Analyse & Kritik 01/2011 (© Lucius & Lucius, Stuttgart) S. 349–364

Julian Culp/Heiner Schumacher

Reciprocity in Economic Games*

Abstract: The evidence of laboratory experiments of behavioral economists shows that individuals behave reciprocally. These data put into question the pure self-interest thesis of human motivation of the homo oeconomicus model and call for alternative models. Focusing on the explanation of reciprocal behavior in Trust Games, this article proposes two directions that economists and other social scientists might want to consider in order to establish a more solid foundation for economic theory. First, it presents models that economic theorists developed to explain the laboratory evidence of reciprocal behavior. It highlights that all of these models subscribe to the Humean view that desires are at the source of any human motivation and suggests an alternative Kantian model where reasons have the capacity to motivate human action. Second, it emphasizes that a supplementary examination of the social background conditions would illuminate the analysis of the findings because of the connection between the 'local' and society-wide demands of reciprocity.

1. Introduction

For a long time the belief that *all* individuals are, or can assumed to be, motivated by pure self-interest has enjoyed the status of an axiom, and even a dogma, in economics.[1] Since the 1980s, however, the results of experimental economists began increasingly to challenge this fundamental motivational assumption of standard economic theory. In numerous laboratory experiments throughout the world, test subjects participated in various types of games that allowed them to gain real pecuniary payoffs. They turned out to behave in ways that cannot be properly explained by the pure self-interest thesis of human behavior. In games like the Prisoner's Dilemma, the Trust Game, the Ultimatum Game and the Dictator Game, *some* players consistently chose actions and strategies that are described more properly as altruistic, fair, or reciprocal behavior rather than purely self-interested behavior: despite the fact that there was a dominant strategy for purely self-interested players that was easily identifiable, players repeatedly did not choose it. This result has prevailed under varying

* The authors would like to thank Nicole Hassoun, Michael Münch and Timothy Waligore for comments on a penultimate version of this article.

[1] Purely self-interested behavior may under certain circumstances be also described as 'selfish', and thus as morally doubtful, but not every purely self-interested behavior, such as taking one's favorite path through the woods, is selfish (cf. Leist 2005, 160).

setups of the experiment. Thus, other motives must be considered in order to explain the altruistic, fair or reciprocal behavior of the players in the laboratory experiments (cf. Fehr/Schmidt 2006 for an overview of the empirical evidence and the novel theoretical explanations). Other types of empirical research, of course, also put into question the pure self-interest assumption. Anthropological research, for instance, has done so by emphasizing the role of reciprocity in social relationships (cf. Mauss 1990[1923]). This article examines the phenomenon of reciprocity, but focuses exclusively on the laboratory evidence of reciprocal behavior in Trust Games.

The aim of this article is to suggest two paths that economists and other social scientists might want to take so as to create robust models of human behavior for economic theory. In particular, it sheds light on the motivation to behave reciprocally as well as on the relation between the demands of reciprocity in 'local' interactions and the society at large. *Sections 2* through *6* highlight that explanatory models usually assume that human motivation always has a desire at its source and propose to further investigate other models where reasons have the capacity to motivate human action. More specifically, in *section 2* this article will show why exactly the pure self-interest thesis fails to explain the laboratory evidence of the players' actual behavior in Trust Games. *Sections 3* and *4* introduce explanatory models that economists created in order to theoretically account for the evidence of the Trust Games. In *section 5* these models are shown to subscribe to a Humean theory of motivation and a related Neo-Humean Model of Practical Reason. *Section 6* introduces an alternate model, the Kantian Model of Practical Reason, and then elucidates how this latter model explains reciprocal behavior in Trust Games. *Section 6* finally also elaborates how such a model could be integrated into rational choice theory. *Section 7* draws attention to the fact that (reciprocal) behavior in 'local' interactions, i.e. interactions among two or a few people that are part of the same society, may be affected by the particular shape of the social background conditions of such local interactions. In particular, it points to the significance of the society-wide demands of reciprocity for the analysis of the evidence of reciprocal behavior. Therefore, it also suggests to further investigate the social background conditions of the laboratory experiments. *Section 8* comes to a conclusion.

2. Experimental Evidence for Reciprocal Behavior

The most frequently used experimental game to analyze reciprocal behavior is the Trust Game (or Investment Game). Two agents, a first-mover and a second-mover, play the Trust Game. At the beginning of the game, the first-mover decides how much of an initial endowment she wants to send to the second-mover. The amount sent is tripled before reaching the second-mover. Then the second-mover decides how much of the received amount she wants to send back to the first-mover. The final payoff of the first-mover is the amount she did not send to the second mover *plus* the amount the second-mover sends back to her; the final payoff of the second-mover is the amount received *minus* the

amount she sent back to the first-mover. Economists usually refer to the amount sent by the first-mover as a measure of 'trust', and the amount returned by the second-mover as a measure of 'trustworthiness' or 'reciprocity'.[2]

If we assume that both agents are purely self-interested (i.e. that agents only care about their personal payoffs), then the Trust Game involves a social-dilemma. The total sum of the players' final payoffs is maximal if and only if the first-mover sends her entire endowment to the second-mover (since each monetary unit (MU) sent from the first-mover adds two MUs to the total payoff, as described above). However, if the second-mover is purely self-interested, she has no incentive to send anything back. Given that she keeps everything for herself, it is rational for the first-mover not to send anything to the second-mover. Standard rational choice theory therefore predicts that agents keep their initial endowment and no amounts are sent between first- and second-mover. This outcome is clearly inefficient for the players.

The first economists who analyzed the Trust Game experimentally with student subjects were Berg, Dickhaut and McCabe (1995). What they found differs substantially from the prediction that is based on pure self-interest. On average, first-movers send 50 percent of their endowment, which indicates a significant degree of 'trust'. The average amount returned by second-movers is 95 percent of what was sent by first-movers.[3] Therefore, many second-movers reciprocate the first-movers' investments. This finding was confirmed in many subsequent studies that analyzed the Trust Game experimentally (see Camerer 2003 for a summary).

Reciprocal behavior has also been observed in other experimental games with structures similar to the Trust Game. In Fehr, Kirchsteiger and Riedl (1993), first-movers, labeled as 'firms', offer a fixed wage to second-movers, labeled as 'workers'. After accepting such an offer, a second-mover then can exert costly effort. The first-mover's payoff increases in effort, while the second-mover's payoff decreases in effort. Any effort exerted by the second-mover leads to a gain in social efficiency. Once again, standard rational choice theory would predict that in the absence of reputation-effects second-movers do not exert any effort, and hence first-movers offer the lowest possible wage. However, it turned out that first-movers offered—on average—'generous' wages and second-movers' effort increases—on average—in the offered wage. This provides additional evidence that many individuals behave reciprocally.

3. Early Explanations for Reciprocal Behavior in Economics

The experimental evidence on reciprocal behavior rejects the hypothesis of pure self-interest. It provoked economic theorists to come up with models that are

[2] In the following the article adopts these meanings of the terms 'trust' and 'reciprocity', or 'reciprocal behavior', in order to refer to these types of behavior of the players in the Trust Game.

[3] The variation is large: half of the subjects return either nothing or very little.

compatible with the experimental data. An obvious alternative to pure self-interested preferences that rank outcomes only according to one's own payoffs, are 'altruistic preferences' that also include the payoffs of others. However, such preferences have been found to be not very robust. In particular, they cannot account for 'negative reciprocity', which means that many individuals are willing to sacrifice their own payoffs in order to punish the purely self-interested behavior of others.[4] For example, Fehr and Gächter (2000) show experimentally that individuals punish free riders in a public-good game, although this punishment is costly for them and does not create future benefits. In response, theorists invented the notion of 'inequity-aversion' (Fehr/Schmidt 1999; Bolton/Ockenfels 2000). Inequity-averse individuals have two objectives: to maximize their own payoffs and to minimize the difference between their own and others' payoffs. The weight of these two objectives is determined by certain fixed parameters. Note that inequity-averse individuals may be willing to sacrifice their own payoffs in order to reduce the payoff differential. Hence, they exhibit both positive and negative reciprocity.

Altruistic and inequity-averse preferences rank the outcomes only according to their payoff distribution. There is, however, substantial evidence that individuals also take into account the intentions of their opponents (Rabin 1993; Falk/Fischbacher 2006): they seem to have a desire to reward 'kind' and to punish 'unkind' behavior. Players assess the kindness of individuals not simply by how much they give, but by their apparent intentions or dispositions. So judgments about the kindness of the other players often depend on the possible options available to them. For example, if the first-mover's endowment in the Trust Game is small, then the second-mover will not interpret the first-mover's behavior as unkind when the first-mover sends a small amount of money. On the contrary, if the first-mover has a large endowment and only sends a small fraction of it, then the second-mover may interpret this behavior as unkind.

In order to analyze strategic behavior in games where players have concerns for intentions, one has to formalize preferences on the domain of material *and* psychological payoffs. Psychological payoffs are derived from the beliefs about other agents' behavior. Consider, for example, a second-mover who beliefs that the first-mover will share half of her endowment. If she receives less, the second-mover may feel disappointed. Her utility then consists of the material payoff (the money the second-mover keeps for herself) and the psychological payoff (the feeling of disappointment).

Psychological payoffs are non-standard in economics and game theory, and there are, probably infinitely, many ways to define them (cf. Geanakoplos/Pearce/Stacchetti 1989; Rabin 2003; Dufwenberg/Kirchsteiger 2004; Falk/Fischbacher 2006; Charness/Dufwenberg 2006). Using psychological payoffs, it is possible to include perceived kindness into formal models and to rationalize the experimental data. However, theories with psychological payoffs are very complex and difficult to apply to games with a richer structure than the Trust Game.

[4] 'Positive reciprocity', by contrast, means the costly rewarding of not purely self-interested behavior by others.

4. Recent Approaches in Economics

A number of classical and contemporary thinkers hypothesized that the desire to be esteemed by others is a basic source of motivation (cf. Brennan/Pettit 2004).[5] Ellingsen and Johannesson (2008) formalize social esteem as the other's belief about one's own 'type'. They construct a model that assumes two types of agents, relatively altruistic and relatively self-interested ones. An agent's type is not transparent to others, but agents can signal their type through actions: by behaving reciprocally they can signal that they are the altruistic type and change the other's belief. An agent's utility increases when the other agent attaches a higher probability to the possibility that she is an altruistic type. Hence, reciprocal behavior may increase the agent's social esteem.

Related to the notion of social esteem is the concept of 'self-image'. According to psychologists and sociologists, many individuals have a desire to maintain conformity between their actions and values. This can have an impact on behavior. For example, Batson (1998, 286) writes: "The ability to pat oneself on the back and feeling good about being a kind, caring person, can be a powerful incentive to help." Bénabou and Tirole (2006) formalize this idea. In their model, agents are uncertain about their true preferences, but they can learn about them from previous actions. Reciprocal behavior then serves as a proof for an agent that she is a good person.

Alternatively, individuals may also have a desire to conform to the behavior of others. Sliwka (2007) models the interaction between a principal and a pool of agents. The principal, the first-mover, can trust or control the agents, the second-movers. Trusting is the principal's optimal choice if most agents will in fact reciprocate this trust. Some agents are 'conformists'. Their utility is maximal if they act in the same manner as the majority of agents. When the principal trusts the agents, this can be a signal for conformists that a majority of the agents behaves reciprocally. Otherwise, it would not pay off for the principal to trust. This in turn leads conformists to reciprocate the principal's trust.

5. The Neo-Humean Model of Practical Reason as Explanation for Reciprocal Behavior

This section will now illustrate how the models employed to explain reciprocal behavior in Trust Games are generally based on a Neo-Humean Model of Practical Reason (NHMPR). First it characterizes the NHMPR and its understanding of the relation between reasons for action and motivation in some more detail. Then it will show the particular ways in which the previously introduced models subscribe to the NHMPR.

---

[5] The desire for social esteem is distinct from the desire for social recognition. The relevance of social recognition in Trust Games is not analyzed in this article, but may, indeed, play a very relevant role (cf. Leist 2005).

5.1 The Neo-Humean Model of Practical Reason[6]

The NHMPR is fundamentally shaped by Hume's theory of motivation (cf. Smith 1987 for an exposition and defense of the Humean theory of motivation). Hume famously claims that "reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them" (Hume 1978[1739], 415). While reason plays an instrumental role in directing the action of an agent, it has no independent motivational force at all. Reason's only task is to determine true statements about means-ends relationships. Hume's account of motivation—that an action is never motivated by reason alone, but always has a desire at its source—is a basic premise of the NHMPR. This premise, together with the *motivational requirement* that any consideration that purports to be a reason for action has to be capable of motivating the agent,[7] leads to the conclusion that any reason for action must be derived from a desire. According to the NHMPR, reasons for actions are at least based upon a desire and a certain means-end belief, i.e. a belief about which actions cause the realization of given ends. Furthermore, this *desire-belief model* of practical reason holds that a rational action consists in performing the action that optimally fulfills the given ends of an agent that are constituted by her desires (cf. Elster 1985 on the optimality condition).[8] Thereby, the NHMPR also fulfills the *normative requirement* that reasons for action must justify an action. The reasons for action not only explain—through the identification of specific motives, i.e. the motivating reasons of the agent—why an agent pursued an action, but they also justify why it was rational to pursue it. So within the NHMPR an action can be justified as rational from the first-person standpoint of the agent by demonstrating that the action was the optimal means for realizing her desires.

The NHMPR defends an internalist view about reasons for action. Following Williams (1981), the debate between internalism and externalism departs from the question of how to interpret the statement 'having a reason to x' (x simply stands for some verb of action). Internalism holds that if it is true that there is a reason for person P to x, then person P has a motive that is served or furthered by x–ing. This implies that if person P has no such motive, it cannot have a reason to x. Externalism, by contrast, claims that the statement 'a person has a reason to x' does not have as truth condition that person P has a motive that is served or furthered by x–ing. Internalism about reasons is driven by the idea that if reasons for actions are to explain why an agent pursues a particular action, then it is necessary that those reasons constitute a motive for the agent to act in that particular way. As the NHMPR derives reasons for actions from the desires of the agents to whom the reasons apply, it is thereby guaranteed that these reasons provide a motive for the agent. By contrast, externalism

---

[6] The next paragraph draws on Gosepath 1999.

[7] In an influential article Korsgaard labeled this motivational requirement the "*internalism requirement*" (Korsgaard 1986, 11).

[8] Note that this way of construing the NHMPR goes beyond mere *instrumental rationality*. It represents *prudential rationality* or "Zweckrationalität" (Weber [1985]1921, 13) as the reasons for action that guide an agent in her behavior reflect a weighting of the alternative ends she has.

about reasons for action does not hold that if person P has reasons to x, then person P also has a motive to x. The upshot is that one cannot say that person P x–ed because of the reasons that person P had to x. For person P may not have had a motive to x since the 'external' reasons that applied to person P did not imply a motive to x. Thus, it might be impossible to *explain* the x–ing of person P, because such an explanation would require the identification of the presence of a motive (cf. Korsgaard 1986, 11). Hence, many dismiss externalism on the ground that it lacks an account of the reasons for action. This issue will be taken up again in *section 6*.

5.2 Early and Recent Economic Models to Explain Reciprocal Behavior and the NHMPR

The following paragraphs elucidate that early and recent economic models of explaining reciprocal behavior subscribe to the NHMPR. All these models hold that reasons for action are based on desires.[9]

Some of the early models in economics explain the reciprocal behavior of some agents in the Trust Game by referring to the 'altruistic preferences' and the 'inequity aversion' of these agents. Here, the reciprocal behavior represents the appropriate means to realize certain desires of the agent, be it the preference for another player's higher payoff or the preference for a more equal payoff distribution. In this way, the 'altruistic preference' and the 'inequity aversion' model subscribe to the NHMPR, as it is the agent's desire for a specific payoff distribution that together with a certain means-ends belief motivate and serve as a justification for her to pursue a reciprocal action. Even in the intention-based model of reciprocity, where some second-movers are said to take into account the intentions of the first-movers, it is also a desire—the desire to reward kind and punish unkind behavior—that ultimately motivates the second-movers' behavior. Thus, this model also endorses the NHMPR as a desire to reward kind and punish unkind behavior and a belief about which action is to realize this desire explain and justify the reciprocal behavior.

The recent economic models of reciprocal behavior also tend to follow the NHMPR. In these models as well the second-mover is led eventually by a desire to act reciprocally. In the social esteem model the agent aspires social esteem and therefore acts reciprocally so as to gain the esteem of the first-mover. So the desire for social esteem, in conjunction with the belief that reciprocal behavior in the Trust Game leads to social esteem, cause and justify the action. The 'self-image' model explains reciprocal behavior by pointing to the desire to act consistently with the self-image that one has of oneself so as to reduce the difference between one's values and actions. This means that reciprocal action results from an agent's desire to conform to her self-image and her belief that this action is an effective means to fulfill this desire. Hence the NHMPR operates here as well. Finally, Sliwka's approach to explaining reciprocal behavior in the

[9] In a different terminology, this is to say that all these models perceive economics' approach to choice as the *optimization of a utility function* and therefore subscribe to a NHMPR. We thank an anonymous reviewer for pointing this out to us.

Trust Game also belongs to the category of the NHMPR: The second-movers, the agents, act reciprocally because of their desire to behave in conformity with the majority of the pool of second-movers and the belief that the first-mover's, the principal's, trust reveals that the majority of the second-movers will reciprocate the trust.

6. The Kantian Model of Practical Reason as Explanation for Reciprocal Behavior

The NHMPR can be contrasted with an alternative Kantian Model of Practical Reason (KMPR), which has so far not been given due consideration in the theoretical discussion of the phenomenon of reciprocal behavior in Trust Games. After presenting the KMPR and clarifying how it explains reciprocal behavior in Trust Games, this section elaborates how experimental economists and other social scientists can incorporate it into their theories.

6.1 The Kantian Model of Practical Reason

The KMPR rejects the idea that reasons for action must have certain desires at their source. If reasons for action are not based on desires, then the KMPR—or so it seems—holds an externalist view on reasons for action. Externalism subscribes to the thesis that justificatory reasons for action, as external reasons, do not need to be tied to the "*subjective motivational set*" (Williams 1981, 102) of the agent. It also affirms that the truth of normative reasons does not need to be capable of motivating the agent to whom the reasons apply to act accordingly. This leads to the peculiar situation where, as outlined above, an agent has normative reasons for an action, although these normative reasons are insufficient to motivate the action and thus are not capable of explaining the action. Hence it is difficult to see how externalism about reasons could underlie a model of practical reason that takes into account the normative and the descriptive dimensions of practical reason.

Therefore it is important to realize that there is conceptual space for the KMPR within an internalist account of reasons for action understood as the claim that 'having a reason to x' means that the person has a motive that is served or furthered by x–ing.[10] The KMPR simply denies that serving or furthering a motive by x–ing means that the motivation was generated by a desire. Indeed, the KMPR can argue that reasons for action need not be derived from a desire, although these reasons are to motivate the agent. Thus the KMPR claims to fulfill the *motivational requirement*—i.e. that reasons for action are to be capable of motivating. Hence, the crucial difference between the NHMPR and the KMPR is not whether or not reasons for action are to motivate an agent—both models can endorse this claim. Rather, the difference between these models lies in how they respond to the question as to what comprises the *source* of the

[10] The KMPR is construed here as a conception of practical rationality as 'weak internalism', as Gosepath 1992, 231, defines it.

motivation (Gosepath 1999, 16). The KMPR, as opposed to the NHMPR, puts forward the thesis that *reasons* can be at the source of a motivation for an action, but nevertheless can grant that an agent always satisfies a desire when performing it. In other words, the KMPR does not need to deny that a desire is *present* when an agent pursues an action, but argues that the desire itself is not the motive that is the motivating reason for the action.

Nagel cashed out this idea very clearly by distinguishing between "motivated" and "unmotivated" desires (Nagel 1978, 29). Motivated desires are those desires that a person has *after* a certain process of deliberation. These desires are not simply given to an agent in the sense that the agent merely perceives to have certain desires. Rather, these desires emerge from a process of reasoning—the result of which is the presence of a certain desire that consequentially will also motivate the agent to act in its pursuit. An agent, for instance, may have the motivated desire to prepare for an exam as she realizes that this will increase her chances of passing it. Unmotivated desires, on the other hand, are desires that the agent simply contemplates without further reflection. The desire to drink water, for instance, might simply occur to an agent although the agent did not think about whether or not she was thirsty in the first place. And more importantly, a "motivated" desire may, but need not be itself motivated by further desires of a different kind. Thereby some desires of the agent's "*subjective motivational set*" can generate—via "a sound deliberative route" (Williams 1995, 35)—other desires that then count as "motivated" desires. It is characteristic of the KMPR, however, that not all "motivated" desires are reducible to other desires. Rather, some desires are motivated by reasons which themselves are not dependent upon any particular desire of the agent to whom the reasons apply.[11]

6.2 Explaining Reciprocal Behavior with the KMPR

The KMPR enables to explain reciprocal behavior differently than the NHMPR and thus may constitute the basis for further explanatory accounts of the laboratory data. It can explain the reciprocal behavior in the following way. The second-mover acts reciprocally, because she thinks that reasons apply to her that do not allow her to act otherwise. Some second-movers, after all, may hold that reciprocating the trust of the first-mover in the Trust Game is a norm that "no one could reasonably reject" under such circumstances and thus constitutes a binding reason to act accordingly (Scanlon 1981, 110; cf. also Forst 1994, 64). Thereby, reciprocal behavior is considered to be a principled commitment to a particular type of action that does not have a desire of the agent at its source. In other words, the second-mover may recognize certain (moral) reasons that lead her to the judgment that it is her moral obligation to behave reciprocally.

While a second-mover who believes that she has an obligation to act reciprocally may form a desire to act reciprocally, an explanation that would consider this desire to be fundamental would fail to identify the actual *cause* for the

[11] Alternately, one may hold the view that desires only rarely provide reasons for action. This view maintains that almost all actions are to be explained with reference to motivating reasons that do not depend on desires (cf. Scanlon 1999, ch. 1).

reciprocal action. The explanation must start with the reasons that the second-mover perceived as (moral) reasons that could not be reasonably rejected and move from there to the desire which the agent formed in response to these reasons. Using Nagel's terminology the desire to act reciprocally is a "motivated" desire which itself is not necessarily reducible to other pre-existing desires of the "*subjective motivational set*" of the second-mover. Therefore the KMPR opens up a further explanatory approach of the reciprocal behavior encountered in the Trust Games that so far has received little attention.[12]

Note that this account of the motivation to behave reciprocally differs significantly from the treatment of social norms in the economic literature (cf. Young 2007 for an overview). Social norms refer to customary rules of behavior that *all* members of a given social group comply with. They effectively coordinate human interaction. Reciprocal behavior in Trust Games does not constitute a social norm, because—as mentioned above—there exists no unanimity of such behavior. Moreover, the account of the KMPR contains a very particular notion of the motivation to behave reciprocally, whereas many different types of motivations are compatible with the observance of social norms.

If one endorses the view that reasons are capable of motivating an agent, then one is prompted to clarify how such a view could be incorporated into a theory of rational choice which takes the preferences of agents as given. The rest of this section points to recent research in this area that takes up exactly this challenge.

6.3 A Reason-based Theory of Preference Formation and Experimental Strategies

Economic theory usually gives no answer as to why an agent has certain preferences, i.e. desires. It ignores in particular or takes as given the reasons that drive an agent's actions. Dietrich and List (2010) advance rational choice theory by developing a reason-based account of preference formation. In their model, an agent's preferences over payoffs or states of the world depend upon the reasons that motivate her. This implies that an agent's preferences can change when her motivating reasons come to change. This in turn allows for the possibility that agents have, to use Rawls's terminology, "the capacity to form, to revise and rationally to pursue a conception of one's rational advantage or good" (Rawls 2005[1993], 19; cf. also Dietrich/List 2010, 2). Such a reason-based theory of rational choice thus serves well to account for the insight of the KMPR that some players may be motivated to behave reciprocally because of certain reasons. It represents a genuine alternative to the early and recent models of economic theorists that all follow the NHMPR.

Dietrich and List's (2010) recent approach has two attractive features. First of all, the relationship between motivating reasons and preferences is modeled in such a way that it allows for a parsimonious representation of these preferences

[12] See, however, Peakock/Schefcyk/Schaber 2005, 194f., on the necessity to allow for motivations that are not compatible with "preference-satisfying" behavior in order to explain altruism.

even if they differ across the agent's motivational states. With this representation economists can use standard methods to analyze strategic interaction in economic games.[13] Second, the framework can also be used to analyze the relationship between an agent's normative reasons and the preferences she ought to have—instead of the relationship between her motivating reasons and her actual preferences. This enables making a distinction between a reason-based explanation and a reason-based justification of choices.

Moreover, experimental economics can show that a change in motivating reasons can trigger a change in behavior. A good example is the recent work of Benjamin, Choi and Fisher (2010). They study the impact of religious norms on economic behavior, by making these norms more salient to a randomly selected sample of student subjects. In the psychological literature, this technique has come to be known as 'priming'. The experimental results indicate that for some primed subjects the religious norm becomes a motivating reason. They show, for instance, that primed Protestants contribute more to a public good than their non-primed counterparts. This provides not only evidence for the fact that individuals' preferences are not stable, but also demonstrates how their behavior changes with motivating reasons. So, apparently, the motivation for this behavior has a reason rather than a desire at its source.

So far the article argued that early and recent models to explain reciprocal behavior rely on the NHMPR. Furthermore, it has suggested that the consideration of the KMPR might be an attractive basis for further attempts to explain the reciprocal behavior in Trust Games. The next section makes another, second proposal as to how models that explain reciprocal could be amended.

7. Reciprocal Behavior and Social Background Conditions

It is a striking feature of the laboratory experiments that they are interpreted in abstraction from the socio-cultural contexts in which the data are collected. This is remarkable because the effect of culture on economic outcomes is empirically well-established (cf. Guiso et al. 2006 for a summary; cf. Knack/Kneefer 1997 on the economic impact of different degrees of 'self-reported trust'). Moreover, there is also evidence that shows that behavior in economic games varies significantly across economically different countries (cf. Cardenas/Carpenter 2008) and culturally distinct groups (cf. Henrich et al. 2004). Thus, socio-cultural contexts apparently affect the results of the laboratory experiments. This has led some to make the point that "structural explanations" have to be supplemented to the analysis of the behavior in the laboratory experiments (Leist 2005, 168–170). Structural explanations take into account the socio-cultural context, i.e. cultural conventions, economic conditions, political constellations and legal norms, within which the experiment occurs, in order to properly explain the

[13] This could be done as in an economic model with "state-dependent preferences". In economics, different "states" refer to different "states of nature" that may change an agent's preferences, e.g. good or bad weather, high or low inflation, good or bad health. In contrast, different motivational states refer to what reasons are currently salient to the agent.

behavior in that experiment. Only with the use of such structural explanations can one clarify how the specific set of particular types of social relations influence the actions and motives of the agents in settings like the Trust Game.

In a similar vein, this section develops the significance of the social background conditions for the analysis of reciprocal behavior in Trust Games. The term 'social background conditions' refers to the major political and economic institutions, like the political constitution or the system of property rights, which Rawls dubs the "basic structure of society" (cf. Rawls 2005[1993], lect. 7). To do so, it illustrates that even in abstraction from the social background conditions the general concept of reciprocity allows for a variety of interpretations of what reciprocity substantively demands in the Trust Game. Then it explains how society-wide demands of reciprocity can affect the demands of reciprocity in local interactions of the type of the Trust Game generally and, more specifically, in the Trust Games in the laboratory experiments. The importance of the social background conditions suggests that the local analysis of reciprocal behavior needs to be complemented by an additional analysis of the social background conditions.

7.1 Alternative Understandings of Reciprocity's Demands

The general concept of reciprocity, i.e. "to return good in proportion to the good we receive" (cf. Becker 1986, 3), leaves fully unspecified the substantive account of proportionality by reference to which reciprocal behavior could be identified. To illustrate, consider three alternative points of view about what reciprocity demands of the second-mover in the Trust Game. From the first point of view, reciprocity demands that the second-mover returns the same amount of MUs that the first-mover sent (before it was tripled) back to the first-mover. The rationale is that because the tripling of the amount sent by the first-mover is not caused by the choice of the first-mover, but is a specific feature of the setup of the Trust Game, the effect of the tripling is therefore not attributable to the first-mover and does not need to be reciprocated. On the second point of view, a splitting of the amount of MUs that the second-mover receives in addition to her initial endowment is the proper understanding of reciprocity's demands. The idea here is that since it is only due to the first-mover's choice to send MUs that the second-mover receives more MUs than her initial endowment, an equal sharing of the MUs that the second-mover receives expresses the demands of reciprocity. From yet another point of view, reciprocity requires that the first- and second-mover dispose over the same amount of MUs in the final payoff structure. As from the second point of view, reciprocity demands an equal distribution, but this time the entire stock of MUs is to be shared equally. All three understandings of what reciprocity demands enjoy a certain amount of intuitive plausibility. So the normative question as to what reciprocity substantively demands in local interactions between two individuals cannot be easily answered, if at all.

7.2 The Importance of the Social Background Conditions

This question is further exacerbated, however, when one considers the social background conditions of local interactions. The connection between reciprocal demands in local interactions and the social background conditions is particularly relevant from a Rawlsian perspective of how to conceive of society and justice. Rawls understands society as a system of cooperation and justice as a normative conception determining the fair terms of this system of cooperation. Rawls's conception of justice as fairness, then, establishes standards which "specify an idea of reciprocity [...]: all who do their part as the recognized rules require are to benefit as specified by a public and agreed upon standard" (Rawls 2001, 6). Therefore, the notion of reciprocity is the central normative source of Rawls's conception of justice as fairness. Accordingly, as Gibbard puts it nicely, for Rawls "justice is fairness in exchange, but on a grand scale: it is fairness in the terms governing a society-wide system of reciprocity" (Gibbard 1991, 266). Rawls's conception of justice can hence also be labeled as 'Justice as Reciprocity' (Gibbard 1991; cf. also Buchanan 1990). Social justice means that society's members reciprocate the other members' abidance to the rules of fair terms of cooperation by supporting these rules themselves.[14] From a Rawlsian perspective, the entire system of cooperation that a society constitutes is subject to the demands of reciprocity.[15]

Now, if the society-wide demands of reciprocity are not met and thus, from a Rawlsian perspective, background injustice exists, then certain demands of reciprocity in local interactions that would otherwise hold can be invalidated. Consider, for instance, the relationship between a firm and a worker. The firm may pay its worker a generous wage and expect that the worker must reciprocate the benefits of the generous wage by increasing her efforts to enhance her performance. If the social background conditions are just, the worker may, indeed, have such an obligation of reciprocity. However, if the social background conditions are unjust, the worker may *not* have this obligation. The justification may be that not reciprocating in a local interaction may be in fact a legitimate means to gain that which one is entitled to—all things considered. In a society with very disparate opportunities, for instance, disadvantaged members, who have had to carry more burdens than others in order to become employed, may legitimately not reciprocate the generous wage of their employer. So certain normative deficits of society, for instance in the educational system or the labor market, may result in unequal opportunities for equally talented and motivated members of society and thereby influence the demands of reciprocity in the firm-worker relationship. They can eventually void the demands of reciprocity in certain local interactions. In such a way the validity of the demands of reciprocity in a local interaction can be conditional upon the particular influence of the social background conditions.

---

[14] As Rawls 2005[1993], 17, states clearly: "[T]he two principles of justice [...] formulate an idea of reciprocity."

[15] The last paragraph benefited from Lister 2011.

In the case of the Trust Game, therefore, the analysis of reciprocal behavior may have to be complemented with a careful examination of the specific characteristics of the social background conditions of the laboratory experiments. If the social background conditions are indeed very unjust, then the failure of the second-mover to reciprocate the trust of the first-mover could also represent a partial realization of that which she is entitled to. The second-mover may occupy a disadvantaged position in an unjust society and legitimately increase her final payoff by not reciprocating the first-mover's trust. Such a deliberation may be considered to be problematic in the case where the first-mover may equally belong to a disadvantaged group of an unjust society. If that is the case, then the demands of reciprocity may equally apply as in interactions among members of a fully just society. Nevertheless, a careful consideration of the justice of the social background conditions remains important, because the second-mover's uncertainty of the first-mover's group membership (because of the anonymity among the experimental subjects) may justify the second-mover's unwillingness to reciprocate the trust. After all, if the society in which the Trust Game is played is very unjust and if the second-mover clearly belongs to a very disadvantaged group of society, she may expect with a sufficiently high probability that the first-mover belongs to a privileged group of society and failing to reciprocate may therefore be justified.

The consequence of this discussion of the demands of reciprocity in local interactions and their dependency on a society-wide system of reciprocity is that the analysis of reciprocal behavior in Trust Games can gain in explanatory power from further and complementary analyses of the social background conditions of the laboratory experiments. Even if the experiments secure anonymity among the players and endow both players with an equal amount of MUs, this does not guarantee a level-playing field. Rather, at least *some* players may permissibly view the game as an opportunity to get what is due to them in light of broader considerations of society-wide reciprocity.

8. Conclusion

The findings of experimental economists seriously challenge the pure self-interest thesis of human motivation and continue to evoke new explanatory models to explicate the laboratory data that are generated by a broad variety of economic games. This article showed that both early and recent economic models to explain reciprocal behavior in Trust Games rely on the NHMPR. The NHMPR is distinct in that it claims that a desire is always at the source of a motivation. This model was contrasted to the KMPR that ascribes to reasons the capacity to motivate action. The KMPR can also explain the reciprocal behavior in Trust Games and recently efforts have been made to develop a reason-based theory of rational choice. So the KMPR may be an attractive alternative to the NHMPR. Finally, the article highlighted the relation between local and society-wide demands of reciprocity and suggested that further examinations of the social background conditions of the laboratory experiments would prove valu-

able. So the article proposed two directions that economists and social scientists might want to further consider so as to establish a more solid foundation for modern economic theory.

Bibliography

Batson, C. D. (1998), Altruism and Prosocial Behavior, in: Gilbert, D. T./S. Fiske/G. Lindzey (eds.), *Handbook of Social Psychology, Vol. 2*, New York, 282–316

Becker, L. (1986), *Reciprocity*, Chicago

Bénabou, R./J. Tirole (2006), Incentives and Prosocial Behavior, in: *American Economic Review* 96(5), 1652–1678

Benjamin, D. J./J. Choi/G. W. Fisher (2010), Religious Identity and Economic Behavior, in: *NBER Working Paper* No. 15925, 1–33

Berg, J./J. Dickhaut/K. McCabe (1995), Trust, Reciprocity, and Social History, in: *Games and Economic Behavior* 10, 122–142

Bolton, G./A. Ockenfels (2000), ERC: A Theory of Equity, Reciprocity, and Competition, in: *American Economic Review* 90(1), 166–193

Brennan, G./P. Pettit (2004), *The Economy of Esteem*, Oxford

Buchanan, A. (1990), Justice as Reciprocity versus Subject-Centered Justice, in: *Philosophy & Public Affairs* 19(3), 227–252

Camerer, C. (2003), *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton

Cardenas, J./J. Carpenter (2008), Behavioral Development Economics: Lessons from Field Labs in the Developing World, in: *Journal of Development Studies* 44(3), 311–338

Charness, G./M. Dufwenberg (2006), Promises and Partnership, in: *Econometrica* 74(6), 1579–1601

Dietrich, F./C. List (2010), A Reason-based Theory of Rational Choice, URL: http://personal.lse.ac.uk/list/PDF-files/ReasonBasedRCT.pdf (March 2011, draft version as of 23 June 2010), 1–33 [forthcoming in: *Noûs*]

Dufwenberg, M./G. Kirchsteiger (2004), A Theory of Sequential Reciprocity, in: *Games and Economic Behavior* 47(2), 268–298

Ellingsen, T./M. Johannesson (2008), Pride and Prejudice: The Human Side of Incentive Theory, in: *American Economic Review* 98(3), 990–1008

Elster, J. (1985), The Nature and Scope of Rational-Choice Explanation, in: LePore, E./B. P. McLaughlin (eds.), *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, Oxford, 60–72

Falk, A./U. Fischbacher (2006), A Theory of Reciprocity, in: *Games and Economic Behavior* 54(2), 293–315

Fehr, E./G. Kirchsteiger/A. Riedl (1993), Does Fairness Prevent Market Clearing? An Experimental Investigation, in: *Quarterly Journal of Economics* 108(2), 437–459

— /K. Schmidt (1999), A Theory of Fairness, Competition, and Cooperation, in: *Quarterly Journal of Economics* 114(3), 817–868

— /S. Gächter (2000), Cooperation and Punishment in Public Goods Experiments, in: *American Economic Review* 90(4), 980–994

— /K. Schmidt (2006), The Economics of Fairness, Reciprocity and Altruism—Experimental Evidence and New Theories, in: Kolm, S.-G./J. M. Ythier, *Handbook of the Economics of Giving, Altruism and Reciprocity*, Amsterdam, 615–691

Forst, R. (1994), *Kontexte der Gerechtigkeit*, Frankfurt; (2002), *Contexts of Justice*, Berkeley

Geanakoplos, J./D. Pearce/E. Stacchetti (1989), Psychological Games and Sequential Rationality, in: *Games and Economic Behavior* 1, 60–79

Gibbard, A. (1991), Review: Constructing Justice, in: *Philosophy & Public Affairs* 20(3), 264–279

Gosepath, S. (1992), *Aufgeklärtes Eigeninteresse*, Frankfurt

— (1999), Praktische Rationalität, in: Gosepath, S. (ed.), *Motive, Gründe, Zwecke*, Frankfurt, 7–53

Guiso, L./P. Sapienza/L. Zingales (2006), Does Culture Affect Economic Outcomes?, in: *Journal of Economic Perspectives* 20(2), 23–48

Henrich, J. et al. (2004) (eds.), *Foundations of Human Sociality. Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, Oxford

Hume, D. (1978[1739]), *Treatise of Human Nature*, Selby-Bigge, L. A. (ed.), rev. Nidditch, P. H., Oxford

Knack, S./P. Keefer (1997), Does Social Capital Have an Economic Payoff? A Cross-Country Investigation, in: *The Quarterly Journal of Economics* 112(4), 1251–1288

Korsgaard, C. (1986), Skepticism about Practical Reason, in: *Journal of Philosophy* 83(1), 5–25

Leist, A. (2005), Social Relations Instead of Altruistic Punishment, in: *Analyse & Kritik* 27, 158–171

Lister, A. (2011), Justice as Fairness and Reciprocity, *this issue*

Mauss, M. (1990[1923]), *The Gift: The Form and Reason for Exchange in Archaic Societies*, London

Nagel, T. (1978), *The Possibility of Altruism*, Princeton

Peacok, M./M. Schefczyk/P. Schaber (2005), Altruism and the Indispensability of Motives, in: *Analyse & Kritik* 27, 188–196

Rabin, M. (1993), Incorporating Fairness into Game Theory and Economics, in: *American Economic Review* 83(5), 1281–1302

Rawls, J. (2005[1993]), *Political Liberalism*, New York

— (2001), *Justice as Fairness*, ed. by E. Kelly, Cambridge/MA

Scanlon, T. (1981), Contractualism and Utilitarianism, in: Sen, A./B. Williams, *Utilitarianism and Beyond*, Cambridge, 103–128

— (1999), *What We Owe to Each Other*, Cambridge/MA

Sliwka, D. (2007), Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes, in: *American Economic Review* 97(3), 999–1012

Smith, M. (1987), The Humean Theory of Motivation, in: *Mind* 96, 36–61

Weber, M. (1985[1921]), *Wirtschaft und Gesellschaft*, ed. by J. Winckelmann, Tübingen

Williams, B. (1981), Internal and External Reasons, in: *Moral Luck*, Cambridge, 101–113

— (1995), Morality and the Obscurity of Blame, in: *Making Sense of Humanity*, Cambridge, 35–45

Young, P. (2007), Social Norms, in: *Department of Economics Discussion Paper Series* 307, University of Oxford