Chapter 8

THE ECONOMICS OF FAIRNESS, RECIPROCITY AND ALTRUISM – EXPERIMENTAL EVIDENCE AND NEW THEORIES

ERNST FEHR

Institute for Empirical Research in Economics, University of Zurich, Bluemlisalpstrasse 10, CH-8006 Zurich, Switzerland
e-mail: efehr@iew.unizh.ch

KLAUS M. SCHMIDT

Department of Economics, University of Munich, Ludwigstrasse 28, D-80539 Muenchen, Germany
e-mail: klaus.schmidt@Lrz.uni-muenchen.de

Contents

Abstract 2
Keywords 2
1. Introduction and overview 3
2. Empirical foundations of other-regarding preferences 7
  2.1. Other-regarding behavior in simple experiments 7
  2.2. Other-regarding preferences or irrational behavior 14
  2.3. Neuroeconomic foundations of other-regarding preferences 17
3. Theories of other-regarding preferences 22
  3.1. Social preferences 23
    3.1.1. Altruism 24
    3.1.2. Relative income and envy 25
    3.1.3. Inequity aversion 25
    3.1.4. Hybrid models 28
  3.2. Interdependent preferences 30
    3.2.1. Altruism and spitefulness 31
  3.3. Models of intention based reciprocity 33
    3.3.1. Fairness equilibrium 33
    3.3.2. Intentions in sequential games 35
    3.3.3. Merging intentions and social preferences 36
    3.3.4. Guilt aversion and promises 37
  3.4. Axiomatic approaches 38
4. Discriminating between theories of other-regarding preferences 39

Handbook of the Economics of Giving, Altruism and Reciprocity, Volume 1
Edited by Serge-Christophe Kolm and Jean Mercier Ythier
Copyright © 2006 Elsevier B.V. All rights reserved
DOI: 10.1016/S1574-0714(06)01008-6

4.1. Who are the relevant reference actors? 40
4.2. Equality versus efficiency 42
4.3. Revenge versus inequity reduction 46
4.4. Does kindness trigger rewards? 48
4.5. Maximin preferences 50
4.6. Preferences for honesty 52
4.7. Summary and outlook 53
5. Economic applications 55
5.1. Cooperation and collective action 55
5.2. Endogenous formation of cooperative institutions 59
5.3. How fairness, reciprocity and competition interact 62
5.4. Fairness and reciprocity as a source of economic incentives 66
6. Conclusions 69
References 70

Abstract

Most economic models are based on the *self-interest hypothesis* that assumes that material self-interest exclusively motivates *all* people. Experimental economists have gathered overwhelming evidence in recent years, however, that systematically refutes the self-interest hypothesis, suggesting that concerns for altruism, fairness, and reciprocity strongly motivate many people. Moreover, several theoretical papers demonstrate that the observed phenomena can be explained in a rigorous and tractable manner. These theories then induced a first wave of experimental research which offered exciting insights into both the nature of preferences and the relative performance of competing fairness theories. The purpose of this chapter is to review these developments, to point out open questions, and to suggest avenues for future research. We also discuss recent neuroeconomic evidence that is consistent with the view that many people have a taste for mutual cooperation and the punishment of norm violators. We further illustrate the powerful impact of fairness concerns on cooperation, competition, incentives, and contract design.

Keywords

behavioral economics, other-regarding preferences, fairness, reciprocity, altruism, experiments, incentives, contracts, competition

JEL classification: C7, C9, D0, J3

1. Introduction and overview

Many influential economists, including Adam Smith (1759), Gary Becker (1974), Kenneth Arrow (1981), Paul Samuelson (1993) and Amartya Sen (1995), pointed out that people often do care for the well-being of others and that this may have important economic consequences. However, most economists still routinely assume that material self-interest is the *sole* motivation of *all* people. This practice contrasts sharply with a large body of evidence gathered by experimental economists and psychologists during the last two decades. This evidence indicates that a substantial percentage of the people are strongly motivated by other-regarding preferences and that concerns for the well-being of others, for fairness and for reciprocity, cannot be ignored in social interactions. One purpose of this chapter is to review this evidence, suggest how it can be best interpreted, and how it should be modeled. We take up this task in Section 2, where we describe the most important experiments that have radically changed the views of many experimental economists over the last two decades. Section 2 also describes recent neuroeconomic experiments that combine the tools of experimental economics with non-invasive brain imaging methods of modern neuroscience to better understand how the brain generates other-regarding behavior.[1]

In hindsight, it is ironic that experiments have proven to be critical for the discovery and the understanding of other-regarding preferences because experimental economists were firmly convinced for several decades that other-regarding motives only had limited impact. They believed that the self-interest assumption provides a good description for most people's behavior. At best, other-regarding behavior was viewed as a temporary deviation from the strong forces of self-interest. Vernon Smith discovered in the 1950s that experimental markets quickly converge to the competitive equilibrium if subjects trade a homogeneous good and all aspects of the good are fully contractible [Smith (1962)]. Hundreds of experiments have since confirmed the remarkable convergence properties of experimental markets [see Davis and Holt (1993), for example]. The equilibrium in these experiments is computed assuming that *all* players are *exclusively* self-interested. Therefore, the quick convergence to equilibrium was interpreted as a confirmation of the self-interest hypothesis.

However, the bargaining and cooperation experiments described in Section 2 below illustrate that this conclusion was premature because a large percentage of the subjects in these experiments – some of which involve fully representative subject pools for whole countries – exhibit other regarding behavior that the self-interest hypothesis cannot rationalize in any reasonable way. Subjects in these experiments have to make simple decisions in situations where the self-interested choice is salient and easy to understand. Thus, if they deviate from the self-interested choice, we can conclude that they

[1] Readers who are interested in the role of reciprocity and altruism at the workplace and, more generally, in cooperative endeavours, should consult the excellent Handbook Chapters 21 and 22 by Putterman and Rotemberg. Kolm, provides an interesting discussion of the concept of reciprocity that differs from the preference based theories dealt with in our chapter.

exhibit some form of other-regarding preference. Given this evidence, the real question is no longer whether many people have other-regarding preferences, but under which conditions these preferences have important economic and social effects and what the best way to describe and model these preferences is.

However, the evidence from competitive market experiments remains. How can we reconcile the fact that the self-interest model predicts behavior in competitive experimental markets with fully contractible goods very well while it completely fails in the simple experiments described in Section 2 below? Some of the recently developed models of other-regarding preferences that are described and discussed in some detail in Section 3 provide a solution to this puzzle; they show that competition may completely remove the impact of other-regarding preferences. Thus, the fact that we do not observe other-regarding behavior in certain competitive markets does not mean that other-regarding preferences are absent. Instead, rational individuals will not express their other-regarding preferences in these markets because the market makes the achievement of other-regarding goals impossible or infinitely costly. However, a large amount of economic activity takes place outside competitive markets – in markets with a small number of traders, in markets with informational frictions, in firms and organizations, and under contracts which are neither completely specified nor enforceable. Models based on the self-interest assumption frequently make very misleading predictions in these environments, while models of other-regarding preferences predict much better. These models thus provide fresh and experimentally confirmed insights into important phenomena like the persistence of non-competitive wage premiums, the incompleteness of contracts and the absence of explicit incentive schemes, the allocation of property rights, the conditions for successful collective action, and the optimal design of institutions.

One of the exciting aspects of this development is that the newly developed theories of other-regarding preferences were tested in a new wave of experiments, sometimes before they were even published. This led to important insights into the power and the limits of different models which will be discussed in Section 4. These experiments also show that it is possible to discriminate between different motivational assumptions, answering one important objection to this research program. There has always been a strong convention in economics of not explaining puzzling observations by changing assumptions on preferences. Changing preferences is said to open Pandora's Box because everything can be explained by assuming the "right" preferences. We believe that this convention made sense in the past when economists did not have the tools to examine the nature of preferences in a scientifically rigorous way. However, due to the development of experimental techniques these tools are now available. In fact, one purpose of this paper is to show that the past decade has yielded both progress on and fascinating new insights into the nature of other regarding preferences.

While many people are strongly concerned about others' well-being, fairness, and reciprocity, we consider it equally important to stress that the available experimental evidence suggests that there are also many subjects who behave quite selfishly even when they are given a chance to affect other people's well-being at a relatively small cost. One of the exciting insights of some of the newly developed theoretical models is that the interaction between fair and selfish individuals is key to understanding the observed behavior in strategic settings. These models explain why almost all people behave as if they are completely selfish in some strategic settings, while the same people will behave as if driven by fairness in others.

We describe several examples that show the economic importance of other-regarding preferences in different settings in the final part of the paper, Section 5. Among other things, we provide evidence indicating that other-regarding preferences are decisive for explaining collective action and multi-lateral cooperation. We present, in particular, recent evidence showing that if individuals can choose between an institution allowing mutual punishment of non-cooperative behavior or one which rules out mutual punishment, they converge to a behavioral equilibrium in which the selfish and the other-regarding types unanimously prefer the punishment institution. Moreover, punishment of free riders actually occurs and drives the behavior in the punishment institution towards a state in which full cooperation and no punishment occurs. The threat of punishment alone suffices to generate full cooperation. This experiment constitutes a powerful example suggesting that other-regarding preferences have shaped many of our cooperative institutions. In addition, we document that other-regarding preferences have deep effects on outcomes in markets with moral hazard problems, while the interaction between selfish and fair-minded subjects in markets with fully contractible goods generates outcomes that are close to the competitive prediction. Finally, we report how other-regarding preferences influence voting behavior in taxation games. These examples, although important, provide only a glimpse into the full range of possibilities how other-regarding preferences shape social and economic interactions including, perhaps, some of our most fundamental institutions. The examples also show that the main reason why other-regarding preferences are important lies in the fact that even a minority of other-regarding people may generate powerful cooperation incentives for selfish people.

To set the stage for the discussion of the following sections we give an informal and intuitive definition of several types of other-regarding preferences that received a lot of attention in the recent literature that tries to explain behavior in economic experiments. In Section 3 we define these preferences in a formal and more rigorous way. The theoretical literature on other-regarding preferences has focused on three departures from the standard self-interest model. In addition to the material resources allocated to him a person may also care about: (i) The material resources allocated to other agents in a relevant reference group. (ii) The fairness of the behavior of relevant reference agents. (iii) The "type" of the reference agents, i.e. whether the agents have selfish, altruistic, spiteful, or fair minded preferences.

Consider first the case where the utility function of an individual also depends on the material resources that other agents in a relevant reference group receive. A typical example is *altruism*. Altruism is a form of *un*conditional kindness; that is, a favor given does not emerge as a response to a favor received [Andreoni (1989), Andreoni and Miller (2002), Cox, Sadiraj and Sadiraj (2001), Charness and Rabin (2002)]. In technical terms, altruism means that the first derivate of the utility function of an individual with respect to the material resources received by any other agent is always strictly positive. Thus, an altruist is willing to sacrifice own resources in order to improve the well being of others. The opposite case is *envy* or *spitefulness*. A spiteful person *always* values the material payoff of relevant reference agents negatively. Such a person is, therefore, always willing to decrease the material payoff of a reference agent at a personal cost to himself [Bolton (1991), Kirchsteiger (1994), Mui Vai-Lam (1995)] irrespective of both the payoff distribution and the reference agent's fair or unfair behavior. Therefore, spiteful preferences represent the antisocial version of other-regarding preferences. A conditional form of altruism and/or envy is *inequity aversion* [Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Charness and Rabin (2002)]. An individual is inequity averse if, in addition to his material self-interest, his utility increases if the allocation of material payoffs becomes more equitable. Thus, an inequity averse person may value additional material resources allocated to a reference agent positively or negatively, depending on whether the allocation becomes more or less equitable. Obviously, the definition of equity is very important in these models. In the context of experimental games equity is usually defined as equality of monetary payoffs. However, departures from equality have been defined differently. They can be measured in terms of the income differences between the individual and all relevant reference agents, or in terms of the difference between the individual and the least well-off in his reference group, or in terms of the individual's relative share of the overall surplus.

The case where preferences depend on the fair or unfair *behavior* of other agents has also received much attention in the literature and is often called reciprocity. A reciprocal individual, as we define it here, responds to actions he perceives to be kind in a kind manner, and to actions he perceives to be hostile in a hostile manner [Rabin (1993), Segal and Sobel (2004), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2005)]. Thus, preferences do not only depend on material payoffs but also on intentions, i.e. on beliefs about why an agent has chosen a certain action. This cannot be modeled by using conventional game theory but requires the tools of psychological game theory [Geanakoplos, Pearce and Stacchetti (1989)].

Finally, preferences may depend on the type of opponent [Levine (1998)]. According to type-based reciprocity, an individual behaves kindly towards a "good" person (i.e, a person with kind or altruistic preferences) and hostilely towards a "bad" person (i.e. a person with unkind or spiteful preferences). Note that it is the "type" of a person and not the "intention" of his action that affects preferences in this case. Therefore, type-based reciprocity can be modeled using conventional game theory.

It is important to emphasize that it is not the expectation of future material benefits that drives reciprocity. Reciprocal behavior as defined above differs fundamentally from "cooperative" or "retaliatory" behavior in repeated interactions that is motivated by future material benefits. Therefore, reciprocal behavior in one-shot interactions is often called "strong reciprocity" in contrast to "weak reciprocity" that is motivated by long-term self-interest in repeated interactions [Gintis (2000), Fehr and Fischbacher (2003)].

Readers who are mainly interested in the experimental evidence that documents the existence of other-regarding preferences should first consult Section 2 and then Section 4 of this chapter. In Section 2, we present a list of simple experiments that indicate the existence and the prevailing patterns of other-regarding preferences. In Section 4, we discuss the most recent evidence in the light of the newly developed models of other-regarding preferences. Readers who are mainly interested in the different models of other-regarding preferences and how they perform relative to the available evidence can directly jump to Section 3 and Section 4. Finally those readers who are mainly interested in the economic impact of other-regarding preferences may directly jump to Section 5.

2. Empirical foundations of other-regarding preferences

2.1. Other-regarding behavior in simple experiments

In the introduction, we referred to the previously held belief of many experimental economists in the validity of the self-interest hypothesis. This "commitment" to the self-interest hypothesis slowly weakened in the 1980s, when experimental economists started studying bilateral bargaining games and interactions in small groups in controlled laboratory settings [see, e.g., Roth, Malouf and Murningham (1981), Güth, Schmittberger and Schwarze (1982)]. One of the important experimental games that eventually led many people to realize that the self-interest hypothesis is problematic was the so-called "ultimatum game" by Güth, Schmittberger and Schwarze (1982). In addition, games like the "dictator game", the "power to take game", the "third party punishment game", the "gift exchange game" and the "trust game" played an important role in weakening the exclusive reliance on the self-interest hypothesis. All these games share the feature of simplicity, enabling the experimental subjects to understand them and therefore making inferences about subjects' motives more convincing. In fact, in all these games one player has a strictly dominant strategy if he is self-interested and this selfish strategy is salient and easy to understand in all cases. Therefore, if this player does not choose his or her selfish strategy, we can infer that he deliberately did not do so, i.e., we can make inferences about his motives.

In the ultimatum game, a pair of subjects has to agree on the division of a fixed sum of money. Person A, the proposer, can make one proposal of how to divide the amount. Person B, the responder, can accept or reject the proposed division. In case of rejection, both receive nothing; in case of acceptance, the proposal is implemented. Under the standard assumptions that (i) both the proposer and the responder are rational *and* care only about how much money they get and (ii) that the proposer knows that the responder is rational and selfish, the subgame perfect equilibrium prescribes a rather extreme outcome: the responder accepts *any* positive amount of money and, hence, the proposer gives the responder the smallest money unit, ε, and keeps the rest to himself.

A robust result in the ultimatum game, across hundreds of experiments, is that the vast majority of the offers to the responder are between 40 and 50 percent of the available surplus. Moreover, proposals offering the responder less than 20 percent of the surplus are rejected with probability 0.4 to 0.6. In addition, the probability of rejection is decreasing in the size of the offer [see, e.g., Güth, Schmittberger and Schwarze (1982), Camerer and Thaler (1995), Roth (1995), Camerer (2003) and the references therein]. Apparently, many responders do not behave in a self-interest maximizing manner. In general, the motive indicated for the rejection of positive, yet "low", offers is that subjects view them as unfair. A further robust result is that many proposers seem to anticipate that low offers will be rejected with a high probability. A comparison of the results of dictator games and ultimatum games suggests this. The responder's option to reject is removed in a dictator game; the responder must accept any proposal. Forsythe et al. (1994) were the first to compare the offers in ultimatum and dictator games. Self-interested proposers should allocate nothing to the Recipient in the dictator game. In experiments, proposers typically dictate allocations that assign the Recipient on average between 10 and 25 percent of the surplus, with modal allocations at 50 percent and zero. These allocations are much less than proposers' offers in ultimatum games, although most players do offer something. Comparing dictator with bilateral ultimatum games shows that fear of rejection is *part* of the explanation for proposers' generous offers, because they do offer less when rejection is precluded. But many subjects offer something in the dictator game, so fear of rejection is not the entire explanation. The considerably lower offers in the dictator game suggest that many proposers apply backwards induction. This interpretation is also supported by the surprising observation of Roth et al. (1991), who showed that the modal offer in the ultimatum game tends to maximize the proposer's expected income.[2]

The "power to take game", invented by Bosman and van Winden (2002), is another tool that has proven useful in understanding punishment behavior. Both the proposer and the responder are endowed with some income in this game. Subjects may have earned this income, as in Bosman and van Winden (2002), or the experimenter may have allocated the money to the subjects as in Bosman, Sutter and van Winden (2005). The proposer can set a take or "theft" rate $t \in [0, 1]$ which is the fraction of the responder's endowment that will be transferred to the proposer. The responder is then informed of the take rate and can destroy part or all of his income. Thus, if the responder destroys his or her whole income nothing is transferred to the proposer. If the responder destroys only a fraction d, $d \in [0, 1]$, of his income, the proposer receives a share of $t(1 - d)$ of the responder's pre-destruction income. In contrast to the ultimatum game, the power to take game allows the punishment behavior to vary continuously with the take rate. The evidence indicates that the destruction rate is roughly $d = 0.5$ for take rates

[2] Suleiman (1996) reports the results of ultimatum games with varying degrees of veto power. In these games a rejection meant that λ percent of the cake was destroyed. For example, if $\lambda = 0.8$, and the proposer offered a $9 : 1$ division of \$10, a rejection implied that the proposer received \$1.8 while the responder received \$0.2. Suleiman reports that proposers' offers are strongly increasing in λ.

around $t = 0.8$, regardless of whether the initial endowment was earned through effort or exogenously allocated by the experimenter. However, the destruction rate is higher for lower take rates if the initial endowment is given to the subjects without effort, whereas the destruction rate is higher for takes rates above 0.8 if the endowment was earned through effort. This indicates that the way the initial endowment is allocated to the subjects matters because it seems to affect their feelings of entitlement. Hoffman, McCabe and Smith (1996b) also reported that feelings of entitlement may be important for punishment behavior in the context of the ultimatum game.

The responders' feelings may be hurt if he or she receives an unfairly low offer in the ultimatum game. Thus, pride or motives to retain self-respect may drive a rejection. Therefore, the question arises whether people would also be willing to punish violations of social or moral norms if they themselves are not the victim of the norm violation. A game that is particularly suited to examine this question is the so-called third party punishment Game [Fehr and Fischbacher (2004)]. The three players in this game are denoted A, B, and C. A and B play a simple dictator game. Player A, the proposer, receives an endowment of S tokens of which he can transfer any amount to player B, the Recipient. B has no endowment and no choice to make. Player C has an endowment of $S/2$ tokens and observes player A's transfer. Player C can then assign punishment points to player A. Player C incurs costs of 1 token and player A is charged 3 tokens for each punishment point player C assigns to player A. Since punishment is costly, a self-interested player C will never punish. However, if there is a sharing norm, player C may well punish player A if A gives too little.

In fact, in the experiments conducted by Fehr and Fischbacher (2004), where $S = 100$, player A was rarely punished if he transferred 50 or more tokens to player B. If he transferred less than 50 tokens, roughly 60 percent of players C punished A and the less A transferred, the stronger was the punishment. If nothing was transferred, A received on average 14 punishment points, reducing A's income by 42 tokens. Thus, if nothing was transferred player A earned (on average) more money in this setting than if he transferred the fair amount of 50. However, if player C was himself the recipient in another dictator game unrelated to that played between A and B, C punished more. All transfer levels below 50 were on average punished so strongly in this case that it was no longer in player A's self-interest to transfer less than 50. It seems that if C is himself a recipient, he is more able to empathize with B if B receives little and thus increase the punishment imposed on A. Finally, if third party punishment is compared to second party punishment (i.e. if B can punish A), it turns out that second party punishment is significantly stronger than is third party punishment. Note that this does not necessarily mean that third party punishment is less effective in sustaining social norms because third parties are often more numerous than second parties.

Dictator games measure pure altruism. Interesting companion games are the trust game [Berg, Dickhaut and McCabe (1995)] and the gift exchange game [Fehr, Kirchsteiger and Riedl (1993)]. In a trust game, both an Investor and a Trustee receive an amount of money S from the experimenter. The Investor can send between zero and S to the Trustee. The experimenter then triples the amount sent, which we term y, so that the Trustee has $S + 3y$. The Trustee is then free to return anything between zero and $S + 3y$ to the Investor. The Investor's payoff is $S - y + z$ and that of the Trustee is $S + 3y - z$ where z denotes the final transfer from the Trustee to the Investor. The trust game is essentially a dictator game in which the Trustee dictates an allocation, with the difference, however, that the Investor's initial investment determines the amount to be shared.

In theory, self-interested Trustees will keep everything and repay $z = 0$. Self-interested Investors who anticipate this should transfer nothing, i.e., $y = 0$. In experiments in several developed countries, Investors typically invest about half the maximum on average, although there is substantial variation across subjects. Trustees tend to repay roughly y so that trust is not or only slightly profitable. The amount Trustees repay increases on average with y if the change in the Investors' transfer is sufficiently high; the Trustees do not necessarily pay back more if the increase in y is modest.

In the gift exchange game, there is again a proposer and a responder. The proposer offers an amount of money $w \in [\underline{w}, \overline{w}]$, $\underline{w} \geqslant 0$, which can be interpreted as a wage payment, to the responder. The responder can accept or reject w. In case of a rejection, both players receive zero payoff; in case of acceptance, the responder has to make a costly "effort" choice $e \in [\underline{e}, \bar{e}]$, $\underline{e} > 0$. A higher effort level increases the proposer's monetary payoff but is costly to the responder. A selfish responder will always choose the lowest feasible effort level $\underline{e}$ and will, in equilibrium, never reject any w. Therefore, if the proposer is selfish and anticipates the responder's selfishness the subgame perfect proposal is the lowest feasible wage level $\underline{w}$. The main difference between the gift exchange game and the trust game is that in the trust game it is the first mover's action that increases the available surplus, while in the gift exchange game it is the second mover who can increase the surplus.

The gift exchange game captures a principal-agent relation with highly incomplete contracts in a stylized way. Several authors have conducted variants of the gift exchange game.[3] All of these studies report that the mean effort is, in general, positively related to the offered wage which is consistent with the interpretation that the responders, on average, reward generous wage offers with generous effort choices. However, as in the case of the ultimatum and the trust game, there are considerable individual differences among the responders. While a sizeable share of responders (frequently roughly 40 percent, sometimes more than 50 percent) typically exhibit a reciprocal effort pattern, a substantial fraction of responders also always make purely selfish effort choices or choices which seem to deviate randomly from the self-interested action. Despite the presence of selfish responders, the relation between average effort and wages can be sufficiently steep to render a high wage policy profitable which may induce proposers to pay wages far above $\underline{w}$. Evidence for this interpretation comes from Fehr, Kirchsteiger

[3] See, e.g., Fehr, Kirchsteiger and Riedl (1993, 1998), Charness (1996, 2000), Fehr and Falk (1999), Gächter and Falk (1999), Falk, Gächter and Kovács (1999), Hannan, Kagel and Moser (1999), Brandts and Charness (2004) and Fehr, Klein and Schmidt (2004).

and Riedl (1998), who embedded the gift exchange game into an experimental market.[4] In addition, there was a control condition where the experimenter exogenously fixed the effort level. Note that the responders can no longer reward generous wages with high effort levels in the control condition. It turns out that the average wage is substantially reduced when the effort is exogenously fixed.

The facts observed in the games mentioned above are now well established and there is little disagreement about them. However, questions remain about which factors determine and change the behavior in these games. For example, a routine question in discussions is whether a rise in the stake level will eventually induce subjects to behave in a self-interested manner. Several papers examine this question [Hoffman, McCabe and Smith (1996a), Fehr and Tougareva (1995), Slonim and Roth (1997), Cameron (1999)]; the surprising answer is that relatively large increases in the monetary stakes did little or nothing to change behavior. Hoffman, McCabe and Smith (1996a) could not detect any effect of the stake level in the ultimatum game. Cameron (1999) conducted ultimatum games in Indonesia and subjects in the high stake condition could earn the equivalent of three months' income in this experiment. She observed no effect of the stake level on proposers' behavior and a slight reduction of the rejection probability when stakes were high. Slonim and Roth (1997) conducted ultimatum games in Slovakia. They found a small interaction effect between experience and the stake level; the responders in the high-stake condition (with a 10-fold increase in the stake level relative to the low stake condition) reject somewhat less frequently in the final period of a series of one-shot ultimatum games. Fehr and Tougareva (1995) conducted gift exchange games (embedded in a competitive experimental market) in Moscow. They did not observe an interaction effect between stake levels and experience. The subjects earned, on average, the equivalent amount of the income of one week in one of their conditions, while they earned the equivalent of a ten weeks' income in another condition. Despite this large difference in the stake size, neither the proposers' nor the responders' behavior shows significant differences across conditions.

Of course, it is still possible that there may be a shift towards more selfish behavior in the presence of extremely high stakes. However, the vast majority of economic decisions for most people involve stake levels well below three months' income. Thus, even if other-regarding preferences played no role at all at stake levels above that size, these preferences would still play a major role in many economically important domains.

[4] When interpreting the results of gift exchange games it is important to stress that – depending on the concrete form of the proposer's payoff function – gift exchange is more or less likely to be profitable for the proposer. In Fehr, Kirchsteiger and Riedl (1993, 1998), the proposer's payoff function is given by $x^P = (v - w)e$ and effort is in the interval [0.1, 1]. With this payoff function the proposer cannot make losses and paying a high wage is less costly if the agent chooses a low effort level. In contrast, in Fehr, Klein and Schmidt (2004) the payoff function used is $x^P = ve - w$ which makes it more risky for the principal to offer a high wage. Indeed, while paying high wages was profitable for the principal in the experiments of Fehr, Kirchsteiger and Riedl, it did not pay off in Fehr, Klein and Schmidt. This difference in performance is predicted by the theory of inequity aversion by Fehr and Schmidt (1999) that is discussed in more detail in Section 3. For a further discussion of gift exchange games in competitive environments see also Section 5.3.

Another important question is to what degree the behavior of students is respresentative for the general population. All the experiments mentioned above were predominantly conducted with students as experimental subjects. Two representative data sets recently addressed this question – one from Germany [Fehr et al. (2002)] and one from the Netherlands [Bellemare and Kröger (2003)]. In both cases, the authors conducted (modified) trust games and in both cases, certain demographic variables affected how the game is played, but these effects do not change the general pattern observed in the experiments with students. In particular, the trustees' back transfers are increasing in the investors' transfer and a large share (79 percent in the Fehr et al. study) of the trustees pays back money. Likewise, 83 percent of the investors transfer positive amounts; roughly 60 percent of them transfer 50% or more of their endowment. Moreover, the proposers' and responders' behavior remains constant, regardless of whether the players' endowment in the trust game is € 10 or € 100.

Among the demographic variables, age seems to be important. Both studies find that people above the age of 60 give less than middle-aged individuals when in the role of an investor. However, both studies also find that the elderly tend to give back more, ceteris paribus, when in the role of a trustee. Fehr et al. also report that subjects who experienced a divorce from their partner during the last year and people who favor none of the parliamentary parties in Germany (i.e. those who feel that they are not represented by the major political parties) pay back significantly less when in the role of a trustee. Furthermore, people who report that they are in good health give back significantly more. The most important result these studies provide, however, is that only very few individual level demographic variables seem to matter for behavior. This suggests that it is possible to detect meaningful behavioral patterns with student subject pools that are representative for a more general subject pool, at least for the trust game.

To what extent does culture affect behavior in these experiments? We define culture in terms of subjects' preferences and their beliefs about others' behavior. For example, in the context of the ultimatum game cultural differences may be reflected in different rejection rates for the same offer or in different beliefs about the rejection rate. In the past, many researchers took subjects' nationality as a proxy for culture. Nationality may be a very imperfect measure for culture in modern nations, however, because different cultures may coexist within the same country. Cohen and Nisbett (1994) provide evidence, for example, indicating that individuals who grew up in the American South have a culture of honour whereas Northerners do not have such a culture. Having said this, comparing subjects' behavior across different continents may nevertheless yield interesting insights. Roth et al. conducted ultimatum games in Japan, Israel, Slovenia, and the USA. Their results indicate somewhat lower rejection rates and lower offers in Japan and Israel compared to the US and Slovenia. Whereas the modal offers remain at 50% of the surplus throughout a ten period experiment with randomly assigned partners in the latter two countries, the modal offer converges to 40% in Israel and to two modes in Japan at 40% and 45%, respectively. The relatively low offers in Israel are also associated with relatively low rejection rates, indicating that a lower proposal in Israel was a rational choice for a self-interested proposer.

Buchan, Croson and Dawes (2002) conducted trust games in China, Japan, South Korea, and the USA. They find significant differences in investors' and in trustees' behavior across countries. American and Chinese Investors transfer significantly more money than do their Japanese and Korean counterparts. Moreover, Chinese and Korean trustees send back a significantly higher proportion of their money than do American and Japanese subjects. Thus, Chinese subjects exhibit relatively high levels of trust (as indicator by investors' behavior) and reciprocation (as indicated by trustees' behavior) whereas Japanese subjects show relatively little trust and little reciprocation. The picture is more mixed for US and Korean subjects. Americans show a relatively high level of trust but a low level of reciprocation, whereas the Koreans show little trust but exhibit high levels of reciprocation.

The study by Henrich et al. (2001) and Henrich et al. (2004) documented the perhaps largest differences across cultures. This study reports the results of ultimatum game experiments conducted in 15 small scale societies located in 5 different continents. The subjects in the cross cultural studies previously discussed were university students; one could therefore argue that, despite national differences, they all share much in common. They probably all have above-average skills, probably stem from higher income families and, perhaps most importantly, share an academic learning environment. This provides a sharp contrast to the Henrich et al. study, where subjects come from vastly different cultures. For example, the Ache from Paraguay practice extreme forms of egalitarianism in which big game is shared equally among the tribe members. Others, like the Au and the Gnau from Papua New Guinea obey norms of competitive gift giving: accepting gifts, even unsolicited ones, obliges one to reciprocate at some future time to be determined by the giver. Acceptance of gifts also establishes a subordinate position between the giver and the receiver. Therefore, large gifts are frequently rejected in this society because of the fear associated with the unspecific commitments.

Henrich et al. observe vastly different proposer behavior across cultures. For example, among the Machiguenga, who live in Peru, the average offer is only 26%, among the Gnau it is 38%, among the Ache it is 51%, while it even reaches 58% among the Lamelara, who are whale hunters on an Island in the Pacific Ocean. Likewise, there are also strong differences regarding rejection rates across several cultures. However, since most offers were around 50% in several societies, few rejections are observed, rendering the analysis of rejection behavior impossible in these societies. Similar to the two representative studies in Germany and the Netherlands, only few, if any, individual level variables predict individual behavior in the experiment. Two group level variables, however, explain a large share of the cross cultural variation in behavior: the more the resources in a society are acquired through market trading and the higher the potential payoffs to group cooperation that are associated with the environment in which the society lives, the higher are the offers in the ultimatum game. For example, groups of 20 and more individuals have to cooperate in order to catch a whale and after the catch, they have to solve a difficult distribution problem: who gets which part of the whale. The Lamaleras have developed an extremely elaborate set of norms that determine in detail who gets what [Alvard (2004)]. These elaborate cooperation and distribution practices may well spill over to the experimental context and induce subjects to make egalitarian offers. In contrast to the Lamelara, the Machiguenga in Peru exhibit little cooperation in production outside narrow family boundaries [Henrich and Smith (2004)]. They are also at the lower end of the spectrum with regard to market integration. It seems plausible that the absence of cooperation norms manifests itself in low offers in the ultimatum game. A third piece of telling evidence comes from the competitive gift giving societies in Papua New Guinea. Among the Au and the Gnau, a significant number of proposers offered *more* than 50% of the surplus, only to have these offers rejected in many cases. Thus, deeply seated social norms again seem to affect behavior in the experiment.

2.2. Other-regarding preferences or irrational behavior

While there is now little disagreement regarding the facts reported above, there is still some disagreement about their interpretation. In Section 3, we will describe several recently developed theories of altruism, fairness, and reciprocity that maintain the rationality assumption but change the assumption of purely selfish preferences. Although opinions about the relative importance of different motives behind other-regarding behavior differ somewhat (see Section 4), it is probably fair to say that most experimental researchers believe that some form of other-regarding preferences exists. However, some interpret the behavior in these games as elementary forms of bounded rationality. For example, Roth and Erev (1995) and Binmore, Gale and Samuelson (1995) try to explain the presence of fair offers and rejections of low offers in the ultimatum game with learning models that are based on purely pecuniary preferences, which assume that the rejection of low offers is not very costly for the responders who therefore only learn very slowly not to reject such offers. The rejection of offers, however, is quite costly for the proposers, who thus quickly realize that low offers are not profitable. Moreover, since proposers quickly learn to make fair offers, the pressure on the responders to learn to accept low offers is greatly reduced. This gives rise to very slow convergence to the subgame perfect equilibrium – if there is convergence at all. The simulations of Roth and Erev and Binmore, Gale and Samuelson show that it often takes thousands of iterations until play comes close to the standard prediction.

In our view, there can be little doubt that learning processes are important in real life as well as in laboratory experiments. There are numerous examples where subjects' behavior changes over time and it seems clear that learning models are prime candidates for explaining such dynamic patterns. We believe, however, that attempts to explain the basic facts in simple games, such as the ultimatum game, the third party punishment game, or the trust game, in terms of learning models that assume completely selfish preferences are misplaced. The responders' decisions, in particular, are so simple in these games that it is difficult to believe that they make systematic mistakes and reject money or reward generous offers, even though their true preferences would require them not to do so. Moreover, the above cited evidence from Roth et al. (1991), Forsythe et al. (1994), Suleiman (1996) and Fehr, Kirchsteiger and Riedl (1998) suggests that many proposers anticipate responders' actions surprisingly well. Thus, at least in these simple two-stage games, many proposers seem to be quite rational and forward looking.

It is also sometimes argued that the behavior in these games is due to a social norm [see, Binmore (1998), for example]. In real life, so the argument goes, experimental subjects make the bulk of their decisions in repeated interactions. It is well known that the rejection of unfair offers or the rewarding of generous offers in repeated interactions can be sustained as an equilibrium among purely self-interested agents. According to this argument, subjects' behavior is adapted to repeated interactions and they tend to apply behavioral rules that are appropriate in the context of repeated interactions *erroneously* to laboratory one-shot games.

We believe that this argument is half right and half wrong. The evidence from the cross-cultural experiments in 15 different small scale societies strongly suggests that social norms of cooperation and sharing have an impact on game playing behavior. Indeed, the very fact that the behavior in the experiment captures relevant aspects of real life behavior is the main reason why such experiments are interesting; if they did not tell us something about how people behave in real life, the external validity of the experiments could be called into question. However, the fact that social norms affect subjects' behavior in the experiment does not at all mean that they are inappropriately applying repeated game heuristics when they play one-shot games. In fact, the evidence suggests that subjects are well aware of the difference between one-shot interactions and repeated interactions where their reputation is at stake. Subjects in the experiments by Andreoni and Miller (1993), Engelmann and Fischbacher (2002), Gächter and Falk (2002), Fehr and Fischbacher (2003), Seinen and Schram (2006) exhibit much more cooperative behavior or punish much more if the probability of repeatedly meeting the same subject increases or if they can acquire a reputation.

Fehr and Fischbacher (2003), for example, conducted a series of ten ultimatum games in two different conditions. Subjects played against a different opponent in each of the ten iterations of the game in both conditions. The proposers knew nothing about the past behavior of their current responders in each iteration of the *baseline condition*. Thus, the responders could not build up a reputation for being "tough" in this condition. In contrast, the proposers knew the full history of their current responders' behavior in the *reputation condition*, i.e., the responders could build up a reputation for being "tough". A reputation for rejecting low offers is, of course, valuable in the reputation condition because it increases the likelihood of receiving high offers from the proposers in future periods.

Therefore, if the responders understand that there is a pecuniary payoff from rejecting low offers in the reputation condition, one should generally observe higher acceptance thresholds in this condition. This is the prediction of an approach that assumes that subjects are rational and not only care for their own material payoff but also have a preference for punishing unfair offers: only the punishment motive plays a role in the baseline condition, while the punishment motive and the self interest motive influence rejection behavior in the reputation condition. If, in contrast, subjects do not understand the logic of reputation formation and apply the same habits or cognitive heuristics to both conditions, there should be no observable systematic differences in responder behavior across conditions. Since the subjects participated in both conditions, it was possible to observe behavioral changes at the individual level. It turns out that the vast majority (slightly more than 80 percent, $N = 72$) of the responders *increase* their acceptance thresholds in the reputation condition relative to the baseline condition.[5] Moreover, the changes in rejection behavior occur almost instantaneously when subjects move from the baseline condition to the reputation condition or vice versa. Thus, the data refutes the hypothesis that subjects do not understand the strategic differences between one-shot play and repeated play.

Therefore, instead of assuming that simple decisions that deviate systematically from self-interest reflect merely a form of erroneous application of rules of thumb, it seems more reasonable to assume that the prevailing social norms affect subjects' preferences. After all, the elaborate cooperation and distribution norms practiced by the whale hunters in Indonesia, or the gift giving norms among the Au and the Gnau in Papua New Guinea have been in place for decades if not centuries. They represent deep seated social practices that are likely to affect subjects' preferences. As these social practices are rather stable, the associated preferences inherit this stability. If a subject rejects a low offer in an anonymous one-shot ultimatum game because he or she is upset by the offer, the subject's emotional reaction to the situation probably drives the behavior. Anger, after all, is a basic emotion and the prevailing fairness norms are likely to be reflected in the emotional response to a greedy offer. Recent papers by Fehr and Gächter (2002), Bosman and van Winden (2002) and Ben-Shakhar et al. (2004) provide evidence for the involvement of anger in punishment behavior.

The view that emotions are important determinants of other-regarding behaviors, however, does not imply that these behaviors are irrational. If I feel bad if I let a greedy proposer go unpunished, and if punishing him makes me feel good, I simply have a taste for punishing a greedy proposer. From a choice theoretic viewpoint, this taste does not differ from my taste for chocolate or lobster. In fact, there is strong experimental evidence suggesting that the demand for altruistic giving and for punishment increases if its price decreases [Eckel and Grossman (1996), Andreoni and Vesterlund (2001) in QJE, Anderson and Putterman (2006)]. In addition, evidence from dictator games [Andreoni and Miller (2002)] also shows that most subjects' preferences for giving in a dictator game obey the generalized axiom of revealed preferences, implying that the preferences can be represented by a utility function. Finally, Andreoni, Castillo and Petrie (2003) have shown that the responder's behavior in a modified ultimatum game, in which the responder could shrink the available pie continuously, can be represented by convex fairness preferences.

[5] The remaining subjects, with one exception, exhibit no significant change in the acceptance threshold. Only one out of 70 subjects exhibits a significant decrease in the threshold relative to the baseline. Note that if a subject places a very high value on fairness, the acceptance threshold may already be very high in the baseline condition so that there is little reason to change the threshold in the reputation condition. Identical thresholds across conditions are, therefore, also compatible with a social preference approach. Only a decrease in the acceptance threshold is incompatible with theories of social preferences.

The above arguments suggest that there is no reason for treating other-regarding preferences differently than other types of preferences. This means that we can apply the standard tools of economics and game theory to this area, enabling us to explain a great deal of behavior in the games described above. For example, why do in Forsythe et al. (1994) the proposers give so much less in the DG compared to the UG? Why do the proposers in the control condition with exogenously fixed effort [Fehr, Kirchsteiger and Riedl (1998)] make such low wage offers? Why do subjects punish less if the price of punishing is higher? Why do subjects reject higher offers if they can gain a reputation for being a tough bargainer compared to a situation where no reputation can be acquired? All these questions can be answered if one assumes that subjects are rational and care both for their own *and* others' payoffs. The problem with the alternative approach, which invokes some form of bounded rationality, is that at least so far it cannot explain these important behavioral variations across different games.

Most of the experiments that we consider in the rest of this paper are fairly simple. Therefore, we restrict attention in the following to approaches that maintain the assumption of rationality and ignore the potential role of learning.[6]

2.3. Neuroeconomic foundations of other-regarding preferences

Recently, some experimental economists and psychologists have begun combining non-invasive brain imaging techniques with behavioral experiments. Brain imaging techniques like Positron Emission Tomography (PET) and functional Magnetic Resonance Imaging (fMRI) enable researchers to examine the brain networks involved in decision making. This means, for example, that subjects' empathic feelings for others are not limited to measurement by self-reports or by making inferences about their motives from observed behavior, but are also possible in terms of brain activity. Likewise, if it is true that subjects derive utility from punishing others for behaving unfairly or from mutual cooperation in a trust game, the researcher should find traces of these hedonic rewards by examining the activity in the brain's reward network. Note that this kind of brain evidence may also help discriminate between an approach that assumes that other-regarding motives drives other-regarding behavior and one that assumes that subjects simply do not understand the differences between one-shot games and repeated interactions. If the first approach is correct, we should observe hedonic responses in reward related brain areas when subjects cooperate or punish others for violations of widely accepted social norms. An approach that assumes that subjects are selfish but confuse one-shot with repeated interactions predicts no such activation. In the following we describe several studies which suggest that subjects indeed experience positive hedonic responses when they cooperate or punish norm violators. Some of the studies also indicate that subjects suffer themselves merely by observing others in distress.

[6] There are a few models that combine other regarding preferences and learning, e.g. Cooper and Stockman (1999) and Costa-Gomes and Zauner (1999).

Singer et al. (2004a) recently published an intriguing paper on the neural basis of empathy for pain in others. The study of empathy is insofar important as empathic concern for others is likely to be an important determinant of other-regarding preferences. Singer's work is based on a neuroscientific model of empathy suggested by Preston and de Waal (2002). According to this model, observing or imagining another person in a particular emotional state automatically activates a representation of that state in the observer with its associated automatic and somatic responses. The term "automatic" in this case refers to a process that does not require conscious and effortful processing but which can nevertheless be inhibited or controlled. Singer et al. recruited couples who were in love with each other for their study; empathy was assessed "in vivo" by bringing both woman and man into the same scanner environment. More specifically, brain activity was assessed in the female partner while painful stimulation was applied either to her own or to her partner's right hand via electrodes attached to the back of the hand. The male partner was seated next to the MRI scanner and a mirror system allowed her to see both hands, hers and that of her partner, lying on a tilted board in front of her. Flashes of different colors on a big screen behind the board pointed to either hand, indicating which of them would receive the painful stimulation and which would be subject to the non-painful stimulation. This procedure enabled the measurement of pain-related brain activation when pain was applied to the scanned subject (the so-called "pain matrix") or to her partner (empathy for pain). The results suggest that some but not the entire "pain matrix" was activated when empathizing with the pain of others. Activity in the primary and secondary somato-sensory cortex was only observed when receiving pain. These areas are known to be involved in the processing of the sensory-discriminatory components of our pain experience, that is, they indicate the location of the pain and its objective quality. In contrast, the bilateral anterior insula (AI) and the rostral anterior cingulate cortex (ACC) were activated when subjects either received pain or a signal that a loved one experienced pain. These areas are involved in the processing of the affective component of pain, that is, how unpleasant the subjectively felt pain is. Thus, both the experience of pain to oneself and the knowledge that a loved partner experiences pain activate the same affective pain circuits, suggesting that if a loved partner suffers pain, our brains also make us suffer from this pain. These findings suggest that we use representations reflecting our own emotional responses to pain to understand how the pain of others feels. Moreover, our ability to empathize may have evolved from a system which represents our own internal feeling states and allows us to predict the affective outcomes of an event for both ourselves and for others.

The results of the Singer et al. (2004a) study further suggest that the empathic response is rather automatic and does not require active engagement of some explicit judgments about others' feelings. The scanned subjects did not know that the experiment was about empathy; they were merely instructed to do nothing but observe the flashes that indicate either pain to the subject or the loved partner. The analysis also confirmed that the ability to empathize is heterogeneous across individuals; standard empathy questionnaires and the strength of the activation in the affective pain regions (AI and ACC) when the partner received pain were used to assess this heterogeneity. Interestingly, individual heterogeneity measured by the empathy questionnaire was highly correlated with individual differences that were measured by brain activation in those areas that process the affective component of pain (i.e. AI and ACC). Thus, neural evidence and questionnaire evidence on empathy mutually reinforce each other.

Does empathy also extend to unknown persons? The results of three recent studies indicate that empathic responses are also elicited when scanned subjects do not know the person in pain. Activity in the ACC and AI has also been observed when subjects witness still pictures depicting body parts involved in possibly painful situations [Jackson, Meltzoff and Decety (2005)] or videos showing a needle stinging in the back of a hand [Morrison et al. (2004)]. In a new paper Singer et al. (2006) investigated whether the level of empathic response in the ACC and AI can be modulated by the fact whether the subject likes or dislikes the "object of empathy". In this study, actors are paid to pretend to be naive subjects participating in two independent experiments, one on "social exchange" and the other on the "processing of pain". In the first experiment, the two confederates repeatedly play a modified trust game in the position of the trustee with the scanned subject. One actor plays a fair strategy and usually reciprocates trusting first mover choices with cooperation; the other actor plays unfairly and defects in response to first mover cooperation most of the time. Behavioral and neuronal findings of a previous imaging study which revealed aversion and fondness reported verbally as well as emotion-related brain activation in response to faces of people who had previously cooperated or defected [Singer et al. (2004b)] indicate that the subjects like fair players and dislike unfair ones. In the second part of the experiment, all three players participate in a pain study that expands the approach by [Singer et al. (2004a)]. One actor sits on each side of the scanner, enabling the scanned subject to observe flashes of different colours indicating high or low pain stimulation to his/her hand or to those of the fair or unfair players. The evidence from Singer et al. (2006) suggests empathy-related activation in the ACC and AI when observing the unfamiliar but likeable person receiving painful stimulation. However, men who observe that the unfair trustee receives pain do not show any empathy related activation in AI and ACC.

An important prerequisite for neuroeconomic studies is the existence of neuroscientific knowledge about the key components of the brain's reward circuits. Fortunately, many recent studies have shown that an area in the midbrain, the striatum, is a key part of reward-related neural circuits. Single neuron recording in non-human primates [Schultz (2000)] and neuroimaging studies with humans using money as a reward medium [Knutson et al. (2001), Delgado et al. (2003), O'Doherty et al. (2004)] clearly support this hypothesis. This knowledge about the brain's reward network enables neuroeconomists to ask intriguing questions. For example, some men's brains show no empathic concern for an unfair subject who receives pain. Do they perhaps even enjoy this experience? The results of Singer et al. (2006) exactly indicate this. Instead of activating empathy related networks like the ACC and AI, the men (but not the women) show activation in the striatum (the Nucleus Accumbens, NACC)! Moreover, men who reported more anger about others' behavior in self-reports collected after the experiment exhibit higher activation in the NACC. As a higher intensity of anger is probably associated with a higher relief if the unfair subject is punished, this finding further supports the hypothesis that the passive observation of the punishment of unfair subjects is associated with positive hedonic feelings.

This raises the question whether reward related brain areas are also activated if subjects can punish unfair behavior themselves or when they even have to pay for punishing the unfair subject. de Quervain et al. (2004) answered this question in a recent study. These authors modified the trust game by including a punishment opportunity for the investor. In this game, the investor had the opportunity of punishing the trustee after observing whether the trustee reciprocated the investor's trust by assigning up to 20 punishment points to the trustee. The monetary consequences of the punishment depended on the treatment conditions and will be explained below. The investor's brain was scanned with PET when he received information about the trustee's decision and when he decided whether to punish the trustee.

de Quervain et al. (2004) hypothesized that the opportunity to punish an unfair partner will activate the striatum. In particular, if the investor punishes the trustee because he anticipates deriving satisfaction from punishing, one should observe activation predominantly in those reward-related brain areas that are associated with goal-directed behavior. There is strong evidence from single neuron recording in non-human primates [Schultz (2000)] that the dorsal striatum is crucial for the integration of reward information and behavioral information in the sense of a goal-directed mechanism. Several recent neuroimaging studies support the view that the dorsal striatum is implicated in processing rewards resulting from a decision [Knutson et al. (2001), Delgado et al. (2003), O'Doherty et al. (2004)]. The fact that the dorsal striatum also responds to expected monetary gains in a parametric way is of particular interest from an economic viewpoint: if subjects successfully complete a task that generates monetary rewards, the activation in the dorsal striatum increases as the expected monetary gain grows. Thus, if the investor's dorsal striatum is activated when punishing the trustee, one has a strong piece of evidence indicating that punishment is rewarding.

To examine the activation of striatal areas during the decision to punish, subjects' brains were mainly scanned in those trust game trials in which the trustee abused the investor's trust. In the condition termed "costly" (C), the punishment was costly for both players. Every punishment point assigned to the trustee cost experimental $1 for the investor and reduced the trustee's payoff by experimental $2. In the condition termed "free" (F), punishment was not costly for the investor. Every punishment point assigned to the trustee cost nothing for the investor while the trustee's payoff was reduced by $2. In a third condition, which we call "symbolic" (S), punishment had only a symbolic (and no pecuniary) value. The punishment points assigned cost neither player anything. Thus, the investor could not reduce the trustee's payoff in this condition.

The hypothesis that punishment is rewarding predicts that the contrast F–S will show the activation of reward related brain areas after the investor's trust has been abused. The rationale behind this prediction is that the investor is likely to have a desire to punish the trustee both in the F and the S condition because the trustee intentionally abused the investor's trust, but the investor cannot really hurt the trustee in the S condition. Thus, the purely symbolic punishment in the S condition is unlikely to be satisfactory because the desire to punish the defector cannot be fulfilled effectively, and in the unlikely case that symbolic punishment is satisfactory, it is predicted to be less so than punishment in the F condition.

The F–S contrast is ideal for examining the satisfying aspects of effective punishment because – except for the difference in the opportunity to punish effectively – everything else remains constant across conditions. However, costly punishment should also generate satisfaction from an economic viewpoint. If there is indeed a taste for punishing defectors and if subjects actually punish because the cost of punishing is not too high, the act of punishment is analogous to buying a good. Rational subjects buy the good as long as the marginal costs are below the marginal benefits. Thus, an economic model based on a taste for punishment predicts that punishment in the C condition should also be experienced as satisfactory, implying that reward related areas will also be activated in the C–S condition.

Questionnaire and behavioral evidence indicates that investors indeed had a strong desire to punish the defectors. In fact, almost all subjects punished maximally in the F condition, while most subjects still punished in the C condition, albeit at a lower level. This reduction in the level of punishment makes sense because punishment was costly in the C condition. Most importantly, however, the dorsal striatum was strongly activated in both the F–S contrast and the C–S contrast, indicating that punishment is experienced as satisfactory. Moreover, the data show that those subjects in the C condition who exhibit higher activations in the dorsal striatum also punish more. This positive correlation can be interpreted in two ways: first, the higher level of punishment could cause the increased activation of the dorsal striatum, i.e., the higher satisfaction. Second, the greater anticipated satisfaction from punishing could cause the higher level of punishment, i.e., the activation in the striatum reflects – in this view – the anticipated satisfaction from punishing. It would be reassuring from an economic viewpoint if the second interpretation were the correct one because it relies on the idea that the anticipated rewards from punishing drive the punishment decision.

de Quervain et al. (2004) provide two pieces of evidence in favor of the second hypothesis. The first piece of evidence is related to the C–F contrast. Subjects face a nontrivial trade off in the C condition between the benefits and costs of punishing, whereas the decision is much simpler in the F condition because no costs exist. Thus, certain parts of the prefrontal cortex (Brodmann areas 10 and 11), which are known to be involved in integrating the benefits and costs for the purpose of decision-making, should be more strongly activated in the C condition than in the F condition. This is in fact the case. The second piece of evidence is based on the observation that most subjects punished maximally in the F condition. Thus, the differences in striatum activation across these subjects cannot be due to different levels of punishment. However, if different striatum activations reflect differences in the anticipated satisfaction from punishment, those subjects who exhibit higher striatum activations in the F condition (although they punish at the same maximal level) should be willing to spend more money on punishment in the C condition. The data again supports this prediction.

Neuroeconomic evidence also suggests that subjects derive special hedonic rewards from mutual cooperation with other human beings. This finding is insofar relevant as many trustees do reciprocate first mover choices in trust games and many subjects also cooperate in simultaneously played one-shot prisoners' dilemmas. One of the first neuroeconomic studies [Rilling et al. (2002)] reports activations in the striatum when subjects experience mutual cooperation with a human partner compared to mutual cooperation with a computer partner. Thus, despite the fact that the subject's monetary gain is identical in both situations, mutual cooperation with a human partner seems to be experienced as a more rewarding outcome, indicating that extra benefits from mutual cooperation extend beyond mere monetary gain. Unfortunately, however, the Rilling et al. study is based on a repeated prisoners' dilemma. A repeated dilemma game involves a host of other confounding influences which might shed doubt on the interpretation of brain activations in terms of other-regarding preferences. A recent paper based on a simplified trust game solved this problem [Rilling et al. (2002)]. The authors again show that the mutual cooperation outcome with a human partner generates higher striatum activation than does the mutual cooperation outcome with a computer partner. Moreover, the mutual cooperation outcome with a human partner also generates higher activations than does earning the same amount of money in a trivial individual decision-making task. A further study shows that the mere viewing of faces of people who previously cooperated in a version of the trust game activates reward related areas [Singer et al. (2004b)], thus indicating the special hedonic qualities of mutual cooperation. This result suggests that people derive more utility from interactions with cooperative people not just because they can earn more money in these interactions but because these interactions are rewarding per se.

3. Theories of other-regarding preferences

The experimental evidence sketched in Section 2 has provoked several theoretical attempts to explain the observed behavior across different experiments within the rational choice framework. Three different approaches can be distinguished:

1. Models of "social preferences" assume that a player's utility function not only depends on his own material payoff, but may also be a function of the allocation of resources within his reference group, i.e. a player may also be concerned about the material resources other people receive. Furthermore, several models assume that people differ. Some people seem to be quite strongly concerned about how they compare to other people, while others seem to be mainly self-interested. Given these social preferences, all agents are assumed to behave rationally, meaning that the well known concepts of traditional utility and game theory can be applied to analyze optimal behavior and to characterize equilibrium outcomes in experimental games.
2. Models of "interdependent preferences" assume that people are concerned about their opponent's "type". Suppose that each player may be either a selfish type or

a (conditionally) altruistic type. If an altruistic player knows that he interacts with another altruistic player, his preferences are altruistic and he is willing to be generous. If however, he knows that he deals with a selfish opponent, his preferences become selfish, too. Thus, whether player 1's preferences are altruistic or selfish depend on player 2's preferences and vice versa.

3. The third class of models deals with "intention based reciprocity". This approach assumes that a player cares about his opponent's intentions. If he feels that the opponent wanted to treat him kindly, he wants to return the favor and be nice to his opponent as well. If he feels that his opponent has hostile intentions, he wants to hurt his opponent. Thus, a player's interpretation of his opponent's behavior is crucial in this approach. Note that it is not the "type" of a player but rather his *intention* that is kind or hostile. Thus, in a given situation there may be an equilibrium in which a player has kind intentions, but there may also be a second equilibrium in which he has hostile intentions. Traditional game theory cannot capture this phenomenon; the framework of psychological game theory is needed.

Almost all models of these three approaches start out by making some fairly specific assumptions about the players' utility functions. Alternatively, one could start from a general preference relation and ask which axioms are necessary and sufficient to generate utility functions with certain properties. Axiomatic approaches are discussed at the end of this section.

Before we discuss the different approaches in detail, a word of caution is required. Many of the models under consideration here use terms such as "fairness", "equity", "altruism" or "reciprocity" that have been debated for a long time by moral philosophers and economists and that can be interpreted in different ways. Furthermore, some of these models are not entirely clear about what the domain of the theory is and what they want to achieve. In this section we will interpret all of these theories very restrictively. First of all, we view them as *purely positive theories* that try to explain actual human behavior. Thus, we disregard any normative implications the theories may have. Second, we view these models as first attempts *to explain the outcomes of economic experiments*. Typically, subjects enter these experiments as equals, they interact anonymously, and the physical outcome of the experiment is an allocation of monetary payoffs. Thus, for the experiments it is fairly straightforward to give a precise (and hopefully uncontroversial) definition of "altruistic preferences", "equitable allocation", "fair behavior" and the like. Of course, the theories discussed here do have implications for human behavior outside the laboratory as well. In some situations these implications may be very straightforward, but in general there are many important questions that have to be answered before the models can be applied to the "real world". This is a very important next step of this research agenda, but it will not be discussed here.

3.1. Social preferences

Classical utility theory assumes that a decision maker has preferences over allocations of material outcomes (e.g. goods) and that these preferences satisfy some "rationality" or "consistency" requirements, such as completeness and transitivity. However, this fairly general framework is often interpreted much more narrowly in applications, by implicitly assuming that the decision maker only cares about one aspect of an allocation, namely the material resources that are allocated to her. Models of social preferences assume, in contrast, that the decision maker may also care about the material resources allocated to others.

Somewhat more formally, let $\{1, 2, \ldots, N\}$ denote a set of individuals and $x = (x_1, x_2, \ldots, x_N)$ denote an allocation of physical resources out of some set X of feasible allocations. For concreteness we assume in the following that x_i denotes the monetary payoff of person i. The self-interest hypothesis says that the utility of individual i only depends on x_i. We will say that individual i has *social preferences* if for any given x_i person i's utility is affected by variations of x_j, $j \neq i$. Of course, simply assuming that the utility of individual i may be any function of the total allocation is often too general because it yields very few empirically testable restrictions on observed behavior.[7] In the following we will discuss several models of social preferences, each of which assumes that an individual's preferences depend on x_j, $j \neq i$, in a different way.

3.1.1. Altruism

A person is altruistic if the first partial derivatives of $u(x_1, \ldots, x_N)$ with respect to $x_1, \ldots, x_N$ are strictly positive, i.e., if her utility increases with the well being of other people. The hypothesis that (some) people are altruistic has a long tradition in economics and has been used to explain charitable donations and the voluntary provision of public goods.

Clearly, the simplest game for eliciting altruistic preferences is the dictator game (DG). Andreoni and Miller (2002) conducted a series of DG experiments in which one agent could allocate "tokens" between herself and another agent for a series of different budgets. The tokens were exchanged into money at different rates for the two agents and the different budgets. Let $U_i(x_1, x_2)$ denote subject i's utility function representing her preferences over monetary allocations (x_1, x_2).

In a first step, Andreoni and Miller check for violations of the General Axiom of Revealed Preference (GARP) and find that almost all subjects behaved consistently and passed this basic rationality check. Thus, their preferences can be described by (quasi-concave) utility functions. Then Andreoni and Miller classify the subjects into three main groups. They find that about 30 percent of the subjects give tokens to the other party in a fashion that equalizes the monetary payoffs between players. The behavior of 20 percent of the subjects can be explained by a utility function in which x_1 and x_2 are perfect substitutes, i.e., these subjects seem to have maximized the (weighted) sum of

[7] One implication, however, is that if a decision maker can choose between two allocations then his decision should be independent on how the two allocations have been generated. This prediction is refuted by some experiments on variants of the ultimatum game, where the proposer either could or could not influence the allocation of resources. See, e.g., Falk, Fehr and Fischbacher (2003) and Blount (1995) and the discussion in Sections 3.2 and 3.3 below.

the monetary payoffs. However, almost 50 percent of the subjects behaved "selfishly" and did not give any significant amounts to the other party. In a different experiment, they find that a sizeable minority (23 percent) of the subjects behaved spitefully by reducing their opponent's payoff if the opponent was better off then they were. Thus, they seem to have preferences that are non-monotonic in the monetary payoff of their opponent. Andreoni and Miller (2002, p. 750) conclude that many individuals seem to have other-regarding preferences and that the individual choice behavior of subjects in dictator games is consistent with rationality. However, individuals are heterogeneous, and only a minority of subjects can be described as unconditional altruists who have a utility function that is always strictly increasing in the payoff of their opponent.[8]

3.1.2. Relative income and envy

An alternative hypothesis is that subjects are not only concerned about the absolute amount of money they receive but also about their relative standing compared to others. The importance of relative income for a person's well being, of envy and jealousy, and of conspicuous consumption has long been recognized by economists and goes back at least to Veblen (1922).[9] Bolton (1991) formalized this idea in the context of an experimental bargaining game between two players. He assumes that $U_i(x_i, x_j) = u_i(x_i, x_i/x_j)$, where $u(\cdot,\cdot)$ is strictly increasing in its first argument and where the partial derivative with respect to x_i/x_j is strictly positive for $x_i < x_j$ and equal to 0 for $x_i \geqslant x_j$. Thus, agent i suffers if she gets less than player j, but she does not care about player j if she is better off herself. Note that this utility function implies that $\partial U_i/\partial x_j \leqslant 0$, just the opposite of altruism. Hence, while this utility function is consistent with the behavior in the bargaining games considered by Bolton, it neither explains generosity in dictator games and kind behavior of responders in trust games and gift exchange games nor voluntary contributions in public good games. The same problem arises in the envy-approach of Kirchsteiger (1994).

3.1.3. Inequity aversion

The preceding approaches assume that utility is either monotonically increasing or monotonically decreasing in the well being of other players. Fehr and Schmidt (1999) assume that a player is altruistic towards other players if their material payoffs are below an equitable benchmark, but she feels envy when the other players' material payoffs exceed this level.[10] For most economic experiments it seems natural to assume that an

[8] Another, more specific model of heterogeneous altruistic preferences has been developed by Cox, Sadiraj and Sadiraj (2001). They assume that the marginal rate of substitution between own income and the income of the opponent depends on whose income is higher.

[9] See e.g. Kolm (1995) for a detailed discussion and formalization of "envy" in economics.

[10] Daughety (1994) and Fehr, Kirchsteiger and Riedl (1998) also assume that a player values the payoff of reference agents positively, if she is relatively better off, while she values the others' payoff negatively, if she is relatively worse off.

equitable allocation is an equal monetary payoff for all players. Thus, inequity aversion reduces to inequality aversion in these games. Fehr and Schmidt consider the simplest utility function capturing this idea.

$$U_i(x_1, x_2, \ldots, x_N) = x_i - \frac{\alpha_i}{N-1}\sum_{j \neq i} \max\{x_j - x_i, 0\} - \frac{\beta_i}{N-1}\sum_{j \neq i} \max\{x_i - x_j, 0\}$$

with $0 \leqslant \beta_i \leqslant \alpha_i$ and $\beta_i \leqslant 1$. Note that $\partial U_i / \partial x_j \geqslant 0$ if and only if $x_i \geqslant x_j$. Note also that the disutility from inequality is larger if another person is better off than player i than if another person is worse off ($\alpha_i \geqslant \beta_i$).

This utility function can rationalize positive *and* negative actions towards other players. It is consistent with generosity in dictator games and kind behavior of responders in trust games and gift exchange games, *and at the same time* with the rejection of low offers in ultimatum games. It can explain voluntary contributions in public good games *and at the same time* costly punishments of free-riders.

A second important ingredient of this model is the assumption that individuals are heterogeneous. If all people were alike, it would be difficult to explain why we observe that people sometimes resist "unfair" outcomes or manage to cooperate even though it is a dominant strategy for a selfish person not to do so, while fairness concerns or the desire to cooperate do not seem to have much of an effect in other environments. Fehr and Schmidt show that the interaction of the distribution of types with the strategic environment explains why very unequal outcomes are obtained in some situations while very egalitarian outcomes prevail in others. For example, even a population that consists *only* of very fair types (high α's and β's) cannot prevent very uneven outcomes in certain competitive environments (see, e.g., the ultimatum game with proposer competition in Section 5.3) because none of the inequity averse players can enforce a more equitable outcome through her own actions. In contrast, a small fraction of inequity averse players in a public good game with punishment is sufficient to credibly threaten that free riders will be punished, inducing selfish players to contribute to the public good.

Fehr and Schmidt choose a distribution for α and β that is consistent with the experimental evidence of the ultimatum game. Keeping this distribution fixed, they show that their model yields surprisingly accurate predictions across many bargaining, market and social dilemma games.[11]

[11] One drawback of the piece-wise linear utility function employed by Fehr and Schmidt is that it implies corner solutions for some games where interior solutions are frequently observed. For example, a decision maker in the dictator game with a Fehr–Schmidt utility function would either give nothing (if her $\beta < 0.5$) or share the pie equally (if $\beta > 0.5$). Giving away a fraction that is strictly in between 0 and 0.5 is optimal only in the non-generic case where $\beta = 0.5$. This problem can be avoided, at the cost of tractability, by assuming non-linear inequity aversion.

Bolton and Ockenfels (2000) independently developed a similar model of inequity aversion. They also show that their model can explain a wide variety of seemingly puzzling evidence such as generosity in dictator, gift exchange and trust games and rejections in the ultimatum game. In their model, the utility function is given by

$$U_i = U_i(x_i, \sigma_i),$$

where

$$\sigma_i = \begin{cases} \dfrac{x_i}{\sum_{j=1}^{N} x_j} & \text{if } \displaystyle\sum_{j=1}^{N} x_j \neq 0, \\ \dfrac{1}{N} & \text{if } \displaystyle\sum_{j=1}^{N} x_j = 0. \end{cases}$$

For any given σ_i, the utility function is assumed to be weakly increasing and concave in player i's own material payoff x_i. Furthermore, for any given x_i, the utility function is strictly concave in player i's share of total income, σ_i, and obtains a maximum at $\sigma_i = 1/N$.[12]

Fehr–Schmidt and Bolton–Ockenfels often yield qualitatively similar results for two-player games, while some interesting differences arise with more than two players. Fehr and Schmidt assume that a player compares herself to each of her opponents separately in this case. This implies that her behavior towards an opponent depends on the income difference towards this person. In contrast, Bolton and Ockenfels assume that the decision maker is not concerned about each individual opponent but only about the average income of all players. Thus, whether $\partial U_i/\partial x_j$ is positive or negative in the Bolton–Ockenfels model does not depend on j's relative position towards i, but rather on how well i does compared to the average. If x_i is below the average, then i would like to reduce j's income even if j has a much lower income than i herself. On the other hand, if i is doing better than the average, then she is prepared to give to j even if j is much better off than i.[13]

[12] This specification of the utility function has the disadvantage that it is not independent of a shift in payoffs. Consider, for example, a dictator game in which the dictator has to divide X Dollars. Note that this is a constant sum game because $x_1 + x_2 \equiv X$. If we reduce the sum of payoffs by X, i.e., if the dictator can take away money from her opponent or give to him out of her own pocket, then $x_1 + x_2 = 0$ for any decision of the dictator and thus we always have $\sigma_1 = \sigma_2 = 1/2$. Therefore, the theory makes the implausible prediction that, in contrast to the game where $x_1 + x_2 = X > 0$, all dictators should take as much money from their opponent as possible. Camerer (2003, p. 111) notes a related problem. Suppose that the ultimatum game is modified as follows: If the responder rejects a proposal, the monetary payoffs are 10 percent of the original offer. In this case the relative shares are the same no matter whether the responder accepts or rejects. Hence, Bolton and Ockenfels predict that the responder will always accept any offer, no matter how unequal it is. These problems do not arise in Fehr and Schmidt's model of inequity aversion.

[13] See Camerer (2003, Section 2.8.5) and Section 4.1 for a more extensive comparison of these two approaches.

3.1.4. Hybrid models

Charness and Rabin (2002) combine altruistic preferences with a specific form of inequity aversion that they call *quasi-maximin preferences*. They start from a "disinterested social welfare function" which is a convex combination of Rawls' maximin criterion and the sum of the monetary payoffs of all players:

$$W(x_1, x_2, \ldots, x_N) = \delta \cdot \min\{x_1, \ldots, x_N\} + (1-\delta) \cdot (x_1 + \cdots + x_N),$$

where $\delta \in (0, 1)$ is a parameter reflecting the weight that is put on the maximin criterion. The first part of the social welfare function represents Rawlsian inequity aversion. The second part reflects altruism based on the idea that each individual's payoff receives the same weight. An individual's overall utility function is then given by a convex combination of his own monetary payoff and the above social welfare function:[14]

$$\begin{aligned} U_i(x_1, x_2, \ldots, x_N) = (1-\gamma)x_i + \gamma\big[&\delta \cdot \min\{x_1, \ldots, x_N\} \\ &+ (1-\delta) \cdot (x_1 + \cdots + x_N)\big]. \end{aligned}$$

In the two player case this boils down to

$$U_i(x_1, x_2) = \begin{cases} x_i + \gamma(1-\delta)x_j & \text{if } x_i < x_j, \\ (1-\gamma\delta)x_i + \gamma x_j & \text{if } x_i \geqslant x_j. \end{cases}$$

Note that the marginal rate of substitution between x_i and x_j is smaller if $x_i < x_j$. Hence, the decision maker cares about the well-being of the other person, but less so if the other person is better off than she is.

Altruism in general and quasi-maximin preferences in particular can explain positive acts to other players, such as generosity in dictator games and kind behavior of responders in trust games and gift exchange games,[15] but it is clearly inconsistent with the fact that subjects try to retaliate and hurt other subjects in some experiments, even if this is costly for them (as in the ultimatum game (UG) or a public good game with punishments). This is why Charness and Rabin augment quasi-maximin preferences by incorporating intention based reciprocity (see Section 3.3.3 below).

Erlei (2004) combines elements of inequity aversion à la Fehr–Schmidt and altruistic preferences à la Charness–Rabin by assuming that

$$U_i(x_1, x_2) = \begin{cases} (1-\sigma_i-\theta_i R)x_i + (\sigma_i+\theta_i R)x_j & \text{if } x_i \leqslant x_j, \\ (1-\rho_i-\theta_i R)x_i + (\rho_i+\theta_i R)x_j & \text{if } x_i \geqslant x_j. \end{cases}$$

[14] Note that Charness and Rabin do not normalize payoffs with respect to N. Thus, if the group size changes, and the parameters δ and γ are assumed to be constant; thus, the importance of the maximin term in relation to the player's own material payoff changes.

[15] However, altruism has some implausible implications even in these games. For example, altruism implies that if the government provides part of the public good (financed by taxes) in a public good context, then every dollar provided by the government "crowds out" one dollar of private, voluntary contributions. This "neutrality property" holds quite generally [Bernheim (1986)]. However, it is in contrast to the empirical evidence reporting that the actual crowding out is rather small. This has led some researchers to include the pleasure of giving (a "warm glow effect") in the utility function [Andreoni (1989)].

In this formulation, σ_i (ρ_i) represents player i's concern for player j's payoff if player i's payoff is larger (smaller, respectively) than player j's. The term $\theta_i R$ models negative reciprocity explicitly. If player j "misbehaved" by taking an action that violates the norms of fairness, R takes the value -1, otherwise it is 0. The parameter $\theta_i \geqslant 0$ measures the importance of this sort of reciprocity as compared to the other elements of the utility function.

Erlei assumes that there are three different types of players: Selfish players have $\sigma_i = \rho_i = \theta_i = 0$, i.e. they only care about x_i. Inequity averse players are characterized by $\sigma_i < 0 < \rho_i < 1$. Altruistic types always put a positive weight on the payoff of their opponent, so $0 < \sigma_i \leqslant \rho_i \leqslant 1$. Erlei applies this model to the games discussed by Charness and Rabin (2002) and by Goeree and Holt (2001). Obviously, the model offers a better predictive fit than do models that only focus on one type of preference. Perhaps more surprisingly, the author shows that direct negative reciprocity (as captured by $\theta_i R$) does not play a significant role in the games he considers.

Cox, Friedman and Gjerstad (2004) suggest another fairly flexible utility function of the form

$$U_i = \begin{cases} \dfrac{1}{\alpha}(x_i^\alpha + \lambda x_2^\alpha) & \text{if } \alpha \neq 0, \\ (x_i \cdot x_j)^\lambda & \text{if } \alpha = 0, \end{cases}$$

where $\alpha \in (-\infty, 1]$ reflects the curvature of indifference curves in the (x_i, x_j) space. The marginal rate of substitution between i's income and j's income in i's utility function is given by

$$MRS = \frac{\partial U_i / \partial x_i}{\partial U_i / \partial x_j} = \lambda^{-1}\left(\frac{x_j}{x_i}\right)^{1-\alpha}.$$

Thus, when $\alpha = 1$, preferences are linear (MRS is constant), when $\alpha < 1$, they are strictly convex. Cobb–Douglas preferences correspond to $\alpha = 0$ and Leontief preferences to $\alpha \to -\infty$. Whether preferences are altruistic or spiteful depends on the parameter $\lambda = \lambda(r)$ that is interpreted as the "emotional state" of player i. This emotional state depends on a reciprocity motive r which is defined as[16] $r(x) = \bar{x}_i(s_j) - x_i^0$, where $\bar{x}_i(s_j)$ is the maximum payoff player i can achieve given strategy s_j of player j and x_i^0 is an appropriate reference payoff. If the maximum payoff player i can achieve given the strategy s_j of his opponent is smaller than this reference payoff, $r(x)$ (and λ) are negative and player i wants to hurt player j.[17]

Cox et al. estimate the parameters of their model separately using the existing experimental data for the mini-ultimatum game [Falk, Fehr and Fischbacher (2003)] and for

[16] Cox, Friedman and Gjerstad (2004) argue that λ may also depend on the social status s of the players, but this seems to be irrelevant in most experiments and the authors do not make any use of s in the applications they consider.

[17] A similar model has been suggested by Sandbu (2002). In his model the marginal rate of substitution between own income and income of the opponent depends on the sets of actions available to the players.

a Stackelberg duopoly game [Huck, Müller and Normann (2001)]. While the model can fit the data of these two games reasonably well, the authors have yet to show that the parameter estimates derived from one game can also explain the data of other games. Furthermore, the model is quite restrictive because it can only be applied to sequential two-person games of perfect information.

Benjamin (2004) considers a model that allows for different types of social preferences. The main innovation in his paper is that utility is not defined on absolute wealth levels but rather on changes in wealth levels. Furthermore, people are loss-averse over their own changes in payoffs, but they do not weight the losses of others more heavily than the gains of others. Benjamin argues that this may explain why it is often considered unfair if a landlord raises rents for existing tenants but not if he raises rents for new tenants. The point is that raising rents on existing tenants causes a gain to the landlord at the expense of the tenant, while a new tenant enters into a transaction in which both parties gain. In models of social preferences that are defined over absolute wealth levels it would not make any difference whether the tenant is old or new.

Benabou and Tirole (2004) develop a model in which people have different degrees of altruism, but are also concerned about their social reputation and self-respect. Thus, people behave altruistically because they are genuinely altruistic, but also because they want to signal to other people (or to themselves) that they are generous. This model has a rich set of implications. In particular, it can explain why monetary incentives may crowd out altruistic behavior. The reason is that the presence of monetary rewards spoils the reputational value of good deeds. These actions are no longer an unambiguous signal of altruism or generosity with explicit rewards (or punishments), however, because they may have been undertaken for the money at stake. Benabou and Tirole apply this model to charitable giving, incentive provision, and multiple social norms of behavior, but they do not try to explain observed behavior in experimental games.

3.2. Interdependent preferences

Models of social preferences assume that players' utility functions depend only on the final allocation of material resources. Thus, if a player has to choose between different allocations, his choice will be independent of how these different allocations came about. This is implausible in some cases. For example, if I have to decide whether to accept or to reject a very unequal allocation, my decision may depend on whether my opponent chose the unfair allocation deliberately, or whether he had no possibility of affecting the allocation.[18]

A possible solution to this problem is to assume that players may be of different types (e.g., altruistic and spiteful types), and that each player's preferences depend on his opponent's type. In such a model my opponent's action affects my utility in two

[18] See, e.g., the experiments on the ultimatum game by Blount (1995) and on the mini-ultimatum game by Falk, Fehr and Fischbacher (2003).

ways. First, it affects my utility directly through its effect on the allocation of material resources. Second, there is an indirect effect if the action conveys information about my opponent's type.

These models are considerably more complex than models of social preferences because they assume that *preferences are interdependent*: my preferences depend on your preferences and vice versa. Several models have been proposed to capture these effects.

3.2.1. Altruism and spitefulness

Levine (1998) considers the utility function

$$U_i = x_i + \sum_{j \neq i} x_j (a_i + \lambda a_j)/(1 + \lambda),$$

where $0 \leqslant \lambda \leqslant 1$ and $-1 < a_i < 1$ for all $i \in \{1, \ldots, N\}$. Suppose first that $\lambda = 0$. In this case, the utility function reduces to $U_i = x_i + a_i \sum_{j \neq i} x_j$. If $a_i > 0$, then person i is an altruist who wants to promote the well being of other people, if $a_i < 0$, then player i is spiteful. While this utility function would be able to explain why some people contribute in public good games and why others reject positive offers in the ultimatum game, it has difficulties explaining why the same person is altruistic in one setting and spiteful in another setting unless the absolute value of a player's a_i is close to zero or the values of the opponent's a_j strongly differs across settings.

Now suppose that $\lambda > 0$. In this case, an altruistic player i (with $a_i > 0$) feels more altruistic towards another altruist than towards a spiteful person. In fact, if $-\lambda a_j > a_i$ player i may behave spitefully herself. In most experiments, where there is anonymous interaction, the players do not know their opponent's parameter a_j and have to form beliefs about them. Thus, any sequential game becomes a signaling game in which beliefs about the other players' types are crucially important for determining optimal strategies. This may give rise to a multiplicity of signaling equilibria.

Levine uses the data from the ultimatum game to calibrate the distribution of a_i and to estimate λ (which he assumes to be the same for all players). He shows that with these parameters the model can reasonably fit the data on centipede games, market games, and public good games. However, because $a_i < 1$, the model cannot explain positive giving in the dictator game.

Rotemberg (2004) suggests a closely related model that focuses on ultimatum and dictator games. He assumes the following utility functions for the proposer and the responder, respectively:

$$U_P = E\left(x_P + a^P x_R\right)^\gamma,$$
$$U_R = x_R + \left[a^R - \xi\left(\hat{a}^P, \underline{a}\right)\right] \cdot x_P.$$

Consider first the responder's utility function which depends on his own income x_R and on that of his opponent x_P. However, the weight with which x_P enters his utility function depends on the difference between his own altruism a^R and a function ξ that depends, in turn, on the responder's estimate of his opponent's altruism, denoted by $\hat{a}^P$, and a minimum level of altruism $\underline{a}$. The function ξ is discontinuous and takes only two values: If $\hat{a}^P \geqslant \underline{a}$, ξ takes the value of 0, if $\hat{a}^P < \underline{a}$ there is a discontinuous jump to $\xi = \bar{\xi} = a^R + 1$. Thus, if the responder believes that the proposer does not satisfy some minimal level of benevolence (that may differ across responders), his preferences turn hostile and he enjoys reducing the proposer's payoff.

Consider now the proposer's utility function that also depends on his own income and on that of the responder weighted with the altruism parameter a^P. The proposer moves first, so he does not learn anything about the responder's type before taking his action. This is why the reciprocity term that is part of the proposer's utility function does not play a role here. However, the outcome of the proposer's decision is risky, because he does not know how the responder will react to it. The parameter γ reflects the proposer's risk aversion. In order to explain the distribution of actual offers in the ultimatum game, Rotemberg assumes that the proposer is risk-loving ($\gamma > 1$). Note that the responder does not face any risk, so his attitudes towards risk are irrelevant.

This model can be fit reasonably well to the data of the ultimatum game. The discontinuity of the function ξ may explain why behavior sometimes changes quite quickly from benevolence to hostility if certain standards of behavior are not met by the opponents. However, it is not clear that the parameter estimates for the ultimatum game yield reasonable predictions if the model is applied to other games. Rotemberg considers only one other game, the dictator game. However, here he imposes the additional assumption that the proposer suffers a utility loss of V if he believes that the responder believes that $a^P < \underline{a}$. This additional assumption is not only ad hoc, it also makes the proposer's payoff a function of the responder's *beliefs* about his type, thus turning the game into a psychological game (see Section 2.3 below).

Gul and Pesendorfer (2005) develop a canonical model of interdependent preferences. For example, they consider reciprocity in the ultimatum game and assume that preferences are linear and of the form

$$U_i = x_i + a_i x_j$$

with

$$a_i = c_0 + \sum_{n=1}^{\infty} c_n \cdot t_n^i \cdot t_{n-1}^j.$$

Here $t^i = (t_0^i, t_1^i, t_2^i, \ldots)$, where t_0^i is normalized to 1, is the type of player i which, together with the type of player j and the sequence of parameters $\{c_0, c_1, \ldots\}$, determines the parameter a_i. The interpretation of the vector t^i is that t_1^i is player i's unconditional level of altruism, irrespective of player j's type. The parameter t_2^i captures the strength of the response to player j's kindness, and so on. Gul and Pesendorfer construct an example with just two types that roughly replicates the main features of the mini-ultimatum game. In particular, it explains that an offer of (80, 20) may be rejected if the responder could have chosen (50, 50), but that it will be accepted if the responder had no choice. This model is very general and quite flexible, but it seems difficult to apply to more complicated games.

3.3. Models of intention based reciprocity

The models considered so far do not allow for the possibility that players care about their opponents' intentions. I may be happy to be kind to my opponent if I believe that he intends to be kind to me – independent of what he actually does. In order to evaluate my opponent's intentions, I not only have to form beliefs about what he is going to do, but also about why he is going to do it. But in order to interpret his behavior, I have to form beliefs about which actions my opponent believes I will take. Thus, for a given action of my opponent, it makes a difference for my utility payoff whether I believe that he takes this action because he believes that I will be kind to him or because he believes that I am going to hurt him. Traditional game theory cannot capture this, as it assumes that outcomes (and not beliefs) determine payoffs. However, Geanakoplos, Pearce and Stacchetti (1989) developed the concept of "psychological game theory" that generalizes traditional game theory by allowing for the possibility that payoffs are a function of players' beliefs. All models discussed in this subsection are based on psychological game theory.

3.3.1. Fairness equilibrium

In a pioneering article, Rabin (1993) modeled intention based reciprocity for simple two-player normal form games. Let A_1 and A_2 denote the (mixed) strategy sets for players 1 and 2, respectively, and let $x_i\colon A_1 \times A_2 \to IR$ be player i's material payoff function.

We now have to define (hierarchies of) beliefs over strategies. Let $a_i \in A_i$ denote a strategy of player i. When i chooses her strategy, she must have some belief about the strategy player j will choose. In all of the following $i \in \{1, 2\}$ and $j = 3 - i$. Let b_j denote player i's belief about what player j is going to do. Furthermore, in order to rationalize her expectation b_j, player i must have some belief about what player j believes that player i is going to do. This belief about beliefs is denoted by c_i. The hierarchy of beliefs could be continued ad infinitum, but the first two levels of beliefs are sufficient for defining reciprocal preferences.

Rabin starts with a "kindness function", $f_i(a_i, b_j)$, which measures how kind player i is to player j. If player i believes that her opponent chooses strategy b_j, then she effectively chooses her opponent's payoff out of the set $[x_j^l(b_j), x_j^h(b_j)]$ where $x_j^l(b_j)(x_j^h(b_j))$ is the lowest (highest) payoff of player j that can be induced by player i if j chooses b_j. According to Rabin, a "fair" or "equitable" payoff for player j, $x_j^f(b_j)$, is just the average of the lowest and highest payoffs (excluding Pareto-dominated payoffs, however). Note that this "fair" payoff is independent of player i's payoff. The kindness of player i towards player j is measured by the difference between the actual payoff she gives to player j and the "fair" payoff, relative to the whole range of feasible payoffs:[19]

$$f_i(a_i, b_j) \equiv \frac{x_j(b_j, a_i) - x_j^f(b_j)}{x_j^h(b_j) - x_j^l(b_j)}$$

with $f_i(a_i, b_j) = 0$ if $x_j^h(b_j) - x_j^l(b_j) = 0$. Note that $f_i(a_i, b_j) > 0$ if and only if player i gives player j more than the "fair" payoff.

Finally, we have to define player i's belief about how kindly player j treats her. This is defined in exactly the same manner, but beliefs have to move up one level. Thus, if player i beliefs that player j chooses b_j and if she believes that player j believes that i chooses c_i, then player i perceives player j's kindness as given by:

$$f_j'(b_j, c_i) \equiv \frac{x_i(c_i, b_j) - x_i^f(c_i)}{x_i^h(c_i) - -x_i^l(c_i)}$$

with $f_j'(b_j, c_i) = 0$ if $x_i^h(c_i) - x_i^l(c_i) = 0$. These kindness functions can now be used to define a player's utility function:

$$U_i(a, b_j, c_i) = x_i(a, b_j) + f_j'(b_j, c_i)\bigl[1 + f_i(a_i, b_j)\bigr],$$

where $a = (a_1, a_2)$. Note that if player j is perceived to be unkind ($f_j'(\cdot) < 0$), player i wants to be as unkind as possible, too. On the other hand, if $f_j'(\cdot)$ is positive, player i gets some additional utility from being kind to player j as well.

While this specification has some appealing properties, it is not consistent. For example, the utility function adds the monetary payoff of player i (measured for example in Dollars) to the kindness function that has no dimension. Note also that by definition the kindness term must lie in the interval $[-1, 0.5]$. Thus, the kindness term becomes less important the higher the material payoffs are. Furthermore, if monetary payoffs are multiplied by a constant (for example, if we move to a different currency) the marginal rate of substitution between money and kindness is affected. Thus, this utility function has very strong cardinal properties which are unappealing.

A "fairness equilibrium" is an equilibrium in a psychological game with these payoff functions, i.e., a pair of strategies (a_1, a_2) that are mutual best responses to each other and a set of rational expectations $b = (b_1, b_2)$ and $c = (c_1, c_2)$ that are consistent with equilibrium play.

Rabin's theory is important because it was the first contribution that precisely defined the notion of reciprocity and explored the consequences of reciprocal behavior. The model provides several interesting insights, but it is not well suited for predictive purposes. It is consistent with rejections in the UG, but many other equilibria exist as

[19] A disturbing feature of Rabin's formulation is that he excludes Pareto-dominated payoffs in the definition of the "fair" payoff, but not in the denominator of the kindness term. Thus, adding a Pareto-dominated strategy for player j would not affect the fair payoff but it would reduce the kindness term.

well, some of which are highly implausible. For example, offers above 50 percent of the surplus are part of an equilibrium even though this is almost never observed in experiments.

The multiplicity of equilibria is a general feature of Rabin's model. If material payoffs are small enough to make psychological payoffs matter, then there is always one equilibrium in which both players are nice to each other and one in which they are hostile. Both equilibria are supported by self-fulfilling prophecies, so it is difficult to predict which equilibrium is going to be played.

The theory also predicts that players do not undertake kind actions unless others have shown their kind intentions. Suppose, for example, that player 1 has no choice but is forced to cooperate in the prisoners' dilemma game. If player 2 knows this, then – according to Rabin's theory – she will interpret player 1's cooperation as "neutral" ($f_2'(\cdot) = 0$). Thus, she will only look at her material payoffs and will defect. This contrasts with models of inequity aversion where player 2 would co-operate irrespective of the reason for player 1's co-operation. We will discuss the experimental evidence that can be used to discriminate between the different approaches in Section 4 below.

3.3.2. *Intentions in sequential games*

Rabin's theory has been defined only for two-person, normal form games. If the theory is applied to the normal form of simple sequential games, some very implausible equilibria may arise. For example, unconditional cooperation by the second player is part of a fairness equilibrium in the sequential prisoners' dilemma. The reason is that Rabin's equilibrium notion does not force player 2 to behave optimally off the equilibrium path.

In a subsequent paper, Dufwenberg and Kirchsteiger (2004) generalized Rabin's theory to N-person extensive form games for which they introduce the notion of a "Sequential Reciprocity Equilibrium" (SRE). The main innovation is to keep track of beliefs about intentions as the game evolves. In particular, it has to be specified how beliefs about intentions are formed off the equilibrium path. Given this system of beliefs, strategies have to form a fairness equilibrium in every proper subgame.[20] Applying their model to several examples, Dufwenberg and Kirchsteiger show that *conditional* cooperation in the prisoners' dilemma game is a SRE. They also show that an offer from the proposer which the responder rejects with certainty can be a SRE in the ultimatum

[20] Dufwenberg and Kirchsteiger also suggest several other deviations from Rabin's model. In particular, they measure kindness "in proportion to the size of the gift" (i.e. in monetary units). This has the advantage that reciprocity does not disappear as the stakes become larger, but it also implies that the kindness term in the utility function has the dimension of "money squared" which again makes the utility function sensitive to linear transformations. Furthermore, they define "inefficient strategies" (which play an important role in the definition of the kindness term) as strategies that yield a weakly lower payoff for all players than some other strategy for all subgames. Rabin (1993) defines inefficient strategies to be those which yield weakly less on the equilibrium path. However, the problem in Dufwenberg and Kirchsteiger (2004) arises with more than two players because an additional dummy player may render an inefficient strategy efficient and might thus affect the size of the kindness term.

game. This is an equilibrium because each player believes that the other party wants to hurt him. However, the equilibrium analysis in this model is very complex, even in these extremely simple sequential games. Furthermore, there are typically multiple equilibria with different equilibrium outcomes, due to different self-fulfilling beliefs about intentions. Some of these equilibria seem highly implausible, but the theory does not offer any formal criteria how to discriminate between "convincing" and "less convincing" equilibria.

3.3.3. Merging intentions and social preferences

Falk and Fischbacher (2006) also generalize Rabin's (1993) model. They consider N-person extensive form games and allow for the possibility of incomplete information. Furthermore, they measure "kindness" in terms of inequity aversion. Player i perceives player j's strategy to be kind if it gives rise to a payoff for player i which is higher than that of player j. Note that this is fundamentally different from both Rabin as well as Dufwenberg and Kirchsteiger, who define j's "kindness" in terms of the feasible payoffs of player i and not in relation to the payoff that player j gets. Furthermore, Falk and Fischbacher distinguish whether player j could have altered an unequal distribution or whether player j was a "dummy player" who is unable to affect the distribution by his actions. The kindness term gets a higher weight in the former case than in the latter. However, even if player j is a dummy player who has no choice to make, the kindness term (which now reflects pure inequity aversion) gets a positive weight. Thus Falk and Fischbacher merge intention based reciprocity and inequity aversion.

Their model is quite complex. At every node where player i has to move, she has to evaluate the kindness of player j which depends on the expected payoff difference between the two players and on what player j could have done about this difference. This "kindness term" is multiplied by a "reciprocation term", which is positive if player i is kind to player j and negative if i is unkind. The product is further multiplied by an individual reciprocity parameter which measures the weight of player i's desire to reciprocate as compared to his desire to get a higher material payoff. These preferences together with the underlying game form define a psychological game à la Geanakoplos, Pearce and Stacchetti (1989). A subgame perfect psychological Nash equilibrium of this game is called a "reciprocity equilibrium".

Falk and Fischbacher show that there are parameter constellations for which their model is consistent with the stylized facts of the ultimatum game, the gift exchange game, the dictator game, and of public good and prisoners' dilemma games. Furthermore, there are parameter constellations that can explain the difference in outcomes if one player moves intentionally or if she is a dummy player. Because their model contains variants of a pure intentions based reciprocity model (like Rabin) and a pure inequity aversion model (like Fehr and Schmidt or Bolton and Ockenfels) as special cases, it is possible to get a better fit of the data, but at a significant cost in terms of the model's complexity.

Charness and Rabin (2002) provide another attempt at combining social preferences with intention based reciprocity. We already described their model of quasi-maximin preferences in Section 3.1.4. In a second step, they augment these preferences by introducing a demerit profile $\rho \equiv (\rho_1, \ldots, \rho_N)$, where $\rho_i \in [0, 1]$ is a measure of how much player i deserves from the point of view of all other players. The smaller ρ_i, the more does player i count in the utility function of the other players. Given a demerit profile ρ, player i's utility function is given by

$$\begin{aligned} U_i(x_1, x_2, \ldots, x_N|\rho) = (1-\gamma)x_i + \gamma\Bigg[&\delta \cdot \min\Big\{x_i, \min_{j \neq i}\{x_j + d\rho_j\}\Big\} \\ &+ (1-\delta) \cdot \left(x_i + \sum_{j \neq i} \max\{1 - k\rho_j, 0\} \cdot x_j\right) - f \sum_{j \neq i} \rho_j x_j\Bigg], \end{aligned}$$

where $d, k, f \geqslant 0$ are three new parameters of the model. If $d = k = f = 0$, this boils down to the quasi-maximin preferences describes above. If d and k are large, then player i does not want to promote player j's well-being. If f is large, player i may actually want to hurt player j.

The crucial step is to endogenize the demerit profile ρ. Charness and Rabin do this by comparing player j's strategy to a "selfless standard" of behavior, which is unanimously agreed upon and exogenously given. The more player j falls short of this standard, the higher is his demerit factor ρ_j.

A "reciprocal fairness equilibrium" (RFE) is a strategy profile and a demerit profile such that each player maximizes his utility function given other players' strategies and given the demerit profile that is itself consistent with the profile of strategies. This definition implicitly corresponds to the Nash equilibrium of a psychological game as defined by Geanakoplos, Pearce and Stacchetti (1989).

The notion of RFE has several drawbacks that make it almost impossible to use for the analysis of even the simplest experimental games. First of all, the model is incomplete because preferences are only defined in equilibrium (i.e., for an equilibrium demerit profile ρ) and it is unclear how to evaluate outcomes out of equilibrium or if there are multiple equilibria. Second, it requires all players to have the same utility functions and agree on a "quasi-maximin" social welfare function in order to determine the demerit profile ρ. Finally, the model is so complicated and involves so many free parameters that it would be very difficult to test it empirically.

Charness and Rabin show that if the "selfless standard" is sufficiently small, every RFE corresponds to a Nash equilibrium of the game in which players simply maximize their quasi-maximin utility functions. Therefore, in the analysis of the experimental evidence, they restrict attention to the much simpler model of quasi-maximin preferences that we discussed in Section 3.1.1 above.

3.3.4. Guilt aversion and promises

Charness and Dufwenberg (2004) argue that people may be willing to help other people because they would feel guilty if they were to let them down. In particular, they would feel guilty if they promised beforehand to help the other party. In order to test this hypothesis, Charness and Dufwenberg conducted several trust game experiments in which one party could send a (free-form) message to the other party before the actual game starts. For example, the second mover could "promise" the first mover that he will reciprocate if the first mover trusts him. The experiments show that these promises significantly increase the probability that the first mover trusts, and second movers who made such a promise are significantly more likely to reciprocate when compared to an experiment without pre-play communication. Of course, pre-play communication is just cheap talk from the point of view of traditional game theory, and should not affect the (unique) equilibrium outcome of this game.

In order to explain the experimental results, Charness and Dufwenberg develop a model of "guilt aversion" using psychological game theory. In this model, players feel "guilt" if they let other players down. More precisely, if player 1 believes that player 2 believes that player 1 will take an action that gives monetary payoff m to player 2, then player 1 feels guilt if he takes an action that gives a payoff of $m' < m$ to player 2. If guilt aversion is sufficiently strong, player 1 may choose an action that is personally costly to him but which benefits player 2 because he does not want to disappoint player 2's belief about his action. As in Rabin's (1993) model, this theory requires that players have second-order beliefs about other players' beliefs and it typically has many equilibria. Pre-play communication and promises can be useful as a coordination device in order to select one of these equilibria. Charness and Dufwenberg also show that guilt aversion can explain tipping behavior, reciprocal effort behavior in the gift exchange game and collusion in oligopolistic markets.

However, the model only focuses on positive reciprocity and cannot explain why people may want to hurt one another. Furthermore, the model shares all of the drawbacks of the other models based on psychological game theory, in particular complexity and multiplicity of equilibria.

3.4. Axiomatic approaches

The models considered so far assume very specific utility functions that are either defined on (lotteries over) material payoff vectors and/or on beliefs about other players' strategies and other players' beliefs. These utility functions are based on psychological plausibility, yet most of them lack an axiomatic foundation. Segal and Sobel (2004) take the opposite approach and ask what kinds of axioms generate preferences that can reflect fairness and reciprocity.

They start by assuming that players have preferences over strategy profiles rather than over material allocations. Consider a given two-player game and let Σ_i, $i \in \{1, 2\}$, denote the space of (mixed) strategies of player i. For any strategy profile $(\sigma_1, \sigma_2) \in \Sigma_1 \times \Sigma_2$, let $v_i(\sigma_1, \sigma_2)$ denote player i's utility function over her own monetary payoff (which is determined by the strategy profile (σ_1, σ_2)), assuming that these "selfish preferences" satisfy the von Neumann–Morgenstern axioms. However, player i's actual preferences are given by a preference relation $\succeq_{i\sigma_j}$ over her own strategies. This preference relation depends of course on the strategy σ_j she expects her opponent to play. Segal and Sobel show that if the preference relation $f_{i\sigma_j}$ satisfies the independence axiom and if, for a given σ_j, player i prefers to get a higher material payoff for herself if the payoff of player j is held constant (called "self interest"), then the preferences $f_{i\sigma_j}$ over Σ_i can be represented by a utility function of the form[21]

$$u_i(\sigma_i, \sigma_j) = v_i(\sigma_i, \sigma_j) + a_{i,\sigma j} v_j(\sigma_{i,} \sigma_j).$$

In standard game theory, $a_{i,\sigma j} \equiv 0$. Positive values of this coefficient mean that player i has altruistic preferences, negative values of $a_{i\sigma j}$ mean that she is spiteful.

The models of social preferences we discussed at the beginning of this chapter, in particular the models of altruism, relative income, inequity aversion, quasi-maximin preferences, and altruism and spitefulness, can all be seen as special cases of a Segal-Sobel utility function. Segal and Sobel can also capture some, but not all, aspects of intention based reciprocity. For example, a player's utility in Rabin's (1993) model not only depended on the strategy her opponent chose, but also on why he chose this strategy. This can be illustrated in the "Battle of the Sexes" game. Player 1 may go to boxing, because she expects player 2 to go to boxing, too (which is regarded as kind behavior by player 2, given that he believes player 1 will go to boxing). Yet, player 2 may also go to boxing, because he expects player 1 to go to ballet (which is regarded as unkind behavior by player 2 if he believes player 1 to go to ballet) and which is punished by the boxing strategy of player 1. This effect cannot be captured by Segal and Sobel, because in their framework preferences are defined on strategies only.

Neilson (2005) provides an axiomatic characterization of the Fehr and Schmidt (1999) model of inequity aversion. He introduces the axiom of "self-referent separability" which requires that if the monetary payoffs of player i and of all other players increase by some constant amount, then player 1's preferences about payoff allocations should not be affected. Neilson shows that this axiom is equivalent to having a utility function that is additively separable in the individual's own material payoff and the payoff differences to his opponents, which is an essential feature of the Fehr-Schmidt model. Furthermore, he shows that in a one-person decision problem under risk the same axiom of "self-referent separability" implies a generalization of prospect theory preferences [Kahneman and Tversky (1979)].

4. Discriminating between theories of other-regarding preferences

Most theories discussed in Section 3 were developed during the last 5–10 years and the evidence to discriminate between these theories is still limited. As we will show, however, the available data do exhibit some clear qualitative regularities which give a first indication of the advantages and disadvantages of the different approaches.

[21] The construction resembles that of Harsanyi's (1955) "utilitarian" social welfare function $\Sigma\alpha_i u_i$. Note, however, that Harsanyi's axiom of Pareto efficiency is stronger than the axiom of self interest employed here. Therefore, the $a_{i\sigma_j}$ in Segal and Sobel may be negative.

4.1. Who are the relevant reference actors?

All theories of other-regarding preferences are based on the idea that actors compare themselves with a set of reference actors or take these actors' payoffs directly into account. To whom do people compare themselves? Who are the relevant reference actors whose payoff is taken into account? There is no ambiguity about who the relevant reference actor is in bilateral interactions; the answer is less clear, however, in multi-person interactions. Most of the theories applicable in the n-person context assume that players make comparisons with all other $n - 1$ players in the game. The only exemption is the theory of Bolton and Ockenfels (BO). They assume that players compare themselves only with the "average" player in the game and do not care about inequities between the other players. In this regard, the BO approach is inspired by the data of Selten and Ockenfels (1998) and Güth and van Damme (1998), which seem to suggest that actors do not care for inequities among the other reference agents. It would greatly simplify matters if this aspect of the BO theory were correct.

One problem with this aspect of the BO approach is that it disenables the theory to explain punishment in the Third-Party Punishment Game [Fehr and Fischbacher (2004)]. Recall that there are three players, A, B, and C in the third party punishment game. Player A is endowed with some surplus S and must decide how much of S to give to B, who has no endowment. Player B is just a dummy player and has no decision power. Player C is endowed with $S/2$ and can spend this money on the punishment of A after he observes how much A gave to B. For any money unit player C spends on punishment the payoff of player A is reduced by 3 units. Note that the total surplus available in this game is $(3/2)S$. Therefore, without punishment, player C is certain to get her fair share $(S/2)$ of the total surplus, implying that the BO model predicts that C will never punish. In contrast to this prediction, roughly 60 percent of the C players punished in this game. This indicates that many players do care about inequities among other players. Further support for this hypothesis comes from Charness and Rabin (2002) who offered player C the choice between the payoff allocations $(575, 575, 575)$ and $(900, 300, 600)$. Because both allocations give player C the fair share of $1/3$ of the surplus, the BO model predicts that player C will choose the second allocation which gives him a higher absolute payoff. However, 54 percent of the subjects preferred the first allocation. Note that the self-interest hypothesis also predicts the second allocation, so one cannot conclude that the other 46 percent of the subjects have BO-preferences. A recent paper by Zizzo and Oswald (2000) also strongly suggests that subjects care about the inequities among the set of references agents.

It is important to note that theories of other-regarding preferences, in which subjects have multiple reference agents, do not necessarily imply that the subjects take actions in favor of *all* other reference agents, even if all other reference agents have the same weight in their utility function. To illustrate this, consider the following three-person UG [Güth and van Damme (1998)]. This game includes a proposer, a responder who can reject or accept the proposal, and a passive Receiver who can do nothing but collect the amount of money allocated to him. The proposer proposes an allocation (x_1, x_2, x_3)

where x_1 is the proposer's payoff, x_2 the responder's payoff and x_3 the Receiver's payoff. If the responder rejects, all three players get nothing, otherwise the proposed allocation is implemented.

It turns out that the proposers allocate substantial fractions of the surplus to the responder in this game but little or nothing to the Receiver. Moreover, Güth and van Damme (1998, p. 230) report that "there is not a single rejection that can clearly be attributed to a low share for the dummy (i.e., the Receiver, FS)". BO take this as evidence in favor of their approach because the proposer and the responder apparently do not take the Receiver's interest into account. However, this conclusion is premature because it is easy to show that approaches with multiple reference agents are fully consistent with the Güth and van Damme data. The point can be demonstrated in the context of the Fehr–Schmidt model. Assume for simplicity that the proposer makes an offer of $x_1 = x_2 = x$ while the Receiver gets $x_3 < x$. It is easy to show that a responder with FS-preferences will never (!) reject such an allocation even if $x_3 = 0$ and even if he is very fair-minded, i.e., has a high β-coefficient. To see this note that the utility of the responder if he accepts is given by $U_2 = x - (\beta/2)(x - x_3)$ which is positive for all $\beta \leqslant 1$, and thus higher than the rejection payoff of zero. A similar calculation shows that it takes implausibly high β-values to induce a proposer to take the interests of the Receiver into account.[22]

The above arguments suggest that the "average" player in a game is not an empirically relevant reference agent. This is particularly important for all games in which subjects may want to punish a particular individual for unfair or morally inappropriate behavior. In all these cases, a model, in which the differences (or the ratio) between a player's own payoff and the group's average payoff is the driving force of the punishment, is not able to predict which individual will be punished. A player who just wants to reduce the difference between his payoff and the group's average payoff does not care about the target of the punishment. Any punishment that reduces this difference, even if it is targeted on cooperative or norm abiding individuals, is equally desirable from the perspective of such a player [see also Falk, Fehr and Fischbacher (2005)].

In general, however, very little is known about the outcome of social comparison processes in games. Therefore, our empirical knowledge about what makes a player a relevant reference agent is very limited. The assumption that all players in a game are relevant reference agents to each other should only be taken as a first approximation and may not be true in some games. It seems reasonable to assume that player A is a relevant reference agent for player B if A can affect B's payoff in a salient way. However, there neither seems to be much theoretical work on this question nor persuasive empirical evidence beyond such general statements. Thus, the question "who are the relevant reference agents" is clearly an important unsolved problem.

[22] The proposer's utility is given by $U_1 = x_1 - (\beta/2)[(x_1 - x_2) + (x_1 - x_3)]$. If we normalize the surplus to one and take into account that $x_1 + x_2 + x_3 = 1$, $U_1 = (\beta/2) + (3/2)x_1[(2/3) - \beta]$. Thus, the marginal utility of x_1 is positive unless β exceeds $2/3$. This means that proposers with $\beta < 2/3$ will give the responders just enough to prevent rejection and, since the responders neglect the interests of the Receivers, nothing to the Receivers.

4.2. *Equality versus efficiency*

Many models of other-regarding preferences are based on the definition of a fair or equitable outcome to which people compare the available payoff allocations. In experimental games, the equality of material payoffs is a natural first approximation for the relevant reference outcome. The quasi-maximin theory of Charness and Rabin assumes instead that subjects care for the total surplus (“efficiency”) accruing to the group. A natural way to study whether there are subjects who want to maximize the total surplus is to construct experiments in which the predictions of both theories of inequality aversion (BO and FS) are in conflict with surplus maximization. This has been done by Andreoni and Vesterlund (2001), Bolle and Kritikos (1998), Andreoni and Vesterlund (forthcoming), Charness and Rabin (2002), Cox (2000) and Güth, Kliemt and Ockenfels (2000). Except for the Güth et al. paper, these papers indicate that a non-negligible fraction of the subjects in dictator game situations is willing to give up some of their own money in order to increase total surplus, even if this implies that they generate inequality that is to their disadvantage. Andreoni and Miller and Andreoni and Vesterlund, for example, conducted dictator games with varying prices for transferring money to the Receiver. In some conditions, the Allocator had to give up less than a dollar to give the Receiver a dollar, in some conditions the exchange ratio was 1 : 1, and in some other conditions the Allocator had to give up more than one dollar. In the usual dictator games, the exchange ratio is 1 : 1 and there are virtually no cases in which an Allocator transfers more than 50 percent of the surplus. In contrast, in dictator games with an exchange ratio of 1 : 3 (or 1 : 2) a non-negligible number of allocators transfer in such a way that they end up with less money than the Receiver. This contradicts the models of Bolton and Ockenfels (2000), of Fehr and Schmidt (1999), and of Falk and Fischbacher (2006) because in these models subjects never take actions that give the other party more than they get in these models. It is, however, consistent with altruistic preferences or quasi-maximin preferences.

What is the relative importance of this kind of behavior? Andreoni and Vesterlund are able to classify subjects in three distinct classes. They report that 44% of their subjects ($N = 141$) are completely selfish, 35 percent exhibit egalitarian preferences, i.e. they tend to equalize payoffs, and 21 percent of the subjects can be classified as surplus maximizers. Charness and Rabin report similar results with regard to the fraction of egalitarian subjects in a simple Dictator Game where the Allocator had to choose between (own, other) allocations of $(400, 400)$ and $(400, 750)$. 31 percent of the subjects preferred the egalitarian and 69 percent the surplus maximizing allocation. Among the 69 percent there may, however, also be many selfish subjects who no longer choose the surplus-maximizing allocation when this decreases their payoff only slightly. This is suggested by the game where the Allocator had to choose between $(400, 400)$ and $(375, 750)$. Here only 49 percent of surplus-maximizing choices were observed. Charness and Rabin also present questionnaire evidence indicating that when the income disparities are greater the egalitarian motive gains weight at the cost of the surplus maximization motive. When the Allocator faces a choice between (400, 400) and (400, 2000), 62 percent prefer the egalitarian allocation.

More recently, Engelmann and Strobel (2004) argued that "efficiency" is an important motive that clearly dominates the desire for equality in 3 player dictator games. For example, the Allocator (who was always player B) could choose between 3 different payoff allocations in one of their games: $(14, 4, 5)$, $(11, 4, 6)$ and $(8, 4, 7)$. Thus B's material payoff was the same in each of the three allocations, but he could redistribute income from the rich person to the poor person. Redistribution has a high efficiency cost in this game because it reduces the rich person's income by 3 units and increases the poor person's income by only 1 unit. Maximin preferences and selfish preferences cannot play a role in this game because the Allocator receives the lowest payoff regardless of the allocation chosen. This game allows, therefore, for a clean examination of how important the equality motive is relative to the "efficiency" motive. Engelmann and Strobel report that 60% of their subjects ($N = 30$) chose the first allocation, i.e., the one with the highest surplus and the highest inequality, and only 33% chose the most egalitarian allocation $(8, 4, 7)$.

However, only students of economics and business administration, which we call for brevity "economists", participated in the Engelmann and Strobel study. These students learn from the very beginning of their studies that surplus maximization is normatively desirable. Therefore, Fehr, Naef and Schmidt (forthcoming) replicated this game with $N = 458$ subjects to examine potential subject pool biases. They find a robust subject pool bias indicating that non-economists ($N = 291$) chose the most egalitarian allocation with the lowest surplus in 51% of the cases whereas economists' probability to choose this allocation was only 26% ($N = 167$). Likewise, the non-economists chose the least egalitarian allocation with the maximal surplus in only 28% of the cases, whereas the economists chose it in 56% of the cases. This result is also important with regard to the interpretation of the results of Charness and Rabin, who also have disproportionately many economists in their subject pool.

Since the evidence in favor of preferences for surplus maximization comes exclusively from dictator games, it is important to ask whether these preferences are likely to play a role in "strategic situations". We define strategic situations to be those in which the potential gift recipients are also capable of affecting the gift givers' material payoffs. This question is important because the dictator game is different from many economically important games and real life situations, because one player is rarely at the complete mercy of another player in economic interactions. It may well be that in situations where *both* players have some power to affect the outcome, the surplus maximization motive is less important than in dictator games or is easily dominated by other considerations. The gift-exchange experiments by Fehr, Kirchsteiger and Riedl (1993, 1998) are telling in this regard because they embed a situation that is like a DG into an environment with competitive and strategic elements.

These experiments exhibit a competitive element because the gift exchange game is embedded into a competitive experimental market. The experiments also exhibit a strategic element because the proposers are wage setters and have to take the responders' likely effort responses into account. Yet, once the responder has accepted a wage offer, the experiments are similar to a dictator game because, for a given wage, the responder essentially determines the income distribution and the total surplus by his choice of the effort level. The gift exchange experiments are an ideal environment for checking the robustness of the surplus maximization motive because an increase in the effort cost by one unit increases the total surplus by five units on average. Therefore, the maximal feasible effort level is, in general, also the surplus maximizing effort level. If surplus maximization is a robust motive, capable of overturning preferences for equality or reciprocity, one would expect that many responders choose effort levels that give the proposer a higher monetary payoff than the responder.[23] Moreover, surplus maximization also means that we should *not* observe a positive correlation between effort and wages because, for a given wage, the maximum feasible effort always maximizes the total surplus.[24]

However, the data supports neither of these implications. Effort levels that give the proposer a higher payoff than the responder are virtually non-existent. In the overwhelming majority of the cases, effort is substantially below the maximally feasible level and the proposer earns a higher payoff than the responder in less than two percent of the cases.[25] Moreover, almost all subjects who regularly chose non-minimal effort levels exhibited a reciprocal effort–wage relation. A related result was observed by Guth, Kliemt and Ockenfels (2003) who also conducted experiments in which dictators face a trade-off between equality and surplus maximization. They report that equality concerns dominate surplus maximization concerns in the sense that dictators never perform transfers such that they earn less than the recipient, even if such transfers would be surplus enhancing. These results are in sharp contrast to the 49 percent of the Allocators in Charness and Rabin who preferred the (375, 750) allocation over the (400, 400) allocation. One reason for the difference across studies is perhaps the fact that it was much cheaper to increase the surplus in the Charness–Rabin example. While the surplus increases in the gift exchange experiments on average by five units, if the responder sacrifices one payoff unit, the surplus increases by 14 units per payoff unit sacrificed in the Charness–Rabin case. This suggests that surplus maximization only gives rise to a violation of the equality constraint if surplus increases are extremely cheap. A second reason for the behavioral difference may be that when both players have some power to affect the outcome, the motive to increase the surplus is quickly crowded out by other considerations. This reason is quite plausible insofar as the outcomes in dictator games themselves are notoriously non-robust.

[23] The responders' effort level may, of course, also be affected by the intentions of the proposer. For example, paying a high wage may signal fair intentions which may increase the effort level. Yet, since this tends to raise effort levels, we would have even stronger evidence against the surplus-maximization hypothesis, if we observe little or no effort choices that give the proposer a higher payoff than the responder.

[24] There are degenerate cases in which this is not true.

[25] The total number of effort choices is $N = 480$ in these experiments, i.e., the results are not an artefact of a low number of observations.

While the experimental results on ultimatum games are fairly robust, the dictator game seems to be a rather fragile situation in which minor factors can have large effects. Cox (2004), e.g., reports that *100 percent* of all subjects transferred positive amounts in his dictator games.[26] This result contrasts sharply with many other games, including the games in Charness and Rabin and many other dictator games. To indicate the other extreme, Eichenberger and Oberholzer-Gee (1998), Hoffman et al. (1994) and List and Cherry (2000) report on dictator games with extremely low transfers.[27] Likewise, in the impunity game of Bolton and Zwick (1995), which is very close but not identical to a dictator game, the vast majority of proposers did not shy away from making very unfair offers. The impunity game differs from the dictator game only insofar as the responder can reject an offer; however, the rejection destroys only the responder's but not the proposer's payoff. The notorious non-robustness of outcomes in situations resembling the dictator game indicates that one should be very careful in generalizing the results found in these situations to other games. Testing theories of other-regarding preferences in dictator games is a bit like testing the laws of gravity with a table tennis ball. In both situations, minor unobserved distortions can have large effects. Therefore, we believe that it is necessary to show that the same motivational forces that are inferred from dictator games are also behaviorally relevant in economically more important games. One way to do this is to apply the theories that were constructed on the basis of dictator game experiments to predict outcomes in other games. With the exemption of Andreoni and Miller (2002) this has not yet been done.

Andreoni and Miller (2002) estimate utility functions based on the results of their dictator game experiments and use them to predict cooperative behavior in a standard public goods game. They predict behavior in period one of these games, where cooperation is often quite high, rather well. However, their predictions differ greatly from final period outcomes, where cooperation is typically very low. In our view, the low cooperation rates in the final period of repeated public good games constitutes a strong challenge for models that rely exclusively on altruistic or surplus-maximizing preferences. Why should a subject with a stable preference for others' payoffs or for those of the whole group contribute much less in the final period compared to the first period? Models of inequity aversion and intention based or type based reciprocity models provide a plausible explanation for this behavior. All of these models predict that fair subjects make their cooperation contingent on the cooperation of others. Thus, if the fair subjects realize that there are sufficiently many selfish decisions in the course of a public goods experiment, they cease to cooperate as well (see also Section 5 below).

[26] In Cox's experiment, both players had an endowment of 10 and the Allocator could transfer his endowment to the Receiver, where the experimenter tripled the transferred amount. The Receiver made no choice.

[27] In Eichenberger and Oberholzer-Gee (1998), almost 90 percent of the subjects gave nothing. In Hoffman et al. (1994), 64 percent gave nothing and 19 percent gave between 1 and 10 percent. In List and Cherry subjects earned their endowment in a quiz. Then they played the DG. Roughly 90 percent of the Allocators transferred nothing to the Receivers.

4.3. Revenge versus inequity reduction

Subjects with altruistic and quasi-maximin preferences do not take actions that reduce other subjects' payoffs; this phenomenon, however, is frequently observed in many important games. Models of inequity aversion account for this by assuming that the payoff reduction is motivated by a desire to reduce disadvantageous inequality. In models of intention based or type based reciprocity subjects punish if they observe an action that is perceived to be unfair or that reveals that the opponent is spiteful. In these models players want to reduce the opponent's payoff irrespective of whether they are better or worse off than the opponent and irrespective of whether they can change income shares or income differences. Furthermore, intention based theories predict that there will be no punishment in games in which no intention can be expressed. Therefore, a clean way to test for the relevance of intentions is to conduct control treatments in which choices are made through a random device or through some neutral and disinterested third party.

Blount (1995) was the first who applied this idea to the ultimatum game. Blount compared the rejection rate in the usual UG to the rejection rates in ultimatum games in which either a computer generated a random offer or a third party made the offer. Because a low offer can neither be attributed to the greedy intentions of the proposer in the random offer condition nor in the third party condition, intention based theories predict a rejection rate of zero in these conditions, while theories of inequity aversion still allow for positive rejection rates. Levine's theory is also consistent with positive rejection rates in these conditions, but his theory predicts a decrease in the rejection rate relative to the usual condition, because low offers made by humans reveal that the type who made the offer is spiteful which can trigger a spiteful response. Blount indeed observes a significant and substantial reduction in the acceptance thresholds of the responders in the random offer condition but not in the third party condition. Thus, the result of the random offer condition is consistent with intention and type based model, while the result of the third party condition is inconsistent with the motives captured by these models. Yet, these puzzling results may be due to some problematic features in Blount's experiments.[28] Subsequently, Offerman (1999) and Falk, Fehr and Fischbacher (2000b) conducted further experiments with offers generated by a random mechanism but without the other worrisome features in Blount. In particular, the responders knew that a rejection affects the payoff of a real, human "proposer" in these experiments. Offerman finds that subjects are 67 percent more likely to reduce the opponent's payoff when the opponent made an intentional low offer compared to a situation where a computer made the low offer.

[28] Blount's results may be affected by the fact that subjects (in two of three treatments) had to make decisions as a proposer *and* as a responder before they knew their actual roles. After subjects had made their decisions in both roles, the role for which they received payments was determined randomly. In one of Blount's treatments deception was involved. Subjects believed that there were proposers, although the experimenters in fact made the proposals. All subjects in this condition were "randomly" assigned to the responder role. In this treatment subjects also were not paid according to their decisions but they received a flat fee instead.

Falk, Fehr and Fischbacher (2000b) conducted an experiment, invented by Abbink, Irlenbusch and Renner (2000), that simultaneously allows for the examination of positive and negative reciprocity. In this game player A can give player B any integer amount of money $g \in [0, 6]$ or, alternatively, she can take away from B any integer amount of money $t \in [1, 6]$. In case of $g > 0$ the experimenter triples g so that B receives $3g$. If player A takes away t, player A gets t and player B loses t. After player B observes g or t, she can pay A an integer reward $r \in [0, 18]$ or she can reduce A's income by making an investment $i \in [1, 6]$. A reward transfers one money unit from B to A. An investment i costs B exactly i but reduces A's income by $3i$. This game was played in a random choice condition and in a human choice condition. It turns out that when the choices are made by a human player A, players B invest significantly more into payoff reductions for all $t \in [1, 6]$. However, as in Blount and Offerman, payoff reductions also occur when a random mechanism determines a hurtful choice.

Kagel, Kim and Moser (1996) provide further support that intentions play a role for payoff-reducing behavior. Subjects bargained over 100 chips in an UG in their experiments. They conducted several treatments that varied the money value of the chips and the information provided about the money value. For example, the proposers received three times more money per chip than the responders in one treatment, i.e., the equal money split required the responders to receive 75 chips. If the responders knew that the proposers were aware of the different money values of the chips, they rejected unequal money splits much more frequently than if the responders knew that the proposers did *not* know the different money values of the chips. Thus, knowingly unequal proposals were rejected at higher rates than unintentional unequal proposals.

Another way to test for the relevance of intention based or type based punishments is to examine behavior in the following two situations [Brandts and Sola (2001), Falk, Fehr and Fischbacher (2003)]. In one treatment, the proposer in a \$10 ultimatum game can choose between an offer of (5, 5) and an offer of (8, 2). In the other treatment the proposer can choose between (8, 2) and (10, 0). If responders do not care about whether the proposer has unfair intentions or is an unfair type, the rejection rate of the (8, 2) offer should be the same across both treatments. However, the information conveyed about the proposer's intention or type is very different across treatments. In the treatment where (5, 5) is the alternative to (8, 2), a proposal of (8, 2) is very likely to indicate that the proposer has unfair intentions or is an unfair type. This information is not conveyed by the (8, 2) proposal if the alternative is the (10, 0) proposal. Thus, if the responders care about the proposer's intention or type, the rejection rate for the (8, 2) offer should be higher in the case where (5, 5) is the available alternative. This prediction is nicely met by the data in Falk, Fehr and Fischbacher (2003): if (5, 5) is the alternative, 45% of the responders reject the (8, 2) offer, while if (10, 0) is the alternative, only 9% of the (8, 2) offers are rejected.

Finally, the relevance of intention based or type based punishments can also be examined by ruling out egalitarian motives as follows: If punishment keeps the relative payoff share or the payoff difference constant or even increases them, egalitarian motives, as modeled by Bolton and Ockenfels and Fehr and Schmidt, predict zero punishment.

Falk, Fehr and Fischbacher (2000a) report the results of ultimatum games that have this feature. In the first (standard) treatment of the ultimatum game the proposers could propose a $(5, 5)$ or an $(8, 2)$ split of the surplus (the first number represents the proposer's payoff). In case of rejection, both players received zero. In the second treatment, the proposers had the same options but a rejection now meant that the payoff was reduced for both players by 2 units. The theory of Bolton and Ockenfels and of Fehr and Schmidt predict, therefore, that there will be no rejections in the second treatment while intention based and type based models predict that rejections will occur. It turns out that the rejection rate of the $(8, 2)$ offer is 56 percent in the first and 19 percent in the second treatment. Thus, roughly one third (19/56) of the rejections are consistent with a pure taste for punishment as conceptualized in intention and type based models.[29] This evidence also suggests that payoff consequences alone are a determinant of the responder's rejection behavior. This conclusion is also supported by the results in Blount (1995) and Falk, Fehr and Fischbacher (2003), who report a significant number of rejections even if a third party makes the offer (as in Blount) or if the proposer is forced to make the $(8, 2)$ offer [as in Falk, Fehr and Fischbacher (2003)].

Taken together, the evidence from Blount (1995), Kagel, Kim and Moser (1996), Offerman (1999), Brandts and Sola (2001) and Falk, Fehr and Fischbacher (2000a, 2000b, 2003) supports the view that subjects want to punish unfair intentions or unfair types. Although the evidence provided by the initial study of Blount was mixed, the subsequent studies indicate a clear role of these motives. However, the evidence is also consistent with the view that egalitarian motives play a non-negligible role.

4.4. Does kindness trigger rewards?

Do intention and type based theories of fairness fare equally well in the domain of rewarding behavior? It turns out that the evidence in this domain is much more mixed. Some experimental results suggest that these motives seldom affect rewarding behavior. Other results indicate some minor role, and a few papers find an unambiguous positive effect of intention or type based reciprocity.

Intention based theories predict that people are generous only if they have been treated kindly, i.e., if the first-mover has signaled a fair intention. Levine's theory is similar in this regard because generous actions are more likely if the first mover is an altruistic type. However, in contrast to the intention based approaches, Levine's approach is also compatible with unconditional giving if it is sufficiently surplus-enhancing.

Neither intention nor type based reciprocity can explain positive transfers in the dictator game. Moreover, Charness (1996), Bolton, Brandts and Ockenfels (1998), Offerman (1999), Cox (2000) and Charness and Rabin (2002) provide further evidence that intentions do not play a big role for rewarding behavior. Charness (1996) conducted gift

[29] Ahlert, Crüger and Güth (1999) also report a significant amount of punishment in ultimatum games where the responders cannot change the payoff difference. However, since they do not have a control treatment it is not possible to say something about the relative importance of this kind of punishment.

exchange games in a random choice condition where a random device determined the proposer's decision and a human choice condition where the proposer made the choice. Intention based theories predict that the responders will not put forward more than the minimal effort level in the random choice condition, irrespective of the wage level, because high wage offers are due to chance and not to kind intentions. Higher wages in the human choice condition indicate a higher degree of kindness and, therefore, a positive correlation between wages and effort is predicted. Levine's theory allows, in principle, for a positive correlation between wages and effort in both conditions, because an increase in effort benefits the proposer much more than it costs the responder. However, the correlation should be much stronger in the human choice condition due to the type-revealing effect of high wages. Charness finds a significantly positive correlation in the random choice condition. Effort in the human choice condition is only slightly lower at low wages and equally high at high wages. This indicates, if anything, only a minor role for intention and type driven behavior. The best interpretation is probably that inequity aversion or quasi-maximin preferences induce non-minimal effort levels in this setting. In addition, negative reciprocity kicks in at low wages which explains the lower effort levels in the human choice condition.

Cox (2004) tries to isolate rewarding responses in the context of a trust game by using a related dictator game as a control condition. Cox first conducts the usual trust game, which provides him with a baseline level of responder transfers back to the proposer. To isolate the relevance of intention driven responses, he then conducts a dictator game in which the distribution of endowments is identical to the distribution of material payoffs after the proposers' choices in the trust game. Thus, the responders face exactly the same distributions of material payoffs in both the trust game and in the dictator game, but the proposers intentionally caused this distribution in the trust game, while the experimenter predetermined the distribution in the dictator game. The motive for rewarding kindness can, therefore, play no role in the dictator game and both intention based theories as well as Levine's theory predict that responders transfer nothing back. If one takes into account that some transfers in the dictator game are likely to be driven by inequity aversion, the difference between the transfers in the dictator game and those in the trust game measure the relevance of intention based theories. Cox's results indicate that transfers in the trust game are roughly by one-third higher than in the dictator game. Thus, intention based reciprocity plays a significant, but not the dominant, role.

The strongest evidence against the role of intentions comes from Bolton, Brandts and Ockenfels (1998). They conducted sequential social dilemma experiments that are akin to a sequentially played Prisoners' Dilemma. In one condition, the first movers could make a kind choice relative to a reference choice. The kind choice implied that – for any second mover choice – the second mover's payoff increased by 400 units at a cost of 100 for the first mover. Then the second mover could take costly actions in order to reward the first mover. In a control condition, the first mover had to make the reference choice, i.e. he could not express any kind intentions. It turns out that second movers reward the first movers even more in the control condition. Although this difference is not significant, the results clearly suggest that intention-driven rewards play no role in this experiment.

The strongest evidence in favor of intentions comes from the moonlighting game of Falk, Fehr and Fischbacher (2000b) described in the previous subsection. They find that players B send back significantly more money in the human choice condition for *all* positive transfers of player A. Moreover, the difference between the rewards in the human choice condition and the random choice condition are also quantitatively important. A recent paper by McCabe, Rigdon and Smith (2003) also reports evidence in favor of intention driven positive reciprocity. They show that if the first-mover makes a kind decision, two-thirds of the second movers also make kind decisions, while only one-third of the second movers make the kind decision if the first mover is forced to make the kind choice.

In the absence of the evidence provided by Falk, Fehr and Fischbacher (2000b) and McCabe, Rigdon and Smith (2003), one would have to conclude that the motive to reward good intentions or fair types is (at best) of minor importance. However, in view of the relatively strong results in the final two papers, it seems wise to be more cautious and to wait for further evidence. Nevertheless, the bulk of the evidence suggests that inequity aversion and efficiency seeking are more important than intention or type based reciprocity in the domain of kind behavior.

4.5. Maximin preferences

The papers by Charness and Rabin (2002) and by Engelmann and Strobel (2004) show that a substantial percentage of the Allocators in multi person dictator games care for the material payoff of the least well-off group member. The relevance of the maximin motive in these games is, for example, illustrated by the dictator game taken from Engelmann and Strobel (2004), in which player B is the dictator who can choose among the following three allocations: (11, 12, 2), (8, 12, 3) and (5, 12, 4). Both surplus maximization as well as the theories by Bolton and Ockenfels and Fehr and Schmidt predict that B will choose the first allocation in this game, whereas a player with maximin preferences chooses the third allocation. In fact, 53% of the players chose the third and only 27% chose the first allocation, indicating the importance of the maximin motive in these games. This game also shows, however, that nonlinear forms of inequity aversion may come close to maximin preferences. This is, for example, the case if the marginal disutility from advantageous inequality strongly increases in the amount of inequality. In this case also an inequity averse player may prefer the third allocation.

Although the maximin motive plays a prominent role in multi person dictator games, there are several papers that cast doubt on the relevance of this motive in strategic games. A salient example is the three-person experiment of Güth and van Damme (1998) that combines an ultimatum and a dictator game. Recall from Section 4.1 that the proposer has to make a proposal (x, y, z) on how to allocate a given sum of money between himself and players two and three in this game. Then the responder has to decide whether to accept or reject the proposal. If he accepts, the proposal is implemented, otherwise all players get zero. Player 3 remains inactive and cannot affect the final outcome. Güth and van Damme report that the proposer allocates only marginal amounts to the passive Receiver and the responder's rejection behavior is seemingly unaffected by the low amounts allocated to the passive Receiver. These observations contradict maximin preferences while they are consistent with the linear Fehr and Schmidt model and the model by Bolton and Ockenfels [see Bolton and Ockenfels (2000) and Section 4.1].

Frechette, Kagel and Lehrer (2003) provide another striking example of the neglect of the weak player's interests in strategic interactions. One player in a group of five can make a proposal on how to allocate a fixed sum of money among the five players in their experiments. Then the players vote on the proposal under the majority rule, i.e., the support of 3 players is sufficient to implement the proposal. In 65% of the cases, the proposals implied that two of the five players received a zero payoff, completely neglecting the interests of members that are not part of the winning coalition. Moreover, such proposals received the support of the majority in most cases. Thus, maximin preferences seem do play little role in this environment.

Finally, the experiments by Okada and Riedl (2005) also indicate that maximin preferences are of little importance in strategic games. In their three person experiments, a proposer could propose an allocation (x, y) to one responder or an allocation (x, y, z) to two responders. If he proposes forming a three person coalition, i.e., making an offer to two responders, the total amount to be distributed among the three players is 3000 points whereas if he only proposes a two person coalition, the total amount to be distributed is an element of the set $\{1200, 2100, 2500, 2800\}$. However, both responders have to accept the proposal (x, y, z) in the case of a three person coalition, whereas only a single responder has to accept the proposal (x, y) in the case of the two person coalition. If one of the responders rejects a proposal, all players receive zero. If only the two person coalition is proposed, the third player automatically receives a payoff of zero. Therefore, proposers with maximin preferences that dominate their self-interest will always propose a three person coalition with $x = y = z$, regardless of the amount available for the two person coalition. In the case of quasi maximin preferences in the sense of Charness and Rabin (2002) the "efficiency" motive puts even more weight on this proposal because the grand coalition produces a larger surplus.

Okada and Riedl report that 90% of the proposer's went for the two-person coalition when the total amount available for the two person coalition is 2500 or 2800. If the available amount for the small coalition is only 2100 still about 40% of the proposers went for the two person coalition. The grand coalition is favored by almost all proposers only in those cases when the small coalition became very inefficient because the available amount shrank to 1200. These regularities in proposers' behavior are predicted by the Fehr and Schmidt and the Bolton and Ockenfels model of inequity aversion.

Given the evidence from the above mentioned papers, it remains to be shown that maximin preferences play a role in strategic games. It seems that dictator games put players in a different frame of mind than strategic games, where the players can mutually affect each others' payoffs. Players in strategic games seem to be much more willing to neglect weak players' interests and to demand fairness or equity mainly for themselves, whereas the dictators seem to care a lot for the interests of the worst-off players in dictator games. This insight may also help in determining when the maximin motive plays a role in naturally occurring environments. In a competitive environment or in an environment where the players view each other as agents behaving strategically, the maximin motive is likely to be not important. However, the maximin motive may be more or even highly relevant in the context of charitable giving or in the context of referenda or elections with a large number of people, where strategic voting is unlikely to occur.

4.6. Preferences for honesty

Three recent papers indicate that a sizeable share of the subjects also care for honesty. Brandts and Charness (2003) show that subjects are more willing to correct unfair outcomes if these outcomes were reached through a lie. Charness and Dufwenberg (2004) show that the second mover in a sequentially played prisoners' dilemma is more willing to reciprocate trusting first mover behavior if the second mover could send a promise to reciprocate before the sequential prisoners' dilemma started. Gneezy (2005) provides direct evidence for dishonesty aversion in a simple but clever dictator game set up as follows: player B is the dictator who can choose among two alternative actions: action a implements payoff allocation $(5, 6)$ and action b implements allocation $(6, 5)$. However, only player A knows the monetary consequences of the two available actions while player B knows nothing about them. Before B chooses, A must send one of two messages to B. Message a is the honest message. It says: "Action a will earn you more money than action b." Message b is the dishonest message. It says: "Action b will earn you more money than action a." Gneezy shows that the vast majority of player B follows A's message, i.e., they choose the action that gives them the higher payoff according to the message. In addition, the vast majority of players A believes that players B will behave in this way. Thus, most players A believed correctly that they could mislead player B by being dishonest. A could gain \$1 at the cost of B by lying.

Gneezy reports that only 36% of the players A were dishonest in the game described above. Moreover, if the monetary consequences of action a were changed to $(5, 15)$, such that A could gain \$1 by imposing a loss of \$10 on B, the lying rate further decreased to 17%. Finally, if action a implied the allocation $(5, 15)$ whereas action b implied the allocation $(15, 5)$, player A could gain \$10 by being dishonest which imposed a cost of \$10 on player B. In this case, 52% of the players A send the wrong message. In a dictator game control experiment in which A had to choose between the allocations mentioned above, player A was much more willing to choose the allocation that favored him. If the alternatives were $(5, 6)$ versus $(6, 5)$ 66% of the A's chose the second allocation. Likewise, if the alternatives were $(5, 15)$ versus $(15, 5)$ 90% of the A's chose the second allocation. Thus, if the favorable outcome could be achieved without a lie, much more players A were willing to choose according to their self interest which documents neat evidence in favor of dishonesty aversion. In addition, dishonesty aversion is affected by the private gains from lying and by the harm imposed on the victim of the lie.

4.7. Summary and outlook

Although most models of other-regarding preferences discussed in Section 3 are just a few years old, the discussion in this section shows that there is already a fair amount of evidence that sheds light on the merits and the weaknesses of the different models. This indicates a quick and healthy interaction between experimental research and the development of new theories. The initial experimental results discussed in Section 2 gave rise to a number of new theories which, in turn, have again been quickly subjected to careful and rigorous empirical testing. Although these tests have not yet led to conclusive results regarding the relative importance of the different motives many important and interesting insights have been obtained. In our view the main results can be summarized as follows:

(1) The average payoff in the group is an empirically invalid reference standard for explaining individual punishment behavior. Approaches that rely on this comparison standard cannot explain important aspects of punishment behavior. Evidence from the Third Party Punishment Game and other games indicates that many subjects compare themselves with other people in the group and not just to the group as a whole or to the group average.

(2) Pure revenge as captured by intention based and type based reciprocity models is an important motive for punishment behavior. Since pure equity models do not capture this motive they cannot explain a significant amount of punishment behavior. While the inequality of the payoffs also is a significant determinant of payoff reducing behavior, the revenge motive seems to be more important in bilateral interactions as illustrated in those experiments where responses to a computerized first-mover choice are compared to the responses to human first mover choices.

(3) In the domain of kind behavior, the motives captured by intention or type based models of reciprocity seem to be less important than in the domain of payoff-reducing behavior. Several studies indicate that inequity aversion or maximin preferences play a more important role here.

(4) In dictator games, a significant share of the subjects prefers allocations with a higher group payoff and a higher inequality within the group over allocations with a lower group payoff and a lower inequality. However, this motive only dominates among economists, while the clear majority of non-economists is willing to sacrifice substantial amounts of the group payoff in order to ensure more equality within the group. Moreover, the relative importance of the motive to increase the group's payoff has yet to be determined for strategic games.

(5) In multi person dictator games, a large share of the subjects cares for the least well-off player's material payoff. However, evidence from several strategic games casts doubt on the relevance of this motive in strategic interactions.

(6) Some recent papers report that a substantial share of the subjects has indicated a preference for honesty.

Which model of other-regarding preferences does best in the light of the data, and which should be used in applications to economically important phenomena? We believe that it is too early to give a conclusive answer to these questions. There is a large amount of heterogeneity at the individual level and any model has difficulties in explaining the full diversity of the experimental observations. The above summary provides, however, some guidance for applied research. In addition to the summary statements above, we believe that the most important heterogeneity in strategic games is the one between purely selfish subjects and subjects with a preference for fairness or reciprocity.

Within the class of inequity aversion models, the evidence suggests that the Fehr and Schmidt model outperforms or does at least as well as the Bolton and Ockenfels model in almost all games considered in this paper. In particular, the experiments discussed in Section 4.1 indicate that people do not compare themselves with the group as a whole but rather with other individuals in the group. The group average is less compelling as a yardstick for measuring equity than are differences in individual payoffs. However, the Fehr and Schmidt model clearly does not recognize the full heterogeneity within the class of fair-minded individuals. Section 4.4 makes it clear that an important part of payoff-reducing behavior is not driven by the desire to reduce payoff-differences, but by the desire to reduce the payoff of those who take unfair actions or reveal themselves as unfair types. The model therefore cannot explain punishing behavior in situations where payoff differences cannot be changed by punishing others. Fairness models exclusively based on intentions [Rabin (1993), Dufwenberg and Kirchsteiger (2004)] can, in principle, account for this type of punishment. However, these models have other undesirable features, including multiple, and very counterintuitive, equilibria in many games and a very high degree of complexity due to the use of psychological game theory. The same has to be said about the intention based theory of Charness and Rabin (2002). It is also worthwhile to point out that intention based reciprocity models cannot explain punishment in the third party punishment game because they are based on bilateral notions of reciprocity. The third party was not treated in an unkind way in this game and will therefore never punish. Falk and Fischbacher (2006) do not share these problems of pure intention models. This is due to the fact that they incorporate equity as a global reference standard. Their model shares however, the complexity costs of psychological game theory.

Even though none of the available theories of other-regarding preferences takes the full complexity of motives at the individual level into account, some theories may allow for better approximations than others, depending on the problem at hand. If, for example, actors' intentions constitute a salient dimension of an economic problem, consideration of some form of intention based reciprocity might be advisable, despite the complexity costs involved. Or, to give another example, a type based reciprocity model in the spirit of Levine (1998) may provide a plausible explanation for third party punishment. The essence of third party punishment is that the punisher is not directly hurt but nevertheless punishes a norm violation. While bilateral notions of reciprocity are unable to explain this kind of punishment type based models provide a natural explanation because norm violations are type revealing. However, the most important message of the evidence presented in Section 2 clearly is that there are many important economic problems where the self-interest theory is unambiguously, and in a quantitatively important way, refuted. Therefore, in our view, it is certainly not advisable to only consider the self-interest model, but to combine the self-interest assumption with the other-regarding motive that is likely to be most important in the problem at hand.

5. Economic applications

5.1. Cooperation and collective action

Free-riding incentives are a pervasive phenomenon in social life. Participation in collective action or in industrial disputes, collusion among firms in oligopolistic markets, the prevention of negative environmental externalities, workers' effort choices under team-based compensation schemes or the exploitation of a common resource are typical examples. In these cases the free rider cannot be excluded from the benefits of collective actions or the public good although he does not contribute. In view of the ubiquity of cooperation problems in modern societies it is crucial to understand the forces shaping people's cooperation. In this section we will show that the neglect of other-regarding preferences may induce economists to largely misunderstand the nature of many cooperation problems. As we will see a key to the understanding of cooperation problems is again the interaction between selfish individuals and individuals with other-regarding preferences.

The impact of other-regarding preferences on cooperation can be easily illustrated for the case of reciprocal or inequity averse individuals. First, reciprocal subjects are willing to cooperate if they are sure that the other people who are involved in the cooperation problem will also cooperate. If the others cooperate – despite pecuniary incentives to the contrary – they provide a gift that induces reciprocal subjects to repay the gift, i.e., reciprocators are conditionally cooperative. Likewise, as we will show below, inequity averse individuals are also willing to cooperate if they can be sure that others cooperate. Second, reciprocal or inequity averse subjects are willing to punish free-riders because free-riders exploit the cooperators. Thus, if potential free-riders face reciprocators they have an incentive to cooperate to prevent being punished.

In the following we illustrate the first claim for the case of inequity averse subjects in a prisoners' dilemma who have utility functions as proposed by Fehr and Schmidt (1999). Table 1 presents the material payoffs in a prisoners' dilemma and Table 2 shows how inequity aversion transforms the material payoffs. Recall that in the two-player case the utility of player i is given by $U_i(x) = x_i - \alpha_i(x_j - x_i)$ if player i is worse off than player j ($x_j - x_i \geqslant 0$), and $U_i(x) = x_i - \beta_i(x_i - x_j)$ if player i is better off than

Table 1
Representation of prisoners' dilemma in terms of material payoffs

	Cooperate (C)	Defect (D)
Cooperate (C)	2, 2	0, 3
Defect (D)	3, 0	1, 1

Table 2
Utility representation of prisoners' dilemma if players are inequity averse

	Cooperate (C)	Defect (D)
Cooperate (C)	2, 2	$0 - 3\alpha, 3 - 3\beta$
Defect (D)	$3 - 3\beta, 0 - 3\alpha$	1, 1

player j ($x_i - x_j \geqslant 0$). For simplicity, Table 2 assumes that both players have the same preferences so that α and β are identical across players.

Table 1 illustrates that if player 2 (the column player) is expected to cooperate, player 1 (the row player) faces a choice between material payoff allocations (2, 2) and (3, 0). The utility of (2, 2) is $U_1(2, 2) = 2$ because there is no inequality. The utility of (3, 0), however, is $U_1(3, 0) = 3 - 3\beta$ because there is inequality that favors the row player. Therefore, player 1 will reciprocate the expected cooperation of player 2 if $\beta > 1/3$. If player 1 defects and player 2 cooperates the payoff of player 2 is $U_2(3, 0) = 0 - 3\alpha$; if player 2 defected instead the utility would be 1. This means that player 2 will always reciprocate defection because cooperating against a defector yields less money and more inequity. Table 2 shows that if $\beta > 1/3$, there are two equilibria: (cooperate, cooperate) and (defect, defect). In utility terms, inequality averse players no longer face a PD. Instead, they face a coordination or assurance game with one efficient and one inefficient equilibrium. If the players believe that the other player cooperates, it is rational for each of them to cooperate, too.

Inequity averse (and reciprocal) players are thus conditional cooperators. They cooperate in response to expected cooperation and defect in response to expected defection. Theories of other-regarding preferences which imply that subjects are conditionally cooperative are, therefore, also consistent with framing effects in the prisoners' dilemma. Ross and Ward (1996) have shown that players achieve higher cooperation rates if the Prisoners' Dilemma is called a "community game" instead of "Wallstreet game". Many people prematurely argue that these effects of framing on cooperation reflect players' irrationality. However, if the game is framed as "community game" it seems plausible that the players are more optimistic about the other players' cooperation, which induces them to cooperate more frequently than in the case were the game is framed as "Wallstreet game". Therefore, the impact of different frames on cooperation behavior is also consistent with the view that the players have stable other-regarding preferences but exhibit different expectations about others' behavior under different frames.

The transformation of the prisoners' dilemma into a coordination game in the presence of reciprocal or inequity averse players can explain one further fact. It has been shown dozens of times that communication leads to much higher cooperation rates in the prisoners' dilemma and in public good games [Sally (1995)]. If all subjects were completely selfish this impact of communication would be difficult to explain. If, however, the game in material terms is in fact a coordination game, communication allows the subjects to coordinate on the superior equilibrium.

If it is indeed the case that the actual preferences of the subjects transform cooperation games into coordination games, the self-interest hypothesis induces economists to fundamentally misperceive the cooperation problems. In view of the importance of this claim it is, therefore, desirable to have more direct evidence on this. Several studies provided evidence in favor of the existence of conditional cooperation during the last few years [Keser and van Winden (2000), Brandts and Schram (2001), Fischbacher, Gächter and Fehr (2001)]. There is a tricky causality issue involved in this question because a positive correlation between an individual's cooperation rate and the individual's belief about others' cooperation rate does not unambiguously prove the existence of conditional cooperation. Perhaps the individual first chooses how much to cooperate and the belief represents merely the rationalization of the chosen cooperation level. This problem has been overcome by Keser and van Winden in the context of a repeated public goods experiment and by Fischbacher, Gächter and Fehr (2001) in the context of a one-shot public goods experiment. Keser and van Winden (2000) show that many subjects adjust their cooperation in period t to move closer to last period's average cooperation rate. This finding suggests that subjects reciprocate to last period's average cooperation of the other group members. Fischbacher, Gächter and Fehr (2001) elicited so-called contribution schedules from their subjects. A contribution schedule stipulates a subject's contribution to every possible level of the average contribution of the other group members in a one-shot experiment. The parameters of the game ensured that a selfish subject will never contribute anything to the public good regardless of the average contribution of the other group members. The surplus maximizing contribution level was given at 20 which was identical to the maximum contribution.

The results of this study show that 50 percent of the subjects are willing to increase their contributions to the public good if the other group members' average contribution increases although the pecuniary incentives always implied full free-riding. The behavior of these subjects is consistent with models of reciprocity (or inequity aversion). However, a substantial fraction of the subjects (30 percent) are complete free-riders who free ride regardless of what the other group members do. 14 percent exhibit a hump-shaped response. They increase their cooperation rate in response to an increase in the average cooperation of others but beyond a cooperation level of 50% of the endowment they start decreasing their cooperation. Yet, taken together there are sufficiently many conditional cooperators such that an increase in the other group members' contribution level causes an increase in the contribution of the "average" individual.

The coexistence of conditional cooperators and selfish subjects has important implications. It implies, e.g., that subtle institutional details may cause large behavioral effects. To illustrate this assume that a selfish and an inequity averse subject are matched in the *simultaneous* prisoners' dilemma and that the subjects' type is common knowledge. Since the inequity averse subject knows that the other player is selfish he knows that the other will always defect. Therefore, the inequity averse player will also defect, i.e., (defect, defect) is the unique equilibrium. This result can be easily illustrated in Table 2 by setting the inequity aversion parameters α and β of one of the players equal to zero. Now consider the *sequential* prisoners' dilemma in which the selfish player first decides whether to cooperate or to defect. Then the reciprocal player observes what the first-mover did and chooses his action. In the sequential case the unique equilibrium outcome is that both players cooperate because the reciprocal second-mover will match the choice of the first-mover. This means that the selfish first-mover essentially has the choice between the (cooperate, cooperate)-outcome and the (defect, defect)-outcome. Since mutual cooperation is better than mutual defection the selfish player will also cooperate. Thus, while in the simultaneous prisoners' dilemma the selfish player induces the reciprocal player to defect, in the sequential prisoners' dilemma the reciprocal player induces the selfish player to cooperate in equilibrium. This example neatly illustrates how institutional details interact in important ways with the heterogeneity of the population.

Since there are many conditional cooperators the problem of establishing and maintaining cooperation involves the management of people's beliefs. If people believe that the others cooperate to a large extent, cooperation will be higher compared to a situation where they believe that others rarely cooperate. Belief-dependent cooperation can be viewed as a social interaction effect that is relevant in many important domains. For example, if people believe that cheating on taxes, corruption, or abuses of the welfare state are wide-spread, they are themselves more likely to cheat on taxes and are more willing to take bribes or to abuse welfare state institutions. It is therefore important that public policy prevents the initial unravelling of civic duties because, once people start to believe that most others engage in unlawful behavior the belief-dependency of individuals' cooperation behavior may render it very difficult to re-establish lawful behavior.

In an organisational context the problem of establishing cooperation among the members of the organisation also involves the selection of the "right" members. A few shirkers in a group of employees may quickly spoil the whole group. Bewley (1999), e.g., reports that personnel managers use the possibility to fire workers mainly as a means to remove "bad characters and incompetents" from the group and not as a threat to discipline the workers. The reason is that explicit threats create a hostile atmosphere and may even reduce the workers' generalised willingness to cooperate with the firm. Managers report that the employees themselves don't want to work together with lazy colleagues because these colleagues do not bear their share of the burden which is viewed as unfair. Therefore, the firing of lazy workers is mainly used to establish internal equity, and to prevent the unravelling of cooperation. This supports the view that conditional cooperation is also important inside firms.

The motivational forces behind conditional cooperation are also likely to shape the structure of social policies that aim at helping the poor [Bowles and Gintis (2000), Wax (2000), Fong, Bowles and Gintis (2005a, 2005b)]. The reason is that the political support for policies favoring the poor depends to a large extent on whether the poor are perceived as "deserving" or as "undeserving". If people believe that the poor are poor because they do not *want* to work hard the support for policies that help the poor is weakened because the poor are perceived as undeserving. If, in contrast, people believe that the poor try hard to escape poverty but that for reasons beyond their control they could not make it, the poor are perceived as deserving. This indicates that the extent to which people perceive the poor as deserving is shaped by reciprocity motives. If the poor exhibit good intentions, i.e., they try to contribute to society's output, or if they are poor for reasons that have nothing to do with their intentions, they are perceived as deserving. In contrast, if the poor are perceived as lacking the will to contribute to society's output, they are perceived as undeserving. This means that social policies that enable the poor to demonstrate their willingness to reciprocate the generosity of society will mobilise greater political support than social policies that do not allow the poor to exhibit their good intentions. Wax (2000) convincingly argues that an important reason for the popularity of President Clinton's 1996 welfare reform initiative was that the initiative appealed to the reciprocity of the people.

5.2. Endogenous formation of cooperative institutions

We argued above that the presence of a selfish subject will induce a reciprocal or inequity averse subject in the simultaneous prisoners' dilemma to defect as well. This proposition also holds more generally in the case of n-person public good games. It can be shown theoretically that even a small minority of selfish subjects induces a majority of reciprocal (or inequity averse) subjects to free-ride in simultaneous social dilemma games [Fehr and Schmidt (1999, Proposition 4)]. In an experiment with anonymous interaction subjects do of course not know whether the other group members are selfish or reciprocal but if they interact repeatedly over time they may learn the others' types. Therefore, one would expect that over time cooperation will unravel in (finitely repeated) simultaneous public goods experiments. This unravelling of cooperation has indeed been observed in dozens of experiments [Ledyard (1995)].

This raises the question of whether there are social mechanisms that can prevent the decay of cooperation. A potentially important mechanism is social ostracism and peer pressure stemming from reciprocal or inequity averse subjects. Recall that these subjects exhibit a willingness to punish unfair behavior or mitigate unfair outcomes and it is quite likely that co-operating individuals view free-riding as very unfair. To examine the willingness to punish free-riders and the impact of punishment on cooperation Fehr and Gächter (2000) introduced a punishment opportunity into a public goods game. In their game there are two stages. Stage one consists of a linear public good game in which the dominant strategy of each selfish player is to free-ride completely although the socially optimal decision requires to contribute the whole endowment to the public good. In stage two, after every player in the group has been informed about the contributions of each group member, each player can assign up to ten punishment points to each of the other group members. The assignment of one punishment point reduces the first-stage income of the punished subject, on the average, by three points but it also reduces the income of the punisher. This kind of punishment mimics an angry group member scolding a free-rider, or spreading the word so the free-rider is ostracised – there is some cost to the punisher, but a larger cost to the free-rider. Note that since punishment is costly for the punisher, the self-interest hypothesis predicts zero punishment. Moreover, since rational players will anticipate this, the self-interest hypothesis predicts no difference in the contribution behavior between a public goods game without punishment and the game with a punishment opportunity. In both conditions zero contributions are predicted.

The experimental evidence completely rejects this prediction.[30] In contrast to the game without a punishment opportunity, where cooperation declines over time and is close to zero in the final period, the punishment opportunity causes a sharp jump in cooperation. Moreover, in the punishment condition there is a steady increase in contributions until almost all subjects contribute their whole endowment. This sharp increase occurs because free-riders often get punished, and the less they give, the more likely punishment is. Cooperators seem to feel that free-riders take unfair advantage of them and, as a consequence, they are willing to punish the free-riders. This induces the punished free-riders to increase cooperation in the following periods. A nice feature of this design is that the actual rate of punishment is very low in the last few periods – the mere threat of punishment, and the memory of its sting from past punishments, is enough to induce potential free-riders to cooperate.

The punishment of free riders in repeated cooperation experiments has also been observed in Yamagishi (1986), Ostrom, Walker and Gardner (1992), Masclet (2003), Carpenter, Matthews and Ong'ong'a (2004), and Anderson and Putterman (2006). In almost all studies the authors report that the possibility to punish causes a strong increase in cooperation rates. Moreover, this increase in cooperation due to punishment opportunities can even be observed in one-shot experiments where the groups are randomly mixed in every period such that no subject ever interacts twice with another subject [Fehr and Gächter (2002)].

More recently, Gürerk, Irlenbusch and Rockenbach (2006) examined whether subjects prefer an institutional environment in which they can punish each other as in Fehr and Gächter (2000) or whether they prefer an institution that rules out mutual punishment by individual actors. In this experiment subjects interacted for a total of 30 periods

[30] In the experiments subjects first participate in the game without a punishment opportunity for ten periods. After this they are told that a new experiment takes place. In the new experiment, which lasts again for ten periods, the punishment opportunity is implemented. In both conditions subjects remain in the same group for ten periods and they know that after ten periods the experiment will be over.

and the final period was known by every participant. At the beginning of each period each of 12 subjects had to indicate the preferred institution. Then the subjects who choose the punishment institution played the public goods game with a subsequent punishment stage whereas the subjects who preferred the institution without punishment just played the public goods game. Regardless of how many subjects joined an institution, the members of the institution as a whole earned always 1.6 tokens from each token contributed to the public good. This feature has the important consequence that for larger groups it is much more difficult to sustain cooperation because the free riding incentive is much stronger. For example, if only 2 subjects join an institution each token that is contributed by a subject provides a private return of 0.8 tokens and a group return of 1.6 tokens because the other subject also earns 0.8 tokens from the contribution. However, if 10 subjects join an institution, the group's overall return from a one unit contribution is still 1.6 tokens, that is, each member of the institution earns only 0.16 tokens from the contribution.

Despite the fact that larger groups faced much stronger free-riding incentives Gürerk et al. report convergence to a single institution. At the beginning roughly 2/3 of the subjects preferred to interact without the mutual punishment opportunity. However, after a few periods cooperation rates became very low under this institution which induced subjects to switch to the punishment institution. In fact, over time the percentage of subjects who preferred the punishment institution rose to more than 90 percent from period 20 onwards and remained stable till the final period. Moreover, from period 15 onwards cooperation rates were very close to 100% under the punishment institution whereas under the no-punishment institution cooperation collapsed completely. Although punishment was frequent in the early periods of the punishment institution because many self-interested subjects also joined and attempted to free ride, little or no punishment was necessary to sustain cooperation in the second half of the experiment. The mere threat of punishment was sufficient to maintain nearly perfect cooperation levels.

These results are indeed remarkable because they can be viewed as the laboratory equivalent of the formation of a proto-state. One of the puzzles of the evolution of cooperation concerns the question why humans are such an extremely cooperative species. Humans seem to be the only species that is able to establish cooperation in large groups of *genetically unrelated strangers*. There are several other species (bees, ants, termites, etc.) which show cooperation in large group of genetically closely related individuals but among humans the average degree of relatedness of individual members of a modern society is close to zero. Of course, in modern societies cooperation is based on powerful institutions (impartial police, impartial judges, etc.) that punish norm violations. However, the existence of these institutions is itself an evolutionary puzzle because their existence constitutes a public good in itself. The experiments by Gürerk et al. suggest that deep seated inclinations to punish free riders and the ability to understand the cooperation enhancing effects of punishment institutions are part of an explanation of these institutions.

5.3. How fairness, reciprocity and competition interact

The self-interest model fails to explain the experimental evidence in many games in which only a few players interact, but it is very successful in explaining the outcome of competitive markets. It is a well-established experimental fact that in a broad class of market games prices converge to the competitive equilibrium [Smith (1982), Davis and Holt (1993)]. This result holds even if the resulting allocation is very unfair by any notion of fairness. Thus, the question arises: If so many people resist unfair outcomes in, say, the ultimatum game or the third party punishment game, why don't they behave the same way when there is competition among the players?

To answer this question we consider the following ultimatum game with proposer competition, that was conducted by Roth et al. (1991) in four different countries. There are $n-1$ proposers who simultaneously offer a share $s_i \in [0,1]$, $i \in \{1,\ldots,n-1\}$, to one responder. The responder can either accept or reject the highest offer $s^{\max} = \max_i\{s_i\}$. If there are several proposers who offered $s^{\max}$, one of them is selected at random with equal probability. If the responder accepts $s^{\max}$, her monetary payoff is $s^{\max}$ and the successful proposer earns $1-s^{\max}$, while all the other proposers get 0. If the responder rejects, everybody gets a payoff of 0.

The prediction of the self-interest model is straightforward: All proposers will offer $s=1$ which is accepted by the responder. Hence, all proposers get a payoff of zero and the monopolistic responder captures the entire surplus. This outcome is clearly very unfair, but it describes precisely what happened in the experiments. After a few periods of adaptation $s^{\max}$ was very close to 1 and all the surplus was captured by the responder. Moreover, this pattern was observed across several different cultures indicating that cultural differences in preferences or beliefs have little impact on behavior under proposer competition.[31]

This result is remarkable. It does not seem to be more fair that one side of the market gets all of the surplus in this setting than in the standard ultimatum game. Why do the proposers let the responder get away with it? The reason is that preferences for fairness or reciprocity cannot have any effect in this strategic setting. To see this, suppose that each of the proposers strongly dislikes receiving less than the responder. Consider proposer i and let $s' = \max_{j\neq i}\{s_j\}$ be the highest offer made by his fellow proposers. If proposer i offers $s_i < s'$, then his offer has no effect and he will get a monetary payoff of 0 with certainty. Furthermore, he cannot prevent that the responder gets s' and that one of the other proposers gets $1-s'$, so he will suffer from getting less than these two. However, if he offers a little bit more than s', say $s'+\varepsilon$, then he will win the competition, receive a positive monetary payoff, and reduce the inequality between himself

[31] The experiments were conducted in Israel, Japan, Slovenia, and the U.S. In all experiments, there were 9 proposers and 1 responder. Roth et al. also conducted the standard ultimatum game with one proposer in these four countries. They did find some small (but statistically significant) differences between countries in the standard ultimatum game which may be attributed to cultural differences. However, there are no statistically significant differences between countries for the ultimatum game with proposer competition.

and the responder. Hence, he should try to overbid his competitors. This process drives the share that is offered by the proposers up to 1. There is nothing the proposers can do about it even if all of them have a strong preference for fairness. We prove this result formally in Fehr and Schmidt (1999) for the case of inequity averse players, but the same result is also predicted by the approaches of Bolton and Ockenfels (2000), Levine (1998) and Falk and Fischbacher (2006).

The ultimatum game with responder competition provides further insights into the interaction between fair minded and selfish actors. Instead of one responder there are now two competing responders and only one proposer. When the proposer has made his offer the two responders simultaneously accept or reject the offer. If both accept, a random mechanism determines with probability 0.5 which one of the responders will get the offered amount. If only one responder accepts he will receive the offered amount of money. If both responders reject, the proposer and both responders receive nil.

The ultimatum game with responder competition can be interpreted as a market transaction between a seller (proposer) and two competing buyers (responders) who derive the same material payoff from an indivisible good. Moreover, as the parties' pecuniary valuations of the good are public information there is a known fixed surplus and the situation can be viewed as a market in which the contract (quality of the good) is enforced exogenously.

If all parties are selfish, competition among the responders does not matter because the proposer is predicted to receive the whole surplus in the bilateral case already. Adding competition to the bilateral ultimatum game has therefore no effect on the power of the proposer. It is also irrelevant whether there are two, three or more competing responders. The self-interest hypothesis thus implies a very counterintuitive result, namely, that increasing the competition among the responders does not affect the share of the surplus that the responders receive. Fischbacher, Fong and Fehr (2002) tested this prediction by conducting ultimatum games with one, two and five responders under a random matching protocol for 20 periods.[32] In every period the proposers and the responders were randomly re-matched to ensure the one-shot nature of the interactions. All subjects knew that after period 20 the experiment would end.

The results of the experiment show that competition has a strong impact on behavior. In the bilateral case the average share is – except for period 1 – always close to 40 percent. Moreover, the share does not change much over time. In the final period the responders still appropriate slightly more than 40 percent of the surplus. In the case of two responders the situation changes dramatically, however. Already in period 1 the responders' share is reduced by 5 percentage points relative to the bilateral case. Moreover, over time responder competition induces a further substantial reduction of the share and in the final period the share is even below 20 percent. Thus, the addition of just one more responder has a dramatic impact on the share of the responders. If we add three

[32] See also Güth, Marchand and Rulliere (1997) and Grosskopf (2003) for experiments with responder competition.

additional responders the share goes down even further. From period 3 onwards it is below 20 percent and comes close to 10 percent in the second half of the session.[33]

The responders' share decreases when competition increases because the rejection probability of the responders declines when there are more competing responders. These facts can be parsimoniously explained if one takes the presence of reciprocal or inequity averse responders into account. Recall that reciprocal responders reject low offers in the bilateral ultimatum game because by rejecting they are able to punish the unfair proposers. In the bilateral case they can always ensure this punishment while in the competitive case this is no longer possible. In particular, if one of the other responders accepts a given low offer, it is impossible for a reciprocal responder to punish the proposer. Since there is a substantial fraction of selfish responders, the probability that one of the other responders is selfish, is higher the larger the number of competing responders. This means, in turn, that the expected non-pecuniary return from the rejection of a low offer is smaller the larger the number of competing responders. Therefore, reciprocal responders will reject less frequently the larger the number of competing responders because they expect that the probability that at least one of the other responders will accept the offer increases with the number of competitors. This prediction is fully borne out by the expectations data. Moreover, these data also indicate that the responders are much less likely to reject a given offer if they believe that one of their competitors will accept the offer.

The previous example illustrates that preferences for fairness and reciprocity interact in important ways with competition. However, this example should not make us believe that sufficient competition will in general weaken or remove the impact of other-regarding preferences on market outcomes. Quite the contrary. In the following we will show that the presence of other-regarding preferences may completely nullify the impact of competition on market outcomes.

To illustrate this argument consider the double auction experiments conducted by Fehr and Falk (1999). Fehr and Falk deliberately chose the double auction as the trading institution because a large body of research has shown the striking competitive properties of experimental double auctions. Fehr and Falk use two treatment conditions: A bilateral condition in which competition is completely removed and a competitive condition. In the competitive condition they embed the gift exchange game into the context of an experimental double auction that is framed in labour market terms. The crucial difference between the competitive condition and the gift exchange game described in Section 2 is that both, experimental firms and experimental workers can make wage bids in the interval [20, 120] because the workers' reservation wage is 20 and the maximum revenue from a trade is 120. If a bid is accepted, a labour contract is concluded and the worker has to choose the effort level. As in the gift exchange game the workers

[33] In the study of Roth et al. (1991) competition led to an even more extreme outcome. However, in their market experiments 9 competing proposers faced only 1 responder and the responder was forced to accept the highest offer.

("responders") can freely choose any feasible effort level. They have to bear effort costs while the firm ("proposer") benefits from the effort. Thus, the experiment captures a market in which the quality of the good traded ("effort") is not exogenously enforced but is chosen by the workers. Workers may or may not provide the effort level that is expected by the firms.

In the competitive condition there are more workers than firms and each firm can only employ one worker. In contrast to the double auction firms in the bilateral condition are exogenously matched with a worker and there is an equal number of firms and workers. The bilateral condition implements the gift exchange game as described in Section 2. In each of the ten periods each firm is matched with a different worker. Firms have to make a wage offer to the matched worker in each period. If the worker accepts he has to choose the effort level. If a worker rejects the firm's offer both parties earn nothing. As in the competitive condition a worker who accepts a wage offer has costs of 20 and the maximum revenue from a trade is 120.

The self-interest model predicts that in both conditions the workers will only provide the minimum effort so that the firms will pay a wage of 20 or 21 in equilibrium. However, we know already from bilateral ultimatum games that firms (proposers) cannot reap the whole surplus, i.e., wages in the bilateral gift exchange game also can be expected to be much higher than predicted by the self-interest model. Moreover, since in the gift exchange game the effort is in general increasing in the wage level firms have an additional reason to offer workers a substantial share of the surplus. The question, therefore, is to what extent competition in the double auction pushes wages below the level in the bilateral condition.

The data reveal the startling result that competition has no long run impact on wage formation in this setting. Only at the beginning wages in the double auction are slightly lower than the wages in the bilateral condition but since workers responded to lower wages with lower effort levels firms raised their wages quickly. In the last five periods firms paid even slightly higher wages in the double auction; this difference is not significant, however. It is also noteworthy that competition among the workers was extremely intense. In each period many workers offered to work for wages that are close to the competitive level of 20. However, firms did not accept such low wage offers. It was impossible for the workers to get a job by underbidding the going wages because the positive effort-wage relation made it profitable for the firms to pay high, non-competitive, wages. This finding is consistent with several field studies that report that managers are reluctant to cut wages in a recession because they fear that wage cuts may hamper work performance [Bewley (1999), Agell and Lundborg (1995), Campbell and Kamlani (1997)].

The positive relation between wages and average effort is the major driving force behind the payment of high – non-competitive – wages in the Fehr and Falk (1999) experiments. On average, it was profitable for the firms in this experiment to pay such high wages. In view of the importance of a sufficiently steep effort-wage relation it is important to ask under which circumstances we can expect the payment of non-competitive wages to be profitable for the proposer. There is evidence indicating that reciprocal effort choices are almost absent if the proposer explicitly threatens to sanction the responder in case of low effort choices [Fehr and Gächter (2002), Fehr and Rockenbach (2003), Fehr and List (2004)]. Likewise, if there is a stochastic relation between effort and output, and the proposer is only informed about output but not effort, the effort wage relation is less steep [Irlenbusch and Sliwka (2005)] than in a situation where effort produces output in a deterministic way. In addition, it seems plausible that if responders do not know the profits of the proposer reciprocity is less likely to occur. In the typical gift exchange experiment full information about the payoffs of the proposer and the responder exists. Therefore, the responder has a clear yardstick which enables him to judge the generosity of the proposer's wage offer. If there is no clear reference point against which the responder can judge the generosity of a given wage offer, it seems easier that self-serving biases affect the responder's behavior, implying that reciprocal effort choices are less frequent. Thus, in the presence of explicit sanctioning threats or when there is a lack of transparency it may not pay for the proposer to offer high wages because reciprocation is weak. Finally, as mentioned in Section 2.1 already, the profitability of high wages also depends on the concrete payoff function of the proposer. In many gift exchange experiments [e.g., Fehr, Kirchsteiger and Riedl (1993) or Fehr and Falk (1999)] the proposer's payoff function is given by $x^P = (v - w)e$ and effort is in the interval $[0.1, 1]$ which makes it less risky to offer high wages than in the case where the proposer's payoff function is given by $x^P = ve - w$. Thus, when interpreting the results of gift exchange experiments it is necessary to investigate the conditions of the experiment carefully. Otherwise, it is difficult to make sense of the data.

5.4. *Fairness and reciprocity as a source of economic incentives*

Perhaps the impact of other-regarding preferences on material incentives is the most important reason why they should be taken seriously by social scientists. This is neatly illustrated by the sequential prisoners' dilemma or the gift exchange game: if there are sufficiently many second movers who reciprocate cooperative first mover choices it is in the self-interest of the first mover to make a cooperative choice. However, simple two-stage games underestimate the power of these preferences in shaping material incentives because in games that proceed beyond just two stages the impact of other-regarding preferences on incentives is greatly magnified. This is illustrated by the work of Fehr, Gächter and Kirchsteiger (1997).

In an extension of a simple two-stage gift exchange experiment these authors examined the impact of giving the employers the option of responding reciprocally to the worker's choice of effort e. In addition to the wage offered in the first stage the employer ("proposer") could also announce a desired effort level $\hat{e}$. In the second stage the workers chose their effort level and in the third stage each employer was given the opportunity to reward or punish the worker after he observed the actual effort. By spending one money unit (MU) on reward the employer could *increase* the worker's payoff by 2.5 MUs, and by spending one MU on punishment the employer could *decrease* the worker's payoff by 2.5 MUs. Employers could spend up to 10 MUs on punishment or on rewarding their worker. The important feature of this design is that if there are only selfish employers they will never reward or punish a worker because both rewarding and punishing is costly for the employer. Therefore, in case that there are only selfish employers there is no reason why the opportunity for rewarding/punishing workers should affect workers' effort choice relative to the situation where no such opportunity exists. However, if a worker expects her employer to be a reciprocator it is likely that she will provide higher effort levels in the presence of a reward/punishment opportunity. This is so because reciprocal employers are likely to reward the provision of $e \geqslant \hat{e}$ and to punish underprovision ($e < \hat{e}$). This is in fact exactly what is observed on the average. If there is underprovision of effort employers punish in 68 percent of the cases and the average investment in punishment is 7 MUs. If there is overprovision employers reward in 70 percent of these cases and the average investment in rewarding is also 7 MUs. If workers exactly meet the desired effort employers still reward in 41 percent of the cases and the average investment into rewarding is 4.5 MUs.

The authors also elicited workers' expectations about the reward and punishment choices of their employers. Hence, they are able to check whether workers anticipate employers' reciprocity. It turns out that in case of underprovision workers expect to be punished in 54 percent of the cases and the expected average investment into punishment is 4 MUs. In case of overprovision they expect to receive a reward in 98 percent of the cases with an expected average investment of 6.5 MUs. As a result of these expectations workers choose much higher effort levels when employers have a reward/punishment opportunity. The presence of this opportunity decreases shirking from 83 percent to 26 percent of the trades, increases exact provision of the desired effort $\hat{e}$ from 14 to 36 percent and increases overprovision from 3 to 38 percent of the trades. The average effort level is increased by almost 50% so that the gap between desired and actual effort levels almost vanishes. An important consequence of this increase in average effort is that the aggregate monetary payoff increases by 40 percent – even if one takes the payoff reductions that result from actual punishments into account. Thus, the reward/punishment opportunity considerably increases the total pie that becomes available for the trading parties.

We believe that the material incentives that are provided by reciprocal principals help solving one of the key problems in many agency relations, which is the problem of incentive provision when there are multiple tasks that an agent has to perform. Because of measurement and verifiability problems it is often not possible to give explicit incentives for all tasks that the agent should care about. It is well known [Holmström and Milgrom (1991), Baker (1992)] that in this situation explicit performance incentives may be harmful because they induce the employees to concentrate only on the rewarded tasks and to neglect the non-rewarded tasks. Holmström and Milgrom show that if a task that cannot be explicitly contracted upon is sufficiently important it may even be better to provide no explicit incentives for any task. Yet, this result presupposes a high degree of voluntary cooperation so that employees are willing to spend some effort even in the absence of any monetary incentives. If the agent is not intrinsically motivated this solution is not viable.

The monetary incentives provided by ex-post rewards or ex-post punishments of reciprocal principals often constitute a superior solution to the multi-tasking problem. The reason is that a principal who decides whether to reward or punish the agent ex post will use subjective performance evaluation, i.e., he will take into account the agent's performance in all observable tasks even if some of them are not verifiable and cannot be contracted upon explicitly. To illustrate this point we consider the experiments conducted by Fehr and Schmidt (2004). In these experiments each principal faces ten different agents in ten one-shot interactions. When an agent agrees to the terms of a contract offered by the principal the agent has to choose the effort level e_1 in task 1 and e_2 in task 2. The revenue of the principal is given by $10e_1e_2$ while the agent's effort cost is an increasing and convex function of total effort $(e_1 + e_2)$. Effort in both tasks can vary between 1 and 10. This set-up ensures that both tasks are important for the principal because the effort levels are complements in his profit function. Both effort levels are observable for both parties but only effort in task 1 is verifiable while effort in task 2 cannot be contracted upon.

In each period the principal can offer to the agent either a piece rate contract that makes pay contingent on effort in task 1 or a so-called bonus contract. The piece rate contract consists of a base wage and a piece rate per unit of effort in task 1. The bonus contract also consists of a base wage. In addition the principal announces that he may pay a bonus after he observed the actual effort levels e_1 and e_2. However, both parties know that the bonus payment is voluntary and cannot be enforced.

Clearly, selfish principals will never pay a bonus. Furthermore, if agents anticipate that principals are selfish they will always choose the minimal effort in the bonus contract. With a piece rate contract the principal, at least, can induce a selfish agent to work efficiently on task 1. Thus, if all subjects are selfish, the piece rate contract is more profitable and more efficient than the bonus contract, even though the agent will only work on task 1 and completely ignore task 2.

If principals behave reciprocally, however, the result is very different. A reciprocal principal is willing to voluntarily pay a bonus if he is satisfied with the agent's performance. This makes it profitable for the agent to spend effort and to allocate his efforts efficiently across *both* tasks. Thus a preference for reciprocity and fairness is a commitment device for the principal to reward the agent for his efforts, even if this cannot be enforced by the courts.[34]

The experiments by Fehr and Schmidt (2004) show that many (but not all) principals pay substantial bonuses. It turns out that the average bonus is strongly increasing in total effort and decreasing in effort differences across tasks. This creates incentives for the agents to spend high effort and to equalize effort levels across tasks. With a piece rate contract, on the other hand, the average effort is always high in the rewarded task but

[34] Note that if the principal is just an efficiency seeker who wants to maximize total surplus he will not pay the bonus. After the agent has chosen his effort levels the bonus is a pure transfer that leaves total surplus unaffected.

close to the minimum level in the non-rewarded task. Thus, the bonus contract induces more efficient effort choices and yields, on average, higher payoffs for both parties. Principals seem to understand this and predominantly (in 81 percent of all cases) choose a bonus contract.

This result also suggests an answer to the puzzling question why many contracts are deliberately left vague and incomplete. Many real world contracts specify important obligations of the contracting parties in fairly vague terms, and they do not tie the parties' monetary payoffs to measures of performance that would be available at a relatively small cost. We believe that the parties often rely on an implicit understanding to reward (or punish) each other that cannot be enforced by the courts but nevertheless works well if the involved parties are motivated by reciprocity and fairness. In an extensive empirical study Scott (2003) provides evidence on deliberately incomplete contracting supporting this claim.

6. Conclusions

The self-interest hypothesis assumes that all people are exclusively motivated by their material self-interest. This hypothesis is a convenient simplification and there are, no doubt, situations in which almost all people behave as if they were strictly self-interested. In particular, for comparative static predictions of aggregate behavior self-interest models may make empirically correct predictions because models with more complex motivational assumptions predict the same comparative static responses. However, the evidence presented in this paper also shows that fundamental questions of social life cannot be understood on the basis of the self-interest model. The evidence indicates that other-regarding preferences are important for bilateral negotiations, for the enforcement of social norms, for understanding the functioning of markets and economic incentives. They are also important determinants of cooperation and collective action and the very existence of cooperative institutions that enforce rules and norms may be due to the existence of other-regarding preferences. The examples that we have given in Section 5 of this chapter do of course not exhaust the potential impact of such preferences on economic and social processes. We did not mention the impact of other-regarding preferences on voting behaviour, tax policy and the demand for redistribution [Fong (2001), Anderhub (2001), Tyran (2004), Riedl and Tyran (2005), Ackert, Martinez-Vazquez and Rider (2004), Fong, Bowles and Gintis (2005a, 2005b), Hahn (forthcoming), Hahn (2004)] and on various aspects of contract economics, the hold-up problem and the optimal allocation of property rights [Anderhub, Gachter and Konigstein (2002), Ellingsen and Johannesson (2004a, 2004b, 2005), Cabrales and Charness (2003), Fehr, Kremhelmer and Schmidt (2004)]. We also did not mention how other-regarding preferences affect trust and may undermine the impact of incentives [Bohnet and Zeckhauser (2004), Bohnet, Frey and Huck (2001), Gneezy and Rustichini (2000), Fehr and Rockenbach (2003), Fehr and List (2004)]. This long list of examples suggests that other-regarding preferences affect social and economic life in many domains. If they are neglected social scientists run the risk of providing incomplete explanations of the phenomena under study or – in the worst case – their explanations may be wrong.

However, although in view of the prevailing modelling practices in economics it is natural to emphasize the existence of a substantial share of subjects with other-regarding preferences, one should not forget the fact that many subjects often show completely selfish behaviors. Moreover, many of the examples we have discussed in Section 5 show that the interaction between self-interested actors and actors with other-regarding preferences may play a key role for the understanding of the outcomes of many experiments. Depending on the strategic environment selfish actors may induce actors with other-regarding preferences to behave as if completely selfish but the converse is also often true: actors with other-regarding preferences induce selfish actors to change their behavior in fundamental ways. In order to fully understand the interaction between selfish and non-selfish actors, social scientists need rigorous formal models of other-regarding preferences. In Section 3 we have documented the current state of the art in this domain. While the current models clearly present progress relative to the self-interest approach the evidence reported in Section 4 also makes it clear that further theoretical progress is warranted. There is still ample opportunity for improving our understanding of other-regarding behavior.

References

Abbink, K., Irlenbusch, B., Renner, E. (2000). "The moonlighting game: An experimental study on reciprocity and retribution". Journal of Economic Behavior and Organization 42 (2), 265–277.

Ackert, L.F., Martinez-Vazquez, J., Rider, M. (2004). "Tax policy design in the presence of social preferences: Some experimental evidence". Discussion paper. Department of Economics and Finance, Kennesaw State University.

Agell, J., Lundborg, P. (1995). "Theories of pay and unemployment: Survey evidence from Swedish manufacturing firms". Scandinavian Journal of Economics 97, 295–308.

Ahlert, M., Crüger, A., Güth, W. (1999). "An experimental analysis of equal punishment games". Mimeo. University of Halle-Wittenberg.

Alvard, M.S. (2004). "The ultimatum game, fairness, and cooperation among big game hunters". In: Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. (Eds.), Foundations of Human Sociality. Oxford University Press, Oxford.

Anderhub, V. (2001). "Tax evasion with earned income – an experimental study". FinanzArchiv 58 (2), 188–206.

Anderhub, V., Gachter, S., Konigstein, M. (2002). "Efficient contracting and fair play in a simple principal – agent experiment". Experimental Economics 5 (1), 5–27.

Anderson, C.M., Putterman, L. (2006). "Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism". Games and Economic Behavior 54, 1–24.

Andreoni, J. (1989). "Giving with impure altruism: Applications to charity and Ricardian equivalence". Journal of Political Economy 97, 1447–1458.

Andreoni, J., Castillo, M., Petrie, R. (2003). "What do Bargainers' preferences look like? Experiments with a convex ultimatum game". American Economic Review 93, 672–685.

Andreoni, J., Miller, J. (1993). "Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence". Economic Journal 103, 570–585.

Andreoni, J., Miller, J. (2002). "Giving according to GARP: An experimental test of the rationality of altruism". Econometrica 70, 737–753.
Andreoni, J., Vesterlund, L. (2001). "Which is the fair sex? Gender differences in altruism". Quarterly Journal of Economics 116, 293–312.
Andreoni, J., Vesterlund, L. (forthcoming). "Which is the fair sex? Gender differences in altruism". Quarterly Journal of Economics.
Arrow, K.J. (1981). "Optimal and voluntary income redistribution". In: Rosenfield, S. (Ed.), Economic Welfare and the Economics of Soviet Socialism: Essays in Honor of Abram Bergson. Cambridge University Press, Cambridge.
Baker, G. (1992). "Incentive contracts and performance measurement". Journal of Political Economy 100, 598–614.
Becker, G.S. (1974). "A theory of social interactions". Journal of Political Economy 82, 1063–1093.
Bellemare, C., Kröger, S. (2003). "On representative trust". Working Paper. Tilburg University.
Ben-Shakhar, G., Bornstein, G., Hopfensitz, A., van Winden, F. (2004). "Reciprocity and emotions: Arousal, self-reports and expectations". Discussion Paper. University of Amsterdam.
Benabou, R., Tirole, J. (2004). "Incentives and prosocial behavior". Mimeo. Princeton University.
Benjamin, D.J. (2004). "Fairness: From the laboratory into the market". Mimeo. Harvard University.
Berg, J., Dickhaut, J., McCabe, K. (1995). "Trust, reciprocity and social history". Games and Economic Behavior X, 122–142.
Bernheim, B.D. (1986). "On the voluntary and involuntary provision of public goods". American Economic Review 76, 789–793.
Bewley, T. (1999). Why Wages Don't Fall during a Recession. Harvard University Press, Harvard.
Binmore, K. (1998). Game Theory and the Social Contract: Just Playing. MIT Press, Cambridge, MA.
Binmore, K., Gale, J., Samuelson, L. (1995). "Learning to be imperfect: The ultimatum game". Games and Economic Behavior 8, 56–90.
Blount, S. (1995). "When social outcomes aren't fair: The effect of causal attributions on preferences". Organizational Behavior and Human Decision Processes LXIII, 131–144.
Bohnet, I., Frey, B.S., Huck, S. (2001). "More order with less law: On contract enforcement, trust, and crowding". American Political Science Review 95 (1), 131–144.
Bohnet, I., Zeckhauser, R. (2004). "Trust, risk and betrayal". Journal of Economic Behavior & Organization 55, 467–484.
Bolle, F., Kritikos, A. (1998). "Self-centered inequality aversion versus reciprocity and altruism". Mimeo. Europa-Universität Viadrina.
Bolton, G.E. (1991). "A comparative model of bargaining: Theory and evidence". American Economic Review 81, 1096–1136.
Bolton, G.E., Brandts, J., Ockenfels, A. (1998). "Measuring motivations for the reciprocal responses observed in a simple dilemma game". Experimental Economics 3, 207–221.
Bolton, G.E., Ockenfels, A. (2000). "A theory of equity, reciprocity and competition". American Economic Review 100, 166–193.
Bolton, G., Zwick, R. (1995). "Anonymity versus punishment in ultimatum bargaining". Games and Economic Behavior 10, 95–121.
Bosman, R., Sutter, M., van Winden, F. (2005). "The impact of real effort and emotions in the power-to-take game". Journal of Economic Psychology 26, 407–429.
Bosman, R., van Winden, F. (2002). "Emotional hazard in a power-to-take-experiment". Economic Journal 112, 147–169.
Bowles, S., Gintis, H. (2000). "Reciprocity, self-interest, and the welfare state". Nordic Journal of Political Economy 26, 33–53.
Brandts, J., Charness, G. (2003). "Truth or consequences: An experiment". Management Science 49, 116–130.
Brandts, J., Charness, G. (2004). "Do labour market conditions affect gift exchange? Some experimental evidence". Economic Journal 114 (497), 684–708.

Brandts, J., Schram, A. (2001). "Cooperation and noise in public goods experiments: Applying the contribution function approach". Journal of Public Economics 79, 399–427.
Brandts, J., Sola, C. (2001). "Reference points and negative reciprocity in simple sequential games". Games and Economic Behavior 36, 138–157.
Buchan, N.R., Croson, R.T.A., Dawes, R.M. (2002). "Swift neighbors and persistent strangers: A cross-cultural investigation of trust and reciprocity in social exchange". American Journal of Sociology 108, 168–206.
Cabrales, A., Charness, G. (2003). "Optimal contracts, adverse selection & social preferences: An experiment". Discussion Paper. Department of Economics, University of Santa Barbara.
Camerer, C.F. (2003). Behavioral Game Theory, Experiments in Strategic Interaction. Princeton University Press, Princeton.
Camerer, C.F., Thaler, R.H. (1995). "Ultimatums, dictators and manners". Journal of Economic Perspectives 9, 209–219.
Cameron, L.A. (1999). "Raising the stakes in the ultimatum game: Experimental evidence from Indonesia". Economic-Inquiry 37 (1), 47–59.
Campbell, C.M., Kamlani, K. (1997). "The reasons for wage rigidity: Evidence from a survey of firms". Quarterly Journal of Economics 112, 759–789.
Carpenter, J.P., Matthews, P.H., Ong'ong'a, O. (2004). "Why punish? Social reciprocity and the enforcement of prosocial norms". Journal of Evolutionary Economics 14 (4), 407–429.
Charness, G. (1996). "Attribution and reciprocity in a labor market: An experimental investigation". Mimeo. University of California at Berkeley.
Charness, G. (2000). "Responsibility and effort in an experimental labor market". Journal of Economic Behavior and Organization 42, 375–384.
Charness, G., Dufwenberg, M. (2004). "Promises and partnerships". Mimeo. University of California at Santa Barbara.
Charness, G., Rabin, M. (2002). "Understanding social preferences with simple tests". Quarterly Journal of Economics 117, 817–869.
Cohen, D., Nisbett, R. (1994). "Self-protection and the culture of honor – Explaining southern violence". Personality and Social Psychology Bulletin 20, 551–567.
Cooper, D.J., Stockman, C.K. (1999). "Fairness, learning, and constructive preferences: An experimental investigation". Mimeo. Case Western Reserve University.
Costa-Gomes, M., Zauner, K.G. (1999). "Learning, non-equilibrium beliefs, and non-pecuniary payoff uncertainty in an experimental game". Mimeo. Harvard Business School.
Cox J.C. (2000). "Trust and reciprocity: Implications of game triads and social contexts". Mimeo. University of Arizona at Tucson.
Cox, J.C. (2004). "How to identify trust and reciprocity". Games and Economic Behavior 46 (2), 260–281.
Cox, J.C., Friedman, D., Gjerstad, S. (2004). "A tractable model of reciprocity and fairness". Mimeo. University of Arizona.
Cox, J.C., Sadiraj, K., Sadiraj, V. (2001). "Trust, fear, reciprocity and altruism". Mimeo. University of Arizona.
Daughety, A. (1994). "Socially-influenced choice: Equity considerations in models of consumer choice and in games". Mimeo. University of Iowa.
Davis, D., Holt, Ch. (1993). Experimental Economics. Princeton University Press, Princeton.
de Quervain, D.J.F., Fischbacher, U., Treyer, V., Schelthammer, M., Schnyder, U., Buck, A., Fehr, E. (2004). "The neural basis of altruistic punishment". Science 305, 1254–1258.
Delgado, M.R., Locke, H.M., Stenger, V.A., Fiez, J.A. (2003). "Dorsal striatum responses to reward and punishment: effects of valence and magnitude manipulations". Cognitive Affective Behavioral Neuroscience 3, 27–38.
Dufwenberg, M., Kirchsteiger, G. (2004). "A theory of sequential reciprocity". Games and Economic Behavior 47, 268–298.
Eckel, C.C., Grossman, P.J. (1996). "The relative price of fairness: Gender differences in a punishment game". Journal of Economic Behavior and Organization 30, 143–158.

Eichenberger, R., Oberholzer-Gee, F. (1998). "Focus effects in dictator game experiments". Mimeo. University of Pennsylvania.

Ellingsen, T., Johannesson, M. (2004a). "Is there a hold-up problem?". Scandinavian Journal of Economics 106 (3), 475–494.

Ellingsen, T., Johannesson, M. (2004b). "Promises, threats and fairness". Economic Journal 114 (495), 397–420.

Ellingsen, T., Johannesson, M. (2005). "Sunk costs and fairness in incomplete information bargaining". Games and Economic Behavior 50 (2), 155–177.

Engelmann, D., Fischbacher, U. (2002). "Indirect reciprocity and strategic reputation building in an experimental helping game". Working Paper No. 132. Institute for Empirical Research in Economics, University of Zurich.

Engelmann, D., Strobel, M. (2004). "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments". American Economic Review 94, 857–869.

Erlei, M. (2004). "Heterogeneous social preferences". Mimeo. Clausthal University of Technology.

Falk, A., Fehr, E., Fischbacher, U. (2000a). "Informal sanctions". Working Paper No. 59. Institute for Empirical Research in Economics, University of Zurich.

Falk, A., Fehr, E., Fischbacher, U. (2000b). "Testing theories of fairness – Intentions matter". Working Paper No. 63. Institute for Empirical Research in Economics, University of Zurich.

Falk, A., Fehr, E., Fischbacher, U. (2003). "On the nature of fair behavior". Economic Inquiry 41, 20–26.

Falk, A., Fehr, E., Fischbacher, U. (2005). "Driving forces behind informal sanctions". Econometrica 73, 2017–2030.

Falk, A., Fischbacher, U. (2005). "A theory of reciprocity", Games and Economic Behavior. Submitted for publication.

Falk, A., Fischbacher, U. (2006). "A theory of reciprocity". Games and Economic Behavior 54, 293–315.

Falk, A., Gächter, S., Kovács, J. (1999). "Intrinsic motivation and extrinsic incentives in a repeated game with incomplete contracts". Journal of Economic Psychology.

Fehr, E., Falk, A. (1999). "Wage rigidity in a competitive incomplete contract market". Journal of Political Economy 107, 106–134.

Fehr, E., Fischbacher, U. (2003). "The nature of human altruism". Nature 425, 785–791.

Fehr, E., Fischbacher, U. (2004). "Third party punishment and social norms". Evolution and Human Behavior 25, 63–87.

Fehr, E., Fischbacher, U., Rosenbladt, B., Schupp, J., Wagner, G. (2002). "A nation-wide laboratory – Examining trust and trustworthiness by integrating behavioral experiments into representative surveys". Schmollers Jahrbuch 122, 519–543.

Fehr, E., Gächter, S. (2002). "Do incentive contracts undermine voluntary cooperation". Working Paper No. 34. Institute for Empirical Research in Economics, University of Zurich.

Fehr, E., Gächter, S. (2000). "Cooperation and punishment in public goods experiments". American Economic Review 90, 980–994.

Fehr, E., Gächter, S., Kirchsteiger, G. (1997). "Reciprocity as a contract enforcement device". Econometrica 65, 833–860.

Fehr, E., Kirchsteiger, G., Riedl, A. (1993). "Does fairness prevent market clearing? An experimental investigation". Quarterly Journal of Economics CVIII, 437–460.

Fehr, E., Kirchsteiger, G., Riedl, A. (1998). "Gift exchange and reciprocity in competitive experimental markets". European Economic Review 42, 1–34.

Fehr, E., Klein, A., Schmidt, K.M. (2004). "Contracts, fairness, and incentives". CESifo Working Paper No. 1215. Munich.

Fehr, E., Kremhelmer, S., Schmidt, K.M. (2004). "Fairness and the optimal allocation of property rights". Mimeo. University of Munich.

Fehr, E., List, J.A. (2004). "The hidden costs and returns of incentives – trust and trustworthiness among CEOs". Journal of the European Economic Association 2 (5), 743–771.

Fehr, E., Naef, M., Schmidt, K.M. (forthcoming). "The role of equality and efficiency in social preferences". American Economic Review.

Fehr, E., Rockenbach, B. (2003). "Detrimental effects of sanctions on human altruism". Nature 422, 137–140.
Fehr, E., Schmidt, K.M. (1999). "A theory of fairness, competition and co-operation". Quarterly Journal of Economics 114, 817–868.
Fehr, E., Schmidt, K.M. (2004). "Fairness and incentives in a multi-task principal-agent model". Scandinavian Journal of Economics 106, 453–474.
Fehr, E., Tougareva, E. (1995). "Do high monetary stakes remove reciprocal fairness? Experimental evidence from Russia". Mimeo. Institute for Empirical Economic Research, University of Zurich.
Fischbacher, U., Fong, C., Fehr, E. (2002). "Fairness and the power of competition". Working Paper No. 133. Institute for Empirical Research in Economics, University of Zurich.
Fischbacher, U., Gächter, S., Fehr, E. (2001). "Are people conditionally cooperative? Evidence from a public goods experiment". Economics Letters 71, 397–404.
Fong, C.M. (2001). "Social preferences, self-interest, and the demand for redistribution". Journal of Public Economics 82, 225.246.
Fong, C.M., Bowles, S., Gintis, H. (2005a). "Behavioral motives for income redistribution". Australian Economic Review 38, 285–297.
Fong, C.M., Bowles, S., Gintis, H. (2005b). "Reciprocity and the welfare state". In: Gintis, H., Bowles, S., Boyd, R., Fehr, E. (Eds.), Moral Sentiments and Material Interests: On the Foundations of Cooperation in Economic Life. MIT Press, Cambridge.
Forsythe, R.L., Horowitz, J., Savin, N.E., Sefton, M. (1994). "Fairness in simple bargaining games". Games and Economic Behavior 6, 347–369.
Frechette, G.R., Kagel, J.H., Lehrer, S.F. (2003). "Bargaining in legislatures: An experimental investigation of open versus closed amendment rules". American Political Science Review 97 (2), 221–232.
Gächter, S., Falk, A. (1999). "Reputation or reciprocity?". Working Paper No. 19. Institute for Empirical Research in Economics, University of Zürich.
Gächter, S., Falk, A. (2002). "Reputation and reciprocity: consequences for the labour relation". Scandinavian Journal of Economics 104, 1–26.
Geanakoplos, J., Pearce, D., Stacchetti, E. (1989). "Psychological games and sequential rationality". Games and Economic Behavior 1, 60–79.
Gintis, H. (2000). "Strong reciprocity and human sociality". Journal of Theoretical Biology 206, 169–179.
Gneezy, U. (2005). "Deception: The role of consequences". American Economic Review 95 (1), 384–394.
Gneezy, U., Rustichini, A. (2000). "A fine is a price". Journal of Legal Studies 29 (1), 1–17.
Goeree, J., Holt, Ch. (2001). "Ten little treasures of game theory and ten intuitive contradictions". American Economic Review 91, 1402–1422.
Grosskopf, B. (2003). "Reinforcement and directional learning in the ultimatum game with responder competition". Experimental Economics 6, 141–158.
Gul, F., Pesendorfer, W. (2005). "The canonical type space for interdependent preferences". Mimeo. Princeton University.
Gürerk, O., Irlenbusch, B., Rockenbach, B. (2006). "The competitive advantage of sanctioning institutions". Science 312, 108–111.
Güth, W., Kliemt, H., Ockenfels, A. (2000). "Fairness versus efficiency – An experimental study of mutual gift-giving". Mimeo. Humboldt University of Berlin.
Guth, W., Kliemt, H., Ockenfels, A. (2003). "Fairness versus efficiency: An experimental study of (mutual) gift giving". Journal of Economic Behavior and Organization 50 (4), 465–475.
Güth, W., Marchand, N., Rulliere, J.-L. (1997). "On the reliability of reciprocal fairness – An experimental study". Discussion Paper. Humbold University Berlin.
Güth, W., Schmittberger, R., Schwarze, B. (1982). "An experimental analysis of ultimatium bargaining". Journal of Economic Behavior and Organization III, 367–388.
Güth, W., van Damme, E. (1998). "Information, strategic behavior and fairness in ultimatum bargaining: an experimental study". Journal of Mathematical Psychology 42, 227–247.
Hahn, V. (2004). "Reciprocity and voting". Discussion paper. Department of Economics, University of Heidelberg.

Hahn, V. (forthcoming). “Fairness and voting”. Social Choice and Welfare.

Hannan, L., Kagel, J., Moser, D. (1999). “Partial gift exchange in experimental labor markets: Impact of subject population differences, productivity differences and effort requests on behavior”. Mimeo. University of Pittsburgh.

Harsanyi, J. (1955). “Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility”. Journal of Political Economy 63, 309–321.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. (2004). Foundations of Human Sociality – Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies. Oxford University Press, Oxford.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R. (2001). “In search of homo economicus: behavioral experiments in 15 small-scale societies”. American Economic Review 91, 73–78.

Henrich, J., Smith, N. (2004). “Comparative experimental evidence from Machiguenga, Mapuche, Huinca, and American Populations”. In: Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. (Eds.), Foundations of Human Sociality. Oxford University Press, Oxford.

Hoffman, E., McCabe, K., Shachat, K., Smith, V. (1994). “Preferences, property right, and anonymity in bargaining games”. Games and Economic Behavior 7, 346–380.

Hoffman, E., McCabe, K., Smith, V. (1996a). “On expectations and monetary stakes in ultimatum games”. International Journal of Game Theory 25, 289–301.

Hoffman, E., McCabe, K., Smith, V. (1996b). “Social distance and other-regarding behavior”. American Economic Review 86, 653–660.

Holmström, B., Milgrom, P. (1991). “Multi-task principal-agent analyses”. Journal of Law, Economics, and Organization 7, 24–52.

Huck, S., Müller, W., Normann, H.-T. (2001). “Stackelberg beats cournot: On collusion and efficiency in experimental markets”. Economic Journal 111, 749–766.

Irlenbusch, B., Sliwka, D. (2005). “Transparency and reciprocity and employment relations”. Journal of Economic Behavior and Organization 56, 383–403.

Jackson, P.L., Meltzoff, A.N., Decety, J. (2005). “How do we perceive the pain of others? A window into the neural processes involved in empathy”. Neuroimage 24, 771–779.

Kagel, J.H., Kim, Ch., Moser, D. (1996). “Fairness in ultimatum games with asymmetric information and asymmetric payoffs”. Games and Economic Behavior 13, 100–110.

Kahneman, D., Tversky, A. (1979). “Prospect theory: An analysis of decision under risk”. Econometrica 47, 263–291.

Keser, C., van Winden, F. (2000). “Conditional cooperation and voluntary contributions to public goods”. Scandinavian Journal of Economics 102, 23–39.

Kirchsteiger, G. (1994). “The role of envy in ultimatum games”. Journal of Economic Behavior and Organization 25, 373–389.

Knutson, B., Fong, G.W., Adams, C.M., Varner, J.L., Hommer, D. (2001). “Dissociation of reward anticipation and outcome with event-related fMRI”. Neuroreport 12, 3683–3687.

Kolm, S.-Ch. (1995). “The economics of social sentiments: The case of envy”. Japanese Economic Review 46, 63–87.

Ledyard, J. (1995). “Public goods: A survey of experimental research”. In: Roth, A., Kagel, J. (Eds.), Handbook of Experimental Economics. Princeton University Press, Princeton.

Levine, D. (1998). “Modeling altruism and spitefulness in experiments”. Review of Economic Dynamics 1, 593–622.

List, J., Cherry, T. (2000). “Examining the role of fairness in bargaining games”. Mimeo. University of Arizona at Tucson.

Masclet, D. (2003). “Monetary and nonmonetary punishment in the voluntary contributions mechanism”. American Economic Review 93 (1), 366–380.

McCabe, K.A., Rigdon, M.L., Smith, V.L. (2003). “Positive reciprocity and intentions in trust games”. Journal of Economic Behavior and Organization 52, 267–275.

Morrison, I., Lloyd, D., die Pellegrino, G., Roberts, N. (2004). "Vicarious responses to pain in anterior cingulate cortex: Is empathy a multisensory issue?". Cognitive, Affective, and Behavioral Neuroscience 4, 270–278.

Mui, V.-L. (1995). "The Economics of envy". Journal of Economic Behavior and Organization 26, 311–336.

Neilson, W. (2005). "Axiomatic reference dependence in behavior toward others and toward risk". Mimeo. Department of Economics, Texas A&M University.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., Dolan, R.J. (2004). "Dissociable roles of ventral and dorsal striatum in instrumental conditioning". Science 304, 452–454.

Offerman, T. (1999). "Hurting hurts more than helping helps: The role of the self-serving Bias". Mimeo. University of Amsterdam.

Okada, A., Riedl, A. (2005). "Inefficiency and social exclusion in a coalition formation game: Experimental evidence". Games and Economic Behavior 50 (2), 278–311.

Ostrom, E., Walker, J., Gardner, R. (1992). "Covenants with and without a sword – self-governance is possible". American Political Science Review 86, 404–417.

Preston, S.D., de Waal, F.B.M. (2002). "Empathy: its ultimate and proximate bases". The Behavioral and Brain Sciences 25, 1–71.

Rabin, M. (1993). "Incorporating fairness into game theory and economics". American Economic Review 83 (5), 1281–1302.

Riedl, A., Tyran, J.-R. (2005). "Tax liability side equivalence in gift-exchange labor markets". Journal of Public Economics 89 (11–12), 2369–2382.

Rilling, J.K., Gutman, D.A., Zeh, T.R., Pagnoni, G., Berns, G.S., Kilts, C.D. (2002). "A neural basis for social cooperation". Neuron 35, 395–405.

Ross, L., Ward, A. (1996). "Naive realism in everyday life: Implications for social conflict and misunderstanding". In T. e. a. Brown, Values and Knowledge 103.

Rotemberg, J. (2004). "Minimally acceptable altruism and the ultimatum game". Mimeo. Harvard Business School.

Roth, A.E. (1995). "Bargaining experiments". In: Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics. Princeton University Press, Princeton.

Roth, A.E., Erev, I. (1995). "Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term". Games and Economic Behavior 8, 164–212.

Roth, A.E., Malouf, M.W.K., Murningham, J.K. (1981). "Sociological versus strategic factors in bargaining". Journal of Economic Behavior and Organization 2, 153–177.

Roth, A.E., Prasnikar, V., Okuno-Fujiwara, M., Zamir, S. (1991). "Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study". American Economic Review 81, 1068–1095.

Sally, D. (1995). "Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992". Rationality and Society 7, 58–92.

Samuelson, P.A. (1993). "Altruism as a problem involving group versus individual selection in economics and biology". American Economic Review 83, 143–148.

Sandbu, M.E. (2002). "A theory of set-dependent fairness preferences". Mimeo. Harvard University.

Schultz, W. (2000). "Multiple reward signals in the brain". Nature Reviews Neuroscience 1, 199–207.

Scott, R. (2003). "A theory of self-enforcing, indefinite agreements". Columbia Law Review 108, 1641–1699.

Segal, U., Sobel, J. (2004). "Tit for tat: Foundations of preferences for reciprocity in strategic settings". Mimeo. University of California at San Diego.

Seinen, I., Schram, A. (2006). "Social status and group norms: Indirect reciprocity in a helping experiment". European Economic Review 50, 581–602.

Selten, R., Ockenfels, A. (1998). "An experimental solidarity game". Journal of Economic Behavior and Organization 34, 517–539.

Sen, A. (1995). "Moral codes and economic success". In: Britten, C.S., Hamlin, A. (Eds.), Market Capitalism and Moral Values. Edward Eldar, Aldershot.

Singer, T., Kiebel, S.J., Winston, J.S., Kaube, H., Dolan, R.J., Frith, C.D. (2004b). "Brain responses to the acquired moral status of faces". Neuron 41, 653–662.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R.J., Frith, C.D. (2004a). "Empathy for pain involves the affective but not sensory components of pain". Science 303, 1157–1162.

Singer, T., Seymour, B., O'Doherty, J.P., Stephan, K.E., Dolan, R.J., Frith, C.D. (2006). "Empathic neural responses are modulated by the perceived fairness of others". Nature 439, 466–469.

Slonim, R., Roth, A.E. (1997). "Financial incentives and learning in ultimatum and market games: An experiment in the Slovak Republic". Econometrica 65, 569–596.

Smith, A. (1759). The Theory of Moral Sentiments.

Smith, A. (1982). The Theory of Moral Sentiments. (1759) edition reprinted. Liberty Fund, Indianapolis.

Smith, V.L. (1962). "An experimental study of competitive market behavior". Journal of Political Economy 70, 111–137.

Suleiman, R. (1996). "Expectations and fairness in a modified ultimatum game". Journal of Economic Psychology 17, 531–554.

Tyran, J.-R. (2004). "Voting when money and morals conflict: An experimental test of expressive voting". Journal of Public Economics 88 (7–8), 1645–1664.

Veblen, T. (1922). The Theory of the Leisure Class – An Economic Study of Institutions. George Allen Unwin, London. First published 1899.

Wax, A.L. (2000). "Rethinking welfare rights: Reciprocity norms, reactive attitudes, and the political economy of welfare reform". Law and Contemporary Problems 63, 257–298.

Yamagishi, T. (1986). "The provision of a sanctioning system as a public good". Journal of Personality and Social Psychology 51, 110–116.

Zizzo, D., Oswald, A. (2000). "Are people willing to pay to reduce others' income". Mimeo. Oxford University.