Theories of Fairness and Reciprocity - Evidence and Economic Applications*

Ernst Fehr[a)]
University of Zurich and CEPR

Klaus M. Schmidt[b)]
University of Munich and CEPR

Published in: Mathias Dewatripont, Lars Peter Hansen and Stephen J Turnovsky (2003): Advances in Economics and Econometrics, Econometric Society Monographs, Eighth World Congress, Volume 1, pp 208 – 257.

Abstract: Most economic models are based on the *self-interest hypothesis* that assumes that *all* people are *exclusively* motivated by their material self-interest. In recent years experimental economists have gathered overwhelming evidence that systematically refutes the self-interest hypothesis and suggests that many people are strongly motivated by concerns for fairness and reciprocity. Moreover, several theoretical papers have been written showing that the observed phenomena can be explained in a rigorous and tractable manner. These theories in turn induced a new wave of experimental research offering additional exciting insights into the nature of preferences and into the relative performance of competing theories of fairness. The purpose of this paper is to review these recent developments, to point out open questions, and to suggest avenues for future research.

JEL classification numbers: C7, C9, D0, J3.
Keywords: Behavioral Economics, Fairness, Reciprocity, Altruism, Experiments, Incentives, Contracts, Competition

---

* We would like to thank Glenn Ellison for many helpful comments and suggestions and Alexander Klein and Susanne Kremhelmer for excellent research assistance. Part of this research was conducted while the second author visited Stanford University and he would like to thank the Economics Department for its great hospitality. Financial support by Deutsche Forschungsgemeinschaft through grant SCHM-1196/4-1 is gratefully acknowledged. Ernst Fehr also gratefully acknowledges support from the Swiss National Science Foundation (project number 1214-05100.97) and the Network on the Evolution of Preferences and Social Norms of the MacArthur Foundation and the EU-TMR Research Network ENDEAR (FMRX-CTP98-0238).
a) Ernst Fehr, Institute for Empirical Research in Economics, University of Zurich, Bluemlisalpstrasse 10, CH-8006 Zurich, Switzerland, email: efehr@iew.unizh.ch.
b) Klaus M. Schmidt, Department of Economics, University of Munich, Ludwigstrasse 28, D-80539 Muenchen, Germany, email: klaus.schmidt@Lrz.uni-muenchen.de.

Contents

Non-technical summary

Most economic models are based on the *self-interest hypothesis* that assumes that *all* people are *exclusively* motivated by their material self-interest. In recent years experimental economists have gathered overwhelming evidence that systematically refutes the self-interest hypothesis. The evidence suggests that many people are strongly motivated by other-regarding preferences and that concerns for fairness and reciprocity cannot be ignored in social interactions. Moreover, several theoretical papers have been written showing that the observed phenomena can be explained in a rigorous and tractable manner. Some of these models shed new light on problems that have puzzled economists for a long time like, e.g., the persistence of non-competitive wage premia, the incompleteness of contracts, the allocation of property rights, the conditions for successful collective action and the optimal design of institutions. These theories in turn induced a new wave of experimental research offering additional exciting insights into the nature of preferences and into the relative performance of competing theories of fairness. The purpose of this paper is to review these recent developments, to point out open questions, and to suggest avenues for future research. Furthermore, we will argue that it is not only necessary but also very promising for mainstream economics to take the presence of other-regarding preferences into account.

Why are economists so reluctant to give up the self-interest hypothesis? One reason is that this hypothesis has been quite successful in providing accurate predictions in some economic domains. For example, models based on the self-interest hypothesis make very good predictions for competitive markets with standardized goods. This has been shown in many carefully conducted market experiments. However, a large amount of economic activity is taking place outside of competitive markets – in markets with a small number of traders, in markets with informational frictions, in firms and organizations, and under incompletely specified and incompletely enforceable contracts. In these environments models based on the self-interest assumption frequently make very misleading predictions. An important insight provided by some of the newly developed fairness models is that they show why in competitive environments with standardized goods the self-interest model is so successful and why in other environments it is refuted. In this way the new models provide fresh and experimentally confirmed insights into important phenomena like, e. g., non-clearing markets or the wide-spread use of incomplete contracts.

We consider it as important to stress that the available experimental evidence also suggests that many subjects behave quite selfishly even when they are given a chance to affect other peoples well being at a relatively small cost. However, there are also many people who are strongly motivated by fairness and reciprocity and who are willing to reward or punish other people at a considerable cost to themselves. 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 the understanding of the observed behavior in strategic settings. These models explain why in some strategic settings almost all people behave as if they are completely selfish, while in others the same people will behave as if they are driven by fairness.

1 Introduction

Most economic models are based on the *self-interest hypothesis* that assumes that *all* people are *exclusively* motivated by their material self-interest. 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. Yet, so far, these opinions have not had much of an impact on mainstream economics. In recent years experimental economists have gathered overwhelming evidence that systematically refutes the self-interest hypothesis. The evidence suggests that many people are strongly motivated by other-regarding preferences and that concerns for fairness and reciprocity cannot be ignored in social interactions. Moreover, several theoretical papers have been written showing that the observed phenomena can be explained in a rigorous and tractable manner. Some of these models shed new light on problems that have puzzled economists for a long time like , e.g., the persistence of non-competitive wage premia, the incompleteness of contracts, the allocation of property rights, the conditions for successful collective action and the optimal design of institutions. These theories in turn induced a new wave of experimental research offering additional exciting insights into the nature of preferences and into the relative performance of competing theories of fairness. The purpose of this paper is to review these recent developments, to point out open questions, and to suggest avenues for future research. Furthermore, we will argue that it is not only necessary but also very promising for mainstream economics to take the presence of other-regarding preferences into account.

Why are economists so reluctant to give up the self-interest hypothesis? One reason is that this hypothesis has been quite successful in providing accurate predictions in some economic domains. For example, models based on the self-interest hypothesis make very good predictions for competitive markets with standardized goods. This has been shown in many carefully conducted market experiments. However, a large amount of economic activity is taking place outside of competitive markets – in markets with a small number of traders, in markets with informational frictions, in firms and organizations, and under incompletely specified and incompletely enforceable contracts. In these environments models based on the self-interest assumption frequently make very misleading predictions. An important insight provided by some of the newly developed fairness models is that they show why in competitive environments with standardized goods the self-interest model is so successful and why in other environments it is refuted. In this way the new models provide fresh and experimentally confirmed insights into

important phenomena like, e. g., non-clearing markets or the wide-spread use of incomplete contracts.

We consider it as important to stress that the available experimental evidence also suggests that many subjects behave quite selfishly even when they are given a chance to affect other peoples well being at a relatively small cost. However, there are also many people who are strongly motivated by fairness and reciprocity and who are willing to reward or punish other people at a considerable cost to themselves. 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 the understanding of the observed behavior in strategic settings. These models explain why in some strategic settings almost all people behave as if they are completely selfish, while in others the same people will behave as if they are driven by fairness.

A second reason for the reluctance to give up the self-interest hypothesis is methodological. There is 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 sophisticated tools to examine the nature of preferences in a scientifically rigorous way. However, due to the development of experimental techniques this is no longer true. In fact, one purpose of this paper is to show that much progress and fascinating new insights into the nature of fairness preferences have been made in the past decade. While there is still much to be done this research clearly shows that it is possible to discriminate between theories based on different preference assumptions. Therefore, in view of the facts, the new theoretical developments, the importance of fairness concerns in many economic domains, and in view of the existence of rigorous experimental techniques that allow us to examine hitherto unsolvable problems in a scientific manner, we believe that it is time to recognize that a substantial fraction of the people is also motivated by fairness concerns. People do not only differ in their tastes for chocolate and bananas but also along a more fundamental dimension. They differ with regard to how selfish or fair-minded they are, and this does have important economic consequences.

The rest of this paper is organized as follows. Section 2 provides many real life examples indicating the relevance of fairness considerations and reviews the experimental evidence. It shows that the self-interest model is refuted in many important situations and that a substantial number of people seem to be strongly concerned about fairness and behave reciprocally. Section 3 surveys different theoretical approaches that try to explain the observed phenomena. In the

meantime there is also a large and growing literature on the evolutionary origins of reciprocity (see, e.g., Bowles and Gintis 1999, Gintis 2000, Sethi and Somananthan forthcoming and 2000). We do not discuss and review this literature in our paper. Section 4 discusses the wave of new experiments that have been conducted in order to discriminate between these theories. Section 5 explores the implications of fairness driven behavior in various economic applications and offers some directions for future research. Section 6 concludes.[1] In view of the length of our paper it is also possible to read the paper selectively. For example, readers who are already familiar with the basic evidence and the different fairness theories may go directly to the new evidence in Section 4 and the economic applications in Section 5.

2 Empirical Foundations of Fairness and Reciprocity

2.1 Where does Fairness matter?

The notion of fairness is frequently invoked in families, at the workplace, and in people's interactions with neighbors, friends and even strangers. For instance, our spouse becomes sour if we do not bear a fair share of family responsibilities. Our children are extremely unhappy and envious if they receive less attention and gifts than their brothers and sisters. We do not like those among our colleagues who persistently escape doing their share of important yet inconvenient departmental activities.

Fairness considerations are, however, not restricted to our personal interactions with others. They shape the behavior of people in important economic domains. For example, employee theft and the general work morale of employees is affected by the perceived fairness of the firm's policy (Bewley 1999;.Greenberg 1990). The impact of fairness and equity norms may render direct wage cuts unprofitable (Agell and Lundborg 1995; Kahneman, Knetsch and Thaler 1986). Firms may, therefore, be forced to cut wages in indirect ways, e.g., by outsourcing activities. Fairness concerns may thus influence decisions about the degree of vertical integration. They may also severely affect the hold-up problem as demonstrated by Ellingsen and Johannesson (2000). Debates about the appropriate income tax schedule are strongly affected by notions of merit and fairness (Seidl and Traub 1999). The amount of tax evasion is likely to be affected by the perceived fairness of the tax system (Andreoni, Erard and Feinstein 1998; Alm,

[1] In the meantime there is also a large and growing literature on the evolutionary origins of reciprocity (see, e.g., Bowles and Gintis 1999, Ginitis 2000, Sethi and Somananthan forthcoming and 2000). We do not discuss and review this literature in our paper.

Sanchez, de Juan 1995; Frey and Weck-Hanneman 1984). Public support for the regulation of private industries depends on the perceived fairness of the firms' policies (Zajac 1995). Compliance with contractual obligations, with organizational rules and with the law in general is strongly shaped by the perceived fairness of the allocation of material benefits and by issues of procedural justice (Fehr, Gächter and Kirchsteiger 1997, Lind and Tyler 1988). The functioning of incentive-compatible mechanisms has been shown to depend on fairness considerations (Andreoni and Varian 1999). The solution of collective action problems like, e.g., rules regulating the access to common pool resources, critically depends on the fairness of the allocation of the costs and benefits of the rules (Ostrom 1990 and 2000; Falk, Fehr and Fischbacher 2000c). The erosion of public support for the welfare state in the US in the last two decades has probably also much to do with deeply entrenched notions of reciprocal fairness (Bowles and Gintis 2000). Many people cease to support public programs that help the poor if they have the impression that the poor do not attempt to bear their share of a society's obligations.

Thus, real world examples where fairness concerns are likely to matter abound. Nevertheless, in the following we concentrate on clean experimental studies because in most real life situations it is impossible to unambiguously isolate the impact of fairness motives. A skeptic may always argue that the notion of fairness is only used for rhetorical purposes that disguises purely self-interested behavior in an equilibrium of a repeated game. Therefore, we rely on experimental evidence of human decision making. In these experiments real subjects make decisions with real monetary consequences in carefully controlled laboratory settings. In particular, the experimenter can implement one-shot interactions between the subjects so that long-term self-interest can be ruled out as an explanation for what we observe. As we will see, in some experiments the monetary stakes involved are quite high – amounting up to the income of three months' work. In the experiments reviewed below subjects do not know each others' identity, they interact anonymously and, sometimes, even the experimenter cannot observe their *individual* choices.

2.2 Experimental Evidence

In hindsight, it is a bit ironical that experiments have proven to be critical for the discovery and the understanding of fairness-driven behavior because for several decades experimental economists were firmly convinced that fairness motives would not matter much. At best, fair behavior was viewed as a temporary deviation from the strong forces of self-interest. In the 1950s

Vernon Smith discovered that under relatively weak conditions experimental markets quickly converge to the competitive equilibrium.[2] Since then the remarkable convergence properties of experimental markets have been confirmed by hundreds of experiments (see, e. g., Davis and Holt 1993). For these experiments the equilibrium is computed under the assumption that *all* players are *exclusively* self-interested. Therefore, the quick convergence to equilibrium has been interpreted as a confirmation of the self-interest hypothesis. We will see later in this paper that this conclusion was premature because, as the newly developed models of fairness (see Section 3 and Section 5.1) show, convergence to standard competitive predictions can occur even if agents are very strongly concerned about fairness.

This strong commitment to the self-interest hypothesis slowly weakened in the 1980s when experimental economists started to study 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 ultimately led many people to realize that the self-interest hypothesis is problematic was the so-called Ultimatum Game invented by Güth, Schmittberger and Schwarze (1982). In addition, the Gift Exchange Game, the Trust Game, the Dictator Game and Public Good Games played an important role in weakening the exclusive reliance on the self-interest hypothesis. All these games share the feature of simplicity. Because they are so simple, they are easy to understand for the experimental subjects and this makes inferences about subjects' motives more convincing.

In the Ultimatum Game (UG) 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 the case of rejection, both receive nothing; in the 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.

A robust result in the UG, across hundreds of experiments, is that proposals offering the Responder less than 20 percent of the available 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, and the references

[2] Smith's results were eventually published in the Journal of Political Economy in 1962, after time consuming debates with the referees. It is also ironical that Smith's initial aim was „to do a more credible job of rejecting competitive price theory" than Chamberlin (1948).

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. This is suggested, for example, by the comparison of the results of Dictator Games (DG) and Ultimatum Games. In a DG the Responder's option to reject is removed – the Responder must accept any proposal. Forsythe et al. (1994) were the first who compared the offers in UGs and DGs. They report that offers are substantially higher in the UG which suggests that many Proposers do apply backwards induction. This interpretation is also supported by the surprising observation of Roth, Prasnikar, Okuno-Fujiwara and Zamir, 1991, who showed that the modal offer in the UG tends to maximize the expected income of the Proposer.[3]

The UG shows that a sizeable fraction of Responders is willing to punish behavior that is perceived as unfair. In contrast, the Gift Exchange Game (GEG) indicates that a substantial fraction of the Responders are willing to reward actions that are perceived as generous or fair. The first GEG has been conducted by Fehr, Kirchsteiger and Riedl (1993). In the GEG the Proposer offers an amount of money $w \in [\underline{w}, \bar{w}]$, $\underline{w} \geq 0$, 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 $e \in [\underline{e}, \bar{e}], \underline{e} > 0$. The monetary payoff for the Proposer is $x^P = ve - w$ while the Responder's payoff is $x^R = w - c(e)$ where v denotes the marginal value of effort for the Proposer and $c(e)$ the strictly increasing effort cost schedule.[4] Under the standard assumptions (i) and (ii) above the Responder will always choose the lowest feasible effort level $\underline{e}$ and will, in equilibrium, never reject any w. Therefore, the subgame perfect proposal is the lowest feasible wage level $\underline{w}$.

The GEG captures a principal-agent relation with highly incomplete contracts in a stylized way. Variants of the GEG have been conducted by several authors.[5] 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

[3] Suleiman (1996) reports the results of UGs 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.

[4] In some applications of this game the Proposer's payoff was given by $x^P = (v - w)e$. This formulation rules out that Proposers can make losses when they offer generously high wages. Likewise, in some applications of the GEG the Responder did not have the option to reject w. Thus, the Proposer just sent w while the Responder choose an effort level. Under the standard assumptions of rationality and selfishness the subgame perfect equilibrium is, however, not affected by these differences.

[5] See, e. g., Fehr, Kirchsteiger and Riedl (1993, 1998), Charness (1996, 2000), Fehr and Falk, (1999), Gächter and Falk (1999), Falk, Gächter and Kovacs (1999), Hannan, Kagel and Moser (1999) and Brandts and Charness (1999).

choices. However, as in the case of the UG, there are considerable individual differences among the Responders. While there typically is a sizeable fraction of Responders (frequently roughly 40 percent, sometimes more than 50 percent) who exhibit a reciprocal effort pattern, there is also a substantial fraction of Responders who always make purely selfish effort choices or whose choices seem to deviate randomly from the self-interested action. Despite the presence of selfish Responders the relation between average effort and wages is in general sufficiently steep to render a high wage policy profitable. This induces Proposers to pay wages far above $\underline{w}$. Evidence for this interpretation comes from Fehr, Kirchsteiger and Riedl who embedded the GEG into an experimental market. In addition to the embedded GEG – there was a control condition in which the effort level was exogenously fixed by the experimenter. Note that in the control condition the Responders can no longer reward generous wages with high effort levels. It turns out that the average wage is substantially reduced when the effort is exogenously fixed.

Another important game that did much to change the exclusive reliance on the self-interest hypothesis was the Trust Game (TG), first studied by Berg, Dickhaut and McCabe (1995). In a TG a Proposer receives an amount of money *y* from the experimenter, and then can send between zero and *y* to the Responder. The experimenter then triples the amount sent, which we term *z*, so that the Responder has *3z*. The Responder is then free to return anything between zero and 3*z* to the Proposer. It turns out that many Proposers send money and that many Responders give back some money. Moreover, there is frequently a strong correlation between *z* and the amount sent back at the individual as well as at the aggregate level (see e.g., Miller 1997, Fahr and Irlenbusch 2000, Cox 2000).

Finally, we briefly consider the evidence on Public Good Games (PGGs). Like the GEG the PGG is not only important because it provides interesting insights into the nature of non-pecuniary motivations but it also captures the essence of numerous real world situations. There is by now a huge experimental literature on PGGs (see Ledyard, 1995, Dawes and Thaler 1988 for surveys). In the typical experiment there are *n* players who simultaneously decide how much of their endowment to contribute to a public good. Player *i*'s monetary payoff is given by $x_i = y_i - g_i + m\Sigma g_j$ where y_i is player i's endowment, g_i her contribution, *m* the monetary payoff per unit of the public good and Σg_j the amount of the public good provided by all players. The unit payoff *m* obeys $m < 1 < nm$. This ensures that it is a dominant strategy to contribute nothing to the public good although the total surplus would be maximized if all players contributed their whole endowment.[6] In many experiments the PGG is repeated for about 10 periods where in each period the group composition

[6] Typically, endowments are identical and $n \leq 10$ but there are also experiments with a group size of 40 and 100 (Isaac, Walker and Williams 1994).

changes randomly. If we restrict attention to behavior in the final period (in order to abstract from repeated games or learning effects) it turns out that roughly 75 percent of all subjects contribute nothing to the public good and the rest contributes very little.[7]

If one adds to the PGG the opportunity to punish other group members the contribution pattern changes radically (Fehr and Gächter, 2000). In a PGG with a punishment option there are two stages. Stage one is identical to the above described PGG. At 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 players. The assignment of one punishment point reduces the first-stage income of the punished subject by 3 points on average but it also reduces the income of the punisher according to a strictly increasing and convex cost schedule. 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 that nobody will contribute, i.e., there should be no difference in the contribution behavior between the usual PGG and a PGG with a punishment opportunity. The experimental evidence is, however, completely at odds with this prediction. While in the usual PGG cooperation is close to zero in the final period, the punishment opportunity causes, on average, stable cooperation rates around 75 percent of subjects' endowment.[8] The reason for these huge differences in contribution behavior is that in the punishment condition many cooperators punish the free-riders. The more a subject deviates from the average contribution of the other group members the more it is punished. Thus, the willingness to punish "unfair" behavior is not restricted to the UG.

The above mentioned facts in the UG, the GEG, the TG and the PGG are now well established and there is little disagreement about them. But there are, of course, questions about which factors change the behavior in these games. For example, a question that routinely comes up in discussions with economists is whether a rise in the stake level will eventually induce subjects to behave in a self-interested manner. There are several papers examining this question (Hoffman McCabe and Smith 1995, Fehr and Tougareva 1995, Slonim and Roth 1998, Cameron 1999). The surprising answer is that relatively large increases in the monetary stakes did nothing or little to change behavior. Hoffman, McCabe and Smith could not detect any effect of the stake level in their

[7] At the beginning of a repeated PGG subjects contribute on average between 40 and 60 percent of their endowment but towards the end contributions are typically very low. This pattern may be due to repeated game effects. Another plausible reason for the decay of cooperation is that many subjects are conditional cooperators as shown by Croson (1999), Fischbacher, Gächter and Fehr (1999) and Sonnemans, Schram and Offerman (1999). Conditional cooperators cease to cooperate once they notice that selfish subjects take advantage of their cooperation.

[8] If the same subjects are allowed to stay together for ten periods the cooperation rate even climbs to 90 percent of subjects' endowments *in the final period*. In Fehr and Gächter (2000) the group size was $n = 4$. Recently, Carpenter (2000) showed that with a group size of $n = 10$ subjects achieve almost *full* co-operation even with a random group composition over time.

UGs. Fehr and Tougareva conducted GEGs (embedded in a competitive exerimental market) in Moscow. In one condition the subjects earned, on average, the equivalent amount of the income of one week in the experiment. In another condition they earned the equivalent of a ten weeks' income. Despite this large difference in the stake size there are no significant differences across conditions in the behavior of both the Proposers and the Responders. Slonim and Roth conducted UGs in Slovakia. They found a small interaction effect between experience and the stake level. In the final period of a series of one-shot UGs the Responders in the high-stake condition (with a 10-fold increase in the stake level relative to the low stake condition) seem to be willing to reject a bit less frequently. Fehr and Tougareva also allowed subjects to repeat the game (with randomly matched partners). They found no such interaction effects. Cameron conducted UGs in Indonesia and – in the high stake condition - subjects could earn the equivalent of three months' income in her experiment. She observed no effect of the stake level on Proposers' behavior and a slight reduction of the rejection probability when stakes were high.

Of course, it is still possible that in the presence of extremely high stakes there may be a shift towards more selfish behavior. However, for large segments of the population this is not the economically relevant question. For almost all people the vast majority of their decisions involves stake levels well below three months' income. Thus, even if fairness-driven behavior would play no role at all at stake levels above that size, fairness concerns would still play a major role in many economically important domains.

2.3 Interpretation of the Evidence

While there is now little disagreement regarding the facts, there is still disagreement about the interpretation of these facts. In Section 3 we will describe several recently developed theories of fairness that maintain the rationality assumption but change the assumption of purely selfish preferences. Some researchers have, however, reservations about changes in the motivational assumptions and prefer, instead, to 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 UG by learning models that are based on purely pecuniary preferences. These models are based on the idea that the rejection of low offers is not very costly for the Responder and, therefore, the Responders learn only very slowly not to reject such offers. The rejection of offers is, however, quite costly for the Proposers. Therefore, Proposers learn more quickly that it does not pay to make low offers. Moreover, since Proposers quickly learn to make fair offers, the pressure on the Responders to learn accepting 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 the behavior of subjects changes over time and it seems clear that learning models are prime candidates to explain such dynamic patterns. We believe, however, that attempts to explain the basic facts in such simple games as the UG, the GEG and the TG in terms of learning models that assume completely selfish preferences are misplaced. The decisions of the Responders, 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 although their true preferences would require them not to do so. Moreover, the above cited evidence from Roth et al. (1991) Forsythe et al (1995), Suleiman (1996) and Fehr, Kirchsteiger and Riedl (1998) suggests that many Proposers do anticipate Responders' actions surprisingly well. Thus, at least in these simple two-stage games, many Proposers seem to be quite rational and forward looking.

Sometimes it is also argued that the behavior in these games is due to a social norm (see, e. g., Binmore 1998). In real life, so the argument goes, experimental subjects make the bulk of their decisions in repeated interactions. It is well known that in repeated interactions the rejection of unfair offers or the rewarding of generous offers can be sustained as an equilibrium. According to this argument, notions of fairness perform the function of selecting a particular equilibrium among the infinitely many equilibria that typically exist in long-term interactions. Subjects' behavior is, therefore, 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. This argument essentially boils down to the claim that subjects cannot rationally distinguish between one-shot and repeated interactions. One problem with this argument – apart from claiming that subjects make systematic mistakes – is that it cannot explain the huge behavioral variations across one-shot games. Why do in Forsythe et al. (1995) 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 so low wage offers? Why is there so much defection in the final round of PGGs while in the presence of a punishment opportunity a high level of co-operation can be achieved? Invoking some kind of social norm cannot explain this behavior unless one is willing to assume that different social norms apply to these different situations. A second problem with the above argument is that there is compelling evidence that in repeated interactions experimental

subjects do behave very differently compared to one-shot situations. In Gächter and Falk (1999) it is shown that the Responders in GEGs put forward much higher effort levels if they can stay together with the same Proposer.[9] In fact, experimental subjects who participate in one-shot GEGs frequently complain after the experiment that the experimenter ruled out repeated interactions because that would have enabled them, so the subjects' claim, to develop a much more trustful and efficient relation with their partner. All this indicates that experimental subjects are well aware of the difference between one-shot interactions and repeated interactions.

The above arguments suggest that an approach that combines bounded rationality with purely selfish preferences does not provide a satisfactory explanation of the facts observed in UGs, GEGs, TGs and PGGs. In our view, there remain two plausible approaches to account for the facts. One approach is to maintain the assumption of rationality at least for the analysis of these simple games and to assume, in addition, that some players are not only motivated by pecuniary forces. The other approach is, to combine models of learning with models that take into account non-selfish motives. In the following we focus on the first approach because there has been much progress in this area in recent years, while the second approach is still in its infancy.[10]

3 Theories of Fairness and Reciprocity

This section surveys the most prominent recent attempts to explain the experimental evidence sketched in Section 2 within a rational choice framework. Two main approaches can be distinguished. The first approach assumes that at least some agents have "social preferences", i.e., the utility function of these agents does not only depend on the own material payoff but also on how much the other players receive. Given these social preferences all agents are assumed to behave perfectly rational and 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. The second approach focuses on "intention-based reciprocity". This approach assumes that a player cares about the intentions of her opponent. If she feels treated kindly, she wants to return the favor and be nice to her opponent. If she feels treated badly, she wants to hurt her opponent. Thus, in this approach it is crucial how a player interprets the behavior of the other players. This cannot be captured by traditional game theory but requires the framework of psychological game theory.

[9] Andreoni and Miller (1993) also report that in Prisonners' Dilemmas increases in the probability of staying together or meeting the same partner again increase cooperation rates.

[10] An exemption is the recent paper by Cooper and Stockman (1999) that combines reenforcement learning with a model of social preferences and the paper by Costa-Gomes and Zauner (1999).

The starting point of both of these approaches is to make rather specific assumptions on the utility functions of the players. Alternatively, one could start from a general preference relation and ask what kind of axioms are necessary and sufficient to generate utility functions with certain properties. Axiomatic approaches are discussed at the end of this section.

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, in almost all applications this fairly general framework is interpreted much more narrowly 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 how much material resources are allocated to others.

Somewhat more formally, let $\{1,2,...,N\}$ denote a set of individuals and $x=(x_1,x_2,...,x_N)$ denote an allocation of physical resources out of some set X of feasible allocations, where x_i denotes the material resources allocated to person i. The self-interest hypothesis says that the utility of individual i depends on x_i only. 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 too general because it does not yield any empirically testable restrictions on observed behavior. In the following we will discuss several models of social preferences, each of which assumes that the preferences of an individual 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 $u(x_1,...,x_N)$ with respect to $x_1,...,x_N$ are strictly positive, i.e., if her utility increases with the well being of other people.[11] The hypothesis that people are altruistic has a long tradition in economics and has been used to explain charitable donations and the voluntary provision of public goods (see, e.g., Becker, 1974).

[11] The Encyclopaedia Britannica (1998, 15th edition) defines an altruistic agent as someone who feels the obligation "to further the pleasures and alleviate the pains of other people". Note that our definition of altruism differs somewhat from the definition used in moral philosophy, where "altruism" requires a moral agent to be concerned *only* about the welfare of others and not about his own happiness.

Clearly, the simplest game to elicit altruistic preferences, is the Dictator Game. Adreoni and Miller (2000) 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 Adreoni 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. Then they 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 the monetary payoffs. However, there are also almost 50 percent of the subjects who behaved "selfishly" and did not give any significant amounts to the other party. Andreoni and Miller (2000, p.23) conclude that altruistic behavior exists and that it is consistent with rationality, but also that individuals are heterogeneous.

Charness and Rabin (2000) consider a specific form of altruism which they call *quasi-maximin preferences*. They start from a "disinterested social welfare function" which is a convex combination of Rawls' maximin criterion and a utilitarian welfare function:

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

where $\delta\in(0,1)$ is a parameter reflecting the weight that is put on the maximin criterion. The utility function of an individual is then given by a convex combination of his own monetary payoff and the above social welfare function:[12]

$$U_i(x_1,x_2,...,x_N)=(1-\gamma)x_1+\gamma[\delta\cdot min\{x_1,...,x_N\}+(1-\delta)\cdot(x_1+...+x_N)] \ .$$

In the two player case this boils down to

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

[12] 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, the importance of the maximin term in relation to the player's own material payoff changes.

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 giving in Dictator Games, voluntary contributions in Public Good Games, and the kind behavior of Responders in trust and Gift Exchange Games,[13] but it is clearly inconsistent with the fact that in some experiments subjects try to retaliate and hurt other subjects even if this is costly for them (as in the ultimatum game or a public good game with punishments). This is why Charness and Rabin augment quasi-maximin preferences by incorporating reciprocity (see Section 3.2.3 below).

3.1.2 Relative Income and Envy

An alternative hypothesis is that subjects are concerned not only about the absolute amount of money they receive but also about their relative standing compared to others. This "relative income hypothesis" has a long tradition in economics and goes back at least to Veblen (1922). Bolton (1991) formalized this idea in the context of an experimental bargaining game between two players and assumed 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 $x_i \geq x_j$. 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 $\partial U_i/\partial x_j \leq 0$, just the opposite of altruism. Hence, while this utility function is consistent with the behavior in the bargaining games considered by Bolton, it fails to explain giving in dictator, gift exchange and trust games or voluntary contributions in public good games. The same problem arises in the envy-approach of Kirchsteiger (1994).

[13] However, even in these games altruism has some implausible implications. For example, in a public good context, altruism implies that if the government provides part of the public good (financed by taxes) 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 lead some researchers to include the pleasure of giving (a "warm glow effect") in the utility function (Andreoni, 1989).

3.1.3 Inequity Aversion

The preceding approaches assumed 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 material payoffs of the other players exceed this level.[14] In most experiments it is natural to assume that an equitable allocation is an equal monetary payoff for all players. Fehr and Schmidt consider the simplest utility function capturing this idea.

$$U_i(x_1,...,x_N) = x_i - [\alpha_i/(N-1)]max\Sigma_{j\neq i}\{x_j-x_i,0\} - [\beta_i/(N-1)]max\Sigma_{j\neq i}\{x_i-x_j,0\}.$$

with $\beta_i \leq \alpha_i$ and $\beta_i \leq 1$. Note that $\partial U_i/\partial x_j \geq 0$ if and only if $x_i \geq x_j$. 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 ($\alpha_i \geq \beta_i$).

This utility function can rationalize positive *and* negative actions towards other players. It is consistent with giving in dictator, gift exchange and trust games, *and* with the rejection of low offers in ultimatum games. It can also explain voluntary contributions in public good games *and* the costly punishment 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 in other environments fairness concerns or the desire to cooperate do not seem to have much of an effect. Fehr-Schmidt show that the interaction of the distribution of types with the strategic environment explains why in some situations very unequal outcomes are obtained while in other situations very egalitarian outcomes prevail. For example, in certain competitive environments (see, e.g., the ultimatum game with Proposer competition in Section 5.1) even a population that consists *only* of very fair types (high α*'s* and β*'s*) cannot prevent very uneven outcomes. The reason is that none of the inequity averse players can enforce a more equitable outcome through her own actions. In contrast, in a public good game with punishment, a small fraction of inequity averse players is sufficient to credibly threaten that free riders will be punished which induces selfish players to contribute to the public good.

[14] 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.

Using the data that is available from many experiments on the ultimatum game, Fehr and Schmidt calibrate the distribution of α and β in the population. Keeping this distribution constant, they show that their model yields quantitatively accurate predictions across many bargaining, market and co-operation games.[15]

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 like, e.g., giving in DGs and GEGs and rejections in UGs. In their model the utility function is given by

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

where

$$\sigma_i = \begin{cases} \frac{x_i}{\sum_{j=1}^{N} x_j} & if \ \sum_{j=1}^{N} x_j \neq 0 \\ \frac{1}{N} & if \ \sum_{j=1}^{N} x_j = 0 \end{cases}$$

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 $\sigma_i = 1/N$.[16] Bolton and Ockenfels do not pin down a specific functional form, so their utility function is more flexible. However, this also makes it more difficult to get closed form solutions and quantitative predictions for the outcomes of many experiments. It also imposes less discipline on the researcher not to adjust the utility function to a specific set of data.

For two-player-games Fehr-Schmidt and Bolton-Ockenfels often yield qualitatively similar results. With more than two players there are some interesting differences. In this case

---

[15] 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, in the dictator game, a decision maker 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 $\beta = 0.5$. However, this problem can be avoided by assuming non-linear inequity aversion.

[16] 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. A related problem has been noted by Camerer (1999, p. 61). Suppose that the ultimatum game is modified as follows: If the Responder rejects a proposal the Proposer receives a small amount $\varepsilon > 0$ while the Responder receives zero. In this game the rejection of a *positive* offer implies $\sigma = 0$ while acceptance implies $\sigma > 0$. Thus, the Responder never rejects any positive offer no matter how small $\varepsilon > 0$.

Fehr and Schmidt assume that a player compares herself to each of her opponents separately. 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 as 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*.[17]

3.1.4 Altruism and Spitefulness

Levine (1998) offers a different solution to explain giving in some games and punishing in others. Consider the utility function

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

where $0\leq\lambda\leq 1$ and $-1<a_i<1$ for all $i\in\{1,...,N\}$. Suppose first that $\lambda=0$. In this case the utility function reduces to $U_i=x_i+a_i\ \Sigma_{j\neq i}x_j$. 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 some (other) people reject positive offers in the ultimatum game, it cannot explain why the same person who is altruistic in one setting is spiteful in another. To deal with this problem, 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 the parameter a_j of their opponents 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 to determine 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* and to estimate λ (which is assumed to be the same for all players). He shows that with these parameters the model can reasonably fit the data on centipete games, market games, and public good games. However, because $a_i<1$, the model cannot explain positive giving in the dictator game.

[17] See Camerer (1999) and Section 4.1 for a more extensive comparison of these two approaches.

3.2 Models of Intention-Based Reciprocity

Models of social preferences share a common weakness. They assume that players are only concerned about the distributional consequences of their acts but not about the intentions that lead their opponents to choose these acts. To see that this may be a problem consider the following two "mini-ultimatum games" in which the strategy set of the Proposer is restricted. In the first condition the Proposer can choose between a 50:50 and an 80:20 split. In the second condition the Proposer must choose between an 80:20 and a 20:80 division of the pie. All theories that look only at the distributional consequences must predict that if a Responder rejects the 80:20 split in the first condition, then she must also reject this offer in the second condition. However, in the second condition a fair division of the pie was not feasible and so the Responder may be more inclined to accept this offer as compared to the first treatment where the Proposer could have split the pie evenly but chose not to do so. In fact, Falk, Fehr and Fischbacher (2000a) report that the 80:20 split is rejected significantly less often under the second condition.[18] This is inconsistent with any theory of social preferences that rely only on preferences over income distributions.

3.2.1 Fairness Equilibrium

In a pioneering article, Rabin (1993) starts from the observation that our behavior is often a reaction to the (expected) *intentions* of other people. If we feel that another person has been kind to us, we often have a desire to be kind as well. If we feel that somebody wanted to hurt us, we often have the desire to retaliate even if this is personally costly.

In order to model intentions explicitly, Rabin departs from traditional game theory and adopts the concept of "psychological game theory" that had been introduced by Geanakoplos, Pearce and Stacchetti (1989). In psychological game theory, utilities do not only depend on terminal-node payoffs but also on players' beliefs. Rabin restricts attention to 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: $A_1 \times A_2 \rightarrow IR$ be player *i*'s material payoff function.

[18] This criticism does not necessarily apply to Levine (1998). In his model, offering 80:20 may be interpreted as a signal that the Proposer is spiteful if the 50:50 split was available, and may be differently interpreted if the 50:50 split was not available. However, if a player knows the type of her opponent, her behavior is independent of what the opponent does to her and of why he does it to her.

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 to be chosen by player *j*. 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 to define 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 chooses effectively her opponents 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 the payoff of player *i*. 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]

$$f_i(a_i,b_j) \equiv [x_j(b_j,a_i)-x_j^f(b_j)]/[x_j^h(b_j)-x_j^l(b_j)]$$

with $j=3-i$ and $f_i(a_i, b_j)=0$ 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 kind she is being treated by player *j*. 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:

$$f_j'(b_j,c_i) \equiv [x_i(c_i,b_j)-x_i^f(c_i)]/[x_i^h(c_i)-x_i^l(c_i)]$$

with $j=3-i$ and $f_j(b_j, c_i)=0$ 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:

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

[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.

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. Note also, that the kindness terms have no dimension and that they must lie in the interval $[-1,0.5]$. Thus, the utility function is sensitive to positive affine transformations. Furthermore, the kindness term becomes less and less important the higher the material payoffs are.

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 mutually 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 made the notion of reciprocity precise 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 there exist many other unreasonable equilibria including equilibria in which the Responders receives more than 50 percent of the pie. The multiplicity of equilibria is a general feature of Rabin's model. If material payoffs are sufficiently small so that psychological payoffs matter, then there are always multiple equilibria. In particular, there is one equilibrium in which both players are nice to each other and one in which they are nasty. 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 in the prisoners' dilemma player 2 has no choice but is forced to cooperate. If player 1 knows this, then - according to Rabin's theory - she will interpret player 2's cooperation as "neutral" $(f_2'(\cdot)=0)$. Thus, she will only look at her material payoffs and will defect. This contrasts with models 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.

3.2.2 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, in the sequential prisoners' dilemma, unconditional cooperation of the second player is

part of a "fairness" equilibrium. 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 (1998) 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 is a SRE. They also show that it can be a SRE in the ultimatum game that the Proposer makes an offer that is rejected by the Responder with certainty. This is an equilibrium because both players believe that the other party wants to hurt them. However, even in these extremely simple sequential games the equilibrium analysis is fairly complex, and there are typically many equilibria with different equilibrium outcomes due to different self-fulfilling beliefs about intentions.

3.2.3 Merging Intentions and Social Preferences

Falk and Fischbacher (1999) also generalize Rabin (1993). They consider *N*-person extensive form games and allow for the possibility of incomplete information. Furthermore, they measure "kindness" in terms of inequity aversion. A strategy of player *j* is perceived to be kind by player *i* if it gives rise to a payoff for player *i* which is higher than the payoff of player *j*. Note that this is fundamentally different from Rabin and Dufwenberg and Kirchsteiger who define "kindness" in relation to the feasible payoffs of player *i* and not in relation to the payoff that player *j* gets. Furthermore, Falk and Fischbacher distinguish whether an unequal distribution could have been altered by player *j* or whether player *j* was a "dummy player" who is unable to affect the distribution by his actions. In the former case the kindness term gets a higher weight than in the latter. However, even if player *j* is a dummy player who has no choice to make, the kindness term

[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 as those which yield weakly less on the equilibrium path. However, with more than two players in Dufwenberg and Kirchsteiger (1998) the problem arises that an additional dummy player may render an inefficient strategy efficient and might thus affect the size of 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 á 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 and 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 complexity of the model.

Another attempt to combine social preferences with intention based reciprocity is due to Charness and Rabin (1999). We described their model of quasi-maximin preferences in Section 3.1.1 already. In a second step they augment these preferences by introducing a demerit profile $\rho \equiv (\rho_1,...,\rho_N)$, 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

$$U_i(x_1,x_2,...,x_N|\rho)=(1-\gamma)x_i+\gamma[\delta\cdot min\{x_i,min_{j\neq i}\{x_j+d\rho_j\}\}$$

$$+(1-\delta)\cdot(x_i+\Sigma_{j\neq i}max\{1-k\rho_j,0\}\cdot x_j)-f\Sigma_{j\neq i}\rho_j x_j]$$

where $d,k,f\geq 0$ 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 the well being of player j. 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 an unanimously agreed upon, exogenously given "selfless

standard" of behavior. 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 is maximizing 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 a 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 it 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 that all players 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, then 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 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 (1999) take the opposite approach and ask what kind of axioms generate preferences that can reflect fairness and reciprocity.

Their starting point is to assume 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 $(\sigma_1,\sigma_2) \in \Sigma \times \Sigma_1$ let $v_i(\sigma_1,\sigma_2)$ denote player *i*'s material payoff function, assuming that these "selfish preferences" satisfy the von Neumann-Morgenstern axioms. However, the actual preferences of player *i* are given by a preference relation $\Delta_{i,\sigma j}$ over her own strategies. Note that this preference relation depends on the

strategy chosen by player *j*. Segal and Sobel show that if the preference relation $\Delta_{i,\sigma j}$ 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 (self interest), then the preferences $\Delta_{i,\sigma j}$ 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 spitelful.

Note that the coefficient $a_{i,\sigma j}$ depends on σ_j. Therefore, whether a player is altruistic or spiteful may depend on the strategy chosen by her opponent, so there is scope to model reciprocity. In order to do so, Segal and Sobel introduce an additional axiom, called "reciprocal altruism". This axiom requires that when player *j* chooses a strategy σ_j which player i likes better than some other strategy σ_j', then player *i* prefers strategies that give a higher payoff to player *j*. Segal and Sobel show that this axiom implies that the coefficient $a_{i\sigma j}$ varies with σ_j such that (other things being equal) the coefficient increases if and only if player j chooses a "nicer" strategy.

The models of social preferences that 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, in Rabin's (1993) model a player's utility did not only depend on the strategy chosen by her opponent, but also on why he has chosen 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 kind of player 2 given that he believes player 1 to go to boxing). Yet, she may also go to boxing, because she expects player 2 to go to ballet (which is unkind of player 2 if he believes player 1 to go to boxing) 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 (2000) 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 payoff differences between player *i* and any subset of all other players remain

[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.

constant, then the preferences of player *i* should not be affected by the magnitude of these differences. 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. Neilson also offers a full axiomatic characterization of the more specific functional form used by Fehr and Schmidt.

4 Discriminating between Theories of Fairness

Most theories discussed in Section 3 have been developed during the last few 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 that give a first indication of the advantages and disadvantages of the different theories.[22]

4.1 Who are the relevant Reference Actors?

All theories of fairness and reciprocity are based on the idea that actors compare themselves with a set of reference actors. To whom do people compare themselves? In bilateral interactions there is no ambiguity about who the relevant reference actor is. In multi-person interactions, however, the answer is less clear. Most of the theories that are 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 renders the theory unable to explain the punishment pattern in the public good game with punishment. Remember that in this experiment the assignment of one punishment point reduces the income of the punished member by 3 points. The theory of BO predicts that punishing subjects are indifferent between punishing a free-rider and punishing a cooperator. All that matters is whether punishment brings the income of the punishing subject closer to the average income in the group and for this purpose the

[22] This section rests to a large extent on joint work of one of the authors with Armin Falk and Urs Fischbacher (Falk, Fehr, Fischbacher 2000a and 2000b, henceforth FFF). In particular, the organization of this section according to the questions below and many of the empirical results emerged from this joint project.

punishment of a cooperator is equally good as the punishment of a defector. Yet, in contrast to this indifference prediction the cooperators predominantly punish the defectors.

To further test the BO-model, Fehr and Fischbacher (2000) conducted the following Third-Party Punishment Game. There are three players, A, B, and C. Player A is endowed with 100 experimental currency units and must decide how much of the 100 units to give to B who has no endowment. Player B is just a dummy player and has no decision power. Player C has an endowment of 50 units 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.[23] Note that without punishment player C is certain to get her fair share of the total surplus (50 out of 150 units). Therefore, BO predict that C will never punish. In contrast to this prediction players A are, however, punished a lot. The less player A gives to B the more C punishes A. For example, if A gives nothing his income is reduced by roughly 30 percent. This indicates that many players do care about inequities among other players. Further support for this hypothesis comes from Charness and Rabin (2000) 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, BO predict 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 in which fair-minded subjects have multiple reference agents do not necessarily imply that fair subjects take actions in favor of *all* other reference agents. To illustrate this, consider the following three-person UG (Güth and van Damme 1998). In this game there is 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 in this game the Proposers allocate substantial fractions of the surplus to the Responder but little or nothing to the Receiver. Moreover, Güth and van Damme (p. 230)

[23] In the experimental instructions the value laden term „punishment" was not used. The punishment option of player C was described in neutral terms by telling subjects that player C could "assign points" to player A that reduced the incomes of A and C in the way described above.

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 $\beta \leq 1$, 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.[24]

4.2 Equality versus Efficiency

Many models of fairness are based on the definition of a fair or equitable outcome to which people compare the available payoff allocations. In experimental games a natural first approximation for the relevant reference outcome is the equality of material payoffs. The quasi-maximin theory of Charness and Rabin assumes instead that subjects care for the total surplus 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 Miller (2000), Bolle and Kritikos (1998), Andreoni and Vesterlund (forthcoming), Charness and Rabin (2000), Cox (2000) and Güth, Kliemt and Ockenfels (2000). Except for the Güth et al. paper, these papers indicate that in DG-situations a non-negligible fraction of the subjects 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 DGs 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 DGs the exchange ratio is 1:1 and

[24] The Proposers 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.

there are virtually no cases in which an Allocator transfers more than 50 percent of the surplus. In contrast, in DGs with an exchange ratio of 1:3 (or 1:2) a non-negligible number of subjects makes transfers such that they end up with less money than the Receiver. This contradicts BO, FS, and Falk and Fischbacher because in these models fair subjects never take actions that give the other party more than they get. 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 DG 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 DG 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.

The evidence cited in the papers mentioned above indicates that surplus maximization is a relevant motive *in DGs*. This motive has not been included in the prevailing models of inequity aversion but it would be straightforward to do this. It should also be remembered that *any* positive transfer in DGs is incompatible with intention based reciprocity models, *irrespective of the exchange rate*. We would like to stress, however, that the DG is different from many economically important games and real life situations, because in economic interactions it is rarely the case that one player is at the complete mercy of another player. 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 DGs. 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 GEG is embedded into a competitive experimental market. The experiments also exhibit a strategic element because the Proposers are wage setters and have to take into account the likely effort responses of the Responders. Yet, once the Responder has accepted a wage offer, the experiments are similar to a DG 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 to check the robustness of the surplus maximization motive because an increase in the effort cost by one unit increases, on average, the total surplus by five units. 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 inequity aversion, one would expect that many Responders choose effort levels that give the Proposer a higher monetary payoff than the Responder.[25] 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.[26]

However, neither of these implications is supported by the data. 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 in less than two percent of the cases the Proposer earns a higher payoff than the Responder.[27] Moreover, almost all subjects who regularly chose non-minimal effort levels exhibited a reciprocal effort-wage relation. These numbers 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 gives rise to a violation of the equality constraint only 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

[25] 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.
[26] There are degenerate cases in which this is not true.
[27] 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.

by other considerations. This reason is quite plausible insofar as the outcomes in DGs themselves are notoriously non-robust.

While the experimental results on UGs, GEGs or PGGs are fairly robust, the DG seems to be a rather fragile situation in which minor factors can have large effects. Cox (2000), e. g., reports, that in his DGs *100 percent* of all subjects transferred positive amounts.[28] This result contrasts sharply with many other games, including the games in Charness and Rabin and many other DGs. To indicate the other extreme, Eichenberger and Oberholzer (1998), Hoffman, McCabe, Shachat and Smith (1994) and List and Cherry (2000) report on DGs with extremely low transfers.[29] Likewise, in the Impunity Game of Bolton and Zwick (1995), which is very close but not identical to a DG, the vast majority of Proposers did not shy away from making very unfair offers. The Impunity Game differs from the DG 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 DG indicates that one should be very careful in generalizing the results found in these situations to other games. Testing theories of social preferences in DGs is a bit like testing the law 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 DGs are also behaviorally relevant in economically more important games. One way to do this is to apply the theories that have been constructed on the basis of DG-experiments to predict outcomes in other games. With the exemption of Andreoni and Miller (2000) this has not yet been done.

Andreoni and Miller (2000) estimate utility functions based on the results of their DG-experiments and use them to predict co-operation behavior in a standard PGG. They predict behavior in period one of these games, where co-operation is often quite high, rather well. However, their predictions are far away from final period outcomes, where co-operation is typically very low. In our view the low co-operation 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 the payoff of others or the payoff 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

[28] In Cox's experiment both players had an endowment of 10 and the Allocator could transfer his endowment to the Receiver where the transferred amount was trippled by the experimenter.

[29] In Eichenberger and Oberholzer (1998) almost 90 percent of the subjects gave nothing. In Hoffman et al. (1992) 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.

provide a plausible explanation for this behavior. All of these models predict that fair subjects make their co-operation contingent on the co-operation of others. Thus, if the fair subjects realize that there are sufficiently many selfish decisions in the course of a PGG experiment, they cease to cooperate as well.

4.3 Revenge versus Inequity Reduction

Subjects with altruistic and quasi-maximin preferences do not take actions that reduce other subjects' payoffs. Yet, this 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 intention-based reciprocity models and in Levine (1998) 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 in games in which no intention can be expressed there will be no punishment. 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 UG. Blount compared the rejection rate in the usual UG to the rejection rates in UGs in which either a computer generated a random offer or a third party made the offer. Because in the random offer condition and the third party condition a low offer cannot be attributed to the greedy intentions of the Proposer, 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 models 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 Bount's experiments.[30] Subsequently, Offermann (1999) and FFF (2000b) conducted further experiments with computerized offers but without the other worrisome features in Blount. In particular, in these experiments the Responders knew that a rejection affects the payoff of a real, human "Proposer". Offerman finds that subjects are 67 percent more likely to reduce the opponent's payoff when the opponent made an intentional hurtful choice compared to a situation where a computer made the hurtful choice.

FFF (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 the computer makes a hurtful choice.

Kagel, Kim and Moser (1996) provide further support that intentions play a role for payoff-reducing behavior. In their experiments subjects bargained over 100 chips in an UG. They conducted several treatments that varied the money value of the chips and the information provided about the money value. For example, in one treatment the Proposers received three times more money per chip than the Responders, i.e., the equal money split requires that the Responders receive 75 chips. If the Responders know that the Proposers know the different money values of the chips they reject unequal money splits much more frequently than if the Responders know that the Proposers do *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 situations in which the subjects cannot increase their relative share or decrease payoff

[30] 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 in fact the experimenters 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.

differences. FFF (2000a) report the results of UGs and PGGs with punishment that have this feature. In the first (standard) treatment of the UG 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 a 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 BO- as well as the FS-model predict, therefore, that there will be no rejections in the second treatment while intention-based and type-based models predict that punishments 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 (19/57) of the rejections are consistent with a pure taste for punishment as conceptualized in intention- and type-based models.[31]

FFF (2000a) also report the results of PGGs with punishment in which the punishing subjects could not change the payoff difference between themselves and the punished subject. In one of their treatments subjects had to pay one money unit in order to reduce the payoff of another group member by one unit. Thus, BO and FS both predict that there will be no punishment at all in this condition. In a second treatment investing one unit into punishment reduced the payoff of the punished group member by three units.

FFF report that 51 percent of all subjects (N = 93) cooperate which is still compatible with both BO and FS. However, another 51 percent of all cooperators punish the defectors. They invest on average 4.8 money units into punishment. Thus, 25 percent of the subjects punish free-riding which is incompatible with BO and FS. To evaluate the relative importance of this amount of punishment we have to compare these results with the results of the second condition. In the second condition 61 percent of all subjects (N = 120) cooperate and 59 percent of them punish the defectors (by imposing a punishment of 5.7 on average). Thus, the overall percentage of subjects who punish the defectors in the second condition is 36 percent. This suggests that a rather large fraction (i.e., 25/36) of the overall amount of punishment is not consistent with BO and FS.

Taken together the evidence from Blount (1995), Offerman (1999) and FFF (2000b) indicates that the motive to punish unfair intentions or unfair types plays an important role. 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 also suggests that inequity aversion plays an additional, non-negligible role. The evidence from the experiments in FFF (2000a)

[31] Ahlert, Crüger and Güth (1999) also report a significant amount of punishment in UGs 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.

suggests that many subjects who reduce the payoff of other players do not have the desire to change the equitability of the payoff allocation. Instead, a large fraction of these subjects seems to be driven by the desire to punish, i.e., a desire to hurt the other player. It is worthwhile to point out that this desire to hurt the other players, while consistent with intention- and type based models of reciprocity, does not necessarily constitute evidence in favor of these models. The reason is that the desire to reduce the payoff of other players may also be triggered by an unfair payoff allocation per se. [32]

4.4 Does Kindness trigger Rewards?

Do intention- and type-based theories of fairness 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 rewarding behavior is almost unaffected by these motives. Other results indicate some minor role and only one paper finds 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 reveals that she 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 DG. Moreover, Charness (1996), Bolton, Brandts and Ockenfels (1998), Offerman (1999), Cox (2000) and Charness and Rabin (2000) provide further evidence that intentions do not play a big role for rewarding behavior. Charness (1996) conducted GEGs in a random choice condition and a human choice condition. Intention-based theories predict that in the random choice condition the Responders will not put forward more than the minimal effort level irrespective of the wage level because high wage offers are due to chance and not to kind intentions. In the human choice condition higher wages 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

[32] Assume that fair subjects have the following utility function: $u_i = x_i + \alpha_i[1/(n\text{-}1)][\Sigma_{j \neq i} \beta(x_i - x_j)v(x_j)]$, where α_i measures the strength of player i's non-pecuniary preference, and $v(\pi_j)$ is an increasing function of player j's material payoff. $\beta(x_i - x_j)$ is positive if $x_i - x_j > 0$ and negative if $x_i - x_j < 0$. Thus, a state of inequality triggers the desire to reduce or increase the other players' payoff. In this regard the above utility function is similar to the preference assumption in FS. Yet, in contrast to FS, the aim of player i is no longer the reduction of the payoff difference. Instead, player i just wants to reduce or increase the other player's payoff depending on the sign of β.

correlation between wages and effort in both conditions, because an increase in effort benefits the Proposer much more than they cost 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. In the human choice condition effort 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 (2000) tries to isolate rewarding responses in the context of a TG by using a related DG as a control condition. In the TG Cox observes a baseline level of Responder transfers back to the Proposer. To isolate the relevance of intention-driven responses he conducts a DG in which the distribution of endowments is identical to the distribution of material payoffs after the Proposers' choices in the TG. Thus, both in the TG and in the DG the Responders face exactly the same distributions of material payoffs but in the TG this distribution has been caused intentionally by the Proposers while in the DG the distribution is predetermined by the experimenter. In Cox' DG the motive of rewarding kindness can, therefore, play no role and 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 DG are driven by inequity aversion or quasi-maximin preferences, the difference between the transfers in the DG and the transfers in the TG measure the relevance of intention- or type-based theories. Cox' results indicate that these theories play only a minor or no role in this context. In one condition there is no difference in transfers between the TG and the DG and in another condition transfers in the DG are lower by only one third.

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 Prisonners' Dilemma. In one condition the first movers could make a kind choice relative to a baseline choice. The kind choice implied that – for any choice of the second mover- the payoff of the second mover 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 could only make the baseline choice, i.e. he could not express any kind intentions. It turns out that second movers reward the first movers even more in this 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 FFF (2000b) described in the previous subsection. FFF find that for *all* positive transfers of player A, players B send back significantly more money in the human choice condition. 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 (2000) also reports evidence in favor of intention driven positive reciprocity. They show that after a nice choice of the first-mover two thirds of the second movers make nice choices, too, while if the first mover is forced to make the nice choice only one third of the second movers make the nice choice.

In the absence of the evidence provided by FFF and McCabe et al. 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 Summary and Outlook

Although most fairness models 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 relative performance 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) Evidence from the Third Party Punishment Game and the PGG with punishment indicates that many subjects do compare themselves with other people in the group and not just to the group as a whole or to the group average.

2) There is a non-negligible number of subjects in DGs whose behavior is consistent with surplus maximization. However, the relative quantitative importance of this motive in

economically relevant settings has yet to be determined and surplus maximization alone cannot account for many robust regularities in other games.

3) Pure revenge as captured by reciprocity models is an important motive for payoff-reducing behavior. In some games like the PGG with punishment it seems to be the dominant source of payoff-reducing behavior. Since pure equity models do not capture this motive they cannot explain a significant amount of payoff-reducing behavior.

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

Which model of fairness does best in the light of the data and which one 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 of fairness has difficulties in explaining the full diversity of the experimental observations. The evidence suggests, however, some tentative answers to these questions. In our view the most important heterogeneity is the one between purely selfish subjects and fair-minded subjects. The success of the BO-model and the FS-model in explaining a large variety of data from bargaining, co-operation and market games is partly due to this recognition. Within the class of these equity models the evidence suggests that the FS-model does better. 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 to measure equity than differences in individual payoffs.

However, the FS-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 underestimates the amount of punishing behavior in situations where the cost of punishment is relatively high compared to the payoff-reductions that can be achieved by punishing. Fairness models that are exclusively based on intentions (Rabin 1993, Dufwenberg and Kirchsteiger 1998) can, in principle, account for this type of punishment. Yet, these models have other undesirable features - including multiple, and very counterintuitive, equilibria in many games and a very high degree of complexity that is due to the use of psychological game theory. The same has to be said about the intention-based theory of Charness and Rabin (2000). Falk and Fischbacher (1999) is

not plagued by the multiple equilibrium problem as much as the 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 can take into account the full complexity of motives at the individual level, some theories may allow for better approximations than others. The evidence presented in Section 2 shows clearly that there are many important economic problems for which the self-interest theory is unambiguously, and in a quantitatively important way, refuted. The recent papers by BO and FS show that one can account for the bulk of this evidence by models that explicitly take into account that there are selfish and fair-minded individuals. Although we believe that it is desirable to tackle the heterogeneity within the class of fair-minded subjects in parsimonious and tractable models, we also believe that the heterogeneity between selfish and fair types is more important. In fact, in the following section we will show that the FS-model provides surprisingly good qualitative and quantitative predictions in important economic domains. Thus, even if we do net yet have a fully satisfactory model of fair behavior, one can probably go a long way with simple models that take into account the interaction between selfish and fair types.

5 Economic Applications

5.1 Competition and Fairness – When Does Fairness Matter?

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.[33] 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, why don't they behave the same way when there is competition among the players?

To answer this question consider the following ultimatum game with Proposer competition, that was conducted by Roth, Prasnikar, Okuno-Fujiwara, and Zamir (1991) in four different countries. There are *n-1* Proposers who simultaneously offer a share $s_i \in [0,1]$, $i \in \{1, ..., n-1\}$, 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

[33] See e.g. Smith (1962) and Davis and Holt (1993).

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.[34]

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 in this strategic setting preferences for fairness or reciprocity cannot have any effect. To see this, suppose that each of the Proposers strongly dislikes to get 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, get a positive monetary payoff, and reduce the inequality between himself 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) and Levine (1998).

Does this mean that sufficiently strong competition will always wipe out the impact of fairness? The answer to this question is negative because fairness matters much more in market games in which the execution of contracts cannot be completely determined at the stage where the parties conclude the contracts. Labor markets are a good example. A labor contracts is highly incomplete, because it cannot enforce the level of effort provided by the employee who chooses his effort level after the contract has been signed. These contractual features are captured by the Gift Exchange Game (GEG) in an experimental setting.

[34] 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 signifant) 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.

When the GEG is embedded into a competitive experimental market, as e.g. in Fehr, Kirchsteiger and Riedl (1998, 1998), wages turn out to be systematically higher than the competitive equilibrium wage predicted by the self-interest model. There is also no tendency for wages to decrease over time. The reason for this stable wage premium is the effort behavior of the Responders: On average, effort levels are increasing with wages which provides an incentive for the firms to pay a wage premium. If, however, the effort level is fixed exogenously by the experimenter, the firms do not shy away from pushing down wages to the competitive level. FS and BO can explain this pattern in a straightforward manner. When effort is endogenous, inequity averse Responders respond to high wages with high effort levels in order to prevent an unequal distribution of the surplus from trade. This induces all firms (including purely selfish ones) to pay a wage premium because it is profitable to do so. When effort is exogenous this mechanism does not work and competition drives down wages to the competitive level.

5.2. Endogenous Incomplete Contracts

If fairness concerns affect the behavior of economic agents in so many situations, then it should also be taken into account in the design of incentive schemes. Surprisingly, hardly any theoretical and very little empirical or experimental work has been done to study the impact of fairness on incentive provision. Standard contract theory neglects this issue and assumes that all agents are only interested in their own material payoffs. Over the past two decades this theory has been highly successful in solving fairly complicated contractual problems and in designing very sophisticated mechanisms and incentive schemes. This gave rise to many important and fascinating insights, and the methods developed there have been applied in almost all areas of economics. However, standard contract theory still finds it difficult to explain the simplicity and incompleteness of many contracts that we observe in the real world. In particular, it cannot explain why the parties' monetary payoffs are often not tied to measures of performance that would be available at a relatively small cost. For example, the salary of a teacher or a university professor is rarely contingent on students' test scores, teaching ratings, or citations. These performance measures are readily available and easily verifiable, so one has to conclude that these contracts are deliberately left incomplete.[35]

[35] The literature on incomplete contracts acknowledges contractual incompleteness, but most of this literature simply assumes that no long-term contingent contracts are feasible and does not attempt to explain this premise. See, e.g., Grossman and Hart (1986) or Hart and Moore (1990) and Section 5.3 below. There is a small literature on endogenous incomplete contracts. Some papers in this literature, e.g. Aghion, Dewatripont and Rey (1994), Nöldeke and Schmidt (1995) or Edlin and Reichelstein (1996), show that in some situations a properly designed incomplete contract can implement the first best, so there is no need to write a more complete contract. Some other papers, e.g.

In a recent paper, Fehr, Klein and Schmidt (2000) take a fresh look at contractual incompleteness by taking concerns for fairness and reciprocity into account. They report on several simple principal-agent experiments in which the principal was given a choice whether to offer a "complete" contract or a less complete one. In the first experimental design an agent had to pick an effort level between 1 and 10 (at a monetary cost to herself) that is perfectly observed by a principal and can be verified (at a small fixed cost) to the courts. The principal can try to induce the agent to spend effort by imposing a fine on the agent that is enforced by the courts if she works too little. However, the fine is bounded above so that the highest implementable effort level ($e^*=4$) falls short of the first best efficient action ($e^{FB}=10$). In this contractual environment principal agent theory predicts that the principal should use the maximal fine in order to induce the agent to choose $e^*=4$, and that he should offer a fixed wage that holds the agent down to her reservation utility. If the agent complies with the contract, the principal can capture roughly 30 percent of the first best surplus for himself while the agent gets nothing.

There are two alternatives to this "incentive contract". In one treatment the principal could choose to offer a "trust contract" which does without a fine and simply pays a generous fixed wage up front to the agent asking her to reciprocate by spending a higher level of effort. However, effort cannot be enforced with this contract. In a second treatment the principal could offer a "bonus contract", which specifies a fixed wage, a desired level of effort, and an announced bonus payment if the effort is to the principal's satisfaction. However, both parties know that the bonus cannot be enforced and is left at the discretion of the principal. The trust and the bonus contract are clearly less complete than the incentive contract. Because the experiments carefully rule out any repeated interactions between the parties, both types of contracts are, according to standard principal agent theory, doomed to fail. Given the fixed wage, a pure self-interested agent will not spend any effort. Similarly, a principal who is only interested in his own income will never pay a bonus, so a rational agent should never put in any effort.

If concerns for fairness and reciprocity are taken into account, the predictions are less clear cut. Consider again the optimal incentive contract (as suggested by principal agent theory). This contract aims at a rather unfair distribution of the surplus. If the agent is concerned about this, there are two ways how she could punish the principal. First, as in an ultimatum game, she could simply reject the contract in which case both parties get a payoff of zero. A second, and more interesting, punishment strategy is to accept the contract and to shirk. Note that if the

---

Che and Hausch (1998), Segal (1999) and Hart and Moore (1999) show that, although an incomplete contract does not implement the first best, a more complete contract is of no value to the parties because it is impossible to get closer to the efficiency frontier.

incentive compatibility constraint is just binding, then the cost of shirking to the agent is zero and independent of the fixed wage offered by the principal. Thus, if the principal offers a somewhat higher wage, that gives a positive (but still "unfair") share of the surplus to the agent, the agent can punish the principal by accepting the wage and shirking (at zero cost to herself). Hence, concerns for fairness and reciprocity suggest that the principal has to offer a fairly generous wage in order to get the agent to accept and to work, which makes the incentive contract less attractive.

On the other hand, concerns for fairness and reciprocity improve the performance of trust and bonus contracts. A fair agent will reciprocate to a generous wage offer in a trust contract by putting in a higher effort level voluntarily. Similarly, a fair principal will reciprocate to a high effort level by paying a generous bonus, making it worth the agent's while to spend more effort. Unfortunately, however, on such a general level it is impossible to make any clear cut predictions about the relative performance of the three types of contracts. Is the incentive contract going to be outperformed by the trust and/or the bonus contract? Induces the bonus contract a higher level of effort than the trust contract or rather the other way round?

In order to obtain quantitative predictions for the experiments, Fehr, Klein and Schmidt (2000) apply the model of inequity aversion by Fehr and Schmidt (1999) to this moral hazard problem. Most other models of fairness or intention-based reciprocity would probably yield similar results and we want to stress that these experiments were not designed to discriminate between different notions of fairness. The main advantage of our model of inequity aversion is just its simplicity, which makes it straightforward to apply to these games. However, Fehr, Klein and Schmidt (2000) have to make a few additional assumptions. In particular, they assume for simplicity that there are only two types of subjects, "selfish" players who are only interested in their own material payoffs, and "fair" players who are willing to give up own resources in order to achieve a more equal payoff distribution. Furthermore, in rough accordance with the experimental results of many ultimatum and dictator games, they assume that 60 percent of the population are selfish and 40 percent are fair.

With these assumptions it is a straightforward exercise to analyse the different types of contracts and to obtain the following predictions:

1. *Trust Contracts:* Fair agents will reciprocate to high wage offers by putting in an effort level that equalizes payoffs, while selfish agents will choose the minimum effort level of 1. Thus, a higher wage offer will, on average, induce a higher level of effort. However, it can be shown that if less than 2/3 of all agents are fair, paying a higher wage does not raise the principal's expected profit. Therefore, with 40 percent fair agents, the trust contract is not going to work.

2. *Incentive Contracts:* For the same reason as in the trust contract it does not pay for the principals to elicit higher average effort levels by paying generous wages. Thus, both selfish and fair principals impose the highest possible fine to induce the agent to choose e = 4. However, while the fair principals share the surplus arising from e = 4 equally with the agent, selfish principals propose unfair contracts that give them the whole surplus. They anticipate that the fair agents reject these contracts, but because the 60 percent selfish agents accept these contracts, this strategy is still profitable.

3. *Bonus Contracts:* Selfish principals always pay a bonus of zero but fair principals pay a bonus that divides the surplus equally between the principal and the agent. Therefore, the bonus is on average increasing with the agent's effort. Moreover, the relation between the effort and the average bonus is sufficiently steep to induce a selfish agent to put it an effort level of 7. However, the fair agent chooses an effort level of only 1 or 2 (depending on the fixed wage). The reason for this surprising result is that the fair agent is not only concerned about her expected monetary payoff, but that she suffers in addition from the inequality that arises if a selfish principal does not pay the bonus. Nevertheless, on average, the bonus contract implements a higher level of effort (*e=5.2*) and yields a higher payoff for the principal than both, the incentive contract and the trust contract.[36]

What are the experimental results? Each experiment had 10 periods, in each of which each principal was matched randomly and anonymously with a different agent. In the first treatment, where principals could choose between a trust and an incentive contract, roughly 50 percent of the principals chose a trust contract and 50 percent chose an incentive contract in period 1. However, the fraction of incentive contracts rose quickly and after period 5 roughly 80 percent of all contractual choices were incentive contracts. Those principals who offered a trust contract paid generous wages to which some agents reciprocated by putting in a high effort level. However, in 64 percent of all trust contracts the agents chose *e=1*. Thus, on average, principals incurred considerable losses when they proposed trust contracts. The incentive contracts did better, but they did much less well than predicted by standard principal agent theory. They also did less well than predicted by the model of inequity aversion. The reason is that at the beginning many principals offered incentive contracts with fairly high wages that were not incentive

[36] The analysis of the bonus contract is complicated by the fact that the principal has to move twice. He offers the terms of the contract at the first stage of the game and he has to choose his bonus payment at the last stage. Thus, his contract offer may reveal some information about his type. However, it can be shown that there is no separating equilibrium in this game and that all pooling equilibria have the properties described above. Furthermore, if we assume that a higher wage offer is not interpreted by the agent as a signal that she faces the selfish principal with a higher probability, then there is a unique pooling equilibrium. See Fehr, Klein and Schmidt (2000).

compatible. In these cases 62 percent of the agents shirked imposing considerable losses on principals. On the other hand, those principals who offered incentive compatible incentive contracts with low wages did fairly well. Principals learnt to properly design incentive contracts over time. The fraction of incentive compatible contracts increased from only 10 percent in period 1 to 64 percent in period 10.

In the second treatment the principal had to choose between a bonus contract and an incentive contract. From the very beginning the bonus contract was much more popular than the incentive contract and accounted for roughly 90 percent of all contractual choices. Many principals did not pay a bonus, but a significant fraction reciprocated generously to higher effort levels. The average bonus was, therefore, strongly increasing in the effort level which made it worthwhile for the agents to put forward rather high effort levels. The average effort level was 5.2, which is significantly higher than the average effort of 2.5 induced by incentive contracts. The bonus contract is not only more efficient than the incentive contract, it also yields on average a much higher payoff to the principal and a moderately higher payoff to the agent. These results are clearly inconsistent with the self-interest model while the model of inequity aversion explains them surprisingly well.[37]

Our experiments demonstrate that quite powerful incentives can be given by a very incomplete bonus contract. The bonus contract relies on reciprocal fairness as an enforcement device. It does better than the more complete incentive contracts *because* it is incomplete and thus leaves more freedom to the parties to reciprocate. This enforcement mechanism is not perfect and, depending on the payoff structure and the fraction of reciprocal types in the population, it can fail. In fact, we have seen that the trust contract, in which the principal has to pay the "bonus" unconditionally in advance, is not viable in the set up of our experiments. Yet, the performance of the bonus contract suggests that the effect of reciprocal fairness, that has been neglected in contract theory so far, is important for optimal contractual design and should be taken into account.

[37] In a second experimental design, Fehr, Klein and Schmidt (2000) consider a multi-task principal agent model inspired by Holmström and Milgrom (1991). In this experiment the agents have to choose two separate effort levels ("tasks"), e_1 and e_2, both of which are observable by the principal but only e_1 is verifiable and can be contracted upon. The principal can choose between a piece-rate contract that rewards the agent for his effort spent on task 1 and a bonus contract that announces a voluntary bonus payment if the agent's effort on both tasks is to the principal's satisfaction. The overwhelming majority of principals opted for the bonus contract which induced the agents to spend, on average, a considerable amount of effort and to allocate total effort efficiently across tasks. Those principals that chose a piece-rate contract, induced the agents to concentrate all of their total efforts on task 1, which is very inefficient. Again, these results are inconsistent with the self-interest model, but they can be nicely explained by the Fehr-Schmidt model of inequity aversion.

5.3 The Optimal Allocation of Ownership Rights

Consider two parties, A and B, who are engaged in a joined project (a "firm") to which they have to make some relationship specific investments today in order to generate a joint surplus in the future. An important question that has received considerable attention in recent years is who should own the firm. In a seminal paper, Grossman and Hart (1986) argue that ownership rights allocate residual rights of control on the physical assets that are required to generate the surplus. For example, if A owns the firm, then he will have a stronger bargaining position than B in the renegotiation game in which the surplus between the two parties is shared ex post, because he can exclude B from using the assets which makes B's relationship specific investment less productive. Grossman and Hart show that there is no ownership structure that implements first best investments, but some ownership structures do better than others and there is a unique second best optimal allocation of ownership rights.

A common feature of most incomplete contract models is that joint ownership cannot be optimal.[38] This result is at odds with the fact that there are many jointly owned companies, partnerships or joint ventures. Furthermore, the argument neglects that reciprocal fairness may be an important enforcement mechanism to induce the involved parties to invest more under joint ownership than otherwise predicted. In order to test this hypothesis, Fehr, Kremhelmer and Schmidt (2000) conducted a series of experiments on the optimal allocation of ownership rights. The experimental game is a grossly simplified version of Grossman and Hart (1986): There are two parties, A and B, who have to make investments, $a, b \in \{1, ..., 10\}$, respectively, in order to generate a joint surplus $v(a,b)$. Investments are sequential: B has to invest first, his investment level b is observed by A, who has to invest thereafter. We consider two possible ownership structures: Under A-ownership, A hires B as an employee and pays her a fixed wage w. In this case monetary payoffs are $v(a,b)-w-a$ for A and $w-b$ for B. Under joint ownership, each party gets half of the gross surplus minus his or her investment cost, i.e. $0.5v(a,b)-a$ for A and $0.5v(a,b)-b$ for B. The gross profit function has been chosen such that maximal investments are efficient, i.e.

[38] To see this note that in the renegotiation game in which the surplus is shared each party gets its reservation utility plus a fixed fraction (50 percent, say) of the joint surplus in excess of the sum of the reservation utilities. Now consider A-ownership. If A invests, then his investment increases not only the joint surplus but also his reservation utility (i.e., what he could get out of the firm without B's collaboration). On the other hand, if B invests, then her investment increases only the joint surplus, but it does not improve her reservation utility. The reason is that the investment requires access to the firm in order to be productive. Hence, without the firm B's investment is useless. This is why A will invest more than B under A-ownership. Consider now joint ownership. If both parties own the firm jointly, then each of them can prevent the other from using the assets. Hence neither A's nor B's investment affects their respective reservation utilities. Therefore, A's investment incentives are reduced while B's investment incentives do not improve. Hence, joint ownership is Inferior.

$a^{FB}=b^{FB}=10$, but if each party gets only 50 percent of the marginal return of their investments, then it is a dominant strategy for a purely self-interested player to choose the minimum level of investment, $\underline{a} = \underline{b} = 1$. Finally, in the first stage of the game, A can decide whether to be the sole owner of the firm and make a wage offer to B, or whether to have joint ownership.

The prediction of the self-interest model is straightforward. Under A-ownership B has no incentive to invest and will choose $b=1$. On the other hand, A is full residual claimant on the margin, so she will invest efficiently. Under joint ownership each party gets only 50 percent of the marginal return which is not sufficient to induce any investments. Hence in this case B's optimal investment level is unchanged, but A's investment level is reduced to $\underline{a}=1$. Thus, A-ownership outperforms joint ownership and A should hire B as an employee.

In the experiments just the opposite happened. Party A chose joint ownership in more than 80 percent (187 out of 230) of all observations and gave away 50 percent of the gross return to B. Moreover, the fraction of joint ownership contracts increased from 74 percent in the first two periods to 89 percent in the last two periods. With joint ownership B-players chose on average an investment level of 8.9 and A responded with an investment of 6.5 (on average). On the other hand, if A-ownership was chosen and A hired B as an employee, B's average investment was only 1.3, while all A-players chose an investment level of 10. Furthermore A-players earned much more on average if they chose joint ownership rather than A-ownership.

These results are inconsistent with the self-interest model, but it is straightforward to explain them with concerns for fairness. Applying the Fehr-Schmidt (1999) model of inequity aversion gives again fairly accurate quantitative predictions. Thus, the experimental results and the theoretical analysis suggest that joint ownership may do better than A-ownership because it offers more scope for reciprocal behavior. Subjects seem to understand this and predominantly choose this ownership structure.

6 Conclusions

The self-interest model has been very successful in explaining individual behavior on competitive markets, but it is unambiguously refuted in many situations in which individuals interact strategically. The experimental evidence on, e.g., ultimatum games, dictator games, gift exchange games, and public good games, demonstrates unambiguously that many people are not only maximizing their own material payoffs, but that they are also concerned about social comparisons, fairness, and the desire to reciprocate.

We have reviewed several models that try to take these concerns explicitly into account. A general lesson to be drawn from these models is that the assumption that some people are fair-minded and have the desire to reciprocate does not imply that these people will always behave "fairly". In some environments like, e.g. in competitive markets or in public good games without punishment, fair-minded actors will often behave as if they are purely self-interested. Likewise, a purely self-interested person may often behave as if he is strongly concerned about fairness like, e.g., the Proposers who make fair proposals in the ultimatum game or generous wage offers in the gift exchange game. Thus, the behavior of fair-minded and purely self-interested actors depends on the strategic environment in which they interact and on their beliefs about the fairness of their opponents. The analysis of this behavior is not trivial and it is helpful to develop theoretical tools to better understand what we observe.

Some of the models reviewed above focus solely on preferences over income distributions and ignore the fact that people often care about the intentions behind the actions of their opponents. Some other papers focus only on intention-based or type-based reciprocity and ignore the fact that some people are bothered by unfair distributions even if their opponent could not do anything about it. It seems natural to try to combine these two motivations in a single model as has been done by Falk and Fischbacher (1998) and Charness and Rabin (2000). However, we believe that the cost of doing so is high. These models are rather complicated, they rely on psychological game theory and it is difficult to apply them even to very simple experimental games. Moreover, Charness and Rabin, in particular, is plagued with multiple equilibria and has much more free parameters than all other models. On the other hand, simple models of social preferences, like Bolton and Ockenfels' (2000) ERC-model or our own (1999) model of inequity aversion, fit the data on large classes of games fairly well. They use standard game theory, they have fewer parameters to be estimated, and it is fairly straightforward to get clear-cut qualitative and quantitative predictions.

The main advantage of these simple models is that they can easily be applied to other fields in economics. For more than 20 years experimental economists concentrated on simple experimental games in order to better understand what drives economic behavior. However, very few of the insights that have been gained had any impact on how economists interpret the world. We feel that it is now time to change this. Many phenomena in situations in which people interact strategically cannot be understood by relying on the self-interest model alone. Our examples from contract theory and the theory of property rights illustrate that models of reciprocal fairness can be fruitfully applied to important and interesting economic questions, yielding predictions that are much closer to what we observe in many situations of the real world and in carefully controlled

experiments than the predictions of the self-interest model. There are many other areas in which fairness models are likely to generate interesting new insights - be it the functioning of labor markets or questions of political economy, be it the design of optimal mechanisms or questions of compliance with organizational rules and the law.

We hope that this is just the beginning. There is no shortage of important questions to which the newly developed tools and insights can be applied.

References

Abbink, K., Bernd Irlenbusch, and Elke Renner, (2000). "The Moonlighting Game. An Experimental Study on Reciprocity and Retribution." *Journal of Economic Behavior and Organization,* forthcoming.

Agell, Jonas and Per Lundborg, 1995. "Theories of Pay and Unemployment: Survey Evidence from Swedish Manufacturing Firms", *Scandinavian Journal of Economics* 97, 295-308.

Ahlert, Marlies, Arwed Crüger and Werner Güth, 1999. "An Experimental Analysis of Equal Punishment Games", mimeo, University of Halle-Wittenberg.

Alm, James, Isabel Sanchez and Ana de Juan, 1995. "Economic and Noneconomic Factors in Tax Compliance", *Kyklos* 48, 3-18.

Andreoni, James 1989. "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence." *Journal of Political Economy* 97, 1447-1458.

Andreoni, James, Brian Erard and Jonathan Feinstein, 1998. "Tax Compliance", *Journal of Economic Literature* 36, 818-860.

Andreoni, James and Miller, John, 1993. "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence", *Economic Journal* 103, 570-585.

Andreoni, James and Miller, John, 2000. "Giving According to GARP: An Experimental Test of the Rationality of Altruism." *Mimeo,* University of Wisconsin and Carnegie Mellon University.

Andreoni, James and Lise Vesterlund, forthcoming. "Which is the fair Sex? Gender Differences in Altruism", *Quarterly Journal of Economics*.

Andreoni, James and Hal Varian, 1999. "Preplay Contracting in the Prisonners' Dilemma", *Proceedings of the National Academy of Sciences* 96, 10933-10938.

Aghion, Philippe, Dewatripont, Matthias and Rey, Philippe, 1994. "Renegotiation Design with Unverifiable Information." *Econometrica* 62, 257-282.

Arrow, Kenneth J., 1981. "Optimal and Voluntary Income Redistribution." In: Rosenfield, Steven (ed), *Economic Welfare and the Economics of Soviet Socialism: Essays in Honor of Abram Bergson,* Cambridge: Cambridge University Press.

Becker, Gary S., 1974. "A Theory of Social Interactions." *Journal of Political Economy* 82, 1063-1093.

Berg, Joyce, John Dickhaut, and Kevin McCabe, 1995. "Trust, Reciprocity and Social History," *Games and Economic Behavior* X, 122-142.

Bernheim, B. Douglas, 1986. “On the Voluntary and Involuntary Provision of Public Goods.” *American Economic Review* 76, 789-793.

Bewley, Truman, 1999. *Why Wages don’t fall during a Recession*, Harvard University Press, Harvard.

Binmore, Kenneth, John Gale and Larry Samuelson, 1995. “Learning to be Imperfect: The Ultimatum Game”, *Games and Economic Behavior* 8, 56-90.

Binmore, Ken, 1998. *Game Theory and the Social Contract: Just Playing*, MIT Press, Cambridge, Massachusetts.

Blount, Sally, 1995. "When Social Outcomes aren’t Fair: The Effect of Causal Attributions on Preferences," *Organizational Behavior and Human Decision Processes* LXIII, 131-144.

Bolle, Friedel and Alexander Kritikos, 1998. “Self-Centered Inequality Aversion versus Reciprocity and Altruism”, mimeo, Europa-Universität Viadrina.

Bolton, Gary E., 1991. “A Comparative Model of Bargaining: Theory and Evidence.” *American Economic Review* 81, 1096-1136.

Bolton, Gary and Rami Zwick, 1995. “Anonymity versus Punishment in Ultimatum Bargaining”, *Games and Economic Behavior* 10, 95-121.

Bolton, Gary E., Jordi Brandts, and Axel Ockenfels, 1998. “Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game”, *Experimental Economics* 3, 207-221.

Bolton, Gary E. and Ockenfels, Axel, 2000. A theory of equity, reciprocity and competition. *American Economic Review* 100, 166-193.

Bowles, Samuel and Herbert Gintis, 1999. “The Evolution of Strong Reciprocity”, mimeo, University of Massachusetts at Amherst.

Bowles, Samuel and Herbert Gintis, 2000. “Reciprocity, Self-Interest, and the Welfare State”, *Nordic Journal of Political Economy* 26, 33-53.

Brandts, Jordi and Gary Charness, 1999. “Gift-Exchange with Excess Supply and Excess Demand”, mimeo, Pompeu Fabra, Barcelona.

Camerer, Colin F., 1999. “Social Preferences in Dictator, Ultimatum and Trust Games.” *Mimeo*. California Institute of Technology.

Camerer, Colin F. and Thaler, Richard H., 1995. Ultimatums, Dictators and Manners. *Journal of Economic Perspectives* 9, 209-19.

Cameron, Lisa A., 1999. “ Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia.” *Economic-Inquiry* 37(1), 47-59.

Carpenter, Jeffrey P., 2000. “Punishing Free-Riders: The Role of Monitoring-Group Size, Second-Order Free-Riding and Coordination”, mimeo, Middlebury College.

Chamberlin, Edward H., 1948. “An Experimental Imperfect Market”, *Journal of Political Economy* 56, 95-108.

Charness, Gary, 1996. "Attribution and Reciprocity in a Labor Market: An Experimental Investigation," mimeo, University of California at Berkeley.

Charness, Gary, 2000. “Responsibility and Effort in an Experimental Labor Market”, *Journal of Economic Behavior and Organization* 42, 375-384.

Charness, Gary, and Rabin, Matthew, 2000. “Social Preferences: Some Simple Tests and a New Model.” *Mimeo*, University of California at Berkeley.

Che, Yeon-Koo and Hausch, Donald B., 1999. “Cooperative Investments and the Value of Contracting.” *American Economic Review* 89(1), 125-47.

Cooper, David J., and Carol Kraker Stockman, 1999. “Fairness, Learning, and Constructive Preferences: An Experimental Investigation”, mimeo, Case Western Reserve University.

Costa-Gomes, Miguel, and Klaus G. Zauner, 1999. “Learning, Non-equilibrium Beliefs, and Non-Pecuniary Payoff Uncertainty in an Experimental Game”, mimeo, Harvard Business School.

Cox, James C., 2000. “Trust and Reciprocity: Implications of Game Triads and Social Contexts”, mimeo, University of Arizona at Tucson.

Croson, Rachel T. A., " Theories of Altruism and Reciprocity: Evidence from Linear Public Goods Games," Discussion Paper, Wharton School, University of Pennsylvania, 1999.

Daughety, Andrew, 1994. “Socially-Influenced Choice: Equity Considerations in Models of Consumer Choice and in Games”, mimeo, University of Iowa.

Davis, Douglas, and Charles Holt, 1993. *Experimental Economics*, Princeton: Princeton University Press.

Dawes, Robyn M., and Richard Thaler, 1988. "Cooperation," *Journal of Economic Perspectives* II, 187-197.

Dufwenberg, Martin and Kirchsteiger, Georg, 1998. “A Theory of Sequential Reciprocity.” Discussion Paper. CentER, Tilburg University.

Edlin, Aaron S. and Reichelstein, Stefan, 1996. “Holdups, Standard Breach Remedies, and Optimal Investment.” *American Economic Review* 86(3), 478-501.

Eichenberger, Rainer and Felix Oberholzer-Gee, 1998. “Focus Effects in Dictator Game Experiments“, mimeo, University of Pennsylvania.

Ellingsen, Tore and Magnus Johannesson, 2000. “Is There a Hold-up Problem?”, Stockholm School of Economics, Working Paper No. 357.

Encyclopaedia Britannica, 1998. *The New Encyclopaedia Britannica*, Volume 1, London, 15th edition.

Fahr, Renè and Bernd Irlenbusch, 2000. “Fairness as a Constraint on Trust in Reciprocity: Earned Property Rights in a Reciprocal Exchange Experiment”, Economics Letters 66, 275-282.

Falk, Armin, Fehr, Ernst, and Fischbacher, Urs, 2000a. “Informal Sanctions”, Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 59.

Falk, Armin, Fehr, Ernst, and Fischbacher, Urs, 2000b. “Testing Theories of Fairness - Intentions Matter”, Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 63.

Falk, Armin, Fehr, Ernst, and Fischbacher, Urs, 2000c. “Appropriating the Commons”, Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 55.

Falk, Armin and Fischbacher, Urs, 1999. “A Theory of Reciprocity.” Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 6.

Falk, Armin, Simon Gächter, and Judith Kovács, 1999. “Intrinsic Motivation and Extrinsic Incentives in a Repeated Game with Incomplete Contracts”, *Journal of Economic Psychology.*

Fehr, Ernst and Armin Falk, 1999. “Wage Rigidity in a Competitive Incomplete Contract Market”, *Journal of Political Economy* 107, 106-134.

Fehr, Ernst and Urs Fischbacher, 2000. “Third Party Punishment”, mimeo, University of Zürich.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl, 1993. „Does Fairness prevent Market Clearing? An Experimental Investigation,“ *Quarterly Journal of Economics* CVIII, 437-460.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl, 1998. „Gift Exchange and Reciprocity in Competitive Experimental Markets“, *European Economic Review* 42, 1-34.

Fehr, Ernst and Klaus M. Schmidt, 1999. “A Theory of Fairness, Competition and Co-operation.” *Quarterly Journal of Economics* 114, 817-868.

Fehr, Ernst, and Simon Gächter, 2000. "Cooperation and Punishment in Public Goods Experiments“, *American Economic Review* 90, 980-994.

Fehr, Ernst, Simon Gächter and Georg Kirchsteiger, 1997. “Reciprocity as a Contract Enforcement Device”, *Econometrica* 65, 833-860.

Fehr, Ernst, Klein, Alexander and Schmidt, Klaus M., 2000. “Endogenous Incomplete Contracts.” *Mimeo*, University of Munich, 2000.

Fehr, Ernst, Kremhelmer, Susanne and Schmidt, Klaus M., 2000. “Fairness and the Optimal Allocation of Property Rights.” *Mimeo*, University of Munich, 2000.

Fehr, Ernst and Tougareva, Elena, 1995: “Do High Monetary Stakes Remove Reciprocal Fairness? Experimental Evidence from Russia.” *Mimeo*. Institute for Empirical Economic Research, University of Zurich.

Fischbacher, Urs, Simon Gächter and Ernst Fehr, 1999. “Are People Conditionally Cooperative? Evidence from a Public Goods Experiment”, Working Paper No. 16, Institute for Empirical Research in Economics, University of Zurich,.

Forsythe, Robert L., Joel Horowitz, N. E. Savin, and Martin Sefton, 1994. "Fairness in Simple Bargaining Games," *Games and Economic Behavior* 6, 347-369.

Frey, Bruno and Hannelore Weck-Hannemann, 1984. “The Hidden Economy as an ‘Unobserved’ Variable”, *European Economic Review* 26, 33-53.

Gächter, Simon and Armin Falk (1999): “Reputation or Reciprocity?“, Working Paper No. 19, Institute for Empirical Research in Economics, University of Zürich.

Geanakoplos, John, Pearce, David, and Stacchetti, Ennio, 1989. “Psychological Games and Sequential Rationality.” *Games and Economic Behavior* 1, 60-79.

Gintis, Herbert, 2000. “Strong Reciprocity and Human Sociality”, Journal of Theoretical Biology 206, 169-179.

Greenberg, Jerald, 1990. “Employee Theft as a Reaction to Underpayment Inequity: The Hidden cost of Pay Cuts”, *Journal of Applied Psychology* 75, 56 –568.

Grossman, Sanford and Hart, Oliver, 1983. “An Analysis of the Principal-Agent Problem, *Econometrica* 51, 7-45.

Güth, Werner, Hartmut Kliemt and Axel Ockenfels, 2000. “Fairness versus Efficiency – An Experimental Study of Mutual Gift-Giving”, mimeo, Humboldt University of Berlin.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze, 1982. "An Experimental Analysis of Ultimatium Bargaining," *Journal of Economic Behavior and Organization* III, 367-88.

Güth, Werner and Eric van Damme, 1998. “Information, Strategic Behavior and Fairness in Ultimatum Bargaining: an Experimental Study”, *Journal of Mathematical Psychology* 42, 227-247.

Hannan, Lynn, John Kagel, and Donald Moser, 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, John, 1955. “Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility”, *Journal of Political Economy* 63, 309-321.

Hart, Oliver and Moore, John, 1990. “Property Rights and the Nature of the Firm”, *Journal of Political Economy* 98, 1119-58.

Hart, Oliver and Moore, John, 1999. “Foundations of Incomplete Contracts.” *Review of Economic Studies* 66, 115-138.

Hoffman, Elisabeth, Kevin McCabe, Keith Shachat, and Vernon Smith, 1994. „Preferences, Property Right, and Anonymity in Bargaining Games”, *Games and Economic Behavior* 7, 346-380.

Hoffman, Elisabeth, Kevin McCabe, and Vernon Smith, 1996. "On Expectations and Monetary Stakes in Ultimatum Games," *International Journal of Game Theory* 25, 289-301.

Holmström, Bengt and Milgrom, Paul, 1991. “Multi-task Principal-Agent Analyses.” *Journal of Law, Economics, and Organization* 7 (Sp.), 24-52.

Isaac, Mark R., James M. Walker, Arlington W. Williams, 1994. “ Group Size and the voluntary Provision of Public Goods”, *Journal of Public Economics* 54, 1-36.

Kagel, John H, Chung Kim and Donald Moser, 1996. “Fairness in Ultimatum Games with Asymmetric Information and Asymmetric Payoffs”, *Games and Economic Behavior* 13, 100-110.

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler, 1986. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review* LXXVI, 728-41.

Kirchsteiger, Georg, 1994. “The Role of Envy in Ultimatum Games”, *Journal of Economic Behavior and Organization* 25, 373-389.

Laffont, Jean-Jacques and Tirole, Jean, 1993. *A Theory of Regulation and Procurement.* Cambridge (Mass.): MIT-Press.

Ledyard, John, 1995. "Public Goods: A Survey of Experimental Research", Chap. 2 in: Alvin Roth and John Kagel (eds.), *Handbook of Experimental Economics.* Princeton: Princeton University Press.

Levine, David, 1998. “Modeling Altruism and Spitefulness in Experiments”, Review of Economic Dynamics 1, 593-622.

Lind, Allan and Tom Tyler, 1988. *The Social Psychology of Procedural Justice*. New York and London: Plenum Press.

List, John and Todd Cherry, 2000. “Examining the Role of Fairness in Bargaining Games”, mimeo, University of Arizona at Tucson.

McCabe, Kevin, Mary Rigdon and Vernon Smith, 2000. “Positive Reciprocity and Intentions in Trust Games”, mimeo, University of Arizona at Tucson.

Miller, Sven (1997): “Strategienuntersuchung zum Investitionsspiel von Berg, Dickhaut, McCabe”, *Diploma thesis*, University of Bonn.

Neilson, William, 2000. “An Axiomatic Characterization of the Fehr-Schmidt Model of Inequity Aversion”, mimeo, Department of Economics, Texas A&M University.

Nöldeke, G., Schmidt, K.M., 1995. Option Contracts and Renegotiation: A Solution to the Hold-Up Problem. *Rand Journal of Economics* 26, 163-179.

Offerman, Theo, 1999. “Hurting hurts more than helping helps: The Role of the self-serving Bias”, mimeo, University of Amsterdam.

Ostrom, Elinor, 1990. “*Governing the Commons – The Evolution of Institutions for Collective Action*”, New York: Cambridge University Press

Ostrom, Elinor, 2000. “Collective Action and the Evolution of Social Norms”, *Journal of Economic Perspectives* 14, 137-158.

Rabin, Matthew, 1993. “Incorporating Fairness into Game Theory and Economics.” *American Economic Review*, 83(5), 1281-1302.

Roth, Alvin E., Michael W. K. Malouf, and J. Keith Murningham, 1981. „Sociological versus strategic Factors in Bargaining“, Journal of Economic Behavior and Organization 2, 153-177.

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

Roth, Alvin E., 1995. "Bargaining Experiments," in: J. Kagel and A. Roth (eds.): *Handbook of Experimental Economics,* Princeton, Princeton University Press.

Roth, Alvin E., and Ido Erev, 1995. "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term," *Games and Economic Behavior* 8, 164-212.

Samuelson, Paul A., 1993. “Altruism as a Problem Involving Group versus Individual Selection in Economics and Biology.” *American Economic Review* 83, 143-148.

Segal, Uzi and Sobel, Joel, 1999. “Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings.” *Mimeo*, University of California at San Diego.

Segal, Ilya, 1999. “Complexity and Renegotiation: A Foundation for Incomplete Contracts.” *Review of Economic Studies* 66(1), 57-82.

Seidl, Christian and Stefan Traub, 1999. “Taxpayers’ Attitudes, Behavior, and Perceptions of Fairness in Taxation”, mimeo, Institut für Finanzwissenschaft und Sozialpolitik, University of Kiel.

Sen, Amartya, 1995. “Moral Codes and Economic Success”, C. S. Britten and A. Hamlin (eds.), Market Capitalism and Moral Values, Edward Eldar, Aldershot.

Selten, Reinhard and Axel Ockenfels, 1998. “An Experimental Solidarity Game”, Journal of Economic Behavior and Organization, 34, 517-539.

Sethi, Rajiv and E. Somananthan, forthcoming. Preference Evolution and Reciprocity, *Journal of Economic Theory*.

Sethi, Rajiv and E. Somananthan, 2000. Understanding Reciprocity, mimeo, Columbia University.

Slonim, Robert, and Alvin E. Roth, 1997. "Financial Incentives and Learning in Ultimatum and Market Games: An Experiment in the Slovak Republic," *Econometrica* 65, 569-596.

Smith, Adam, 1759, reprinted 1982. *The Theory of Moral Sentiments*. Indianapolis: Liberty Fund.

Smith, Vernon L., 1962. "An Experimental Study of Competitive Market Behavior," *Journal of Political Economy* 70, 111-137.

Sonnemans, Joep, Arthur Schram and Theo Offerman, 1999. „Strategic Behavior in Public Good Games – When Partners drift apart“, *Economics Letters* 62, 35-41.

Suleiman, Ramzi, 1996. “Expectations and Fairness in a modified Ultimatum Game”, *Journal of Economic Psychology* 17, 531-554.

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

Zajac, Edward, 1995. “*Political Economy of Fairness*”, Cambridge, Massachusetts: MIT Press.

Zizzo, Daniel and Andrew Oswald, 2000. “Are People Willing to Pay to Reduce Others’ Income”, mimeo, Oxford University.