine-grained enough to support either position in the particular debate over altruism. In order to show that altruistic motivation underlies some type of helping behaviour—such as the one observed in the dictator game— the economists should design experiments that allow them to systematically control for the most compelling 'egoistic' explanatory hypotheses available; they must aim to rule out any possibility that 'self-directed' interpretations are sufficient to fully explain the generous behaviour they elicit. One way to do this would be to apply Batson's methodology, which consists in controlling, one after the other, for all plausible egoistic hypotheses. But this requires painstaking work. Economists would have to control at least for: the selfish desire to get rid of the negative emotion elicited by the perception of social inequity; the fear of feeling guilty for not helping or for disappointing the other player; the expectation of feeling the warm glow of being a caretaker. It is not an easy matter to obtain persuasive evidence on these issues, particularly because egoistic explanations are fully compatible with the presence of motives related to social norms. Subjects might, for example, care about fair distribution and still be acting out of self-interest. This is typically the case of those who expect to feel a warm glow or are afraid of feeling guilty. Thus, identifying the influence of social—or even moral norms—is not enough to refute the egoism hypothesis.

It must also be said that other-regarding helping behaviour is especially difficult to trigger in the context of economic games. Croson and Konow (2009) have found that people are significantly more responsive to lack of fairness than to kindness; they punish more than they reward. Cherry and colleagues (2002) reveal the "other people's money problem": subjects who are given money at the beginning of the game do not attach much importance to it, whereas if they are first required to earn their endowments, there is a dramatic decrease in their inclination to spend it generously in the dictator game (see also Oxoby & Spraggon 2008). Frohlich and colleagues (Frohlich, *et al.* 2001) point to the difficulty in anonymous games of eradicating the subject's doubt that a real person will actually receive the money. There is also the difficulty of preventing the subject's tendency to consider an economic game as a setting in which the main goal is to win money (Bowles 2008, Frohlich, et al. 2001). Finally, anonymous games are often

built in a way that makes them resemble real situations very little. This raises serious questions about the external validity of these studies.[10]

Another, possibly more promising, part of the experimental economics literature on motivation concentrates on the proximate mechanisms underlying *b-altruistic punishing behaviours*—as opposed to helping or rewarding behaviour. For example, in an interesting 2002 paper, Ernst Fehr and Simon Gächter conducted an experiment with two tasks. The subjects were first asked to play a 'public goods game'. In such a game, all participants are free to contribute to a group project and once the group project is realized, every member of the group can benefit from it, even those who did not contribute. In this particular experiment, the public goods game was complemented with a punishing condition: at the end of the game, subjects could punish each of the other players after they were informed about the investments they had made. The results showed that subjects usually adopted b-altruistic punishing behaviour; they sanctioned free-riders—those who benefit from the public good without contributing—even under conditions where it was costly for them and they could not expect any present or future material benefits. In the second part of the experiment, participants were asked to fill in a questionnaire designed to sound out the psychological motivation for b-altruistic punishment. Participants were presented with a hypothetical situation and asked to indicate the intensity of their anger and annoyance towards free-riders. The experimenters hypothesized that strong negative emotions such as anger or disgust were the proximate mechanisms responsible for b-altruistic punishment. This could be confirmed by the participants' responses: 84% of them indicated that they would feel a high intensity of anger if they were cheated by a free-rider. Fehr and Gächter concluded that negative emotions were the proximate mechanisms behind human b-altruistic punishment. These results are a first step in our understanding of human motivation. Here again, however, the data do not allow us to reach definitive conclusions about the primary motives underlying b-altruism. Negative emotions, such as anger, can be associated with different sorts of motives, such as the desire to enforce justice or the desire for personal revenge.

More promising results are to be found in games using a 'third party' condition: a third player, not directly involved in the game, observes the other players and is offered the possibility of using his own money to punish the free-riders. Third party conditions are particularly interesting because subjects are not personally involved in the situation so their motivation cannot stem from the most obvious self-directed goals. It has already been shown that third-party pun-

[10] This remark is particularly relevant for some recent experimental designs that, on top of the anonymity condition, ensure *non*-costly punishment (Croson & Konow 2009).

ishment is a stable behavioural pattern (Fehr & Fischbacher 2004b)[11] and it seems that this behaviour is tightly linked to reputation and group membership. A recent study by Kurzban and colleagues (2007) reveals that the propensity and degree of b-altruistic punishment decreases dramatically once there is no reputation to gain from such an action. In another interesting field study with indigenous groups in Papua New Guinea, Bernhard and colleagues (2006) found a significant influence of group membership on punishing behaviour: the b-altruistic punishers protected in-group victims far more often than out-group victims.

These studies show the existence of some patterns in third party punishment. However, they do not provide the necessary controls for deciding whether some of the punishers—especially those who cannot gain a good reputation and those who protect out-group victims—are motivated by purely other-regarding considerations. To test this hypothesis, one would have to be able to disqualify the most compelling 'egoistic' interpretations for this behaviour: for example, the idea that these punishers might seek the warm glow of being a caretaker of justice,[12] or the idea that they do not really discriminate between their own welfare and those of others—in this case, they cannot help identifying themselves with the betrayed players and, consequently, feeling angry and wanting to punish in revenge, as if they themselves had been the target of the free-riding behaviour.

Again, more carefully designed experiments are needed in order to discriminate between possible hypotheses. Experimental economists interested in the debate over psychological altruism could either take this line of approach or they might explore an alternative investigatory path recently made available thanks to the development of new technologies: some economists have started to use fMRI methods to see what happens in our brain while we make decisions to act b-altruistically. The following sections deal with this approach.

3 Some interesting data from neuroeconomics

Recently a new tool has started being used in the social sciences and particularly in economics: functional neuroimaging, most notably, functional magnetic resonance imaging (fMRI). With help of this 'brain scan' technique, one can obtain rough measures of brain activity while people perform a task. Experimental economic laboratory games are particularly well-suited for

[11] In one of their studies, Fehr and Fischbacher (2004b) observed that 60% of third-party participants punished, although they knew that their economic payoff could not be affected by norm violation and that punishment was costly for them and would yield no future benefit. Similar results are to be found in Kahneman et al. (1986) and Carpenter et al. (2004). It should be noted, however, that third party participants punish significantly less than implicated players (Croson & Konow 2009).

[12] Such an attitude is plausible. In a society in which third party punishment is generally praised by the other members of the group, third party punishers might learn to unconsciously associate this type of behaviour with the psychological reward of being praised. Once such an association is fixed in individuals' psychological makeup, subjects might be motivated to punish altruistically even in situations where there is no obvious reward expected.

the use of this method. The hope is that this new tool can provide additional information on subjects' motivation and reactions in social dilemma situations. This tool is particularly interesting for the debate over psychological egoism, since we are at a loss to understand precisely what motivates people to act b-altruistically. In this section, we will survey and discuss some interesting studies in neuroeconomics that seem relevant to our theoretical investigation.

Recall Fehr and Gächter's experiment on the public goods game (2002): the second part of their experiment consisted in asking the subjects what motivated them to b-altruistically punish free-riders; results pointed to the emotion of anger. This result is supported by a complementary study using brain-imaging as a research tool. Alan Sanfey, James Rilling and colleagues (2003) investigated the neural substrates of the cognitive and emotional processes involved in decision-making. Subjects were asked to play an ultimatum game while their brains were scanned with the fMRI method. In an ultimatum game,[13] participants have to decide either to accept or reject an offer of money made by another player whose task is to distribute a sum of money between the two of them. If the offer is rejected, no one gets anything. Thus, the offer is called an 'ultimatum offer'. The results of the study revealed that players confronted with unfair behaviour—unfair distribution of the sum was proposed by the other player—showed increased activity in the 'anterior insula', a brain area associated with negative emotional feelings. Moreover, the strength of the negative emotional response was correlated with the rejection rate of the unfair offer. In this context, rejection of a distribution offer can be interpreted as b-altruistic punishment, because a possible gain is sacrificed by the player in order to maintain a fairness norm.[14]

In another highly interesting study, Dominique de Quervain, Urs Fischbacher, Ernst Fehr and colleagues (2004) introduced a neuroimaging tool into the experimental design of a trust game followed by a punishment condition. In such a game, two players receive the same amount of money. The first player is asked to decide how much of his money to pass on to the second player—the trustee. All money passed is increased by a multiplication factor of two to four—depending on the game. The trustee then decides how much of this to return to the first player. She is allowed to keep all the money for herself, in which case she would show free-riding behaviour. As in Sanfey's experiment, the subjects were brain-scanned during the game

[13] The ultimatum game was first used as an experimental paradigm by W. Güth in 1980. At that time, fair distribution of outcomes in a cooperation game was considered 'irrational'. Experiments using the ultimatum game draw attention to the crucial role of fairness and pro-sociality in economic behaviour.

[14] In the ultimatum game, a significant number of first players propose an equal share of money (Güth, *et al.* 1982, Roth, *et al.* 1991). Thus, at first glance, one could think that it is an instance of generous b-altruistic behaviour. However, one has to keep in mind that the success of the proposal depends on the second player's willingness to reciprocate. As a matter of fact, unfair offers are generally refused by second players. Therefore, the fair offer is not b-altruistic because it is the best strategy available for first players. Moreover, the strategic component of the game is a good reason to doubt whether, at the psychological level, this fair behaviour can be considered non self-interested. In other words, strategic considerations can very well crowd out impulses toward generosity—especially because first players know that small offers are likely to be rejected (Charness & Gneezy 2008, Croson & Konow 2009, Forsythe, *et al.* 1994).

while they learned that they had been cheated by another player and made the decision whether to punish b-altruistically. Observation of the subjects' neural circuit activation showed that a brain area linked to anticipation of reward—the 'caudate nucleus'—played a prominent role when people decided to punish. Subjects who exhibited stronger activation of the caudate nucleus were ready to incur more personal costs to punish a free-rider in comparison with subjects who exhibited low caudate activation. Moreover, it must be noted that the configuration of the game did not allow the punishers to expect future monetary gains in the course of the game; thus, punishment had to be b-altruistic. Experimenters interpreted the reward-directed mechanism underlying b-altruistic punishment in hedonistic terms—people seek satisfaction—and considered it to be a result of evolution.

At first glance, the analysis of the results in terms of *hedonistic* motives is rather surprising because experimental economists are well known to argue in favour of *b-altruism*. In fact, there is no contradiction here because the concept of psychological egoism does not deny that actions caused by self-interested *motives* can have positive *effects* for others and unfavourable ones for the agent—which is precisely how b-altruism is defined. This allows researchers to argue in favour of psychological egoism while maintaining that the behaviour is b-altruistic. As de Quervain, Fischbacher and colleagues put it:

> Thus, the punishment of defectors is an altruistic act in the biological sense[15] because, typically, it is costly for the punisher and induces the punished individual to defect less in the future interactions with others. However, our results suggest that it is not an altruistic act in the psychological sense. (de Quervain, *et al.* 2004: 1257)

Such a remark makes sense in the light of an argumentative strategy often used by advocates of psychological altruism, which consists in providing paradigmatic examples of actions that are not easily explainable in self-directed terms. De Quervain and colleagues invert this strategy and use an example of an apparently altruistic action, showing that, under closer analysis, it turns out to be self-directed. Such a procedure is often used by advocates of psychological egoism in order to undermine the plausibility of the competing theory.

This paradigmatic way of conceiving of motivation in terms of anticipated pleasure is to be found in many economic studies. De Quervain and colleagues report that people seek satisfaction from re-establishing equity. The same line of thought is applied to generous behaviour: in various studies on the psychological and neural mechanisms underlying reciprocal cooperation,

[15] According to the definitions provided in the first section of this paper, the formula "*biological* sense" should be understood here as "*behavioural* sense". Driven by a noble desire to draw interdisciplinary links, experimental ec-

it has been shown that activation of the brain areas linked with reward processing—including the 'caudate nucleus'—positively reinforce reciprocity and help to resist the temptation to defect (Decety, *et al.* 2004, King-Casas, *et al.* 2005, Rilling, *et al.* 2004, Rilling, *et al.* 2002). Tabibnia and colleagues (2008, 2007) propose the same theoretical approach and go as far as saying that it is regrettable that Sanfey et al. (2003) failed to report activation in the reward regions of the brain.

Let us now focus on the question of whether these studies are meaningful in the context of the altruism debate. Data on the neural basis underlying reciprocal cooperation are only mildly relevant because reciprocal cooperation can hardly be conceived as a case of psychological altruism. These studies do not help to answer the question of whether people can act altruistically when it is costly for them and no future benefit is to be expected. Moreover, it must be noted that most of these studies involve economic games where the goal is to win as much money as possible. In such settings, subjects are inclined to cooperate because they are placed in a setting in which they are expected to behave in way conducive to reaping the long term benefits of social interaction.

Studies on the psychological motivation for b-altruism seem more relevant, provided one adopts an interpretation of pro-social behaviour in terms of norm compliance. De Quervain, Fischbacher and colleagues, for example, suggest that "many people voluntarily incur costs to punish violations of social norms" (p.1254) and they feel "satisfaction" and "relief" (p.1258) when they are able to punish norm violations. They claim that it is precisely this satisfaction that drives them to sustain social norms in their group. The most charitable formal way to understand the causal chain implicitly favoured by the authors is the following[16]:

Understanding that the other player has failed to send money back → Understanding that this behaviour does not uphold the social norm of fairness → Feeling outraged → perception of the unpleasantness of this feeling → Desire to get rid of this unpleasant feeling [primary motive] → Desire to re-establish equity of payoffs as a means of relief from the unpleasant feeling [instrumental motive] → Cost-benefit deliberation: re-establishing equity would be a source of satisfaction (a relief); punishment is a means to achieve it, but is costly → Choice to punish → Punishing act

---

onomists sometimes fail to make important distinctions that would help to avoid unnecessary confusions (more on this in West, *et al.* 2007).

16 There is no clear agreement across economists' writings on the sort of psychological motive underlying b-altruism. One finds: "motive to sanction violations of fairness and cooperation norms" (de Quervain, *et al.* 2004); motive to reestablish equal share ("egalitarian motive": Fowler, *et al.* 2005); or simple retaliation motive (Fehr & Gächter 2005). The problem is that it is often unclear what the authors really mean by these formulae. At times, one can even observe slight changes of meaning from one paper to the other. Our goal is not to try to understand what the

It is worth noting that such an interpretation in terms of social norms[17] is especially favoured by those who consider the establishment of cooperation and other-directed behaviour through the enforcement of social norms to be a specific feature of *human* societies, in contrast with animal societies. As de Quervain and colleagues write: "The ability to develop social norms that apply to large groups of genetically unrelated individuals and to enforce these norms through altruistic sanctions is one of the distinguishing characteristics of the human species (de Quervain, *et al.* 2004: 1258)". On this occasion, the neuroeconomists are very much in line with the neoclassical tradition in economics and the social sciences; they elaborate the notion of human b-altruism precisely through a characteristic that allows them to posit a sharp distinction between humans and animals: the behavioural characteristic of norm compliance fits perfectly with the ideal of the uniqueness of the human being. If it were emotions that we have in common with animals that explained altruism, it would be harder to argue that human altruism was unique. In these words, "Experimental evidence indicates that human altruism is a powerful force and unique in the animal world. […] Human societies represent a large anomaly in the animal world. They are based on a detailed division of labour and cooperation of genetically unrelated individuals in large groups" (Fehr & Fischbacher 2003: 785).

4 An Alternative Interpretation

As we have seen, one can garner support for the psychological egoism thesis by showing that a type of behaviour apparently driven by a non-selfish motive—a desire for justice or to promote fairness —is in fact motivated by self-interested considerations—the desire to be relieved of an unpleasant feeling. Confirmation for this interpretation of altruistic behaviour can be found in fMRI studies that show the activation of a hedonistic reward mechanism just before the decision is taken (de Quervain, *et al.* 2004). This analysis is indirectly strengthened by further fMRI studies on cooperative behaviour that reveal activation of the same reward-directed area (Tabibnia, *et al.* 2008, Tabibnia & Lieberman 2007).[18] According to these studies, the desire to promote fairness is only an instrumental motive. Of course, this presents an interesting argu-

authors have in mind, but rather to provide some useful conceptual distinctions that help frame the controversy over psychological altruism and highlight the most interesting competing explanatory models.

[17] This tendency to concentrate the analysis around the notions of fairness or equity and psychological reward is also to be found in studies on cooperative behaviour (e.g. Tabibnia, *et al.* 2008, Tabibnia & Lieberman 2007).

[18] However, we would like to point out that even if fairness is obviously hedonically valued *in some contexts*—mainly economic contexts with the expectation of future win-win cooperation—, it is not clear whether this is the case in contexts in which it is more plausible to hypothesise the occurrence of altruistic motivation. Thus the relevance of these studies for the altruism debate is unclear.

ment in favour of psychological egoism. Noble actions that seem to be driven by a desire for justice and fairness prove to be self-directed under close analysis.

Unfortunately, there are good reasons to doubt the validity of this analysis because the various studies discussed provide only a few indicators about the mental considerations—conscious or not—underlying the choice to punish b-altruistically. Thus, it is really hard to provide an overall explanation of what happens in people's minds when they exhibit b-altruistic behaviour. In what follows, we propose to unfold and assess the plausibility of various possible interpretations for the same data.

If we rely on the first two studies mentioned in the third section (Fehr & Gächter 2002, Sanfey, *et al.* 2003), it seems clear that negative emotions such as outrage or anger have an important role to play in the motivation for b-altruistic punishment (see also Dawes, *et al.* 2007). These emotional reactions are obviously elicited by the perception of an inequitable split. Unfortunately, Fehr, Sanfey and colleagues' results do not allow us to distinguish whether subjects feel angry because (i) a norm has been scorned—which would be an instance of outrage—or because (ii) they are frustrated by a personal loss—which would be a more basic form of anger.[19]

Under the interpretation (i), the subjects feel outraged because what they observe does not correspond to the social norms of cooperation and equity present in their society; at first glance, this seems to lead them either to the 'noble' motive to reinforce these norms by means of punishing the free-riders—a view favoured by de Quervain and colleagues—or to the desire to punish the free-riders simply because, as wrongdoers, they deserve it. However, to make this explanatory scheme compatible with the hedonistic reward mechanism discovered by fMRI studies, a further motive has to be introduced in the causal chain, namely the desire to get rid of the unpleasantness of being outraged. We then end up with a complex and unintuitive causal chain where the feeling of outrage in itself does not play a crucial motivational role. However, the most parsimonious way to explain the observed correlation between the strength of punishment and the emotion of anger—or outrage—consists in attributing a causal role to this emotion. This possibility is overlooked in the abovementioned causal explanation.

Alternatively, we favour the second and more straightforward interpretation (ii), which is more in tune with studies on emotions, does not contain the additional causal step of an instrumental motive and does not attribute sophisticated thoughts and desires to agents, such as a willingness to re-establish justice. According to this account, subjects are simply frustrated and angry because they expect a more favourable outcome for themselves; consequently, either they

[19] The authors of the studies only use the general notion of anger.

aim to compensate their loss—which could be described as a self-directed equity motive[20]—or they expect the sweet psychological taste of revenge.[21] Both of these motives are self-directed. The first motive however seems weaker because a study by Falk and colleagues (2005) has shown that in a public goods game, altruistic punishment is used even in circumstances where the cost of punishing equates the cost of being punished—that is, punishment does not obviously help to diminish inequity towards oneself. In such a situation, simple forms of self-directed equity motives are unlikely to be elicited and revenge seems to be a more plausible motive.

There are further reasons to favour the more straightforward interpretation for b-altruistic behaviour—at least in contexts where the subjects are victims of free-riding or inequity. It is well established that people are particularly sensitive to unfair shares toward themselves (Croson & Konow 2009) and report desire for revenge after having been victim of free riding behaviour (Carlsmith, *et al.* 2008). Moreover, in a recent study, Dawes, Fowler and colleagues (2007) found that subjects reported feeling angry towards other players with a high income, even when they knew that this high income had been randomly distributed. This emotional reaction is also correlated with engagement in b-altruistic punishment. This means that one can be frustrated by inequity and feel revengeful towards the lucky beneficiary,[22] even if she is clearly not responsible for this situation. An interpretation of the proximate mechanism for b-altruism in terms of the desire to maintain social norms or in terms of morally deserved punishment cannot account for these results.

Even more impressively, a recent study by Yamagishi and colleagues (Yamagishi, *et al.* 2009) shows that people 'punish' in situations similar to the ultimatum game, even when their 'punishment' is only costly for them and not costly—or not a 'punishment' at all—for the free-riders. In one version of this game, free-riders are not even informed about the ineffective 'punishments'. In this case, the motive for revenge should be conceived in a broad sense. Subjects merely seek internal retaliation. Their offended pride makes them angry; since there is no real punishing opportunity, they find some satisfaction in showing to themselves that they are independent enough to refuse submission to inequity.

---

[20] It is important to distinguish between two sorts of equity motives. One falls under the category of non self-directed motive for norm compliance; the second is a self-directed motive to reestablish equitable payoff after having been victim of a free-riding behaviour. It is the latter that applies in this context. Moreover, there is an important distinction between 'equity' and 'equality' that is often neglected in the literature: equity, broadly conceived, refers to what is accepted as 'fair' in a social group, whereas equality refers to the equal distribution of a good. This distinction is of importance because in many circumstances—such as dictator games—some unequal shares are considered equitable.

[21] In a short comment following the publication of de Quervain and colleagues' paper, Brian Knutson (2004) already realised that revenge might also play a role in the explanation of the data. However, he was not aware of the full implications of this element for an analysis of the detailed causal relationships.

To come back to the particular case of de Quervain and colleagues' study, if we use the above line of thought, the most plausible explanation seems to be the following: subjects in a trust game get angry when a player they trust does not fulfil their expectations. They trust him to cooperate and share the benefit he has gained thanks to their own trusting behaviour. As soon as they understand that this is not the case, they start feeling frustrated, get angry, seek revenge and begin to think of possible retaliatory actions. Under this interpretation, the following causal chain would hold:

Understanding that the other player has failed to send money back → Understanding that this behaviour does not match one's expectations → Feeling angry → Desire for revenge [motive] → Cost-benefit reflection: punishment is a means to satisfy the desire for revenge, but it is costly → Choice to punish → Punishing act

It is worth noting here that we are not denying the role of social norms in the decision to punish. We take the content of norms to depend on various psychological biases as well as widespread practices in particular communities (Kahneman*, et al.* 1986). The norms set a framework that influences people's expectations and also their reactions and behaviour. In a social environment where norms of fairness are not widely held and applied, people may not expect fair behaviour from other agents. Thus, greedy behaviour may not cause anger and revenge reactions; subjects may even react negatively towards excessively generous offers![23] These behavioural patterns have in fact been observed in various studies (see Bahry & Wilson 2006, Henrich 2004). Overall, it seems that people's emotional and behavioural tendencies are highly context dependent.

5 Disappointments and hopes

If we favour the second interpretation of de Quervain and colleagues' study—people seek revenge—it becomes clear that no decisive conclusion can be drawn from it concerning the possibility of psychological altruism. If subjects in the experiment react out of anger and seek revenge, the experiment is no longer dealing with actions apparently motivated by the desire for justice or to enforce social norms. It is simply concerned with the motivation underlying retalia-

[22] Interestingly, Carlsmith and colleagues (2008) showed that, although people expect revenge to make them feel better, retaliatory actions sometimes fail to produce the expected hedonistic consequences—this is because punishment seems to encourage thoughts of the offender and rumination on the unpleasant situation experienced.

[23] As Benoît Dubreuil points out (submitted), this form of punishment is not yet well understood. Various interpretations are available, such as distrust against those who seem to show off, or pleasure in hurting those who make themselves look cheap.

tory actions and no advocate of psychological altruism would deny that motivation for those actions is egoistic. The altruism thesis does not assert that all actions are altruistically motivated; it maintains that at least some actions are. In sum, at least the experimental results discussed in this paper favour neither psychological altruism nor psychological egoism.

Let us emphasise here that we are not denying the relevance of these studies for a better understanding of the proximate mechanisms underlying cooperative and b-altruistic behaviour. We have only shown that these studies cannot be easily integrated with the particular debate over psychological altruism. This is bad news for those who want to use this debate as a lever to link economic studies with investigations into human morality. This rather disappointing conclusion leads us to more general considerations on the contribution of experimental and neuroeconomic research for the debate over psychological altruism. We will start with some sceptical remarks before concluding with constructive ideas and hopes for future research.

First, there are good reasons to be sceptical about associating psychological motives with behaviour on the basis of brain scans. The possibility of teasing out selfish motivation from other directed motivation by looking at brain activation is controversial. This technology is immensely helpful in understanding highly modular machineries such as vision or smell. But it is unclear whether we know enough about the physical mechanisms underlying people's thoughts—and even their motivation—to allow for interesting inferences.[24] A mental state or an emotional reaction is never located in one single brain region and each brain region is involved in more than one process (LeDoux 2002). Neuropsychologists are very well aware of the difficulty. For example, Golnaz Tabibnia and Matthew Lieberman admit that "we cannot confidently infer from the observation of increased signal in a region that activity in that region evoked one mental process rather than another" (Tabibnia & Lieberman 2007). This being said, it seems plausible to think that observation of neural activity in particular regions of the brain sometimes enables one to rule out specific hypotheses[25] or to strengthen some others. In the latter case, neuroimaging data could converge with interpretations elaborated in other sciences—such as psychology, philosophy or economics.

In sum, one should carefully avoid overstating neurological findings without being oversuspicious. After all, a brief look back into history of science reminds us that new scientific discoveries can stun people at first and that it often takes a long time for the broader scientific community to admit and integrate new knowledge (Camerer 2008).

[24] See Poldrack (2006) and Henson (2006) on the difficulties linked to invalid inference processes while interpreting fMRI results. See also Vul et al. (2009) for a survey of 55 articles in social neuroscience in which the correlations between brain activity and personality measures seem to be overstated given the limited reliability of fMRI techniques.

To conclude, we propose one possible line of research that could be undertaken in order to help solve the debate. Neuroimaging is a particularly interesting tool since, in principle, it could allow us to address some of the limitations encountered by more traditional psychological enquiries. Consider the example of studies on the 'caudate nucleus'. For the sake of the argument, let us take it for granted that activation of the 'caudate nucleus' is necessarily linked to self-directed thoughts—more precisely, to anticipation of reward for oneself. Under this assumption, brain-imaging methods could be used to scan the brains of subjects who seem to act without seeking reward. We have seen that ultimatum games are not particularly well designed for that, but there are other games where apparently altruistic behaviour occurs: for example, third party games would provide an excellent setting for this kind of investigation.

In section 2, we have seen that classical experimental economics helped to reveal some behavioural patterns in third party punishment. However, existing studies still fail to provide data fine grained enough to decide whether any of the punishers are motivated by purely other-regarding considerations. At least two 'egoistic' interpretations need to be dismissed: either the punishers seek the warm glow of being a caretaker of justice[26] or they do not really discriminate between their own welfare and those of others and consequently feel revengeful. Scanning the brains of the players who seem to behave as 'impartial' justice makers might be a way to control for these explanatory hypotheses. If a brain scan can show that the 'caudate nucleus' is not significantly activated in relation to some types of third party punishment, both of these hypotheses might be ruled out—of course, under the condition that this brain area can truly be associated with reward expectation. B-altruistic punishers would more accurately be described as impartial observers and caretakers of justice and social order, thus, altruists in the psychological sense.

If, on the contrary, brain scans show activation of the 'caudate nucleus', the psychological egoism thesis could not be refuted. A further step would then be to determine whether the same kind of self-directed mechanism is activated before other apparently altruistic actions. One type of behaviour that should obviously be investigated is the motivation to help or reward other players at some personal cost, even in circumstances where no reputation or future gain can be expected. The dictator game would provide an excellent setting for this kind of investigation. Here, the most compelling 'egoistic' hypotheses to control for would be the desire to get rid of the negative emotion elicited by perception of social inequity, the expectation to feel a warm

[25] For example, if a particular type of mental process is hypothesised and previous research in neurology indicates that activity of two or more brain regions underlie this type of mental process, one can test whether these particular brain areas are activated according to the hypothesis. If it is not the case, the hypothesis can be put into question.

[26] Such an attitude is rather plausible. In a society in which third party punishment is generally praised by the other members of the group, third party punishers might learn to unconsciously associate this type of behaviour with the psychological reward of being praised. Once such an association is fixed in individuals' psychological makeup, subjects might be motivated to punish altruistically, even in situations where not obvious reward is to be expected.

glow of being a caretaker,[27] and the fear of feeling guilty for not helping or for disappointing the other player. If a brain scan can show that the 'caudate nucleus' is not significantly activated in relation with generous dictators' choices—that is when dictators transfer up to half of their endowment to the second player—it would be difficult to maintain the two first hypotheses. However, the fear of feeling guilty or of disappointing would still be available alongside the other-regarding interpretation.[28] Further studies would then be needed to discriminate among these explanations.[29]

At this stage of the debate, it seems to us that the most promising way to decide empirically whether psychological altruism or psychological egoism is true is to run a carefully designed series of experiments where all plausible alternative hypotheses are systematically controlled for. Here economists could take inspiration from the excellent work done in social psychology (Batson 1991). Both classical experiments and fMRI studies can be combined depending on the hypothesis tested. Brain imaging studies have the disadvantage—at least nowadays—that fMRI data can easily be over-interpreted, casting doubt on the procedure itself. However, if reliable, such a technique can provide an unprecedented tool for discarding—or providing confirmation for—various hypotheses at once.

6. Conclusion

Debates over biological and behavioural altruism seem to have more or less been settled thanks to excellent studies and recent theoretical achievements. This is not the case for the controversy over psychological altruism. The target of this paper was to discuss the possible contribution of experimental economics and neuroeconomics to this particular debate. After having reviewed some relevant studies in experimental and neuroeconomics, we have argued that these experiments are open to competing interpretations regarding motivation, leaving a question mark over whether the psychological altruism thesis is true. This is not to deny in principle that experiments in experimental economics or neuroeconomics could prove relevant for the philosophical debate. However, either more systematic work needs to be done in classical experimental eco-

[27] There has already been some highly interesting neuroeconomic work conducted on this particular hypothesis (Harbaugh, *et al.* 2007).

[28] Recent studies indicate that the caudate nucleus is not only linked to anticipation of reward but also to anticipation of escaping costly punishment (McCabe 2008, Spitzer, *et al.* 2007). We ask ourselves whether the caudate nucleus is so specifically oriented or whether it is linked to a broader sensitivity towards unwanted situations. If it is activated in anticipation of reward and of relief from unfortunate outcomes—among them the fact of feeling guilty—the two latter egoistic hypotheses could be controlled for with the same experiment.

[29] A complementary path of investigation would consist in making use of experimental games as a way of testing alternative hypotheses on altruistic motivation, such as empathy or norm compliance (for a review of the existing literature see Mayr, *et al.* 2009).

nomics, or new designs should be considered in neuroeconomics—or both. We have attempted to provide some hints for the direction of possible future research in this field.

* Many thanks to the editor, three anonymous referees, Philip Kitcher, Daniel Kelly, Michel Chapuisat, Benoît Dubreuil, and Chloë FitzGerald for correction, advice, and comments on previous versions of this paper.

References

Andreoni, James (1990), "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving", *The Economic Journal*, 100, pp. 464-77.

Bahry, Donna L. & Wilson, Rick K. (2006), "Confusion or Fairness in the Field? Rejections in the Ultimatum Game under the Strategy Method", *Journal of Economic Behavior & Organization*, 60, pp. 37-54.

Batson, C. Daniel (1991), *The Altruism Question: Toward a Social Psychological Answer*. Hillsdale, N.J.: L. Erlbaum.

Bernhard, Helen, Fischbacher, Urs & Fehr, Ernst (2006), "Parochial Altruism in Humans", *Nature*, 442, pp. 912-15.

Bolton, Gary E. & Ockenfels, Axel (2000), "Erc: A Theory of Equity, Reciprocity, and Competition", *The American Economic Review*, 90, pp. 166-93.

Bowles, Samuel (2008), "Policies Designed for Self-Interested Citizens May Undermine "The Moral Sentiments": Evidence from Economic Experiments", *Science*, 320, pp. 1605-09.

Butler, Joseph (1991), "Fifteen Sermons", *in* D.D. Raphael (eds.), *British Moralists, 1650-1800 : Selected and Edited with Comparative Notes and Analytical Index*. Oxford: Clarendon Press, pp. 325-77.

Camerer, Colin F. (2008), "The Potential of Neuroeconomics", *Economics and Philosophy*, 24, pp. 369-79.

Camerer, Colin, Loewenstein, George & Rabin, Matthew, eds. (2003), *Advances in Behavioral Economics*. The Roundtable Series in Behavioral Economics. New York, N.Y.

Princeton, N.J.: Russell Sage Foundation ;

Princeton University Press.

Carlsmith, Kevin M., Wilson, Timothy D. & Gilbert, Daniel T. (2008), "The Paradoxical Consequences of Revenge", *Journal of Personality & Social Psychology*, 95, pp. 1316-24.

Carpenter, Jeffrey P., Matthews, Peter Hans & Ong'ong'a, Okomboli (2004), "Why Punish? Social Reciprocity and the Enforcement of Prosocial Norms", *Journal of Evolutionary Economics*, 14, pp. 407-29.

Charness, Gary & Dufwenberg, Martin (2006), "Promises and Partnership", *Econometrica*, 74, pp. 1579-601.

Charness, Gary & Gneezy, Uri (2008), "What's in a Name? Anonymity and Social Distance in Dictator and Ultimatum Games", *Journal of Economic Behavior & Organization*, 68, pp. 29-35.

Cherry, Todd L., Frykblom, Peter & Shogren, Jason F. (2002), "Hardnose the Dictator", *American Economic Review*, 92, pp. 1218-21.

Cialdini, Robert B., Schaller, Mark, Houlihan, Donald, Arps, Kevin, Fultz, Jim & Beaman, Arthur L. (1987), "Empathy-Based Helping: Is It Selflessly or Selfishly Motivated?" *Journal of Personality and Social Psychology*, 52, pp. 749-58.

Croson, Rachel & Konow, James (2009), "Social Preferences and Moral Biases", *Journal of Economic Behavior & Organization*, 69, pp. 201-12.

Dana, Jason, Cain, Daylian & Dawes, Robyn (2006), "What You Don't Know Won't Hurt Me: Costly (but Quiet) Exit in Dictator Games", *Organizational Behavior and Human Decision Processes*, 100, pp. 193-201.

Dawes, Christopher T., Fowler, James H., Johnson, Tim, McElreath, Richard & Smirnov, Oleg (2007), "Egalitarian Motives in Humans", *Nature*, 446, pp. 794-96.

de Quervain, Dominique J. F., Fischbacher, Urs, Treyer, Valerie, Schellhammer, Melanie, Schnyder, Ulrich, Buck, Alfred & Fehr, Ernst (2004), "The Neural Basis of Altruistic Punishment", *Science*, 305, pp. 1254-58.

Decety, Jean, Jackson, Philip L, Sommerville, Jessica A, Chaminade, Thierry & Meltzoff, Andrew N (2004), "The Neural Bases of Cooperation and Competition: An Fmri Investigation", *NeuroImage*, 23, pp. 744-51.

Doris, John M., Stich, Stephen P. & Roedder, Erica (forthcoming), "Altruism", (eds.), *The Handbook of Moral Psychology*.

Dubreuil, Benoît (submitted), "Emotions and the Punishment of Normative Violations: A Discussion on Righteous Anger, Indignation, Contempt, and Disgust".

Dufwenberg, Martin & Gneezy, Uri (2000), "Measuring Beliefs in an Experimental Lost Wallet Game", *Games and Economic Behavior*, 30, pp. 163-82.

Eckel, Catherine C., Grossman, Philip J. & Johnston, Rachel M. (2005), "An Experimental Test of the Crowding out Hypothesis", *Journal of Public Economics*, 89, pp. 1543-60.

Falk, Armin, Fehr, Ernst & Fischbacher, Urs (2005), "Driving Forces Behind Informal Sanctions", *Econometrica*, 73, pp. 2017-30.

Fehr, Ernst & Fischbacher, Urs (2005a), *in* H. Gintis, *et al.* (eds.), *Moral Sentiments and Material Interests : The Foundations of Cooperation in Economic Life*. Cambridge, Mass.: MIT Press, pp. 151-91.

Fehr, Ernst & Fischbacher, Urs (2005b), "Human Altruism−Proximate Patterns and Evolutionary Origins", *Analyse & Kritik*, 27, pp. 6-47.

Fehr, Ernst & Fischbacher, Urs (2003), "The Nature of Human Altruism", *Nature*, 425, pp. 785-91.

Fehr, Ernst & Fischbacher, Urs (2004a), "Social Norms and Human Cooperation", *Trends in Cognitive Sciences*, 8, pp. 185-90.

Fehr, Ernst & Fischbacher, Urs (2004b), "Third-Party Punishment and Social Norms", *Evolution and Human Behavior*, 25, pp. 63-87.

Fehr, Ernst & Gächter, Simon (2002), "Altruistic Punishment in Humans", *Nature*, 415, pp. 137-40.
Fehr, Ernst & Gächter, Simon (2005), "Human Behaviour: Egalitarian Motive and Altruistic Punishment (Reply)", *Nature*, 433, pp. E1-E2.
Fehr, Ernst & Rockenbach, Bettina (2003), "Detrimental Effects of Sanctions on Human Altruism", *Nature*, 422, pp. 137-40.
Fehr, Ernst & Schmidt, Klaus M. (1999), "A Theory of Fairness, Competition, and Cooperation", *Quarterly Journal of Economics*, 114, pp. 817-68.
Forsythe, Robert, Horowitz Joel, L., Savin, N. E. & Sefton, Martin (1994), "Fairness in Simple Bargaining Experiments", *Games and Economic Behavior*, 6, pp. 347-69.
Fowler, James H., Johnson, Tim & Smirnov, Oleg (2005), "Human Behaviour: Egalitarian Motive and Altruistic Punishment", *Nature*, 433, pp. E1-E1.
Frohlich, N., Oppenheimer, J. & Bernard Moore, J. (2001), "Some Doubts About Measuring Self-Interest Using Dictator Experiments: The Costs of Anonymity", *Journal of Economic Behavior and Organization*, 46, pp. 271-90.
Gigerenzer, Gerd (2008), *Gut Feelings: Short Cuts to Better Decision Making* (2007). London, New York: Penguin Books.
Gintis, Herbert, Bowles, Samuel, Boyd, Robert & Fehr, Ernst, eds. (2005), *Moral Sentiments and Material Interests : The Foundations of Cooperation in Economic Life*. Cambridge, Mass.: MIT Press.
Guala, Francesco (2005), *The Methodology of Experimental Economics*. Cambridge ; New York: Cambridge University Press.
Güth, Werner, Schmittberger, Rolf & Schwarze, Bernd (1982), "An Experimental Analysis of Ultimatum Bargaining", *Journal of Economic Behavior & Organization*, 3, pp. 367-88.
Haley, Kevin J. & Fessler, Daniel M. T. (2005), "Nobody's Watching? Subtle Cues Affect Generosity in an Anonymous Economic Game", *Evolution and Human Behavior*, 26, pp. 245-56.
Hamilton, William D. (1964), "The Genetical Evolution of Social Behaviour. I & Ii", *Journal of Theoretical Biology*, 7, pp. 1-52.
Hamilton, William D. (1970), "Selfish and Spiteful Behaviour in an Evolutionary Model", *Nature*, 288, pp. 1218-20.
Hamilton, William D. (1975), "Social Aptitudes of Man; an Approach from Evolutionary Genetics", *in* R. Fox (eds.), *Biosocial Anthropology*. New York: Wiley, pp. 133-55.
Harbaugh, William T., Mayr, Ulrich & Burghart, Daniel R. (2007), "Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations", *Science*, 316, pp. 1622-25.
Haselhuhn, Michael P. & Mellers, Barbara A. (2005), "Emotions and Cooperation in Economic Games", *Cognitive Brain Research*, 23, pp. 24-33.
Henrich, Joseph & Henrich, Natalie (2006), "Culture, Evolution and the Puzzle of Human Cooperation", *Cognitive Systems Research*, 7, pp. 220-45.
Henrich, Joseph Patrick (2004), *Foundations of Human Sociality : Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press.
Henson, Richard (2006), "Forward Inference Using Functional Neuroimaging: Dissociations Versus Associations", *Trends in Cognitive Sciences*, 10, pp. 64-69.
Hoffman, Elizabeth, McCabe, Kevin & Vernon, L. Smith (1996), "Social Distance and Other-Regarding Behavior in Dictator Games", *The American Economic Review*, 86, pp. 653-60.
Hutcheson, Francis (2004), *An Inquiry into the Original of Our Ideas of Beauty and Virtue : In Two Treatises* (1725). Natural Law and Enlightenment Classics. Indianapolis, Ind.: Liberty Fund.
Kahneman, Daniel (2003), "Maps of Bounded Rationality: Psychology for Behavioral Economics", *The American Economic Review*, 93, pp. 1449-75.
Kahneman, Daniel, Knetsch, Jack & Thaler, Richard (1986), "Fairness as a Constraint on Profit Seeking: Entitlements in the Market", *The American Economic Review*, 76, pp. 728-41.
King-Casas, Brooks, Tomlin, Damon, Anen, Cedric, Camerer, Colin F., Quartz, Steven R. & Montague, P. Read (2005), "Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange", *Science*, 308, pp. 78-83.
Knutson, Brian (2004), "Behavior: Sweet Revenge?" *Science*, 305, pp. 1246-47.
Koch, Alexander K. & Normann, Hans-Theo (2008), "Giving in Dictator Games: Regard for Others or Regard by Others?" *Southern Economic Journal*, 75, pp. 223-31.
Kurzban, Robert, DeScioli, Peter & O'Brien, Erin (2007), "Audience Effects on Moralistic Punishment", *Evolution and Human Behavior*, 28, pp. 75-84.
LeDoux, Joseph E. (2002), *Synaptic Self: How Our Brains Become Who We Are*. New York: Viking.
Lehmann, Laurent & Keller, Laurent (2006), "The Evolution of Cooperation and Altruism; a General Framework and a Classification of Models", *Journal of Evolutionary Biology*, 19, pp. 1365-76.
Mayr, Ulrich, Harbaugh, William T. & Tankersley, Dharol (2009), "Neuroeconomics of Charitable Giving and Philantropy", *in* P.W. Glimcher, *et al.* (eds.), *Neuroeconomics : Decision Making and the Brain*. Amsterdam ; Boston: Elsevier/Academic Press, pp. 303-20.
McCabe, Kevin (2008), "Neuroeconomics and the Economic Sciences", *Economics and Philosophy*, 24, pp. 345-68.
Nagel, Thomas (1970), *The Possibility of Altruism*. Oxford: Clarendon P.

Oxoby, Robert J. & Spraggon, John Michael (2008), "Mine and Yours: Property Rights in Dictator Games", *Journal of Economic Behavior & Organization*, 65, pp. 703-13.

Poldrack, Russell (2006), "Can Cognitive Processes Be Inferred from Neuroimaging Data?" *Trends in Cognitive Sciences*, 10, pp. 59-63.

Rilling, J. K., Sanfey, A. G., Aronson, Jessica A., Nystrom, L. E. & Cohen, J. D. (2004), "Opposing Bold Responses to Reciprocated and Unreciprocated Altruism in Putative Reward Pathways", *Neuroreport*, 15, pp. 2539-43.

Rilling, James. K., Gutman, David A., Zeh, Thorsten R., Pagnoni, Giuseppe, Berns, Gregory S. & Kilts, Clinton D. (2002), "A Neural Basis for Social Cooperation", *Neuron*, 35, pp. 395-405.

Roth, Alvin E., Prasnikar, Vesna, Okuno-Fujiwara, Masahiro & Zamir, Shmuel (1991), "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study", *The American Economic Review*, 81, pp. 1068-95.

Sanfey, Alan G., Rilling, James K., Aronson, Jessica A., Nystrom, Leigh E. & Cohen, Jonathan D. (2003), "The Neural Basis of Economic Decision-Making in the Ultimatum Game", *Science*, 300, pp. 1755-58.

Smith, Kyle D., Keating, John P. & Stotland, Ezra (1989), "Altruism Revisited: The Effect of Denying Feedback on a Victim's Status to Empathic Witness", *Journal of Personality and Social Psychology*, 57, pp. 641-50.

Sober, Elliott & Wilson, David Sloan (1998), *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, Mass.: Harvard University Press.

Spitzer, Manfred, Fischbacher, Urs, Herrnberger, Bärbel, Grön, Georg & Fehr, Ernst (2007), "The Neural Signature of Social Norm Compliance", 56, pp. 185-96.

Stich, Stephen P. (2007), "Evolution, Altruism and Cognitive Architecture: A Critique of Sober and Wilson's Argument for Psychological Altruism", *Biology and Philosophy*, 22, pp. 267-81.

Stocks, Eric L., Lishner, David A. & Decker, Stephanie K. (forthcoming), Altruism or Psychological Escape: Why Does Empathy Promote Prosocial Behavior? (eds.), *European Journal of Social Psychology*. http://dx.doi.org/10.1002/ejsp.561 (Accessed Accessed).

Tabibnia, Golnaz , Satpute, Ajay B. & Lieberman, Matthew D. (2008), "The Sunny Side of Fairness: Preference for Fairness Activates Reward Circuitry (and Disregarding Unfairness Activates Self-Control Circuitry)", *Psychological Science*, 19, pp. 339-47.

Tabibnia, Golnaz & Lieberman, Matthew D. (2007), "Fairness and Cooperation Are Rewarding", *Annals of the New York Academy of Sciences*, 1118, pp. 90-101.

Vul, Edward, Harris, Christine, Winkielman, Piotr & Pashler, Harold (2009), "Puzzlingly High Correlations in Fmri Studies of Emotion, Personality, and Social Cognition", *Perspectives on Psychological Science*, 4, pp. 274-90.

West, S. A., Griffin, A. S. & Gardner, A. (2007), "Social Semantics: Altruism, Cooperation, Mutualism, Strong Reciprocity and Group Selection", *Journal of Evolutionary Biology*, 20, pp. 415-32.

Yamagishi, Toshio, Horita, Yutaka, Takagishi, Haruto, Shinada, Mizuho, Tanida, Shigehito & Cook, Karen S. (2009), "The Private Rejection of Unfair Offers and Emotional Commitment", *Proceedings of the National Academy of Sciences*, 106, pp. 11520-23.