Chapter 2
Game Theory and Fairness Preferences

During the last three decades a lot of attention was given to experimental investigations of the ultimatum game.[1] Contrary to the theoretical "standard" prediction based on maximization of the monetary payoff (responders accepting the smallest possible offer and proposers offering the minimum possible offer), experiments with ultimatum games show that players are typically not simply maximizing their monetary payoff. Instead, responders frequently *reject* offers they perceive as *unfair* and proposers anticipate this by offering a substantial share, usually with modal and median offers between 40 and 50 percent. A good overview of the various experiments done with ultimatum games is given by Camerer (2003, chap. 3, tables 2–5). The following section 2.1 very briefly summarizes further experimental evidence that subjects are not always maximizing material payoffs.

How can this behavior be in line with classical game theory? The theory of games as defined by von Neumann and Morgenstern is based on *preferences*, and these preferences are described by a utility function. The *payoffs* that classical game theory is based on are *utility levels*. However, as Weibull (2004, p. 6) states:[2]

> The formal machinery of non-cooperative game theory does not require that a player's payoff value $u_i(\omega)$ at an end node ω be a function of the material consequences at that node. Indeed, two plays resulting in identical material payoffs to all players may well differ in terms of information sets reached, choices made or not made during play and so on—aspects that may be relevant for players' preferences and hence influence their Bernoulli functions. Standard game theory only requires the *existence* of a Bernoulli function u_i for each (personal) player i. Indeed, several laboratory experiments have convincingly—though perhaps not surprisingly for the non-economist—shown that human subjects' preferences are not driven only by their own monetary payoffs.

Therefore it is *not contradictory* to classical game theory to define a player's utility function not only on his own monetary payoff, but, for example, on his own payoff *and* the monetary payoff of others. This way, it is possible to explain and predict behavior by using classical game theory in conjunction with a new class of

---

[1] See section 1.2 for a short description of the ultimatum game.

[2] In the original text a player's payoff value was denoted by $\pi_i(\omega)$. To keep consistency with the rest of this book this was replaced by $u_i(\omega)$.

social preferences and *social utility functions*. The area of research concerned with social preferences is called *social utility* and is part of behavioral game theory.

Some economists argue that altering the utility function makes it possible to explain anything. For example, Güth (1995, p. 342) remarked:

> Very often this type of research resembles, however, a neoclassical repair shop in the sense that one first observes behaviour for a certain environment and then defines a suitable optimisation or game model which can account for what has been observed.

Comments like this show how behavioral game theory can be misunderstood. The goal is not to explain all kinds of experimental results but to find *adjusted utility functions* on the basis of *psychological analysis*, which help to *explain* and *predict* human behavior.

Based on the idea that a preference for fairness can explain rejections in experimental ultimatum games, two classes of fairness-models evolved: *outcome-based* and *intention-based* models. Outcome-based models treat the intentions that players attribute to one another as unnecessary for predicting behavior and focus solely on the *monetary payoffs*. Section 2.2 summarizes two prominent contributions in this domain: the equity-based approaches by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). In contrast, intention-based approaches, in particular the reciprocity and trust hypothesis, rely on the attribution of *intentions* (i.e., players not only observing their actions but forming beliefs on each other's *motives*) in an essential way. Depending on the available alternatives, identical outcomes may be interpreted in different ways. For outcome-based approaches, this is not the case: since it is only the intrinsic properties of outcomes that drive behavior, the alternatives the players face are irrelevant. Section 2.3 gives a detailed exposition of the reciprocity model that was developed by Falk and Fischbacher (2006) and points out some differences to the models of reciprocity by Rabin (1993) and by Dufwenberg and Kirchsteiger (2004).

This chapter lays some foundations for later chapters: chapter 3 is based on the model of reciprocity by Falk and Fischbacher (2006), while chapters 5 to 7 employ a utility representation that is closely related to the model of Fehr and Schmidt (1999).

2.1 Evidence from Experiments

In most applications of economic models, it is assumed that people maximize their own material well-being without caring for "social" goals. In contrast to this, a large body of experimental evidence suggests that there are other determinants of human behavior.

In experimental *ultimatum games*, most proposers offer between forty and fifty percent of the total pie to the receiver while offers below twenty percent are rarely observed. Small offers are usually rejected, with the rejection rate decreasing in

the amount offered to the responder. Another observation is that responders accept lower offers when they are made by a random device instead of a human.[3]

A similar game is the *best-shot game*, in which the proposer can only choose between two payoff distributions. He can either propose a split that is very advantageous or disadvantageous to himself. Participants in experiments usually accept a higher degree of inequity than in the ultimatum game.[4]

The *dictator game* is similar to the ultimatum game, but the second player has no choice and must accept the first player's offer. Experiments show that proposers in dictator games usually offer less than in the ultimatum games. About 80 percent offer a positive amount while practically nobody offers more than 50 percent.[5]

In the *gift-exchange game*, the first mover (firm) offers a wage w to the second mover (worker). The worker can reject the offer so that both earn nothing, or he can accept. If he accepts, he makes a costly effort decision with a convex cost function $c(e)$. Higher wages yield lower monetary payoffs for the firms and higher ones for workers, while higher effort levels have the reverse effect on payoffs. The standard game-theoretic prediction based on monetary payoff maximization is that workers will invariably choose the lowest possible effort level, since this choice is dominant in a pecuniary sense. In anticipation of this, firms will only make the lowest possible wage offer. In contrast to this, some firms offer wages above the market rate, and a positive relationship between wages and effort levels is usually observed.[6]

In *public-goods games*, agents simultaneously decide whether to pay for the provision of a public good or not. The good is said to be public because every agent, regardless of whether he paid for it or not, enjoys the same benefit from this good. Experiments show that the amount subjects contribute is increasing in their expectation about the contributions of others.[7]

These experimental results suggest that subjects do not only care about their own material payoff but also about something like fairness. Fehr and Schmidt (2006) give a good overview of recent papers written based on these results.

---

[3] Analysis of the ultimatum game was pioneered by Güth et al. (1982). For surveys on ultimatum games, see Thaler (1988), Güth and Tietz (1990), Camerer and Thaler (1995) and Roth (1995).

[4] The best-shot game was introduced by Harrison and Hirshleifer (1989) and Prasnikar and Roth (1992). See also Falk et al. (2003) for experimental results.

[5] See, for example, Forsythe et al. (1994) and Andreoni and Miller (2002).

[6] See, for example, Berg et al. (1995), Gächter and Falk (2002), or Fehr et al. (1996). Falk (2007) reports findings from a field study that demonstrates—in a setting outside the labor market—how gift-giving to donors induces them to reciprocate the gift by donating more.

[7] For surveys on public-goods games, see Ledyard (1995) and Dawes and Thaler (1988). Another interesting study was published by Falk et al. (2005), who find that besides fairness considerations, spitefulness is an important driver of sanctioning behavior in multi-person games.

2.2 Equity-based Approaches

This section introduces two papers that try to reconcile the above-mentioned stylized
facts from experiments with traditional Bayesian equilibrium behavior of optimizing
agents: Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). They describe
models in which agents have preferences that exhibit *inequity aversion*. Agents can
increase their utility by sacrificing their own material payoff if by doing so their
payoff is closer to the payoff the other players receive. Preferences depend only
on the *payoff distribution* but not on a measure for the *intentions* attributed to the
opponents' behavior.

In Fehr and Schmidt (1999), subjects are introduced that dislike inequitable out-
comes, but dislike them less when they are to their own advantage. The utility func-
tion for player $i \in \{1, \ldots, n\}$ is given by

$$u_i(\pi) = \pi_i - \frac{\alpha_i}{n-1} \sum_{j \neq i} \max(\pi_j - \pi_i, 0) - \frac{\beta_i}{n-1} \sum_{j \neq i} \max(\pi_i - \pi_j, 0) \,,$$

or simplified for the two-player case $i \in \{1, 2\}$:

$$u_i(\pi) = \pi_i - \alpha_i \max(\pi_j - \pi_i, 0) - \beta_i \max(\pi_i - \pi_j, 0), \quad i \neq j \,,$$

with $\alpha_i \geq \beta_i$ and $0 \leq \beta_i < 1$. π_i denotes the material payoff to player i; α_i and β_i are
capturing how much the subject suffers from inequitable outcomes, where $\alpha_i \geq \beta_i$
captures the idea that the subject suffers more from inequality that is to his disad-
vantage. The share of individuals with a concern for equity (in the model, part of
the total population consists of purely selfish players), where α_i and β_i are positive
is exogenously given and common knowledge. The assumption that individuals are
heterogenous is an important ingredient of the model.

Despite its simplicity, many stylized facts can be explained. The model is con-
sistent with giving in dictator, trust, and gift-exchange games and with the rejection
of low offers in ultimatum games. However, since the model does not account for
intentions, the model fails to explain why people behave differently when playing
against a random device instead of a real player, or why low offers in a best-shot
game are more readily accepted than in an ultimatum game.

The approach by Bolton and Ockenfels (2000) is similar to this model although
there are some differences in detail. For example, in their model, the subjects com-
pare their material payoff to the material *average* payoff of the group, so there
is a difference regarding the *reference point* for fairness considerations. Another
difference is that the marginal disutility of small deviations from equality is zero.
Therefore, if subjects are non-satiated in their own material payoff, they will never
propose an equal split in the dictator game. The utility function is given by

$$u_i(\pi) = \begin{cases} (\pi_i, \frac{\pi_i}{\sum_{j=1}^{n} \pi_j}) & \text{if } \sum_{j=1}^{n} \pi_j > 0 \\ f(\pi_i, \frac{1}{n}) & \text{if } \sum_{j=1}^{n} \pi_j = 0 \,. \end{cases}$$

This utility function is assumed to be weakly increasing and concave in player i's own material payoff π_i for any given u_i. For any given π_i, the utility function is strictly concave in player i's share of total income. The specific functional form is not fixed, so this utility function is more flexible than the one in the model of Fehr and Schmidt.

The model of Bolton and Ockenfels (2000) can also explain giving in dictator and gift-exchange games as well as rejections in ultimatum games. However, the model fails to explain punishment patterns in the public-goods games,[8] and because the reference point is the average payoff, it cannot explain behavior dependent on inequities among other players.[9]

2.3 Reciprocity

The concept of *reciprocity* captures a motivational force behind human behavior. Reciprocity can be distinguished from simple altruism, which corresponds to unconditional generosity. Positive (negative) reciprocity is the impulse or the desire to be kind (unkind) to those who have been kind (unkind) to us.[10]

A crucial feature of the psychology of reciprocity is obviously that people decide about their actions towards others not only according to the *material consequences* resulting from the actions taken by the latter, but also dependent on the *intentions* attributed to them. One example are people who are motivated by *positive reciprocity*. They differentiate between people who take a generous action by choice and those who are forced to do so.[11]

This section reviews some of the literature which tries to integrate the phenomenon of reciprocal behavior into standard game theory.[12] The focus is on the approach of Falk and Fischbacher (2006). Some important differences to the approaches by Rabin (1993), and Dufwenberg and Kirchsteiger (2004) are pointed out

---

[8] For example, Fehr and Gächter (2000) provide evidence that free riders in public-goods games are punished if there is an opportunity for the other players to do so. The more the free riders negatively deviate from the group standard, the more they are punished—even if punishing is costly to the punisher.

[9] For an example of behavior dependent on inequities among other players, see Charness and Rabin (2002). They let a player C choose between the material payoff allocations (575,575,575) and (900,300,600). In both allocations, player C receives the fair share of $1/3$, so the Bolton and Ockenfels model predicts the second choice as the payoff for player C is higher. However, in experiments, 54% of subjects choose the first allocation.

[10] See, for example, Fehr and Gächter (1998) for more detailed observations regarding reciprocity.

[11] Experimental evidence is given in Falk et al (2003) where the authors performed four different mini-ultimatum games. In each game, the proposer had two choices, one of which was always to offer 20%. The alternatives have been 0%, 20%, 50%, or 80%. The rejection rate of the 20% offer was highest when the alternative was equal division; it was lowest when the only alternative was to offer nothing to the second player. The fact that the rejection rate was not zero in this case suggests that pure equity considerations indeed play a role.

[12] See Klein (2000) for a detailed comparison of literature on reciprocity.

to clarify why the model of Falk and Fischbacher was chosen to be further analyzed in chapter 3.

Rabin was the first who adapted the framework of psychological games of Geanakoplos et al (1989) to suit the phenomenon of reciprocity. He introduced so-called *fairness games* in which a reciprocity payoff is added to the material payoff of the players. This reciprocity payoff can be described as the product of a *kindness* and a *reciprocation term*. The kindness term is positive if a player feels he is treated well. The player then wants to make the reciprocation term positive in order to increase his overall payoff. He can achieve this by behaving nicely in return, as the reciprocation term is defined to be positive when the player chooses a kind action. The motivation for negative reciprocity is modeled analogously. Because the kindness and the reciprocation term depend explicitly on *belief*s, a psychological game is formed.

However, Rabin's model is restricted to simultaneous, two-player normal-form games. This implies a drawback when a sequential game is rewritten in normal form and solved accordingly: Rabin's model cannot not take the sequential structure of the game into account. Therefore, an equilibrium in Rabin's model may allow for non-optimizing behavior at information sets that are not reached. The paper by Dufwenberg and Kirchsteiger (2004) is closely related to Rabin (1993). They generalize Rabin's model to *n*-person extensive-form games of imperfect information. In contrast to Rabin, they impose not only subgame perfection but also *sequential rationality* in non-proper subgames. The main idea behind their extension is to keep track of players' beliefs about the strategy profile being played as the game evolves. Falk and Fischbacher (2006) also extend Rabin's approach, but to extensive-form games of perfect information. They use a more complex utility function that allows for equity concerns *and* for intentions.

Each of these papers has four basic "ingredients": the *kindness term* $\tilde{f}$, the *reciprocation term* f, a *social utility function* u,[13] and an *equilibrium concept*. In the following subsections, each "ingredient" of the approach by Falk and Fischbacher (2006) is discussed, and some relevant differences to the other approaches are briefly pointed out. All discussed models boil down to games with only two players, so the *n*-player case, which is important for market environments, for example, is not discussed.

2.3.1 Kindness Term

Consider a two-player game with strategy sets S_i and S_j for the two players i and j with $\pi_i \colon S_i \times S_j \longrightarrow \mathbb{R}$ describing the expected payoff for player i. The *kindness term* $\tilde{f}_j(\cdot)$ measures the kindness player i experiences from player j's expected actions. It is positive if player j is considered as acting kindly and negative if he is considered as acting unkindly. Player i's kindness term depends on the strategy $a_j \in S_j$

---

[13] The social utility function is usually a sum of the material payoff π and the reciprocity payoff, which is the product of kindness term $\tilde{f}$ and reciprocation term f. For example, $u = \pi + \tilde{f} \times f$.

chosen by player j, player i's belief $b_j \in S_j$ about the strategy a_j, and player i's second-order belief $c_i \in S_i$ about what he believes player j believes he is choosing; or in other words: c_i is player i's belief about b_i. In the following $i \in \{1, 2\}$ and $j = 3 - i$.

Falk and Fischbacher (2006) base their model on extensive-form games with a finite number of stages and with complete and perfect information. They define $\mathcal{N}_i$ as the set of nodes at which player i has to move. Furthermore they define $\pi_i(n, s_i, s_j) \equiv \pi_i(s_i | n, s_j | n)$ as the expected payoff of player i conditional on node $n \in \mathcal{N}_i$, that is, the expected payoff of player i in the subgame starting from node n, given that the strategies s_i and s_j are played.

The kindness term combines the psychological approach with equity considerations. The kindness term $\tilde{f}_j(n)$ in a node $n \in \mathcal{N}_i$ is given by[14]

$$\tilde{f}_j(n) = \vartheta_j(n) \Delta_j(n) \,.$$

The second expression $\Delta_j(n)$ is called the *outcome term*. It is positive if player i experiences a higher payoff than his opponent:

$$\Delta_j(n) = \pi_i(n, b_j, c_i) - \pi_j(n, b_j, c_i)$$

with $\pi_i(n, b_j, c_i)$ being player i's *belief* about the material payoff player j is offering to player i if player j chooses to play a_j and expects player i to choose b_i. In the following, let $(\pi_i^0, \pi_j^0) = (\pi_i(n, b_j, c_i), \pi_j(n, b_j, c_i))$ to simplify notation.

The first component $\vartheta_j(n)$, which is called the *intention factor*, is given by

$$\vartheta_j(n) = \max\{\Omega(\tilde{\pi}_i, \tilde{\pi}_j, \pi_i^0, \pi_j^0) | (\tilde{\pi}_i, \tilde{\pi}_j) \in \Pi_i(n)\} \,.$$

This factor measures how much intention player i attaches to the actions he expects player j to choose and which would result in the payoffs (π_i^0, π_j^0). $\vartheta_j(n)$ is calculated in two steps.

First, (π_i^0, π_j^0) is compared with all the alternatives player j has. These alternatives are denoted by $(\tilde{\pi}_i, \tilde{\pi}_j)$ and are collected in the set of payoff combinations $\Pi_i(n)$. Every comparison is summarized by $\Omega(\tilde{\pi}_i, \tilde{\pi}_j, \pi_i^0, \pi_j^0)$ and results in a value of $\Omega(\cdot)$ between 0 and 1. A value of 1 resembles *full intentionality* while a smaller value of $\Omega(\cdot)$ expresses a lower degree of intention.

In a second step, the *maximum value* of all these comparisons is taken to be the *overall intention* player i attaches to (π_i^0, π_j^0). The main intuition behind this approach is the following: if player j has at least one alternative, where he could give more to player i without suffering any loss himself, then (π_i^0, π_j^0) is considered as fully intentional.

The value of Ω is calculated as follows:

---

[14] $\tilde{f}_j(n)$ corresponds to $\varphi_j(n)$ in their paper.

$$\Omega(\tilde{\pi}_i, \tilde{\pi}_j, \pi_i^0, \pi_j^0) = \begin{cases} 1 & \text{if } \pi_i^0 \geq \pi_j^0 \text{ and } \tilde{\pi}_i < \pi_i^0 & \text{(a)} \\ \varepsilon_i & \text{if } \pi_i^0 \geq \pi_j^0 \text{ and } \tilde{\pi}_i \geq \pi_i^0 & \text{(b)} \\ 1 & \text{if } \pi_i^0 < \pi_j^0, \tilde{\pi}_i > \pi_i^0 \text{ and } \tilde{\pi}_i \leq \tilde{\pi}_j & \text{(c)} \\ \max(1 - \frac{\tilde{\pi}_i - \tilde{\pi}_j}{\pi_j^0 - \pi_i^0}, \varepsilon_i) & \text{if } \pi_i^0 < \pi_j^0, \tilde{\pi}_i > \pi_i^0 \text{ and } \tilde{\pi}_i > \tilde{\pi}_j & \text{(d)} \\ \varepsilon_i & \text{if } \pi_i^0 < \pi_j^0 \text{ and } \tilde{\pi}_i < \pi_i^0. & \text{(e) .} \end{cases}$$

In cases (a) and (b), player i receives a higher payoff than player j. While in case (a), this is interpreted as being fully intentional ($\Omega = 1$), it is not in case (b). This is because in case (a), the alternative which player j could choose would leave him a smaller payoff ($\tilde{\pi}_i < \pi_i^0$), while in case (b), the alternatives would make him even better off ($\tilde{\pi}_i \geq \pi_i^0$). In this latter case, player j's action is not seen by player i as being really generous and therefore the intention factor is set to be $\Omega = \varepsilon_i$. This means that the choice of (π_i^0, π_j^0) in case (b) is not seen as intentional, but evaluated with the individual *pure outcome-concern parameter* $0 \leq \varepsilon_i \leq 1$.[15] Case (b) is also relevant when player j has no alternative to choose from. Then the intuition behind setting $\Omega = \varepsilon_i$ is that kindness or unkindness is perceived to be weak but not zero, regardless of the actual material distribution.

In cases (c), (d), and (e), player i's payoff is smaller than the payoff of player j ($\pi_i^0 < \pi_j^0$). In case (c), player j could have put player i in a better position ($\tilde{\pi}_i > \pi_i^0$) without giving up his relatively better position ($\tilde{\pi}_i \leq \tilde{\pi}_j$). This is evaluated by player i as fully intentional *meanness* of player j by setting $\Omega = 1$. In case (d), player j could also have improved player i's situation ($\tilde{\pi}_i > \pi_i^0$), but at the expense of making himself worse off ($\tilde{\pi}_i > \tilde{\pi}_j$). How player i evaluates this depends on how much player j could have decreased the payoff difference between the player relative to the chosen outcome (π_i^0, π_j^0). The smaller the resulting difference in the alternative, the higher the intention factor player i attributes to this choice. If player j would have to sacrifice a lot to decrease the inequity, then $\Omega = \varepsilon_i$ as above. In this case, ε_i could be interpreted as an *"envy term."* Finally, case (e) captures the situation where player j has no possibility at all to offer a higher payoff to player i, so no intention is visible which results in $\Omega = \varepsilon_i$.

This measure of kindness is robust against additions of supposedly irrelevant alternatives to the game, which does not hold for Rabin's model. Furthermore, Rabin (1993), and Dufwenberg and Kirchsteiger (2004) limit the kindness term to behavior which is driven only by intentions; pure equity consideration cannot play a role in their models. In this regard, the model of Falk and Fischbacher (2006) is more general, as it can encompass not only intention-based behavior, but also pure concern for equity ($\varepsilon_i = 1$).

---

[15] As the parameter ε_i measures player i's *pure* concern for an equitable *outcome*, it can be used to model behavior which is purely intention driven ($\varepsilon_i = 0$), as in Rabin (1993) and Dufwenberg and Kirchsteiger (2004), or a behavior which is purely outcome oriented ($\varepsilon_i = 1$), as in the equity-based approaches by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000).

2.3.2 Reciprocation Term

The *reciprocity payoff* consists of two terms—the *fairness term* as given above, and a *reciprocation term*. The reciprocation term captures player i's reaction to player j's expected behavior and gives an interpretation of player i's *reaction* as either nice or mean. The product of kindness and reciprocation terms are part of the utility function. Therefore, both terms will have the same sign in equilibrium.

   The reciprocation term of Falk and Fischbacher (2006) captures the influence player i's reaction has on player j's payoff:[16]

$$f_i(n,t,b_j,c_i) = \pi_j(\nu(n,t),c_i,b_j) - \pi_j(n,c_i,b_j)$$

with $\nu(n,t)$ denoting the unique node that directly follows node $n \in \mathcal{N}$ on the path to a terminal node $t \in \mathcal{T}$. The reciprocation term $f_i(\cdot)$ describes the *change* in player j's *expected payoff* implied by player i's move when he selects a path on the game tree towards the terminal node t. The second term is player j's expected payoff before this move and the first term is player j's expected payoff after player i's move. Falk and Fischbacher call $f_i(\cdot)$ the *"alteration"* of player j's payoff from $\pi_j(n,c_i,b_j)$ to $\pi_j(\nu(n,t),c_i,b_j)$. The intuition that a positive change in player j's payoff means a "reward" for him is misleading. In fact, it does measure how much player j is pleased or upset, after each realization of player i's mixed strategy. The reciprocation utility player i receives does depend on how much he is able to surprise player j with his action, given player j's beliefs. However, in equilibrium, beliefs are consistent, and therefore it is not possible for player i to surprise player j by his moves. Because of this, the reciprocity payoff $f_i(\cdot)$ in equilibrium is always zero by definition. Still, this concept helps to understand how an equilibrium is reached. In principle, $f_i(\cdot)$ is defined so that player i tries to reward player j for kind moves and tries to punish him for unkind actions.[17]

   This shows the role beliefs play in psychological games. They do not have any deeper meaning, but are a helpful device to construct equilibria.[18]

2.3.3 Utility Function

The total utility that the players maximize is a sum of the material payoff and a reciprocity payoff. In the model of Falk and Fischbacher (2006), the utility for player i at each end node $t \in \mathcal{T}$ is given by

---

[16] $f_i(n,t,b_j,c_i)$ corresponds to $\sigma_i(n,f)$ in their paper.

[17] See Klein (2000) for this explanation.

[18] In the literature, it is not clear if or how beliefs are given or formed before the equilibrium choices are made, or if they are formed according to equilibrium choices after equilibrium is reached. We leave this question open for discussion.

$$u_i(t) = \pi_i(t) + \rho_i \sum_{\substack{n \in \mathcal{N}_i \\ n \to t}} \tilde{f}_j(n) \times f_i(n, t, b_j, c_i) \,. \tag{2.1}$$

As above, the *pure material payoff* $\pi_i(t)$ at the end node t is combined with an expression representing the *reciprocity payoff*. In Falk and Fischbacher's model, this reciprocity payoff is calculated as the sum over all nodes in the game tree on the path from the root node to the terminal node t (that is denoted by $\substack{n \in \mathcal{N}_i \\ n \to t}$). For each node, the fairness expression $\tilde{f}_j(\cdot)$ is multiplied with the reciprocation term $f_i(\cdot)$, and finally, the sum is weighted with the reciprocity parameter ρ_i against the material payoff. If the reciprocity parameter for both players is zero $(\rho_i = \rho_j = 0)$, then the game is reduced to a standard game.

To clarify how the calculations actually work, the following demonstrates how the concept is applied to the ultimatum game.

In a first step, the decision of the responder is analyzed. The responder can either accept or reject the offer. His payoff is the sum of the actual material payoff he receives and his reciprocity payoff. For offers of exactly or more than half of the pie, the responder is in a (weakly) advantageous position. Because the proposer could have offered less, this move is considered as fully intentional and therefore $\Omega = 1$. The outcome term is also positive, because the responder's payoff is weakly higher than the proposer's. Taken together, the kindness term is positive, too. Therefore, the responder tries to reciprocate and always accepts the offer. If the offer is strictly below half of the pie, it is regarded as being fully intentional with $\Omega = 1$, because the proposer could have offered half. This would have improved the responder's situation without putting the proposer at a relative disadvantage. The outcome term is negative because the responder's payoff is smaller than the proposer's. So in this case, the kindness term is negative. Now the receiver's reciprocity payoff is positive when he rejects, as this would result in a negative reciprocation term. Whether he indeed rejects depends on a trade-off between the reciprocity payoff he gains and the material payoff he forgoes by rejecting.

In a second step, the behavior of the proposer can be analyzed. Whenever he offers a positive amount, the responder has the choice to accept or to reject. The responder's behavior is considered fully intentional $(\Omega = 1)$ because rejecting would leave the proposer with a smaller payoff than accepting. For offers of more than half, the outcome term from the proposer's point of view is negative. So in these cases, the utility is decreasing in the amount offered, which in turn leads the proposer to offer half of the pie, at the most. For offers of less than half, the outcome term is positive, so the proposer gains reciprocity utility by offering more to the receiver. Again, the result is a trade-off between reciprocity payoff and material payoff. Note that this definition of the utility function can allow for pure altruism. This is, for example, ruled out in the model by Rabin (1993).

2.3.4 Equilibrium Concepts

The utility function used by Falk and Fischbacher (2006) has the properties as in the definition by Geanakoplos et al. (1989) of a psychological game—all information sets are singletons and beliefs do not depend on the history of play.[19] So the result of Geanakoplos et al. (1989) applies: the concept of subgame perfection is also a suitable equilibrium refinement in psychological games. Falk and Fischbacher (2006) call a subgame perfect psychological Nash equilibrium a *reciprocity equilibrium*. For $\rho_i = \rho_j = 0$, the definition is equivalent to the definition of a subgame perfect Nash equilibrium.

In contrast to the models of Rabin (1993) and Dufwenberg and Kirchsteiger (2004), the model of Falk and Fischbacher (2006) makes *unique* equilibrium predictions for most relevant games. Application to several games shows that it is possible to derive plausible predictions for different games with the same utility function and even the same parameter constellation.

The example of the ultimatum game given in section 2.3.3 can be extended to equilibrium considerations. Because the model allows for backward induction, the responder's behavior is derived first. Recall that the strategy of the responder R depends on a trade-off between material and reciprocity payoff. Calculations show that offers c above a certain threshold $c_0(\rho_R) \leq 1/2$ are always accepted, with $c_0(\rho_R)$ increasing in ρ_R. For offers below this threshold, the responder plays a mixed strategy of randomly choosing between accepting and rejecting with an acceptance-probability $p(c, \rho_R)$, which is increasing in c and decreasing in ρ_R. A higher offer is more likely to be accepted, while a more reciprocal responder rejects the same offer more often.

There are two cases for the proposer P. If he has a relatively low concern for reciprocity, he simply maximizes his expected material payoff, which results in an offer of exactly $c_0(\rho_R)$. This is the smallest offer which assures acceptance by the responder. However, when his concern for reciprocity is relatively high, he can gain additional utility by offering more, so in this case, the equilibrium offer is $c(\rho_P) > c_0(\rho_R)$. Nevertheless, he will always offer at most half of the pie.[20]

Falk and Fischbacher (2006) call a subgame perfect psychological Nash equilibrium a *reciprocity equilibrium*. For a proof of existence in the considered class of games and more details on the following equilibrium derivations, see their paper. The following paragraphs give a formal proof of the equilibrium for the ultimatum game:

Assume that the following equations (2.2) and (2.3) describe the unique reciprocity equilibrium of the ultimatum game: Equilibrium play results in *acceptance probability*

---

[19] The utility function in Falk and Fischbacher (2006) as given here is not continuous. However, Falk and Fischbacher show that a minor technical modification in the definition of Ω guarantees the existence of a reciprocity equilibrium.

[20] This is a result of the considerations at the end of section 2.3.3.

$$p^*(c) = \begin{cases} \min\left\{1, \frac{c}{\rho_R \times (1-2c)(1-c)}\right\} & \text{if } c < \frac{1}{2} \\ 1 & \text{if } c \geq \frac{1}{2} \end{cases} \qquad (2.2)$$

for offer c if $\rho_R \neq 0$, and $p^* \equiv 1$ if $\rho_R = 0$. The share c^* offered by the proposer P to responder R in equilibrium is given by

$$c^* = \max\left\{\frac{3\rho_R + 1 - \sqrt{1 + 6\rho_R + \rho_R^2}}{4\rho_R}, \frac{1}{2} \times (1 - \frac{1}{\rho_P})\right\} \qquad (2.3)$$

for $\rho_P, \rho_R \neq 0$.

First, let p' denote the proposer's belief about acceptance probability p and let p'' denote the responder's belief about p'. Let $\vartheta_P(c)$ be the intentionality factor at the decision node after the proposer P's choice of c. The responder's utility is

$$u_{R_A} = c + \rho_R \times \vartheta_P(c)\, p''\, [c - (1-c)] \times [(1-c) - p''(1-c)]$$

in case she accepts the offer and

$$u_{R_R} = \rho_R \times \vartheta_P(c)\, p''\, [c - (1-c))] \times [0 - p''(1-c)]$$

if she rejects. The former is greater for $c \geq 1/2$, implying acceptance. For $c < 1/2$ define (by setting $u_{R_A} = u_{R_R}$)

$$p''_{\text{crit}} = \frac{c}{\rho_R \vartheta_P(c)(1-2c)(1-c)} \,. \qquad (2.4)$$

Note that $p'' > p''_{\text{crit}}$ would ask for $p = p'' = 0$ in contradiction to $p''_{\text{crit}} \geq 0$ (for $c < 1/2$). Hence, either $p'' < p''_{\text{crit}}$ so that optimal responder behavior in equilibrium (involving consistent beliefs) requires $p = p'' = 1$, or we have $p'' = p''_{\text{crit}}$ so that $p = p'' = p''_{\text{crit}}$ is optimal for consistent beliefs. Optimal responder behavior can thus be summarized by $p^*(c) = \min\{1, p''_{\text{crit}}\}$. If $c < 1/2$, then responder R is disadvantaged and the proposer P's move is considered as fully intentional because $c = 1/2$ would lead to a higher payoff for responder R without making P worse off. Therefore, $\vartheta_P(c) = 1$ in equation (2.4).

The expected utility of a proposer with a correct belief p' determined by $p^*(\cdot)$ is

$$u_P = p^*(c) \times (1-c) + \rho_P \times \vartheta_R(c)\, p^*(c'')\,(1-2c'') \times [p^*(c)c - p^*(c'')c''] \quad (2.5)$$

where c'' denotes the proposer P's second-order beliefs.[21] In equilibrium, we must have $c \leq 1/2$ because u_P is decreasing in c for $c'' \geq 1/2$. For $c > 0$, the responder's move $p^*(\cdot)$ is fully intentional, as she has the option of rejecting the offer, which leads to a smaller payoff for the proposer; so $\vartheta_R = 1$.

---

[21] One can think of c'' as an offer that proposer P conjectures R's response to be based on in order to evaluate her kindness. This kindness renders a particular reciprocation and corresponding actual offer c optimal, which must coincide with c'' in equilibrium.

Now define

$$c_0 = \frac{1 + 3\rho_R - \sqrt{1 + 6\rho_R + \rho_R^2}}{4\rho_R}$$

as the smallest c such that $p^*(c) = 1$. u_P is increasing in c for $c < c_0$; so a rational proposer must choose $c^* \geq c_0$, implying acceptance with probability 1. Setting $p^* = 1$ and $c \geq c_0$ in equation (2.5), one obtains

$$u_P = (1 - c) + \rho_P \times (1 - 2c'') \times (c - c'')$$

and

$$\frac{\partial u_P}{\partial c} = -1 + \rho_P \times (1 - 2c'').$$

So u_P is decreasing in c for

$$c'' > c''_{\text{crit}} = \frac{1}{2}\left(1 - \frac{1}{\rho_P}\right)$$

and increasing for $c'' < c''_{\text{crit}}$. First, consider $c''_{\text{crit}} < c_0$: Since, in equilibrium, $c = c''$, we get $c''_{\text{crit}} < c_0 \leq c = c''$ and u_P is decreasing in c. Then, the optimal proposal is $c^* = c_0 (= \max(c_0, c''_{\text{crit}}))$. Second, consider $c''_{\text{crit}} \geq c_0$: If $c'' > c''_{\text{crit}}$, u_P is decreasing in c and therefore c would have to be chosen equal to c_0 which is, however, incompatible with $c = c''$ because $c'' > c''_{\text{crit}} \geq c_0 = c$. If $c'' < c''_{\text{crit}}$, u_P is increasing in c and therefore c is chosen equal to 1, which is also incompatible with $c = c''$ because $c'' < c''_{\text{crit}} < 1/2 < 1 = c$. Therefore, $c^* = c'' = c''_{\text{crit}} (= \max(c_0, c''_{\text{crit}}))$.

The equilibrium in the dictator game for agents with preferences based on Falk and Fischbacher (2006) is easily derived. The key difference from the ultimatum game is that acceptance is not intentional in the dictator game. So the intentionality factor at the proposer's decision node equals ε_P. Then

$$u_P = (1 - c) + \rho_P \varepsilon_P (1 - c'' - c'')c$$

and the first order condition yields

$$c''_{\text{crit}} = 1/2\left(1 - \frac{1}{\rho_P \varepsilon_P}\right).$$

2.3.5 Explanatory Power

In the reciprocity equilibrium as defined by Falk and Fischbacher (2006), the responder's strategy in the ultimatum game is mixed. The acceptance probability is

monotonically increasing in the amount offered up to a point, called *cutoff point*,[22] from which on the offer is always accepted. The cutoff point is decreasing in the responder's reciprocity parameter and always lower than one half.

The offer of a selfish proposer matches this cutoff point and is therefore also increasing in the responder's reciprocity parameter. If the proposer himself is reciprocal, his offer can be more generous. This depends on how reciprocal he is relative to the responder.

The equilibrium in random-offer ultimatum games, in which the offer is not determined by a human but by a random device, demonstrates the role of intentions: because no intentions are attributed to a random offer, the acceptance probability of a given offer is weakly higher than in the regular ultimatum game. The acceptance probability is not always one, because equity considerations also play a role.

The model of Falk and Fischbacher (2006) correctly predicts the main stylized facts for the gift-exchange game, in which positive reciprocity plays a major role. The worker's effort choice is increasing in the wage paid and in his reciprocity parameter. This is because the higher the wage paid by the company, the higher the kindness term from the worker's point of view. A reciprocal worker tries to react kindly by increasing his effort choice. The results are strictly positive wages.

In analyzing the best-shot game, the dictator game, and the public-goods game the results of Falk and Fischbacher (2006) are also consistent with the experimental evidence discussed above.

2.4 Summary of Fairness Approaches

The models discussed in sections 2.2 and 2.3 fall in two categories: equity-based and intention-based psychological approaches to fairness. One important difference between them is that the former assume that players are either intrinsically fair or intrinsically egoistic or that each player randomizes whether he acts fairly or not. The ex-ante probability of a player being fair is common knowledge. In psychological games, beliefs are built independently of information. A shortcoming of this approach is that beliefs do not have any inherent meaning, and therefore might be formed strategically.

The approaches further differ in what matters to the players: only the outcomes, only the intentions, or both. This is achieved by different definitions of the utility function. The more degrees of freedom the utility function of a fairness game does allow for, the more a model can explain. Unfortunately, equilibrium calculations in psychological games are very complex and sometimes result in ambiguous predictions. Therefore, equity-based models are much easier to handle. However, according to most experimental evidence, human behavior depends on *both* intentions and considerations regarding the distribution of material payoffs.[23]

---

[22] The cutoff point is formally given by $c_0(\rho_R)$, see above.

[23] For example, the mini-ultimatum game experiment conducted by Falk et al (2003) shows this explicitly.

Of all discussed models, only the one developed by Falk and Fischbacher (2006) is able to account for this. Not surprisingly, it is most successful in predicting behavior observed in experiments with computable unique equilibria for many games. With two free parameters ρ and ε for each player's utility function, it is furthermore possible to model pure selfish behavior, pure inequity aversion, or pure intentional reciprocity by simply using limit cases. This makes it a very powerful tool for predicting behavior resulting from a variety of preferences in many different games.

References

Andreoni J, Miller JH (2002) Giving according to GARP: An experimental test of the consistency of preferences for altruism. Econometrica 70:737–753

Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity and social history. Games Econom Behav 10(1):122–142

Bolton GE, Ockenfels A (2000) ERC: A theory of equity, reciprocity, and competition. Am Econ Rev 90(1):166–193

Camerer CF (2003) Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press, Princeton, NJ

Camerer CF, Thaler RH (1995) Ultimatums, dictators, and manners. J Econ Perspect 9(2):209–219

Charness G, Rabin M (2002) Understanding social preferences with simple tests. Q J Econ 117(3):817–869

Dawes RM, Thaler RH (1988) Cooperation. J Econ Perspect 2:187–197

Dufwenberg M, Kirchsteiger G (2004) A theory of sequential reciprocity. Games Econom Behav 47(2):268–298

Falk A (2007) Gift exchange in the field. Econometrica 75(5):1501–1511

Falk A, Fischbacher U (2006) A theory of reciprocity. Games Econom Behav 54(2):293–315

Falk A, Fehr E, Fischbacher U (2003) On the nature of fair behavior. Econ Inq 41(1):20–26

Falk A, Fehr E, Fischbacher U (2005) Driving forces behind informal sanctions. Econometrica 7(6):2017–2030

Fehr E, Gächter S (1998) Reciprocity and economics. The economic implications of homo reciprocans. Eur Econ Rev 42(3–5):845–859

Fehr E, Gächter S (2000) Cooperation and punishment in public goods experiments. Am Econ Rev 90(4):980–994

Fehr E, Schmidt KM (1999) A theory of fairness, competition and cooperation. Q J Econ 114(3):817–868

Fehr E, Schmidt KM (2006) The economics of fairness, reciprocity and altruism—Experimental evidence and new theories. In: Kolm SC, Mercier Ythier J (eds) Handbook on the Economics of Giving, Reciprocity and Altruism, vol 1, Elsevier, Amsterdam, Netherlands, chap 8, pp 615–691

Fehr E, Gächter S, Kirchsteiger G (1996) Reciprocal fairness and noncompensating wage differentials. Journal of Institutional and Theoretical Economics 152(4):608–640

Forsythe R, Horowitz J, Savin N, Sefton M (1994) Fairness in simple bargaining games. Games Econom Behav 6:347–369

Gächter S, Falk A (2002) Reputation and reciprocity: Consequences for the labour relation. Scand J Econ 104(1):1–26

Geanakoplos J, Pearce D, Stacchetti E (1989) Psychological games and sequential rationality. Games Econom Behav 1(1):60–79

Güth W (1995) On ultimatum bargaining experiments—A personal review. J Econ Behav Organ 27(3):329–344

Güth W, Tietz R (1990) Ultimatum bargaining behavior—A survey and comparison of experimental results. J Econ Psychol 11(3):417–449

Güth W, Schmittberger R, Schwarze B (1982) An experimental analysis of ultimatum bargaining. J Econ Behav Organ 3(4):367–388

Harrison GW, Hirshleifer J (1989) An experimental evaluation of weakest link/Best shot models of public goods. J Polit Econ 97(1):201–225

Klein A (2000) Reciprocity, endogenous incomplete contracts, and the role of options in law. PhD thesis, Berlin, Germany: dissertation.de

Ledyard J (1995) Public goods: A survey of experimental research. In: Kagel JH, Roth AE (eds) The Handbook of Experimental Economics, Princeton University Press, Princeton, NJ

Prasnikar V, Roth AE (1992) Considerations of fairness and strategy: Experimental data from sequential games. Q J Econ 107(3):865–888

Rabin M (1993) Incorporating fairness into game theory and economics. Am Econ Rev 83(5):1281–1302

Roth AE (1995) Introduction to experimental economics. In: Kagel JH, Roth AE (eds) The Handbook of Experimental Economics, Princeton University Press, Princeton, NJ, chap 1, pp 3–110

Thaler RH (1988) The ultimatum game. J Econ Perspect 2(4):195–206

Weibull JW (2004) Testing game theory. In: Huck S (ed) Advances in Understanding Strategic Behaviour: Game Theory, Experiments and Bounded Rationality. Essay in Honour of Werner Güth, Palgrave, Basingstoke, UK, pp 85–104