Available online at www.sciencedirect.com
SCIENCE DIRECT®
Games and Economic Behavior 54 (2006) 293–315
GAMES and Economic Behavior
www.elsevier.com/locate/geb

A theory of reciprocity ☆

Armin Falk [a], Urs Fischbacher [b,*]

[a] *Institute for the Study of Labor and University of Bonn, IZA, PO Box 7240, D-53072 Bonn, Germany*
[b] *Institute for Empirical Research in Economics, University of Zurich, Blümlisalpstrasse 10, CH-8006 Zürich, Switzerland*

Received 6 February 2003

Available online 26 April 2005

Abstract

People are reciprocal if they reward kind actions and punish unkind ones. In this paper we present a formal theory of reciprocity. It takes into account that people evaluate the kindness of an action not only by its consequences but also by its underlying intention. The theory is in line with the relevant stylized facts of a wide range of experimental games, such as the ultimatum game, the gift-exchange game, a reduced best-shot game, the dictator game, the prisoner's dilemma, and public goods games. Furthermore, it predicts that identical consequences trigger different reciprocal responses in different environments. Finally, the theory explains why outcomes tend to be fair in bilateral interactions whereas extremely unfair distributions may arise in competitive markets.

JEL classification: C7; C91; C92; D64; H41

Keywords: Reciprocity; Fairness; Cooperation; Competition; Game theory

☆ Financial support by the Swiss National Science Foundation (Project 12-43590.95), the Ludwig Boltzmann Institute for the Analysis of Economic Growth and the Network on the Evolution of Preferences and Social Norms of the MacArthur Foundation is gratefully acknowledged.

* Corresponding author.
E-mail addresses: falk@iza.org (A. Falk), fiba@iew.unizh.ch (U. Fischbacher).

0899-8256/$ – see front matter
doi:10.1016/j.geb.2005.03.001

Kindness is the parent of kindness.
(Adam Smith, 1759)

1. Introduction

In this paper, we develop a formal theory of reciprocity. According to this theory, reciprocity is a behavioral response to perceived kindness and unkindness, where kindness comprises both distributional fairness as well as fairness intentions. There is a large body of evidence which indicates that reciprocity is a powerful determinant of human behavior: experiments and questionnaire studies performed by psychologists and economists as well as an impressive literature in sociology, ethnology and anthropology emphasize the omnipresence of reciprocal behavior (see, e.g., Kahneman et al., 1986; Fehr and Gächter, 2000).[1] In the ultimatum game, for example, low offers are frequently rejected (Güth et al., 1982; Thaler, 1988; Güth, 1995; Camerer and Thaler, 1995; Roth, 1995). In addition, if subjects are given the possibility of sanctioning each other, subjects often sanction defectors, even if sanctioning is costly (Fehr and Gächter, 2000; Carpenter and Matthews, 2003). The reward of kind actions is reported, e.g., in the investment game (Berg et al., 1995) or in the gift-exchange game (Fehr et al., 1993). Some evidence from market experiments, however, *seems* to be incompatible with reciprocal preferences. These experiments typically support the outcome standard economic theory predicts, which assumes selfish preferences. We show below that our theory is capable of reconciling the seemingly contradictory evidence that bilateral interactions may yield distributions, which seem to be "fair" while competitive markets often produce "very unfair" distributions.

According to our theory, a reciprocal action is modeled as the behavioral response to an action that is perceived as either kind or unkind. The central part of the theory is therefore devoted to the question *how people evaluate the kindness of an action*. Two aspects are essential in our model,

(ii) the *consequences* of an action, and
(ii) the actor's underlying *intentions*.

The fact that fair intentions play a major role for the perception of kindness is suggested by several experimental studies (Brandts and Sola, 2001; Falk et al., 2003; McCabe et al., 2003; Offerman, 2002; Greenberg and Frisch, 1972; Goranson and Berkowitz, 1966).[2] In Falk et al. (2000) e.g., second movers could reciprocate first movers' kind or unkind

[1] Importantly, reciprocity means a behavior that cannot be justified in terms of selfish and purely outcome-oriented preferences. To avoid terminological confusion let us, therefore, clarify that reciprocity is sharply distinguished from 'reciprocal altruism' (Trivers, 1971). A reciprocal altruist is only willing to reciprocate if there are future rewards arising from reciprocal actions. In the parlance of game theory this kind of reciprocal action may be supported as an equilibrium strategy in infinitely repeated games (folk theorems) or in finitely repeated games with incomplete information (see Kreps et al., 1982).

[2] For a dissenting view, see Bolton et al. (1998), and Cox (2003).

actions. In a treatment where first movers actually make decisions, we observe strong positive and negative reciprocity. In a treatment where first movers' actions are determined randomly, however, reciprocal responses following the *same* "actions" are significantly weaker. Similarly, Falk et al. (2003) show that in a series of reduced ultimatum games, the exact same offer is rejected at a significantly different rate, depending on the proposers' choice set. A given offer x is rejected at a higher rate if the proposer's action signals an unfair intention (because he could have chosen a more friendly offer) compared to a situation where x signals no intention or even a fair one. Thus, intention matters for the perception of kindness and the corresponding reciprocation. Notice, however, that even in situations where intention is absent, most people still exhibit some reciprocal behavior. In Falk et al. (2000), some second movers punish unfair offers and reward advantageous offers, even if offers were randomly determined. This finding is corroborated by the experiments by Blount (1995) and Charness (2004), who report that reciprocity is weak but not absent in a condition where intention plays no role. In our model, we therefore incorporate *both* the concern for the outcome per se as well as for the underlying intention.

This paper is organized as follows: in the following section we present evidence of a questionnaire study we performed to elicit how people evaluate the kindness of an action. Section 3 introduces the formal model. Section 4 discusses applications. The games we address are the ultimatum game, the gift-exchange game, a reduced best-shot game, competitive market games, the dictator game, the sequential prisoner's dilemma, and the centipede game. Section 5 concludes and discusses how our approach differs from other fairness models of inequity aversion (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999) and reciprocity (Rabin, 1993; Levine, 1998; Dufwenberg and Kirchsteiger, 2004; Charness and Rabin, 2002).

2. How people evaluate kindness: questionnaire evidence

Reciprocity is the behavioral response to a perceived kindness or unkindness. It is therefore crucial to understand how people evaluate the kindness of a particular action. In order to empirically investigate this question we conducted a questionnaire study with 111 subjects from the University of Zurich and the Swiss Federal Institute of Technology Zurich. Each subject i in this study was in a hypothetical bilateral exchange situation with another subject j. Subjects i were asked to indicate how *kind* or *unkind* they perceive different divisions of an endowment of 10 Swiss Francs (roughly 7 $US at that time) where j always divides the pie between herself and i. Subjects could express the kindness or unkindness of a particular outcome by choosing a sign ($+$ or $-$) and a number between 0 and 100.[3] The most unkind was expressed by -100, slightly less unkind was -99, etc. The most kind was $+100$, slightly less kind was $+99$, etc. To understand the nature of i's perception of j's kindness, we systematically varied j's set of alternatives and asked players i how kind they perceived different actions of j for each set of alternatives.

[3] The questionnaire was in German. We used the expression "nett" to elicit kindness and "nicht nett" to elicit unkindness, respectively.

Table 1
Player i's estimation of j's kindness (average values, $n = 111$)

(π_j, π_i)	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)
(0, 10)	+72.3					+79.9	+73.4		+80.3
(1, 9)	+68.0					+73.3	+62.0		+72.5
(2, 8)	+62.0	+75.3		+41.1	+61.2	+61.9	+40.8		+62.2
(3, 7)	+51.4								
(4, 6)	+40.0								
(5, 5)	+29.4	+33.4							+27.9
(6, 4)	−23.2								
(7, 3)	−52.9								
(8, 2)	−71.9	−70.6	−31.5		−47.7	−50.5		−9.1	−60.9
(9, 1)	−84.5					−80.3		−56.4	−82.6
(10, 0)	−95.4					−97.3		−88.8	−97.3

In total subjects were given nine different decision situations, which are summarized in Table 1. For example, the first decision situation is displayed in column (i). In this situation j's choice set contains 11 possible offers, ranging from offering 10 and keeping 0 (denoted by $(\pi_j, \pi_i) = (0, 10)$),[4] offering 9 and keeping 1 $(1, 9)$, and so on up to offering 0 and keeping 10 $(10, 0)$. For each of these 11 possible offers, subjects had to indicate their perceived kindness or unkindness. In columns (ii) to (ix) the action space is smaller than in (i): In column (ii), e.g., j can offer only 2, 5 or 8 to player i, while j can offer only 2 in column (iii) and so on. Table 1 reveals several regularities, which are important for an understanding of how people evaluate kindness and which will be incorporated in our model.

Let us first look at column (i). If j offers 0 to i (and keeps everything for herself), players i perceive this as very unkind on average (last row, -95.4). If j offers 1 (and keeps 9 for herself) this is regarded as less unkind (-84.5) and so on. If j keeps nothing for herself it is viewed as very kind ($+72.3$). Column (i) shows that kindness is monotonically increasing in the offer. The more j is willing to share with i the more kind this is perceived by i. Moreover, an *equitable share* of payoffs seems to be the reference standard to determine what is a fair or unfair offer. This can be inferred from the fact that at the equitable offer of 5 the sign changes from $-$ to $+$, i.e., the perception changes from 'unkind' to 'kind'. This observation will be used to justify equity as a reference standard in our model.

Equity is also used as a reference standard in the inequity aversion models. In these models, however, the perceived kindness of an offer is solely determined by the material outcomes. In contrast to this assumption, the results from our questionnaire clearly indicate that player j's intentions play an important role as well. The signaling of fairness intention rests on two premises:

(i) Player j's choice set actually allows the choice between a fair and an unfair action, and
(ii) j's choice is under her full control.

[4] We use the order (π_j, π_i) because player j is in the situation of a first mover. Player i is in the situation of a responder, since he has the possibility of expressing an emotional response.

From these two premises it immediately follows that in order to evaluate the intentions of a particular action of j, player i takes into account the alternatives j had, i.e., he takes j's strategy set into account. To better understand how j's strategy set influences i's perception of j's kindness we now discuss columns (ii) to (ix). We focus primarily on the "kind" $(2, 8)$ offer and the "unkind" $(8, 2)$ offer. Six observations with respect to how intentions play a role can be derived.

First, if j's strategy set contains only one element, i.e., if j has no alternative to choose from, player i cannot learn much about j's intentions. As a consequence, the perceived kindness or unkindness of the *same* offer is much weaker, compared to a situation where j can choose between fair and unfair offers. This can be seen by comparing the indicated average kindness values for the $(2, 8)$ offer in column (i) and (iv) (62.0 vs. 41.1) and the $(8, 2)$ offer in column (i) and column (iii) (-71.9 vs. -31.5). *Second*, even if j has no alternative and therefore cannot signal any intention, the perceived kindness or unkindness of an offer is *not zero* (see columns (iii) and (iv): $41.1 > 0$ and $-31.5 < 0$). Together observations 1 and 2 reveal that both, intentions and outcomes are important, i.e., pure inequity aversion as well as purely intention driven reciprocity cannot explain the data.

Third, even if j's strategy space is *limited*, a friendly offer $(2, 8)$ is viewed as similarly kind compared to an unlimited strategy space as long as j could have made an offer to i which was less friendly. This means that $(2, 8)$ signals fair intentions if j could have been less friendly (compare the indicated average kindness values for the $(2, 8)$ offer in column (i) (62.0) with the respective values in columns (ii) (75.3), (v) (61.2) and (vi) (61.9)). By the same token, the kindness of the $(2, 8)$ offer is lower if player j does not have the chance to make a less friendly offer (compare column (i) (62.0) with columns (iv) (41.1) and (vii) (40.8)). The intuition for the latter result is straightforward. If j has no chance to behave more "opportunistically," how should i infer that j really wanted to be kind from a friendly action? After all, j took the least friendly action possible. *Fourth*, the perception of an unfriendly offer $(8, 2)$ depends on j's possibility to make a more friendly offer: If j has no chance to make a friendlier offer, $(8, 2)$ is not viewed as very unkind (compare column (i) (-70.6) with columns (iii) (-31.5) and (viii) (-9.1)). The intuition is that you cannot blame a person for being mean if—after all—she did the best she could. *Fifth*, if j *has* the option of making a friendlier offer, the perception of the unfair offer $(8, 2)$ depends on *how much j has to sacrifice in order to make the more friendly offer*. If making a more friendly offer implies that i earns more than j, $(8, 2)$ is still perceived as unkind but not that much. The intuition is that it is not reasonable to demand that the other person is fair to me if this implies that (relative to me) she puts herself in a disadvantageous position. Put differently j's unwillingness to propose an *unfair offer to herself* does not reveal that she wants to be unfair to i (compare column (i) (-70.6) with columns (v) (-47.7) and (vi) (-50.5)). If, however, there is an offer, which is friendlier and does *not* imply that j earns less than i, making the $(8, 2)$ offer is considered as quite unkind (compare column (i) (-71.9) with (ii) (-70.6) and (ix) (-60.9)). *Sixth*, fairness intention perception is *not symmetric* with respect to kindness and unkindness (beyond the absolute values of the perceived kindness). We see this for instance when comparing columns (i) and column (v). In column (i), the kindness of giving 8 is on average 62.0 and the unkindness of giving only 2 is -71.9. Let us now compare the kind $(2, 8)$-offer and the unkind $(8, 2)$-offer in column (i) with the identical offers in column (v) where j can decide only between $(2, 8)$

and $(8, 2)$. If j chooses the kind $(2, 8)$-offer, the perceived kindness is 61.9, which is practically indistinguishable from that displayed in column (i). In case of the unkind $(8, 2)$-offer the perceived unkindness drops significantly to -47.7.

These six observations together with the shown importance of equity as a reference standard will be used to model perceived kindness. This will be done in the next section.

3. The model

Our theory formalizes the concept of reciprocity, which consists of a kind (or unkind) treatment (represented by the *kindness term* φ) and a behavioral reaction to that treatment (represented by the *reciprocation term* σ). Our procedure transforms a standard game into a psychological game, the so-called "reciprocity game." The players' utility in this new game not only depends on the payoffs of the original game but also on the kindness and the reciprocation term. In the following, we derive both terms.

Consider a two-player extensive form game with a finite number of stages and with complete and perfect information. (We develop the theory for the two-player case for notational simplicity. The extension to games with more than 2 players is given in Appendix 2.[5]) Let $i \in \{1, 2\}$ be a player in the game. N denotes the set of nodes and N_i is the set of nodes where player i has the move. Let $n \in N$ be a node of the game. Let A_n be the set of actions in node n. Let F be the set of end nodes of the game. The payoff function for player i is given by $\pi_i : F \to \mathbb{R}$.

Let $P(A_n)$ be the set of probability distributions over the set of actions in node n. Then $S_i = \prod_{n \in N_i} P(A_n)$ is player i's behavior strategy space. Thus, a player's behavior strategy puts a probability distribution on each of the player's decision nodes. Let player j be the other player[6] and let k be one of the players (either i or j). For $s_i \in S_i$ and $s_j \in S_j$ we define $\pi_k(s_i, s_j)$ as k's expected payoff, given strategies s_i and s_j.

Let $s_i \in S_i$ be a behavior strategy. We define $s_i|n$ as the strategy s_i *conditional* on node n. This strategy is simply s_i, except for the fact that the probability of the unique actions leading to n are set to 1 for all nodes n' which precede n and the probabilities of the other actions in nodes n' are set to 0. Furthermore, we define $\pi_k(n, s_i, s_j) := \pi_k(s_i|n, s_j|n)$ as the expected payoff conditional on node $n \in N$, as the expected payoff of player k in the subgame starting from node n, given that the strategies s_i and s_j are played.

Let s_i' denote the *first-order belief* of player i. It captures i's belief about the behavior strategy $s_j \in S_j$, which player j will choose. Similarly, the *second-order belief* s_i'' of player i is defined as i's belief about j's belief about which behavior strategy i will choose. In other words, s_i'' is i's belief about s_j'. Like Rabin (1993), we assume that s_i' is an element of S_j and s_i'' is an element of S_i. A set of beliefs is said to be *consistent*, if $s_i = s_j' = s_i''$ holds for $i \neq j$.

[5] This appendix is available as a pdf-document at: http://www.iew.unizh.ch/home/fischbacher/downloads/fafiA2-3.pdf.

[6] Throughout the paper, we will use the male form for player i (and for first movers). For player j (and second movers) we will use the female form.

3.1. The kindness term φ

The kindness term φ_j is the central element of our theory. It measures how kind a *person i perceives the action by another player j*. As outlined in the introduction as well as in the previous section, there is ample evidence that the perceived kindness of an action depends on the consequence or outcome of that action and the underlying intention. In our theory, the outcome is measured with the *outcome term* Δ_j, where $\Delta_j > 0$ expresses an advantageous outcome for player i and $\Delta_j < 0$ expresses a disadvantageous outcome for i. In order to determine the overall kindness, Δ_j is multiplied with the *intention factor* ϑ_j. This factor is a number between zero and one, where $\vartheta_j = 1$ captures a situation where Δ_j is the result of an action which j completed intentionally, and $\vartheta_j < 1$ implies the fact that j's action was not fully intentional. The kindness term φ_j is simply the product of Δ_j and ϑ_j.

Definition 1. For given strategies and beliefs we define the *kindness term* $\varphi_j(n, s_i'', s_i')$ in a node $n \in N_i$ as:

$$\varphi_j\big(n, s_i'', s_i'\big) = \vartheta_j\big(n, s_i'', s_i'\big)\Delta_j\big(n, s_i'', s_i'\big). \tag{1}$$

In the following we derive both terms (Δ and ϑ) in detail. First, we define the *outcome term*:

$$\Delta_j\big(n, s_i'', s_i'\big) := \pi_i\big(n, s_i'', s_i'\big) - \pi_j\big(n, s_i'', s_i'\big). \tag{2}$$

To interpret this expression, let us fix the intention factor $\vartheta_j(n, s_i'', s_i')$. For a given $\vartheta_j(n, s_i'', s_i')$, the outcome term $\Delta_j(n, s_i'', s_i')$ captures the kindness of player j as perceived by player i: the kindness of j in node n increases, ceteris paribus, as the offers to player i increase. This is expressed in the term $\pi_i(n, s_i'', s_i')$. From j's perspective, j offers $\pi_i(n, s_j', s_j)$ to i. This is the payoff i expects to get if j chooses s_j and if j expects i to choose s_j'. Player i's belief about this offer is $\pi_i(n, s_i'', s_i')$.

The sign of $\Delta_j(n, s_i'', s_i')$ determines whether an action is considered to be kind or unkind. In order to determine the sign of $\Delta_j(n, s_i'', s_i')$, i needs to compare the offer $\pi_i(n, s_i'', s_i')$ with a *reference standard*. The empirical evidence presented in Table 1 as well as many experiments indicate that an *equitable* share of payoffs is a salient and commonly held standard.[7] As long as the payoff of i is larger than that of j, $\Delta_j(n, s_i'', s_i') > 0$, while $\Delta_j(n, s_i'', s_i') < 0$ if i's payoff is lower than that of j. The expression $\pi_j(n, s_i'', s_i')$ models the equitable reference standard; it is i's belief about which payoff j wants to keep for herself. If $\pi_i(n, s_i'', s_i') > \pi_j(n, s_i'', s_i')$ holds, player i thinks that j wants him to get more than j wants for herself, i.e., i believes that j is acting kindly. If, on the other

[7] The idea that equity is a salient reference standard was first developed in the so-called *equity theory*. Beginning in the late sixties social psychologists developed *equity theory* as a special form of *social exchange theory*. Compare, e.g., Adams (1965), and Walster and Walster (1978). See also Loewenstein et al. (1989).

hand, $\pi_i(n, s_i'', s_i') < \pi_j(n, s_i'', s_i')$ holds, i believes that j claims more for herself than she is willing to leave for i. In this case, i perceives j to be unkind.[8]

Let us now derive the *intention factor* ϑ, which models fairness intention. As discussed in Section 2, a player i takes into account j's choice alternatives in order to evaluate the intention of a particular action of j. Let us first state precisely what we mean by alternative payoff combinations. Let S_j^p be the set of pure strategies of j. For given strategies and beliefs we define in a node n:

$$\Pi_i\left(n, s_i''\right) := \left\{\left(\pi_i\left(s_i''|n, s_j^p\right), \pi_j\left(s_i''|n, s_j^p\right)\right) \mid s_j^p \in S_j^p\right\}. \tag{3}$$

$\Pi_i(n, s_i'')$ is a set of payoff combinations. This set contains the payoff combinations player j can induce by choosing a pure strategy s_j^p, given her beliefs about player i's strategy. Since $\Pi_i(n, s_i'')$ is determined from player i's perspective, player i takes his belief into account about which strategy player j believes he will choose, namely $s_i''|n$. In short, $\Pi_i(n, s_i'')$ is player i's belief about all payoff combinations player j considers as her payoff opportunity set.

In the previous section, we presented six observations on how alternatives matter for the perception of kindness. We will now incorporate these observations into the model. For expositional ease, we present in the main text a simplified version of the model, which uses only the first four observations. A version which takes all six observations into account is presented in Appendix A. As a notational simplification for the next equation, we define $\pi_i^0 = \pi_i(n, s_i'', s_i')$ and $\pi_j^0 = \pi_j(n, s_i'', s_i')$, the payoffs that determine the outcome term $\Delta_j(n, s_i'', s_i')$. We define the *intention factor*:

$$\vartheta_j\left(n, s_i'', s_i'\right) = \begin{cases} 1 & \text{if } \pi_i^0 \geqslant \pi_j^0 \text{ and } \exists \tilde{\pi}_i \in \Pi_i(n, s_i'') \text{ with } \tilde{\pi}_i < \pi_i^0, \\ \varepsilon_i & \text{if } \pi_i^0 \geqslant \pi_j^0 \text{ and } \forall \tilde{\pi}_i \in \Pi_i(n, s_i''),\ \tilde{\pi}_i \geqslant \pi_i^0, \\ 1 & \text{if } \pi_i^0 < \pi_j^0 \text{ and } \exists \tilde{\pi}_i \in \Pi_i(n, s_i'') \text{ with } \tilde{\pi}_i > \pi_i^0, \\ \varepsilon_i & \text{if } \pi_i^0 < \pi_j^0 \text{ and } \forall \tilde{\pi}_i \in \Pi_i(n, s_i''),\ \tilde{\pi}_i \leqslant \pi_i^0, \end{cases} \tag{4}$$

where ε_i is an individual parameter with $0 \leqslant \varepsilon_i \leqslant 1$. This parameter is called the pure outcome concern parameter. It is interpreted below.

The term ϑ_j equals 1 if and only if there is any true alternative. If the outcome is advantageous ($\pi_i^0 \geqslant \pi_j^0$), an alternative is true if it results in a lower payoff for player i, i.e., $\tilde{\pi}_i < \pi_i^0$ (third observation). If the outcome is disadvantageous ($\pi_i^0 < \pi_j^0$), an alternative is true if it results in a higher payoff for player i, i.e., $\tilde{\pi}_i < \pi_i^0$ (fourth observation). Note that in the special case where player j has *no* alternative at all (i.e., $|\Pi_i(n, s_i'')| = 1$), $\vartheta_j(n, s_i'', s_i')$ equals ε_i (this corresponds to our first observation). From our second observation, it follows that there are players with $\varepsilon_i > 0$.

The individual parameter ε_i is called the *pure outcome concern parameter*. It measures a player i's *pure* concern for an equitable *outcome*: If, e.g., ε_i is equal to zero, player i considers a particular outcome only as kind or unkind if it was caused intentionally, i.e.,

[8] Note that we talk a bit loosely about the kindness of an *action*. The way we model kindness comprises both the kindness of actions which actually occurred as well as anticipated future actions. Node n reflects the actions which already occurred. The belief s_i' reflects j's anticipated actions.

if the other player had an alternative to act differently. A player with $\varepsilon_i = 0$ has a purely intention driven notion of fairness. A player with $\varepsilon_i = 1$ cares *only* about the consequences of j's action, i.e., intention plays no role.[9]

3.2. The reciprocation term σ

The second ingredient of our theory concerns the formalization of reciprocation. Let us fix an end node f that follows (directly or indirectly) node n. Then $v(n, f)$ denotes the unique node that directly follows node n on the path that leads from n to f.

Notation. Let n_1 and n_2 be nodes. If node n_2 follows node n_1 (directly or indirectly), we denote this by $n_1 \to n_2$.

Definition 2. Let strategies and beliefs be given as above. Let i and j be the two players and n and f be defined as above. Then we define

$$\sigma_i\left(n, f, s_i'', s_i'\right) := \pi_j\left(v(n, f), s_i'', s_i'\right) - \pi_j\left(n, s_i'', s_i'\right) \tag{5}$$

as the *reciprocation term* of player i in node n.

The *reciprocation term* expresses the response to the experienced kindness, i.e., it measures how much i alters the payoff of j with his move in node n. Given i's belief about j's expectations about her payoff in node n (i.e., given $\pi_j(n, s_i'', s_i')$), i can choose an action in node n. The reciprocal impact of this action is represented as the *alteration* of j's payoff from $\pi_j(n, s_i'', s_i')$ to $\pi_j(v(n, f), s_i'', s_i')$ (always from i's perspective). For a given $\pi_j(n, s_i'', s_i')$, i can thus either choose to reward or to punish j. A rewarding action implies a positive, whereas a punishment implies a negative *reciprocation term*.

3.3. The utility function

Having defined the kindness and reciprocation term, we can now derive the players' utility of the transformed "reciprocity game":

Definition 3. Let i and j be the two players of the game and f an end node. We define the utility in the transformed reciprocity game as:

$$U_i\left(f, s_i'', s_i'\right) = \pi_i(f) + \rho_i \sum_{\substack{n \to f \\ n \in N_i}} \varphi_j\left(n, s_i'', s_i'\right) \sigma_i\left(n, f, s_i'', s_i'\right). \tag{6}$$

According to Definition 3, player i's utility in the reciprocity game is the sum of the following two terms: the first term of the sum is simply player i's *material payoff* $\pi_i(f)$.

[9] Thus, saying that a person's ε_i is always equal to 1 means that this person is purely outcome oriented, as suggested by the models of Bolton and Ockenfels (2000), and Fehr and Schmidt (1999). An analysis of the individual kindness assessments discussed in Table 1 reveals that there is substantial heterogeneity with respect to ε_i.

The second term—which we call *reciprocity utility*—is composed of the reciprocity parameter ρ_i, the kindness term $\varphi_j(n, s_i'', s_i')$, and the reciprocation term $\sigma_i(n, f, s_i'', s_i')$.

The *reciprocity parameter* ρ_i is a positive constant. Both parameters ρ_i as well as ε_i are assumed to be common knowledge. It is an individual parameter which captures the strength of player i's reciprocal preferences. The higher ρ_i, the more important is the reciprocity utility as compared to the utility arising from the material payoff. Note that if ρ_i is zero, i's utility is equal to his material payoff. If, in addition, ρ_j is also zero, the reciprocity game collapses into the standard game.

The product of the *kindness* and the *reciprocation term* measures the reciprocity utility in a particular node. If the kindness term in a particular node n is greater than zero, player i can ceteris paribus increase his utility if he chooses an action in that node which increases j's payoff. The opposite holds if the kindness term is negative. In this case, i has an incentive to reduce j's payoff. Since kindness is measured in each node where i has the move, the overall reciprocity utility is the sum of the reciprocity utility in all nodes (before the considered end node), weighted with the reciprocity parameter.

3.4. The reciprocity equilibrium

The introduced preferences form a psychological game (Geanakoplos et al., 1989). In psychological games, the utility of a player i not only depends on the selected strategies of the players but also on beliefs (compare Definition 3). Note, however, that beliefs are not part of the action space. Put differently, beliefs cannot be formed strategically, i.e., they are taken as given. Player i chooses the optimal strategy based on the given beliefs. The additional requirement in a psychological Nash equilibrium as compared to a standard Nash equilibrium is that all beliefs match actual behavior. This means, that an optimal strategy is only part of an equilibrium if the beliefs are also consistent with actual behavior.

Geanakoplos et al. (1989) show that the refinement concept of subgame perfectness can also be applied to psychological Nash equilibria. In our reciprocity game, we call a subgame perfect psychological Nash equilibrium a *reciprocity equilibrium*. If $\rho_i = \rho_j = 0$, the definition of a reciprocity equilibrium is equivalent to the definition of a subgame perfect Nash equilibrium.[10] Note that beliefs are not updated and only initial beliefs enter into the utility function in the concept of Geanakoplos et al. (1989). In sequential move games this creates conceptual problems and yields nonsensical predictions. This is why Rabin (1993) analyzes only normal form games. Applying psychological game theory to sequential games requires a solution to the updating problem. In Dufwenberg and Kirchsteiger (2004), players maximize their utility in each node using updated beliefs. In our model, only initial beliefs enter the utility. This allows the use of the equilibrium concept of Geanakoplos et al. (1989). Our model solves updating in the outcome and the reciprocation term by defining utility components in each node (which are summed up, see

[10] A remark on the existence of reciprocity equilibria: In the present form, a reciprocity equilibrium does not always exist because the function ϑ can be discontinuous. A minor technical modification of ϑ, however, guarantees the existence of a reciprocity equilibrium. For the ease of exposition we delegate the existence proof to Appendix A.

above). Beliefs about actions, which do not belong to the current subgame, are irrelevant for determining these utility components.

4. Applications

In this section we discuss the predictions of our theory in different experimental games. The games under study are the ultimatum game, the gift-exchange game, a reduced best-shot game, market games with proposer or responder competition, the dictator game, the sequential prisoner's dilemma, and the centipede game. Appendix 3 [11] contains the propositions that describe the reciprocity equilibria as well as the corresponding proofs.

4.1. Negative reciprocity: the ultimatum game

In the ultimatum game a first mover ("proposer") receives an amount of money (which we normalize to 1). He has to make an offer c to the second mover ("responder"), where $0 \leqslant c \leqslant 1$. The responder either accepts or rejects the offer. If she accepts, the resulting payoffs are $1 - c$ for the proposer and c for the responder. In case of a rejection, payoffs are zero for both parties. Given the standard assumptions, the outcome according to the subgame perfect Nash equilibrium is ($c = 0$; accept). The ultimatum game has been studied intensively. Overviews of experimental results are presented, e.g., in Güth et al. (1982), Thaler (1988), Güth (1995), Camerer and Thaler (1995), and Roth (1995). The reported behavioral regularities are quite robust and can be summarized as follows:

(i) practically no offers exceed 0.5,
(ii) the modal offers lie in a range between 0.4 and 0.5,
(iii) offers below 0.2 are extremely rare, and
(iv) whereas offers close to 0.5 are practically never rejected, the rejection rate for offers below 0.2 is rather high.

These stylized facts contrast strongly with the standard prediction.

We now state our predictions. Upon acceptance, material payoffs are $1 - c$ for the proposer and c for the responder, respectively. Let p denote the probability that the responder accepts the offer.

Proposition 1. *If ρ_1 and ρ_2 are positive there is a unique reciprocity equilibrium (c^*, p^*) in the ultimatum game as follows*:

$$p^* = \begin{cases} \min\left(1, \frac{c}{\rho_2 \cdot (1-2c)(1-c)}\right) & \text{if } c < \frac{1}{2}, \\ 1 & \text{if } c \geqslant \frac{1}{2}; \end{cases} \tag{7}$$

$$c^* = \max\left[\frac{1+3\rho_2 - \sqrt{1+6\rho_2+\rho_2^2}}{4\rho_2}, \frac{1}{2} \cdot \left(1 - \frac{1}{\rho_1}\right)\right]. \tag{8}$$

[11] This appendix is available as a pdf-document at: http://www.iew.unizh.ch/home/fischbacher/downloads/fafiA2-3.pdf.

If either ρ_1 or ρ_2 is zero p^* and c^* are the limits of the above formulas where ρ_1 and ρ_2 approach zero from above.

If ρ_1 and ρ_2 are both zero, $p^* = 1$ and $c^* = 0$.

Discussion. Equation (7) reveals the conditions that determine the responder's acceptance probability p^* of an offer c in the reciprocity equilibrium: If $c \geqslant 1/2$, the responder accepts the offer irrespective of her concern for reciprocity (compare the second row of Eq. (7)). If $c < 1/2$ the willingness to accept an offer increases with the size of the offer and decreases with the responder's concern for reciprocity, ρ_2 (see the first row of Eq. (7)).

Equation (8) shows the proposer's equilibrium choice of c, which depends on two expressions. While the first expression depends on the *responder's* reciprocal inclination, the second expression depends on the *proposer's* concern for reciprocity. The second expression represents the proposer's *intrinsic* concern for a fair outcome. If ρ_1 is large, he will offer a positive c. The first expression can be interpreted as an *extrinsic* constraint to offer a positive c: this expression corresponds to the offer that maximizes the proposer's material payoff, given the responder's rejection behavior. It is equal to the smallest possible offer that guarantees an acceptance probability of 1. (We call this offer c_0, i.e., $c_0 := (1 + 3\rho_2 - \sqrt{1 + 6\rho_2 + \rho_2^2})/(4\rho_2)$). The expression c_0 increases in ρ_2 and approaches $1/2$ as ρ_2 gets very large. The equilibrium offer c^* is the maximum of the first and the second expression of Eq. (8): This means, for example, that if a selfish proposer plays against a reciprocal responder, he will offer a higher share as ρ_2 increases. If, however, the responder has a very low ρ_2, i.e., he accepts practically any offer, the proposer's concern for an equitable outcome is decisive.[12]

Before we turn to the next game, we will briefly discuss the predictions of our theory for the non-intentional treatment reported by Blount (1995).[13] In her treatment, offers were randomly selected and therefore did not signal any intentions. She finds that the acceptance rate for a given offer is *much higher* than in the "regular" treatment. However, even in the absence of intentions, some subjects reject extremely disadvantageous offers. Our theory predicts exactly these two stylized facts. In the non-intentional treatment, the equilibrium acceptance rate for p^* is given by:

$$p^* = \begin{cases} \min\left(1, \frac{c}{\varepsilon_2 \rho_2 \cdot (1-2c)(1-c)}\right) & \text{if } c < \frac{1}{2}, \\ 1 & \text{if } c \geqslant \frac{1}{2}. \end{cases}$$

Figure 1 depicts the predicted acceptance probabilities in the "regular" ultimatum game (lower graph) and in Blount's treatment (upper graph) for a given ρ_2. As the figure shows, a responder's acceptance probability for low offers is higher if intention is absent. The lower the outcome concern parameter ε_2, the more the upper graph shifts to the left. On the other hand, if a responder is purely outcome oriented, i.e., if $\varepsilon_2 = 1$ holds, she exhibits the *same* behavioral pattern as in the "regular" treatment. As Blount's data reveal, however, many people care for both outcome and intention (i.e., $0 < \varepsilon_i < 1$).

[12] The range of ρ_1- and ρ_2-combinations where the equilibrium offer c^* equals c_0 is given by $\rho_2 \geqslant \rho_1(\rho_1 - 1)/(\rho_1 + 1)$. This holds in particular if $\rho_2 \geqslant \rho_1$.

[13] For similar results and a discussion on the role of intention, see Falk et al. (2000).

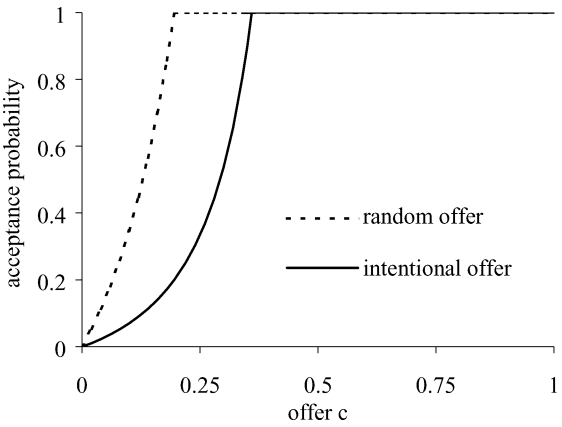

Fig. 1. Acceptance probabilities in the ultimatum game with intensions (lower graph) and without intention (upper graph) depending on the offer, for parameters $\rho_2 = 2$ and $\varepsilon_2 = 0.2$.

4.2. Positive reciprocity: the gift-exchange game

The gift-exchange game is a two-person sequential game. The first mover (called an employer) offers a wage w to a second mover (called a worker). After receiving the wage, the worker makes an effort decision e. Providing effort is costly with $c(e) = \alpha e^2$. Payoff functions are given by $\pi_1 = ve - w$ for employers and $\pi_2 = w - c(e)$ for workers, respectively. Contrary to the standard prediction, the main experimental findings are that

(i) wages and efforts exceed the lowest possible wage, and
(ii) there is a positive wage-effort relation (Fehr et al., 1993; Fehr and Falk, 1999; Gächter and Falk, 2002).

Our model predicts the main stylized facts. In equilibrium a worker's effort choice equals

$$e^* = \begin{cases} 0 & \text{if } \rho_2 = 0, \\ \min\left(1, \frac{-2\alpha - \rho_2 + \sqrt{(2\alpha+\rho_2)^2 + 8\alpha\rho_2^2 w}}{2\alpha\rho_2}\right) & \text{if } \rho_2 > 0. \end{cases}$$

The higher the wage, the higher is the kindness term φ_1. As a result, sufficiently reciprocal workers respond to higher wages with higher efforts. Moreover, for a given wage, efforts increase in workers' reciprocal inclination. In equilibrium, firms pay wages strictly above zero.

Charness (2004) has conducted a non-intentional treatment of the gift-exchange game where a third party or a random mechanism determines the wage. Compared to the "regular" treatment, this leads to a weaker correlation between wages and efforts. Our model explains this pattern: Since the kindness term captures both the concern for outcome *and* intention, the kindness term and therefore the reciprocal reaction is smaller in the random treatment.

4.3. Further games

In this section we briefly discuss the intuition of our theory's predictions in the best-shot game, market games, the sequential public goods game, and the centipede game, which illustrates that the model is applicable for multi-stage games.

4.3.1. Identical outcomes yield different responses: a comparison between best-shot and ultimatum games

Harrison and Hirshleifer (1989), and Prasnikar and Roth (1992) introduced the best-shot game. In this game second movers are willing to accept a higher degree of inequity than in ultimatum games. This finding cannot be reconciled with the inequity aversion models by Bolton and Ockenfels (2000), and Fehr and Schmidt (1999). The difference between the best-shot and the ultimatum game can be explained in terms of intentions, however. We show this with the help of two stylized games, a reduced ultimatum and a reduced best-shot game (Falk et al., 2003). In the reduced best-shot game, a first mover can offer two different payoff distributions (π_1, π_2), namely $(2, 8)$ and $(8, 2)$; the second mover can accept or reject the chosen offer. If the second mover accepts, the offered payoff distribution is implemented. Otherwise, both players receive nothing. The crucial feature of this game (and the richer original best-shot game) is that the first mover can only offer a payoff share that is either very advantageous or very disadvantageous to himself ($8, 2$ or $2, 8$). This distinguishes the game from the reduced ultimatum game, where the first mover can either offer $(8, 2)$ or $(5, 5)$. The best-shot game nicely illustrates the asymmetry in the kindness perception of advantageous and disadvantageous situations (see also our fifth and sixth observation in Section 2). If the choice set is either $(8, 2)$ or $(2, 8)$, offering $(2, 8)$ is perceived as fully kind since there was a choice to be less kind. However, the $(8, 2)$ offer is not perceived as fully unkind since the possible alternative is not reasonable.

According to our theory,[14] the predicted acceptance probability q^* for the unkind offer $(8, 2)$ in the reduced best-shot game is given by $q^* = \min(1, 5/(12\rho_2))$ while it is $q^* = \min(1, 5/(12\varepsilon_2\rho_2))$ in the reduced ultimatum game. In both games, the acceptance probability for the unkind offer decreases in ρ_2. However, for a given ρ_2 the acceptance probability is lower in the ultimatum game compared to the best-shot game. Thus, in the best-shot game a reciprocal second mover is willing to accept a higher degree of inequity. The intuition is that there is no reasonable alternative to $(8, 2)$ in the best-shot game, while there is one in the ultimatum game $(5, 5)$. Thus, depending on the first mover's alternatives, the *same* offer signals different intentions and will therefore be accepted with a different probability. This prediction is confirmed in the experimental study by Falk et al. (2003). The rejection rate of the $(8, 2)$-offer is 27 percent in the reduced best-shot game and 44 percent in the reduced ultimatum game.

4.3.2. Competition

In the preceding games, we only analyzed bilateral interactions. In particular, we restricted our analysis to games without competition. Therefore, we apply our theory to

[14] This prediction is based on the formulation of the intention factor presented in Appendix A. All other predictions presented in this paper are the same, regardless of whether we use the more simple intention factor (main text) or the more complex one (Appendix A).

games with more than 2 players in this section and show how competitive pressure interacts with reciprocal preferences.[15]

It is a well established fact in experimental economics that outcomes converge very well towards the outcome predicted by standard economic theory in competitive markets (Smith, 1982; Davis and Holt, 1993). This holds even in markets where the equilibrium outcome is very "unfair" in the sense that one side of the market reaps almost the whole surplus. In this section, we show that our theory is consistent with why outcomes tend to be "fair" in bilateral institutions while reciprocal subjects' behavior in markets gives rise to equilibria that imply an extremely uneven distribution of the gains from trade.

As an illustration, we analyze a market game with proposer competition which has been studied by Roth et al. (1991). In this game, there are $n-1$ proposers who simultaneously propose an offer $c_i \in [0,1]$ to the responder with $i \in [1,\ldots,n-1]$. These offers are revealed to the responder who has to decide whether to accept or reject the highest offer $c_{\max}$. If more than one proposer offers $c_{\max}$, a random mechanism determines whose offer will be selected. Payoffs are analogous to the ultimatum game, i.e., the proposer whose offer is accepted receives $1-c_{\max}$ and the responder gets $c_{\max}$. A proposer whose offer is not accepted receives a payoff of zero. If the responder rejects $c_{\max}$, no one receives anything. Given standard assumptions, the subgame perfect outcome is that at least two proposers offer $c_{\max}=1$, which the responder accepts. This prediction is largely supported by the experimental data.

Our theory coincides with the standard prediction. In a reciprocity equilibrium in the market game with proposer competition, at least 2 proposers offer $c_{\max}=1$, which the responder accepts. The striking feature of this prediction is that reciprocal proposers will accept a very uneven distribution of the pie. The intuition is that in a competitive market a proposer has no chance to achieve a "fair" outcome: Assume that a reciprocal proposer i refuses to offer more than 0.5 in a market game with two proposers. By infinitesimally overbidding player i's offer, the other proposer can increase his material payoff by a positive amount (because he can increase the winning probability from 0.5 to 1). His reciprocity disutility resulting from the unfair relation to the responder only changes infinitesimally. This means that player i's refusal to propose more than 0.5 is not an effective tool for achieving a "fair" outcome. As a consequence, he tries overbidding the other proposer to get at least a minimal share of the pie. This mutual "overbidding" inevitably leads to the equilibrium prediction.

4.3.3. The dictator game

In the dictator game the first mover (the so-called "dictator") divides an amount of money between himself and a counterpart (the "receiver"). Let 1 be the amount of money and c the share for the receiver and $0 \leqslant c \leqslant 1$. Payoffs are c for the receiver and $1-c$ for the dictator. The dictator game has been studied, e.g., by Forsythe et al. (1994), Hoffman et al. (1996), and Eckel and Grossman (1998). The stylized facts can be summarized as follows.

[15] This requires a twofold extension of our theory: first to games with more than 2 players and second to games with almost perfect information. These extensions are discussed in Appendix 2 (see footnote 5).

(i) Offers larger than half of the pie, i.e., $c > 1/2$ are practically never observed.
(ii) Roughly 80 percent of the offers are between zero and half of the pie, i.e., $0 < c \leqslant 1/2$. However, *compared to the ultimatum game, the distribution of offers shifts towards zero*.
(iii) About 20 percent of the offers are zero.[16]

Our theory predicts a unique reciprocity equilibrium where the dictator offers $c^* = \max\left[0, \frac{1}{2} \cdot \left(1 - \frac{1}{\varepsilon_1 \rho_1}\right)\right]$. If $\varepsilon_1 \rho_1 > 1$, the dictator offers a positive amount of money but even for very high values of $\varepsilon_1 \rho_1$ his offer will never exceed $1/2$. If $\varepsilon_1 \rho_1 \leqslant 1$, the dictator chooses $c = 0$. Comparing the equilibrium offers in the dictator and the ultimatum game note that the equation that determines the offer in the dictator game equals the second expression in Eq. (8) of Proposition 1—if we replace ρ_1 with $\varepsilon_1 \rho_1$. Since $\varepsilon_1 \leqslant 1$, the same person will offer at least as much in the ultimatum as in the dictator game. Thus, consistent with stylized facts (i) to (iii), our theory predicts that dictators offer between zero and half of the pie and that the distribution of offers in the dictator game shifts downwards compared to the corresponding distribution in the ultimatum game.

4.3.4. The sequential prisoner's dilemma and public goods games

In a sequential prisoner's dilemma, player 1 can either cooperate or defect. After observing player 1's choice, player 2 has the same choice. The standard subgame perfect solution is that both players defect. Contrary to this prediction, our theory predicts the following: first, if player 2 is sufficiently reciprocally motivated, there is a positive probability that player 2 rewards player 1's cooperation with cooperation. Second, player 2 always defects if player 1 defected beforehand. Experimental studies of sequential versions of the prisoner's dilemma are reported in Bolle and Ockenfels (1990), and Clark and Sefton (2001). The results of their studies are in line with our predictions. In particular, unconditional cooperation is practically inexistent.

The strategic structure of the prisoner's dilemma is very similar to that of *public goods games*. Most public goods experiments have been conducted as simultaneous move games. In Fischbacher et al. (2001), however, subjects could *conditionally* indicate how many tokens they wanted to contribute to the public good. Despite the fact that the best reply is to provide zero tokens irrespective of the other group members' contributions, subjects contributed more if the contributions of the other group members were higher.[17] This "conditional cooperation strategy" was in most cases specified in such a way that subjects provided less than the group average. This is exactly what our theory predicts. Moreover, our theory replicates the stylized fact that the propensity to cooperate increases in the marginal per capita return of an investment in the public good (see Ledyard, 1995).

[16] It should be noted that the results of the dictator game are not very robust with respect to treatment variations. For example, increasing the social distance among participants of an experiment and the experimenter (double blind treatment) increases the percentage of zero proposals (compare Hoffman et al., 1996).
[17] On conditional cooperation see also Keser and van Winden (2000).

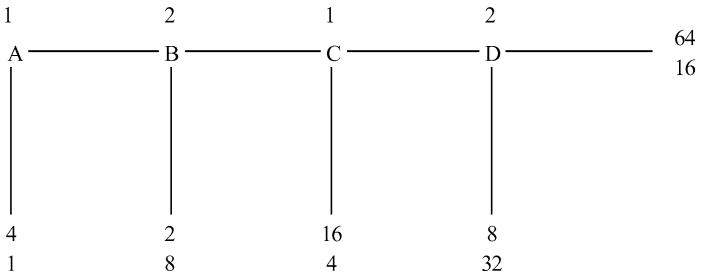

Fig. 2. Game tree of a centipede game.

4.3.5. Multi-move games: the centipede game

We conclude this section with the discussion of a game where players move more than once. The analysis demonstrates that our model is applicable not only to games where players have just one move. As an example, we choose the centipede game. The four move-centipede studied by McKelvey and Palfrey (1992) is shown in Fig. 2. Each player can at each node either "take" 80 percent of a growing pie or "pass." Passing enlarges the pie substantially but players also face the temptation to take the pie. With backward induction selfish players therefore always decide to "take." Contrary to this prediction, McKelvey and Palfrey (1992) report that on average 93 percent of the players pass in A, 62 percent pass in B, 35 percent pass in C, and 25 percent pass in D. These results are remarkable and not only challenge the selfishness hypothesis but also explanations based on inequity aversion. An equity averse player 2 should not decide to pass at D because passing implies a lower material payoff, an increase in inequity, and a switch from an advantageous to a disadvantageous inequity. Reciprocally motivated subjects on the other hand may decide to pass even in the final node since they reward the kindness expressed by their opponent's previous passing decisions. This allows for equilibria where player 1 starts with passing in A. In particular we can show that if $\rho_2 > 1/57$ or $\rho_1 > 1/24$,[18] there is a reciprocity equilibrium where the first mover passes with a strictly positive probability. For example, if player 2 is sufficiently reciprocal, she will pass in D with a high enough probability to make it materially worthwhile for player 1 to pass in C, etc. The same argument holds if player 1 is sufficiently reciprocal.

5. Concluding remarks

In this paper, we have presented a formal theory of reciprocity. According to the theory, people reward kind and punish unkind actions. Kindness comprises both the consequences as well as the intention of an action. Our theory captures the empirical finding that the same consequences of an action are perceived and reciprocated differently, depending on the underlying intention. The theory is also capable of reconciling the puzzling evidence

[18] To make these predictions comparable to the other predictions in the paper, we have to normalize the game (dividing all numbers in the game tree by 64). The thresholds in the normalized game are $\rho_2 > 64/57$ or $\rho_1 > 64/24 = 8/3$.

that very unfair outcomes emerge in competitive experimental markets, while outcomes tend to be fair in bilateral bargaining situations.

Our concept of reciprocity differs both from the inequity aversion models of fairness as well as from other reciprocity approaches. In Bolton and Ockenfels (2000), and Fehr and Schmidt (1999), players reciprocate to reduce inequity. Their concept differs from ours in at least two dimensions. First, they model fairness in a consequentialistic way, i.e., they assume that fairness intentions are irrelevant. Distributive consequences of an action alone trigger reciprocal actions. Second, players reward or sanction only if this reduces inequity. This differs from our model, where people reciprocate perceived kindness or unkindness, i.e., our model predicts rewarding and sanctioning even if this does not reduce inequity. Which of the two approaches better predicts behavior is an empirical question: In Falk et al. (2001), this question is addressed in detail. One of their experiments is a three person prisoner's dilemma with a subsequent sanctioning stage. The cost of sanctioning is such that sanctions do not reduce the inequity between a cooperator and a defector. As a consequence, inequity aversion models predict no sanctions in this situation, while our model does. The results of the experiment support the latter prediction: 46.8 percent of the cooperators punish defectors. This result suggests that reciprocal behavior is mainly driven as a response to kindness—not as a desire to reduce inequity.

Our model also differs from the reciprocity approaches by Rabin (1993), and Dufwenberg and Kirchsteiger (2004). First, these models assume that reciprocity is exclusively intention driven, while outcomes are in our model decisive as well. Second, we assume that payoffs are level-comparable and that the concept of kindness is based on interpersonal comparisons. To derive the kindness of a move, players compare their payoff with that of the other players. This feature distinguishes our model from Rabin (1993), and Dufwenberg and Kirchsteiger (2004) because fairness evaluations are not based upon interpersonal comparisons in their models. In their view, people do not consider whether they have more or less than their opponent(s). Rather it is assumed that they compare the actually chosen outcome with the alternative actions their opponent(s) could have chosen. According to this concept player j's move is considered as unfair by player i not because it leaves player i with less than j gets, but because j could have offered player i a higher payoff.

Acknowledgments

We thank Martin Brown, Ernst Fehr, Simon Gächter, Herb Gintis, Lorenz Götte, Michael Kosfeld, Matthew Rabin, Armin Schmutzler, the associate editor, and our referees for helpful comments and Sally Gschwend and Mirja Koenigsberg for reviewing the manuscript.

Appendix A

A.1. An extension of the intention term

In Section 2, we presented six empirical observations from a questionnaire study on the perception of kindness. In Section 3, we used only observations one to four for modeling

the intention factor. In particular, we did not take the asymmetry of kindness perceptions into account as established in our observations five and six. In this appendix, we present a version of the intention term, which takes all six observations into account. As in Section 3, the intention term equals 1 if there is a true alternative. In this appendix, we define more carefully what is perceived as a true alternative. The function Ω evaluates how "true" an alternative is, i.e., how intentional player j's choice of a given payoff combination (π_i^0, π_j^0) is perceived, given one of j's alternatives $(\tilde{\pi}_i, \tilde{\pi}_j)$. If the choice is fully intentional, Ω equals 1. If the choice is not considered to be fully intentional, Ω is smaller than 1. The function $\Omega : \mathbb{R}^4 \longrightarrow [0,1]$ is defined as follows:

$$\Omega\left(\tilde{\pi}_i, \tilde{\pi}_j, \pi_i^0, \pi_j^0\right) := \begin{cases} 1 & \text{if } \pi_i^0 \geqslant \pi_j^0 \text{ and } \tilde{\pi}_i < \pi_i^0, \\ \varepsilon_i & \text{if } \pi_i^0 \geqslant \pi_j^0 \text{ and } \tilde{\pi}_i \geqslant \pi_i^0, \\ 1 & \text{if } \pi_i^0 < \pi_j^0,\ \tilde{\pi}_i > \pi_i^0 \text{ and } \tilde{\pi}_i \leqslant \tilde{\pi}_j, \\ \max\left(1 - \frac{\tilde{\pi}_i - \tilde{\pi}_j}{\pi_j^0 - \pi_i^0}, \varepsilon_i\right) & \text{if } \pi_i^0 < \pi_j^0,\ \tilde{\pi}_i > \pi_i^0 \text{ and } \tilde{\pi}_i > \tilde{\pi}_j, \\ \varepsilon_i & \text{if } \pi_i^0 < \pi_j^0 \text{ and } \tilde{\pi}_i \leqslant \pi_i^0. \end{cases} \tag{9}$$

The first two rows capture situations where j treated i in a kind way ($\pi_i^0 \geqslant \pi_j^0$). In these situations, the value of Ω depends on whether j could have reduced i's payoff ($\tilde{\pi}_i$ compared to π_i^0) or not. This case is equal to the model in Section 3.

The other three rows represent instances where j puts i in a disadvantageous situation, i.e., where $\pi_i^0 < \pi_j^0$ holds. If j *has* the alternative of improving i's payoff without putting herself in a disadvantageous situation ($\tilde{\pi}_i > \pi_i^0$ and $\tilde{\pi}_i \leqslant \tilde{\pi}_j$), her unkindness is fully intentional. Therefore, Ω is equal to 1 (see our fifth observation). Now suppose that there *is* an alternative of improving i's payoff, but this alternative leads to a disadvantageous situation for j. The more this alternative is disadvantageous for player j, the less reasonable it is considered. As a consequence, the choice of π_i^0 is not considered to be fully intentionally unkind and Ω is equal to $\max(1 - (\tilde{\pi}_i - \tilde{\pi}_j)/(\pi_j^0 - \pi_i^0), \varepsilon_i) \leqslant 1$. The expression $1 - (\tilde{\pi}_i - \tilde{\pi}_j)/(\pi_j^0 - \pi_i^0)$ measures "how much j must put herself into a disadvantageous situation" if she wants to improve i's payoff—related to the reference situation (π_i^0, π_j^0).[19] Finally, if j's only alternative is to choose an even lower payoff for player i, i.e., $\tilde{\pi}_i \leqslant \pi_i^0$, i cannot infer that j wanted to treat him unkindly. Consequently, the action was unintentionally "unkind" yielding $\Omega = \varepsilon_i$ (see our fourth observation).

We use the payoffs that determine the outcome term $\Delta_j(n)$, namely $\pi_i(n, s_i'', s_i')$ and $\pi_j(n, s_i'', s_i')$ as the reference distribution (π_i^0, π_j^0) and define the *intention factor* as:

$$\vartheta_j\left(n, s_i'', s_i'\right) = \max\left\{\Omega\left(\tilde{\pi}_i, \tilde{\pi}_j, \pi_i\left(n, s_i'', s_i'\right), \pi_j\left(n, s_i'', s_i'\right)\right) \mid \left(\tilde{\pi}_i, \tilde{\pi}_j\right) \in \Pi_i\left(n, s_i''\right)\right\}. \tag{10}$$

The maximum-operator guarantees that a particular action is considered to be intentional if there is *any* "true" alternative.

[19] If, e.g., j must disadvantage herself only a little (the numerator is small), the alternative action will ceteris paribus be considered to be rather reasonable. If, however, the numerator is large (in particular, if the numerator is larger than the denominator) Ω is equal to ε_i.

Of course, even this extended definition of the intention term is only an approximation which does not completely capture the full richness of the relationship between alternatives and intentions. For instance, as a referee pointed out to us, the efficiency of alternatives could also play a role for the intentionality of an action.

A.2. Existence of reciprocity equilibria

The existence of a reciprocity equilibrium is not always guaranteed in the presented form of our theory. A game where a reciprocity equilibrium does not exist is shown in Proposition 10 in the Appendix 3.[20]

The reason why the existence of an equilibrium is not guaranteed has to do with the discontinuity of the function Ω. To show this, we define for a (small) positive number λ a continuous approximation Ω^λ for Ω.[21] We set

$$\Omega^\lambda\big(\tilde\pi_i,\tilde\pi_j,\pi_i^0,\pi_j^0\big) := \begin{cases} \min\big(1,\varepsilon_i+\frac{1}{\lambda}(\pi_i^0-\tilde\pi_i)\big) & \text{if } \pi_i^0\geqslant\pi_j^0 \text{ and } \tilde\pi_i<\pi_i^0,\\ \varepsilon_i & \text{if } \pi_i^0\geqslant\pi_j^0 \text{ and } \tilde\pi_i\geqslant\pi_i^0,\\ \min\big(1,\varepsilon_i+\frac{1}{\lambda}(\tilde\pi_i-\pi_i^0)\big) & \text{if } \pi_i^0<\pi_j^0,\ \tilde\pi_i>\pi_i^0 \text{ and } \tilde\pi_i\leqslant\tilde\pi_j,\\ \max\big(\varepsilon_i,\min\big(1-(\tilde\pi_i-\tilde\pi_j)/(\pi_j^0-\pi_i^0), & \text{if } \pi_i^0<\pi_j^0,\ \tilde\pi_i>\pi_i^0 \text{ and } \tilde\pi_i>\tilde\pi_j,\\ \quad\varepsilon_i+\frac{1}{\lambda}(\tilde\pi_i-\pi_i^0)\big)\big) & \\ \varepsilon_i & \text{if } \pi_i^0<\pi_j^0 \text{ and } \tilde\pi_i\leqslant\pi_i^0. \end{cases}$$

Given the continuous variant Ω^λ of Ω, we define a modified kindness function φ^λ and a utility function U^λ. We call a λ-reciprocity equilibrium a subgame perfect equilibrium of the psychological game with utility U^λ. This modification now guarantees the existence of an equilibrium as the following theorem shows:

Theorem 2 (Existence Theorem). *Let Γ be a finite two person extensive form game with complete information. Let $\lambda>0$. Then Γ has a λ-reciprocity equilibrium.*

Proof of the Existence Theorem. Let $n\in N_i$ be a node of the game. Then $S^n=\{(p_a)_{a\in A_n}\mid p_a\in[0,1],\sum_a p_a=1\}$ is the set of mixed strategies in this node. Let $S=\prod_{n\in N}S^n$ be the set of behavior strategy combinations. It includes the strategies of both players. Let $S^{-n}=\prod_{m\neq n}S^m$ be the strategies at all other nodes. Let $s=(s^n,s^{-n})$ be a behavior strategy combination with $s^n\in S^n$ and $s^{-n}\in S^{-n}$. Let s' and s'' be the beliefs of first and second order. We define $V(n,(s^n,s^{-n}),s',s'')$ as the utility U_i^λ conditional on node n, i.e. it is the expected utility of the player who can move in node n given this player knows he is in node n. We now define the best reply correspondence

$$B:S\rightrightarrows S:s\mapsto B(s)\subset S.$$

[20] This appendix is available as a pdf-document at: http://www.iew.unizh.ch/home/fischbacher/downloads/fafiA2-3.pdf.

[21] We have $\Omega(\tilde\pi_i,\tilde\pi_j,\pi_i^0,\pi_j^0)=\lim_{\lambda\to0}\Omega^\lambda(\tilde\pi_i,\tilde\pi_j,\pi_i^0,\pi_j^0)$ for any choice of $\tilde\pi_i,\tilde\pi_j,\pi_i^0,\pi_j^0$.

The component $B^n : S \rightrightarrows S^n$ is defined as

$$B^n(s) := \arg\max_{\tau^n \in S^n} V\left(n, \left(\tau^n, s^{-n}\right), s, s\right).$$

We get

$$B(s) := \left\{\left(b^n\right)_{n\in N} \mid b^n \in B^n(s)\right\}.$$

This definition is the best reply correspondence of the agent-strategic form (see Fudenberg and Tirole, 1991): The player behaves as if there was an agent in every decision node. A behavior strategy that optimizes in the agent-strategic form corresponds to a subgame perfect Nash equilibrium.

As S is the product of convex and compact sets, it is convex and compact. Since S is compact and V is continuous in s^n, $B(s)$ is not empty. Because U^λ is linear in the strategies, $B(s)$ is convex.

We now show that B is upper hemi-continuous: First, we show that V depends continuously on the strategies and beliefs: $\Delta_j(n, s_i'', s_i')$ is continuous and $\Delta_j(n, s_i'', s_i') = 0$ holds for $\pi_i(n, s_i'', s_i') = \pi_j(n, s_i'', s_i')$. Because Ω^λ is bounded (by 1), we get the desired continuity of $\vartheta_j^\lambda(n, s_i'', s_i')\Delta_j(n, s_i'', s_i')$ at $\pi_i(n, s_i'', s_i') = \pi_j(n, s_i'', s_i')$. By construction of Ω^λ, $\vartheta_j^\lambda(n, s_i'', s_i')$ is continuous in strategies and beliefs if $\pi_i(n, s_i'', s_i') \neq \pi_j(n, s_i'', s_i')$. Therefore, this also holds for $\vartheta_j^\lambda(n, s_i'', s_i')\Delta_j(n, s_i'', s_i')$. Hence, $\varphi_j^\lambda(n, s_i'', s_i')$ is continuous. The reciprocation term and the material profit are obviously continuous and therefore V is continuous. If any function $f : X \times Y \to \mathbb{R}$ is continuous, then the best reply correspondence $R : X \rightrightarrows Y : x \mapsto \arg\max_y f(x, y)$ is upper hemi-continuous. Hence, because V is continuous, the best reply correspondence B is upper hemi-continuous.

Therefore, we can apply the fixed point theorem of Kakutani and get an $s^* \in S$ with $s^* \in B(s^*)$. This strategy s^* with first order belief s^* and second order belief s^* now forms a reciprocity equilibrium: The triple (s^*, s^*, s^*) trivially satisfies the consistency of the beliefs. By construction of B each player optimizes his utility in each node—given the beliefs and given the strategies of the other players. This is exactly the definition of the subgame perfect psychological equilibrium. □

References

Adams, J.S., 1965. Inequity in social exchange. In: Berkowitz, L. (Ed.), In: Advances in Experimental Psychology, vol. 2. Academic Press, New York, pp. 267–299.

Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. Games Econ. Behav. 10, 122–142.

Blount, S., 1995. When social outcomes aren't fair: The effect of causal attributions on preferences. Organ. Behav. Human. Dec. Proc. 63, 131–144.

Bolle, F., Ockenfels, P., 1990. Prisoner's dilemma as a game with incomplete information. J. Econ. Psych. 11, 69–84.

Bolton, G., Ockenfels, A., 2000. ERC—A theory of equity, reciprocity and competition. Amer. Econ. Rev. 90, 166–193.

Bolton, G.E., Brandts, J., Ockenfels, A., 1998. Measuring motivations for the reciprocal responses observed in a simple dilemma game. Exper. Econ. 1, 207–220.

Brandts, J., Sola, C., 2001. Reference points and negative reciprocity in simple sequential games. Games Econ. Behav. 36, 138–157.

Camerer, C., Thaler, R., 1995. Ultimatums, dictators, and manners. J. Econ. Perspect. 9, 209–219.
Carpenter, J.P., Matthews, P.H., 2003. Social reciprocity. Working paper series No. 0229. Middlebury College.
Charness, G., 2004. Attribution and reciprocity in a simulated labor market: An experimental investigation. J. Lab. Econ. 22 (3), 665–688.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quart. J. Econ. 117, 817–869.
Clark, K., Sefton, M., 2001. The sequential prisoner's dilemma: Evidence on reciprocation. Econ. J. 111, 51–68.
Cox, J., 2003, Trust and reciprocity: Implications of game triads and social contexts. Working paper. University of Arizona.
Davis, D., Holt, C., 1993. Experimental Economics. Princeton Univ. Press, Princeton.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav. 47, 268–298.
Eckel, C., Grossman, P., 1998. Are women less selfish than men? Evidence from dictator experiments. Econ. J. 108, 726–735.
Falk, A., Fehr, E., Fischbacher, U., 2000. Testing theories of fairness—Intentions matter. Working paper No. 63. Institute for Empirical Research in Economics, University of Zurich.
Falk, A., Fehr, E., Fischbacher, U., 2001. Driving forces of informal sanctions. Working paper No. 59. Institute for Empirical Research in Economics, University of Zurich.
Falk, A., Fehr, E., Fischbacher, U., 2003. On the nature of fair behavior. Econ. Inquiry 41 (1), 20–26.
Fehr, E., Falk, A., 1999. Wage rigidities in a competitive, incomplete contract market. J. Polit. Economy 107, 106–134.
Fehr, E., Gächter, S., 2000. Fairness and retaliation: The economics of reciprocity. J. Econ. Perspect. 14, 159–181.
Fehr, E., Kirchsteiger, G., Riedl, A., 1993. Does fairness prevent market clearing? An experimental investigation. Quart. J. Econ. 108, 437–460.
Fehr, E., Schmidt, K., 1999. A theory of fairness, competition, and cooperation. Quart. J. Econ. 114, 817–868.
Fischbacher, U., Gächter, S., Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. Econ. Letters 71, 397–404.
Forsythe, R., Horowitz, J., Savin, N., Sefton, M., 1994. Fairness in simple bargaining experiments. Games Econ. Behav. 6, 347–369.
Fudenberg, D., Tirole, J., 1991. Game Theory. MIT Press, MA.
Gächter, S., Falk, A., 2002. Reputation or reciprocity? Scand. J. Econ. 104, 1–26.
Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological games and sequential rationality. Games Econ. Behav. 1, 60–79.
Goranson, R.E., Berkowitz, L., 1966. Reciprocity and responsibility reactions to prior help. J. Personality Soc. Psych. 3, 227–232.
Greenberg, M.S., Frisch, D.M., 1972. Effect of intentionality on willingness to reciprocate a favor. J. Exp. Soc. Psych. 8, 99–111.
Güth, W., 1995. On ultimatum bargaining experiments—A personal review. J. Econ. Behav. Organ. 27, 329–344.
Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. J. Econ. Behav. Organ. 3, 367–388.
Harrison, G.W., Hirshleifer, J., 1989. An experimental evaluation of weakest link/best shot models of public goods. J. Polit. Economy 97, 201–225.
Hoffman, E., McCabe, K., Smith, V.L., 1996. Social distance and other-regarding behavior in dictator games. Amer. Econ. Rev. 86, 653–660.
Kahneman, D., Knetsch, J., Thaler, R., 1986. Fairness as a constraint on profit-seeking: entitlements in the market. Amer. Econ. Rev. 76 (4), 728–741.
Keser, C., van Winden, F., 2000. Conditional cooperation and voluntary contributions to public goods. Scand. J. Econ. 102, 23–39.
Kreps, D., Milgrom, P., Roberts, J., Wilson, R., 1982. Rational cooperation in the finitely repeated prisoner's dilemma. J. Econ. Theory 27, 245–252.
Ledyard, J., 1995. Public goods: A survey of experimental research. In: Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics. Princeton Univ. Press, Princeton, pp. 111–252.
Levine, D., 1998. Modeling altruism and spitefulness in experiments. Rev. Econ. Dynam. 1, 593–622.
Loewenstein, G.F., Thompson, L., Bazerman, M.H., 1989. Social utility and decision making in interpersonal contexts. J. Personality Soc. Psych. 57, 426–441.

McCabe, K., Rigdon, M., Smith, V., 2003. Positive reciprocity and intentions in trust games. J. Econ. Behav. Organ. 52, 267–275.
McKelvey, R., Palfrey, T., 1992. An experimental study of the centipede game. Econometrica 60, 803–836.
Offerman, T., 2002. Hurting hurts more that helping helps: The role of the self-serving bias. Europ. Econ. Rev. 46, 1423–1437.
Prasnikar, V., Roth, A.E., 1992. Considerations of fairness and strategy: Experimental data from sequential games. Quart. J. Econ. 107, 865–888.
Rabin, M., 1993. Incorporating fairness into game theory and economics. Amer. Econ. Rev. 83, 1281–1302.
Roth, A., 1995. Bargaining experiments. In: Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics. Princeton Univ. Press, Princeton, pp. 253–348.
Roth, A., Prasnikar, V., Okuno-Fujiwara, M., Zamir, S., 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. Amer. Econ. Rev. 81, 1068–1095.
Smith, A., 1976. The Theory of Moral Sentiments (D.D. Raphael, A.C. Macfie, Eds.). Clarendon Press, Oxford (Original work published 1759).
Smith, V.L., 1982. Microeconomic systems as an experimental science. Amer. Econ. Rev. 72, 923–955.
Thaler, R.H., 1988. Anomalies: The ultimatum game. J. Econ. Perspect. 2, 195–206.
Trivers, R., 1971. The evolution of reciprocal altruism. Quart. Rev. Biol. 46, 35–57.
Walster, E., Walster, G.W., 1978. Equity—Theory and Research. Allyn and Bacon, Boston.