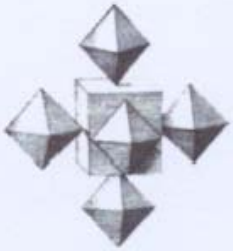

IKERLANAK

SOCIAL PREFERENCES

by

Jaromír Kovárík

2009

Working Paper Series: IL. 36/09

Departamento de Fundamentos del Análisis Económico I

Ekonomi Analisiaren Oinarriak I Saila

University of the Basque Country

Social Preferences - Literature Survey*

Jaromír Kovářík†
University of Alicante

March 6, 2008

Abstract

This paper surveys the theories of social preferences. Social preferences are based on that people not only care about their own well-being, but they have a certain concern with payoffs and/or actions of others. We classify two approaches: distributional and intention-based models, and later discuss models that combine both theories. In order to provide a better illustration of the discussed models, we derive predictions of these models for two classic experimental protocols: ultimatum game and public good game with punishment. These predictions are compared with the stylized facts of these two games.

1 Introduction

Although the traditional definition of utility function allows for a broad type of preferences and the most prominent economists have discussed the importance of non-selfish preferences in economic theory,[1] a large part of applied modeling in economics has restricted, and restricts the attention on simple payoff maximization. Firms are assumed to simply maximize their profits and individuals are supposed to maximize their own material well-being. However, a selfish individual only represents a particular case of a more general set-up. There is a huge number of real-life examples showing that many economic subjects, institutions and humans, satisfy needs different from payoff maximization. The goals of non-governmental organizations, charity donations, long-distance adoption of children from developing countries, parental gifts etc. can serve as examples of other-regarding behavior. Recently, the development of sophisticated experimental techniques allows to directly test the issues of non-selfish behavior in controlled laboratory environment. Both real-life observations and

*I am greatful to Marco Casari, Hubert János Kiss, Giovanni Ponti and Marco van der Leij for their efforts while correcting this paper. Financial support from the Spanish Ministry of Education and Science (AP-2004-1893) is greatfully acknowledged.

†Departmento de Fundamentos del Análisis Económico, University of Alicante, Campus San Vicente, 03080 Alicante, Spain. *E-mail*: jaromirkovarik@merlin.fae.ua.es.

[1] See, for instance, Beker (1974), Harsanyi (1955), Sen (1977) and Smith (1976).

experimental evidence have stimulated a stream of theories, which argue that human decisions can be based on different behavioral grounds than self-interest. Nowadays, these theories are known as *social preferences*.

Traditionally, utility function, $u_i : \Sigma \rightarrow \mathbb{R}$, is a function that assigns a value to each (mixed) strategy profile, such that individual i prefers strategies that assign her higher utilities. This definition does not imply any particular function to represent human preferences. Rather, it allows for a great variety of functional forms to be applied. The preference ordering of a selfish individual is completely independent of payoffs of other involved players, and if two strategy profiles lead to the same payoff the selfish individual is indifferent between the two strategy profiles. Social preferences differ, with respect to self-interest, in that payoffs and/or actions of other players influence the preference ordering of agents. As a consequence, two strategy profiles does not have to lead to the same utility even though they bear the same material payoff.

Most of the existing models of social preferences do not deviate from the above definition of utility function. Agents are still maximizing their utilities, but utility maximization does not coincide with payoff maximization, as it does while working with selfish individuals. Therefore, social preferences require to strictly distinguish between the utility function and material payoff. Throughout the paper, we always denote individual i's utility function u_i and her material payoff π_i. The material payoff function, π_i, is the standard von Neumann-Morgenstern payoff function. Despite the non-selfish nature of agents' preferences, the individuals are always assumed to be *somehow* self-interested. More precisely, if anything else remains unchanged players prefer more material well-being to less.[2]

This paper provides a detailed classification and characterization of existing models of social preferences. We are not the first surveying social preferences. Chapter 2.8 in Camerer (2003), Fehr and Schmidt (2003) and Sobel (2005) also review the existing models. Since Camerer (2003) is a brief description of selected literature, we discuss our contribution with respect to Fehr and Schmidt (2003) and Sobel (2005).

The motivation of this survey is to provide the researchers from any area of social sciences a classification based on the ability of individual models to predict desired behavioral regularities. Therefore, we do not analyze models that have already been extended or generalized, even though they are considered important for the formalization of other theories. Rather, we directly work with the more general version of the model.[3] This is the first difference with respect

[2] The interpretation of 'anything else unchanged' can differ across models. For example, if people are concerned with their own material payoff and their share, the increase of their payoff while remaining the payoff of opponents cannot be interpreted as 'anything else unchanged', since the agents' shares have also changed.

[3] A great example is Rabin (1993), which is considered the father of the formal incorporation of other-regarding motives into game theory. However, his model has been succesfully generalized by Dufwenberg and Kirchsteiger (2004). Hence, even though we discuss Rabin's contribution in the text, the formal analysis is directly performed using the model of Dufwenberg and Kirchsteiger. Similar considerations hold for Bolton and Ockenfels (2000) who generalize Bolton (1991).

to the other surveys.

Furthermore, this survey can be viewed as complementary to Fehr and Schmidt, and Sobel. We review the same models of social preferences, but the analytical tools differ. The main difference is that, due to the motivation of this study, we provide formal predictions of each model in standard experimental games, which allow to see in detail the shortcomings of the corresponding model that do not have to be obvious without formal analysis. Consequently, this survey is slightly more technical than both Fehr and Schmidt, and Sobel.[4]

We follow the literature by distinguishing two types of models: distributional and intention-based. Distributional models assume that only monetary consequences matter. Therefore, material payoffs of opponents can enter in the utility function. On the contrary, intention-based theories focus on why (and which) different actions are taken. Within this stream, *reciprocity* has won main attention. Reciprocal individuals reward kind actions with kindness and punish mean behavior. Both approaches provide explanations for many findings, but each model fails in some aspects. Thus, we also discuss models that combine distributional and intentionalist motivations. Troughout the paper, we contrast and motivate the presented models with behavioral regularities observed in experiments, but the aim of this paper is *not* to review the experimental findings regarding prosocial behavior.[5]

In the analysis, we focus on ultimatum game and public good game with punishment. There are several reasons why these two games have been chosen as a good illustration. Both are widely studied and experimentally tested in the lab, and strategically very simple under the assumption of self interest. In both cases, the option to respond to some previous play gives players the possibility to retaliate opponents' actions what can, in principle, allow for different predictions depending on whether the corresponding model is purely consequentionalist or purely intention-based. Both, furthermore, allows for direct comparison with another, simpler game as the dictator game and the public good game without the punishment option.[6]

Now, we briefly review the results of this paper. Table 1 provides a comprehensive summary of the predictions of the discussed models in three widely studied experimental games: dictator game (DG), ultimatum game (UG) and public good game (PGG) with punishment.[7] The rows correspond to the exist-

[4] Other minor differences are that, in contrast to Sobel (2005), we concentrate exclusively on social preferences, abstracting from any other models and motives that can also predict (apparent) non-selfish behavior, and, in contrast to Fehr and Schmidt (2003), we do not discuss the possible applications of these theories to economic problems.

[5] The interested readers are referred to reviews of Camerer (2003) and Kagel and Roth (1995).

[6] For similar reasons, we could have also chosen trust/investment game (Berg *et al.* (1995)) or gift exchange games (Fehr *et al.* (1993) and (1998)). Nevertheless, we believe that the additional contribution of the formal analysis of these two games would not compensate the cost of the additional space necessary to formalize them.

[7] We also added to the table the dictator game, since it is very simple, leads to clear-cut predictions, and its close relation with the ultimatum game allows to derive the predictions trivially from the ultimatum game results. Nevertheless, we do not provide the formal analysis of this game in the main text.

ing models of social preferences; the columns list the experimental stylized facts for each game.

Put Table 1 around here.

It is worth mentioning that the table is illustrative. Each cell in the table shows whether the corresponding model is cabable to predict the stylized fact, but it does not show how much the predictions match the experimental observations and that there may exist additional, implausible equilibria. Therefore, before any conclusion is made, we recommend to look at the complete analysis in Chapter 3.

Observe that, in spite of a relatively simple strategic situation of the three games in Table 1, no model is able to predict all the behavioral regularities. Distributional models generally fail to predict well in public good game with punishment. Even though Fehr and Schmidt (1999) seem to be an exception and can predict any common contribution scenario, they offer no strong arguments why high contributions are generally observed. On the other hand, the concept of reciprocity fails to predict pure altruism in the dictator game. In spite of the idea of combing both approaches, neither this stream of models does better.

Hence, the first conclusion of our analysis is that we still need a further empirical and theoretical research to propose a sufficiently simple model that can predict all the experimental findings.

Second, to use a specific model, one has to couterweight the technical simplicity of ditributional models, among which Fehr and Schmidt (1999) is the most succesful predicting the observations, against the generality of combinations of distributional motives and intentions that can capture interesting regularites at the cost of mathematical complexity. Contrary to informal arguments of the literature, which stresses the role of intentions, Table 1, jointly with our analysis, suggests that, in spite of its simplicity, Fehr and Schmidt (1999) provide very reasonably good predictions, and some weak points of Fehr and Schmidt‘s model can be overcome by a non-linear version of their model, which we propose in Section 3.1.

The remainder of the paper is organized as follows. The following section discusses two traditionally used experimental games and states the predictions for a self-interested agent. Section 3 is divided into three parts. The first analyzes distributional models. The second part is devoted to intention-based models. In the third part, models combining both approaches are introduced. The last section, Section 4, provides with some remarks and suggestions for future research. All proofs are relegated to the Appendix A and Appendix B contains technical details of Falk and Fischbacher (2004).

2 Ultimatum game, Public Good Game with Punishment, and Self Interest

In this section, we introduce two widely used experimental games, ultimatum game and public good game with punishment, and contrast the model of selfish individual with the empirical evidence.

Ultimatum game is a two-stage two-player game. In the first stage, Proposer proposes a division of a fixed amount of money, say 1 monetary unit, giving to the Responder $c \in [0, 1]$ and leaving $1-c$ for herself. Responder can either accept this division, or reject it. In the former case, the proposed division is realized. Otherwise, both players receive nothing. With selfish agents, the prediction is:

Proposition 1 *In ultimatum game, the unique subgame-perfect equilibrium offer of a self-interested Proposer is zero. Selfish Responder accepts any offer.*

The large experimental evidence on ultimatum game unambiguously rejects this prediction, and the results does not change considerably across studies. An early example is the study of Gűth *et al.* (1982). A more recent experiment, made by Slonin and Roth (1998), confirms these observations for high stakes. Proposers, in general, offer strictly positive amounts of money, on average between 30 and 50% of the stake. Offers of 50% are commonly observed. On the other hand, offers exceeding 50% of the divided amount are extremely rare and never rejected, while offers below 20% are rejected frequently, suggesting that Proposers have to take into account the social concern of Responders, who reject too low offers.

A variation of ultimatum game where Responder can only accept is called the dictator game. In this case, there are no strategic consideration from the part of Proposer. The experimental evidence on this game reports that dictator-game Proposers (Dictators, hereafter) propose significantly less than ultimatum-game Proposers and the average offer is around 20% of the divided sum. Offers of 50% of the stake are still frequently observed in this game (see Camerer (2003)).

The second game, public good game with punishment, is a n-person two-stage game. The first stage coincides with standard public good game without punishment. All players receive an endowment of $y > 1$ monetary units and decide how much to contribute to public good and how much to leave for themselves. Each agent earns $\frac{1}{n} < a < 0$ units of money from each unit contributed to public good. In the second stage, which makes public good games with and without punishment different, each agent observes the vector of all contributions $g = (g_1, ..., g_n)$ and decides whether and who she punishes. Each unit of punishment costs $c \in (0, 1)$. Let p_{ij} denote how much agent i punishes agent j. The material payoff of agent i, therefore, is:

$$\pi_i(g, (p_{ij})_{i,j}) = y - g_i + a \sum_{j \in N} g_j - c \sum_{i \neq j} p_{ij} - \sum_{i \neq j} p_{ji}.$$

Proposition 2 *In public good game with punishment played by n selfish players, the unique equilibrium prediction is no contribution and no punishment.*

Using the assumption of self-interest, Public good game with and without punishment coincide. However, in experiments, we observe completely different behavioral patterns. In last rounds of public good game without punishment, experimental subjects' behavior stabilizes close to full free-riding, as predicted assuming self-interest, while the possibility to punish lead to high contribution levels in public good game with punishment. Figure 1. shows the results of experiment of Fehr and Gachter (2000). Observe the contrast between the public good game with and without punishment. In the upper box, the punishment options causes that more than 80% of individuals contribute fully to the public poll, while around 80% of individuals are close to the Nash equilibrium prediction when the punisment is not possible. The lower box provides similar contrast for the stranger treatment, where players are rematched in each round.

Put Figure 1. around here

Concerning the second stage of the game, contrary to the prediction, people *do* punish deviators. Generally, as the contribution of an individual decreases, the more seriously and frequently is this player punished. Unambiguously, it is the punishment what causes that, in experiments, we observe the convergence to high contribution scenarios in this game.

3 Social preferences

3.1 Distributional Models

The models discussed in this section assume agents to have preferences over the whole final allocation of material well-being. They differ in the way social concern enters into the utility function of an individual.

One of the most studied deviation from self-interest hypothesis is *altruism.* The idea of altruism reaches the very beginning of economics thoughts. Smith (1976) largely discusses altruism in his *Theory of Moral Sentiments* in 1759. Also, Beker (1974) provides an exploration of altruism. Formally, an individual i is altruistic toward an agent j if

$$\frac{\partial u_i(\pi_1, ..., \pi_n)}{\partial \pi_j} > 0. \tag{1}$$

In words, an altruistic individual is positively concerned with the welfare of others. The more others have, the happier he feels. Note that altruism, as described in (1), does not require linearity. Kirchsteiger (1994) uses the opposite idea, envy. An envious agent prefers the others to have a lower material payoff. In this case, the utility function is decreasing in the payoff of others.

An agent with *egalitarian* utility prefers allocations where each individual derives the same utility. Suppose, for example, two agents, i and j. If one unit of a good bears utility 3 to i and utility 1 to j, then, egalitarian agent i prefers allocation of 1 unit for herself and 3 units to agent j, rather than allocation of 2

units for each. Andreoni and Miller (2002) find an evidence for this utility type in their experiment.

Charness and Rabin (2002) develop a multiperson model based on their experimental observations and calibration. They propose the following utility function:

$$U_i(\pi_1, ..., \pi_n) = (1-\lambda)\pi_i + \lambda\left[\delta \min\{\pi_1, ..., \pi_n\} + (1-\delta)(\pi_1 + ... + \pi_n)\right]. \quad (2)$$

with $\lambda \in [0,1]$ and $\delta \in (0,1)$. Subjects tend to help the worst-off player and maximize the social income. Both features form the other-regarding part of the utility function. The value of δ determines the relative power of these two concepts. Then, both other-regarding preferences and self-interest play a role. The power of each is reflected by parameter λ. The higher λ, the less important is self-interest in the model and the higher is the impact of social concerns on individual's behavior. Extreme values, for instance, allow for pure a completely selfish player or a pure social maximizer.

The utility function (2) is based on two separated ideas: *social-welfare maximization* and *maximin criterion*.[8]

The first in proposing the concept of social-welfare maximization was Jeremy Bentham. Therefore, this utility function is sometimes called *Benthamian*. Social-welfare maximizer maximizes the social income. In mathematical terminology, the social-welfare maximizer's utility function is the weighted sum of individual incomes. An evidence of existence of agents with this type of utility function are Andreoni and Miller (2002) and Charness and Rabin (2002), themselves. Charness and Rabin's model assumes equal weights for each player.[9]

The other concept is so-called *maximin criterion*, or *Rawlsian justice*, due to John Rawls (1971). He argues that we reach a perfect justice by maximizing the minimum income in the society that is, $U_i(\pi_1, ..., \pi_n) = \min_j \{\pi_j\}$.

We, now, provide the prediction of the model in the ultimatum game:

Proposition 3 *Consider ultimatum game with Responder and Proposer having utility functions (2). The equilibrium offer of Proposer is* 0 *if* $\lambda_P < \frac{1}{1+\delta_P}$, $\frac{1}{2}$ *if* $\lambda_P > \frac{1}{1+\delta_P}$, *and any* $c \in \left[0, \frac{1}{2}\right]$ *if* $\lambda_P = \frac{1}{1+\delta_P}$. *Responder never rejects.*

Unless $\lambda_P = \frac{1}{1+\delta_P}$, only very fair or very unfair offers are predicted and Responders never reject. Both results are in contrast with intermediate values observed in experiments. In addition, observe that, since Responder never rejects, the second stage is irrelevant and the predictions of ultimatum and dictator games should coincide. We know that it is not the case in the lab.

This results show that main problem of Charness and Rabin's model: the absence of negative emotions. Similar feature arises in the next result:

[8] The utility function, which maximizes the social welfare is sometimes called *utilitarian* in the literature.

[9] Observe that in (2), the weight of i's own and the worst-off player's payoffs are, in fact, larger, due to the other parts of individual's utility function.

Proposition 4 *In equilibrium of public good game with punishment played by players with utility functions (2), there is no punishment in second stage and*

$$g_i = \begin{cases} 0 & if \quad \lambda_i < \frac{1-a}{(n-1)a(1-\delta_i)+\delta_i} \\ y & if \quad \lambda_i > \frac{1-a}{(n-1)a(1-\delta_i)} \\ \max\{g_{-i}\} & \lambda_i \in (\frac{1-a}{(n-1)a(1-\delta)+\delta_i}, \frac{1-a}{(n-1)a(1-\delta_i)}) \end{cases} .$$

For , $\lambda_i = \frac{1-a}{(n-1)a(1-\delta)+\delta_i}$, $g_i \in [0, \max\{g_{-i}\})$ *and for* $\lambda_i = \frac{1-a}{(n-1)a(1-\delta_i)}$, $g_i \in (\max\{g_{-i}\}, y]$. *Therefore, no contribution, full contribution, and k players play* $g_i = y$ *and* $n-k$ *players play* $g_i = 0$ *can be equilibrium outcomes.*

Also in this case, the prediction of Charness and Rabin's model is rejected by observed behavior. Even if it allows for equilibria with high contribution levels, it is independent of punishment. The experimental evidence on public good game with and without punishment shows that punishment option is crucial to induce high contribution levels.

The concept of *inequity aversion,* developed simultaneously by Bolton and Ockenfels (2000) and Fehr and Schmidt (1999), is based on relative income hypothesis. Although all the models were published in last decade the idea that people care about their position in society dates back to Veblen (1922).

Bolton and Ockenfels extend Bolton's (1991) two-person model. Bolton suggests that individuals care about both their absolute and relative payoff. He defines:

$$U_i(\pi_i, \pi_j) = u_i(\pi_i, \frac{\pi_i}{\pi_j}) \tag{3}$$

such that u_i is strictly increasing in the first argument and weakly increasing with respect to the second if $\pi_i < \pi_j$. For $\pi_i \geq \pi_j$, $\frac{\partial u_i(\pi_i, \pi_i/\pi_j)}{\partial(\pi_i/\pi_j)} = 0$. In words, an agent suffers if he gets less, but in the other case, he becomes self-interested. In dictator game, it predicts only zero-offers, but on the other hand, low-ultimatum game offers can be rejected. Nevertheless, (3) is not applicable when more than two players are involved. Therefore, Bolton and Ockenfels propose the following form:

$$U_i = u_i(\pi_i, \varpi_i) \text{ with } \varpi_i = \begin{cases} \frac{\pi_i}{\sum_j \pi j} & \text{if } \sum_j \pi_j > 0 \\ 1/n & \text{if } \sum_j \pi_j = 0 \end{cases} \tag{4}$$

with u_i weakly increasing and concave in material well-being of player i, π_i. The conditions with respect to the second argument are: (a) strict concavity of (4) in ϖ_i for $\sum_j \pi_j > 0$, and (b) for a fixed material payoff, (4) achieves its unique maximum when agent i's material payoff equals the average share, i.e. $\pi_i / \sum_j \pi_j = 1/n$. The fact that this is independent of the distribution among i's opponents is crucial when comparing this model to Fehr and Schmidt, below.

Bolton and Ockenfels define two threshold shares, r_i and s_i, that are crucial in predicting behavior using their model. The former denotes the share that maximizes agent i's utility. It reflects how much the agent is willing

to deviate upwards (to increase her share and, consequently, material payoff) from the equitable share to maximize her utility. In two-player case, it is represented by the offer Proposer makes in dictator game. Formally, $r_i(\sum_j \pi_j) = \arg\max_{\varpi_i} \; u_i(\varpi_i \sum_j \pi_j, \varpi_i)$.

The latter threshold is the lowest share a player still prefers rather than zero for everybody. It, for instance, determines the Responder's rejection threshold in Ultimatum game. In mathematical formulation, $s_i(\sum_j \pi_j)$ is defined such that $u_i(s_i \sum_j \pi_j, s_i) = u_i(0, 1/n)$.

By definition, $r_i(.) \in \left[\frac{1}{n}, 1\right]$ and $s_i(.) \in \left(0, \frac{1}{n}\right]$, and the strict concavity of (4) with respect to ϖ_i ensures both their existence and unicity.

First, we derive the prediction of this model in the ultimatum game:[10]

Proposition 5 *Consider an ultimatum game with Responder and Proposer having the utility function (4). In the unique equilibrium, Proposer offers*

$$\max\{s_R(.), 1 - r_P(.)\}$$

and Responder accepts this offer.

As shown above, Bolton and Ockenfels's model explains positive offers in dictator and ultimatum games. If $1 - r_P(.) < s_R(.)$, dictator-games offer are lower than in case of ultimatum games. Therefore, if most of people in ultimatum-game experiments satisfy this relation, we would, actually, observe that Dictators offer less than Proposers. Furthermore, agents highly concerned with the share can offer the same in both games. Indeed, in experiments, there are subjects who offers half the share in Dictator game, even if Responder cannot do anything but accept.

Let us proceed to a characterization of the optimal behavior in the second game:

Proposition 6 *Consider the public good game with punishment, played by n players, who have the utility function (4). Then:*

1. *No contribution and no punishment is an equilibrium.*
2. *If there is at least an i such that $r_i > \frac{1}{n}$ and $c \geq (n-1)^{-1}$, no common contribution scenario such that $g_i = \overline{g} \in (0, y]$ is an equilibrium.*

The specification of Bolton and Ockenfels's utility function and the fact that if a player punishes she does not mind who she punishes causes that we cannot provide more specific result than Proposition 2. Nevertheless, few conclusions can be made. First, the proof shows that a player can be punished due to a misbehavior of other players. This is clearly inconsistent with experimental evidence. Second, as Proposition 2 shows that, even with low c (relative to n), only zero contribution are made in equilibrium. However, many experimental

[10] The following proposition and its proof is a simplification of the discussion of equilibrium behavior in Bolton and Ockenfels (2000).

studies use $c > (n-1)^{-1}$ and still observe high contributions (Carpenter (2007), Fehr and Gachter (2000), Isaac and Walker (1988)), violating the prediciton of Bolton and Ockenfels's model.

We cannot prove that high cannot be sustained in equilibrium, but the fact that people compare themselves with the average, rather the deviators, and, consequently, any punisher may punish any other player with positive probability, suggests that it will be difficult have a positive contributions in an equilibrium.

Fehr and Schmidt offer a different model of the same idea. Their individuals are concerned about the payoffs of others, relative to their own material well-being. Fehr and Schmidt's type of inequity aversion is self-centered, in the sense that the reference point each individual compares with is her own material payoff.[11] Formally,

$$U_i(\pi_1, ..., \pi_n) = \pi_i - \frac{\alpha_i}{n-1} \sum_{j \neq i} \max\{0, \pi_j - \pi_i\} - \frac{\beta_i}{n-1} \sum_{j \neq i} \max\{0, \pi_i - \pi_j\} \tag{5}$$

with $0 \leq \beta_i \leq \alpha_i$ and $\beta_i \leq 1$. In (5), agents' utility function has three components: agents derive utility from their own material payoffs, they feel envy toward agents who get more than they do, and feel guilt towards those who are worst off. The condition $\beta_i \leq \alpha_i$ reflects that being better off creates less disutility than being worse off. Note that there is no upper limit for α_i what practically allows any level of envy. This model goes along the lines of experimental findings of Loewenstein *et al.* (1989) who discover that their subjects have non-linear utility functions, similar to that proposed by Fehr and Schmidt.

Let us state the prediction of Fehr and Schmidt's model in the Ultimatum game:

Proposition 7 *Consider an ultimatum game played by players with Fehr and Schmidt's utilities. It is a dominant strategy for a Responder to accept any* $c \geq \frac{1}{2}$ *and to reject if* $c < \frac{\alpha_R}{1+2\alpha_R} < \frac{1}{2}$ *and to accept otherwise. Then, in equilibrium,*

$$\textit{Proposer offers } c \begin{cases} = 0.5 & \textit{if} \quad \beta_P > \frac{1}{2} \\ \in \left[\frac{\alpha_R}{1+2\alpha_R}, \frac{1}{2}\right] & \textit{if} \quad \beta_P = \frac{1}{2} \\ = \frac{\alpha_R}{1+2\alpha_R} & \textit{if} \quad \beta_P < \frac{1}{2} \end{cases}.$$

Proof. See Fehr and Schmidt [21], Proposition 1. ■

Proposition 7 shows that Fehr and Schmidt's model can explain positive offers and rejection of low offers in ultimatum game. However, note that all offers between zero and one half are positive only due to Responder's threat of rejection. Thus, in Dictator game, we should only observe offers of either zero or one half. This goes against the evidence, because a significant fraction of subjects gives less than in ultimatun games, being the gift still positive.

[11] Fehr and Schmidt argue that inequity aversion should be based on a neutral reference point that allows to perceive (un)fairness. Since the reference point in their model is agent's own payoff, they prefer to use the term *self-centered* inequity aversion.

Nevertheless, this problematic feature of Fehr and Schmidt can be solved by a slight two-player variation of (5):

$$U_i(\pi_i, \pi_j) = \pi_i - \sum_{j \neq i} \left[\alpha_i \max\{0, \pi_j - \pi_i\} + \beta_i \max\{0, \pi_i - \pi_j\}\right]^2. \quad (6)$$

Proposition 8 *Let players have the utility function (6). Then:*

1. *If Dictator's guilt parameter is high enough, she can offer positive amounts of money. If so, she keeps for herself* $\pi_D = \dfrac{1 + 4\beta_D^2}{8\beta_D^2} \geq \dfrac{5}{8}$.

2. *In ultimatum game, Responder accepts any* $c \geq \frac{1}{2}$. *In equilibrium, Proposer proposes* $c = \max\left\{\dfrac{4\beta_P^2 - 1}{8\beta_P^2}, \dfrac{1 + 4\alpha_R^2 - \sqrt{1 + 8\alpha_R^2}}{8\alpha_R^2}\right\}$ *and Responder accepts this offer.*

Put Figure 2. around here

Figure 2 shows how much Dictator keeps for herself in dictator game as a function of Dictator's parameter of inequity aversion β. This non-linear version of (5) can explain positive giving in dictator game if the Dictator is concerned enough with inequity aversion. In particular, if $\beta_D > \frac{1}{2}$ Dictator does not keep for herself the whole amount. For high values of β_D she can offer the opponent until 40% of the stake divided. Furthermore, if Proposer is non-selfish enough, she can offer high shares in both ultimatum and dictator games, as actually observed in experiments. The behavior of Responders has very similar features under both linear and non-linear specifications, even if the rejection level is slightly lower in the non-linear specification. Thus, the proposed non-linear model provides good predictions in both games.

Before we state the equilibrium prediction of Fehr and Schmidt's model in the public good game with punishment, we have to introduce a concept of a *conditionally cooperative enforcer*. Fehr and Schmidt define a conditionally cooperative enforcer as an individual, who "is suffcientely concerned about inequity (p.841)". The following proposition contains the formal definition.

Proposition 9 *Suppose there is a group of* $n' \in [1, n]$ *conditionally cooperative enforcers with preferences (5) obeying* $a + \beta_i \geq 1$ *and*

$$c < \frac{\alpha_i}{(n-1)(1+\alpha_i) - (n'-1)(\alpha_i + \beta_i)}$$

for all $i \in \{1, ..., n'\}$, *whereas all other players do not care about inequality. Then, the subgame perfect equilibrium is:*

- *In the first stage each player contributes* $g_i = g \in [0, y]$

- *If each player does so, there is no punishment. If a non-enforcer deviate downwards, each enforcer chooses* $p_{ji} = (g - g_i)/(n' - c)$*, while any of the other players do not punish.*

Proof. See Fehr and Schmidt [21], Proposition 4. ■

Although Proposition 9 also provides a condition when high level of cooperation can be sustained in equilibrium, low contribution equilibria are equally likely. Fehr and Schmidt argue that full contribution is a natural focal point, but, using the same argument, no contribution can be considered a natural focal point as well. Hence, neither Fehr and Schmidt nor Bolton and Ockenfels (2000) provide any particularly strong argument why mostly high contribution levels are observed in experiments with punishment option.

Even if Fehr and Schmidt do not allow for all values of model parameter, disregarding their constraints brings on different versions of their model which can provide interesting cases. For example, if $\beta_i < 0 < \alpha_i$ the individual utility can be written as

$$U_i(\pi_1, ..., \pi_n) = \gamma_i \pi_i - \varepsilon_i \sum_{\pi_j > \pi_i} \pi_j - \epsilon_i \sum_{\pi_j < \pi_i} \pi_j$$

what is known in the literature as status-seeking. Depending on the relation of $|\beta_i|$ and $|\alpha_i|$, the individual can treat worse those that are better off. Another case is to simply allow for $\alpha_i < 0 < |\alpha_i| < \beta_i \leq \frac{1}{2}$. In such a case, agent is a social-welfare maximizer with weights on her and her opponents' payoffs being $(1 - \alpha_i - \beta_i)$ and $\frac{(\alpha_i + \beta_i)}{n-1}$, respectively.[12]

Let us, now, discuss the difference between Fehr and Schmidt's and Bolton and Ockenfels's model in more detail. For two-player games, both Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) obtain similar predictions. The two-player version of (5) is:

$$U_1(\pi_1, \pi_2) = \pi_1 - \alpha_1 \max\{0, \pi_2 - \pi_1\} - \beta_2 \max\{0, \pi_1 - \pi_2\}. \tag{7}$$

Note that (7) is two-piece-wise linear. Since it increases for $\pi_2 > \pi_1$ and decreases otherwise, it is concave in 1's material payoff. For a fixed material payoff, the maximum with respect to opponent's payoff is achieved if 1's share is $1/n = 1/2$. Moreover, (7) is strictly increasing in 1's payoff, for a given payoff difference. Similarly to (4), the utility function can only be decreasing in her own payoff if her payoff advantage is too large.

The crucial difference is that the role of share is in (7) replaced by payoff difference. Another difference between the two models is the functional form with respect to inequity part. Recall that Bolton and Ockenfels assume a strict concavity what allows for a unique $s_i(.)$ and $r_i(.)$, while this is not, in general, the case in Fehr and Schmidt's specification, since the relation is linear and, therefore, not strictly concave.

[12] It is important to remark that predictions of Fehr and Schmidt's model and, in particular, Propositions 7 and 9 might change dramatically under these different specifications.

The contrast arises in multiperson games. In Fehr and Schmidt, each player compares herself with each player separately, while, in Bolton and Ockenfels, she compares herself with the average. In this latter case, whether $\partial_{\pi_j} U_i$ is positive or negative does not depend on whether j is worse or better off than i, but exclusively on i's position in comparison with the average, i.e. if π_i is below the average she can even lower payoffs of worse-off opponents. This feature is the crucial difference between the 2 models.

Let us illustrate the difference on a simple example. Consider, first, an agent i who has to choose a division of 6 monetary units among three agents. The three options are (2,2,2), (2,1,3), and (2,0,4), such that the first element is for agent i and the second and third for other players. A Bolton and Ockenfels's individual is completely indifferent which one to choose, while Fehr and Schmidt's agent prefers (2,2,2) to any other and (2,1,3) to (2,0,4).

Engelmann and Strobel (2004) provide a direct experimental comparison of both versions of inequity aversion. Their results argue in favor of Fehr and Schmidt's formalization. This setting performs better in their experiment. However, the general performance of inequity-aversion models in their experiment is very poor in comparison with other preference types.

Moreover, as argued by Charness and Rabin (2002), the complete anonymity among experimental subjects in experimental lab makes the interpersonal comparison based exclusively on material gains from the play. Hence, since this prevents subjects to take into the account other aspects (such as deserveness, sympathy etc.), this feature of experiments makes payoffs and, consequently, payoff differences salient. The role of inequity aversion can, then, be exaggerated in experiments.

The above argumets suggest that something is missing in the model of this section. Hence, we proceed, in the next chapter, with the role of intentions.

3.2 Reciprocity

In this subsection, we try to review literature where people are concerned with the behavior of others and, as a response to this behavior, apply their emotions in the decision process. There are observations that the same people are kind in some situations or toward some individuals, and spiteful in other situations or toward other people. In our case, we focus on the notion of reciprocity. The literature distinguishes two types of reciprocity, negative and positive. Negative reciprocity is the tendency to punish unfair or unkind behavior. Ultimatum game and public good game with punishment are two typical games where negative reciprocity is observed. In ultimatum game, Responders can punish unfair offers by rejecting them. In public good game with punishment, people observe the contributions made by others and eventually punish them for their behavior. As mentioned in Section 2 above, a self-interested agent neither reject nor punish and the evidence contradicts this prediction. Blount (1995) directly tests the hypothesis that people reject in ultimatum game because they find the offers mean. She runs an ultimatum game experiment where the offers are generated by a computer and Responders know it. The potential difference between ulti-

matum game and this setting can point out the role of reciprocity in rejection behavior. Actually, even low offers are rarely rejected. The experimental studies show that the punished subjects are those with the lowest contribution in Public good game with punishment.

Experimental examples of positive reciprocity are gift-exchange and trust game. In the gift exchange game, an Employer offers a wage to a Worker who, consequently, exhibits effort. Given that Employer cannot condition the wage on effort it is optimal for a selfish agent to make the lowest possible effort, independently of the offered wage. In the trust game, Investor can invest an amount by giving it to Receiver. The invested amount is multiplied by three, and Receiver have to decide how much he returns to Investor. Fehr *et al.* (1997), for instance, experimentally test gift exchange game in labor market framework. They find a positive relation between the salary offered and "returned" effort. The higher offer the employer makes (the better the employer treats), the more effort the worker performs (the more the worker repays to the employer). Similarly, in case of trust game, Berg *et al.* (1995) find positive correlation between the money sent by the Investor and the amount returned by the Receiver.[13]

The models discussed in this section use *psychological game theory*, proposed by Geanakoplos *et al.* (1989), and later generalized by Battigalli and Dufwenberg (2005). Psychological game theory introduces personal beliefs into the analysis what, as explained below, allows for direct modeling and evaluation of intentions.

The first attempt to directly model reciprocity is Rabin (1993). His model is built on three stylized facts: people helps those who are being kind with them (positive reciprocity), punish those who are treating them meanly (negative reciprocity) and, as the material stakes rise, the above effects become weaker. He follows Geanakoplos *et al.* (1989) by allowing payoffs to depend on their actions and beliefs, simultaneously. He argues that first and second-order beliefs are sufficient to model intentions and suggests methodology how to deal with two-players normal-form games. Rabin simplifies the analysis of psychological game theory by assuming that $\sigma_i \in \Sigma_i$, $\sigma'_{ji} \in \Sigma_i$ and $\sigma''_{iji} \in \Sigma_i$, such that σ_i denotes strategy chosen by player i, σ'_{ji} player j's beliefs about the action of player i, and σ''_{iji} player i's belief about what player j beliefs i plays.[14]

We proceed directly to the discussion of Dufwenberg and Kirchsteiger (2004), who provide a generalization of Rabin's model for N-persons extensive-form games. Let H be the set of histories leading to subgames, and A_i the set of behavior strategies of i. Then, $a_i(h)$ is a behavioral strategy of player i,

[13] See Section 1 of Rabin (1993) for a more exhaustive discussion of evidence on both negative and positive reciprocity from psychology and economics. Camerer (2003) surveys the experimental evidence on the issue of reciprocity.

[14] In psychological game theory, first-order belief is a probability measure over the space of other players' mixed strategies. So, the set of first order beliefs is: $S'_i :=\triangle (\Sigma_{-i})$. The set of second-order beliefs is defined as $S''_i :=\triangle (\Sigma_{-i} \times S'_i)$. Hence, Rabin (1993) simplifies considerablly the analysis by defining that $S'_i \equiv \Sigma_j$ and so on. The term belief is also used in games with incomplete information and is used to denote the probability of being in a particular node x in an information set (see, for example, Vega Redondo (2003), p.118).

for a given history. The definition of behavioral strategy differs from standard game theory though. In standard game-theoretical approach, it is completely irrelevant for the decision in a particular node what happened in the past or which path lead into the node. The only relevant information for the decision in that node is the information conveyed by the final nodes that follow it. This differs in psychological game theory, where beliefs enter the utility function. Thus, what happens with the beliefs about the play preceding a node is as relevant as what happens as the game unravels afterwards. To deal with this problem, Dufwenberg and Kirchsteiger suggest that beliefs up-date correctly; that is, the actions taken in the past are believed to be played with probability 1. This feature of their model causes that what had been predicted to be done in a particular node may change once this node has been reached; something that cannot happen in standard game theory.

Define a set of efficient strategies as

$$E_i = \left\{ \begin{array}{c} a_i \in A_i \mid \nexists\ a_i' \in A_i \text{ s.t. for all } h \in H,\ (a_j)_{j\neq i} \in \prod_{j\neq i} A_j\ , \\ \text{and } k \in N \text{ it holds that } \pi_k(a_i'(h),(a_j(h))_{j\neq i}) \geq \pi_k(a_i(h),(a_j(h))_{j\neq i}), \\ \text{with strict inequality for some } (h,(a_j)_{j\neq i},k) \end{array} \right\}.$$

Intuitively, a strategy is inefficient if there exists another strategy which, conditional on any history of play and subsequent choices by the others, provides no lower material payoff for any player, and a higher material payoff for some player for some history of play and subsequent choices by the others (Dufwenberg and Kirchsteiger (2004), p.9). Let $(b_{ij})_{j\neq i} \in \prod_{j\neq i} B_{ij} = \prod_{j\neq i} A_j$ be i's beliefs about what strategies her opponents are playing. An agent deduces a kind or mean action of an opponent comparing the realized payoff with

$$\pi_j^{ei}((b_{ij})_{j\neq i}) = \frac{1}{2}\left[\max\left\{\pi_j(a_i,(b_{ij})_{j\neq i}) \mid a_i \in A_i\right\} + \min\left\{\pi_j(a_i,(b_{ij})_{j\neq i}) \mid a_i \in E_i\right\}\right].$$

Thus, the reference point is the average of the highest feasible payoff and the lowest efficient payoff, given beliefs. Define the kindness of player i toward j, at history h, as a function $\kappa_{ij} : A_i \times \prod_{j\neq i} B_{ij} \longrightarrow \mathbb{R}$:

$$\kappa_{ij}(a_i(h),(b_{ij}(h))_{j\neq i}) = \pi_j(a_i(h),(b_{ij}(h))_{j\neq i}) - \pi_j^{ei}((b_{ij}(h))_{j\neq i}).$$

In words, the kindness of agent i toward j is proportional to the kindness expected from j. The sign is determined by the comparison of the actual payoff of player j with the above reference point.

Analogously, the second-order beliefs, $(c_{ijk})_{k\neq j,j\neq i}$, reflect i's beliefs about what j thinks that k plays. Observe that $i = k$ is allowed what states for i's beliefs about j's beliefs about what i herself will do. Then, $\lambda_{iji} : B_{ij} \times \prod_{k\neq j} C_{ijk} \longrightarrow \mathbb{R}$ such that

$$\lambda_{iji}(b_{ij}(h),(c_{iji}(h))_{k\neq j}) = \pi_i(b_{ij}(h),(c_{iji}(h))_{k\neq j}) - \pi_i^{ei}((c_{iji}(h))_{k\neq j})$$

measures how i believes treated by j.

At this level, we can define the utility function $U_i : A_i \times \prod_{j \neq i}(B_{ij} \times \prod_{k \neq j} C_{ijk}) \to \mathbb{R}$ such that

$$U(a_i(h), (b_{ij}(h), (c_{iji}(h))_{k \neq j})_{i \neq j}) =$$

$$\pi_i(a(h)) + \sum_{j \in N \setminus i} Y_{ij} \cdot \kappa_{ij}(a_i(h), (b_{ij}(h))_{j \neq i}) \cdot \lambda_{iji}(b_{ij}(h), (c_{ijk}(h))_{k \neq j}) \qquad (8)$$

where Y_{ij} is an exogenously given non-negative number for each $j \neq i$, measuring how sensitive is i regarding player j.

Dufwenberg and Kirchsteiger's model introduces the idea of *conditional cooperation*. In sequential games, people can decide whether to treat kindly or meanly on basis of past actions of their opponents. Conditional cooperation is commonly observed in experiments. Experimental evidence on public good game without punishment option suggests that conditional cooperation can be a possible explanation of gradual decreasing of contributions over time.[15]

Since the above model is very different from the preceding, Dufwenberg and Kirchsteiger define a concept of *sequential reciprocity equilibrium* (SRE, p. 278). First, for a given $a_i(h) \in A_i$ and h, let $A_i(h, a)$ be the set of strategies i may use if she behaves according to the behavioral strategy $a_i(h)$ in all $h' \neq h$, but is free to use any strategy at h. With this notation at hand, the equlilibrium concept can be stated as follows:

Definition 10 *The profile $a^* = (a_i^*)_{i \in N}$ is a SRE if for all $i \in N$ and for each history $h \in H$, it holds that:*

(a) $a_i^*(h) \in \arg \max_{a_i \in A_i(h, a^*)} U_i(a_i, (b_{ij}(h), (c_{ijk}(h))_{k \neq j})_{j \neq i})$,

(b) $b_{ij} = a_j^*$ *for al* $j \neq i$,

(c) $c_{ijk} = a_k^*$ *for all* $j \neq i, k \neq j$.

The first part of the definition is parallel to the standard equilibrium concepts: all players play the best response to strategies of all the other players. The part *(b)* and *(c)* requires the beliefs to match in equilibrium. Thus, in equilibrium, beliefs have to be consistent with the actual play in the game.

The definition allows us to state the predictions of Dufwenberg and Kirchsteiger's model in the ultimatum game:[16]

Proposition 11 *In a SRE of Ultimatum game played by Proposer and Responder with utility functions (8), Responders:*

[15] See Fischbacher *et al.* (2001) for a direct test of this hypothesis.

[16] Dufwenberg and Kirchsteiger (1998) sketch an informal proof of their prediction in a slightly different ultimatum game in the working paper version. Here, we provide the formal proof for our specification of the game in Appendix A.

- *accept any* $c > \frac{Y_R}{Y_R+2}$
- *reject any* $c < \frac{2+3Y_R-\sqrt{4+12Y_R+Y_R^2}}{4Y_R}$ *and*
- *may both accept or reject any* $c \in \left[\frac{2+3Y_R-\sqrt{4+12Y_R+Y_R^2}}{4Y_R}, \frac{Y_R}{Y_R+2}\right]$.

There exist equilibria, where Proposers offer the lowest acceptable offer. For for Y_P *high enough and* $Y_R > 0$ *, there exist equilibria, in which the offer is rejected.*

As shown in Proposition 11, very high and very low offers, respectively, are accepted and rejected with probability 1, and there is a range of values, for which any strategy of Responder is optimal. This is due to the belief consistency condition in Definition 10. These offers are characterized by self-fulfilling prophecies: if negative beliefs prevail, offers from this range will be rejected, while positive beliefs may lead to the acceptance of the same offer.

Concerning the behavior of Proposer, since Responder can never be nice to Proposer,[17] the latter maximizes his utility, offering the lowest acceptable offer. On the other hand, if Proposer's psychological part of the utility is important enough to overweight her selfishness, she may prefer zero for the Responder, if she expects rejection, and state a low enough offer, which would be rejected with probablity 1. Such offer exists for any $Y_R > 0$.

Observe that this model offer no insight into the dictator-game evidence. Since the Responder can be neither kind nor mean in this game, Proposer simply maximizes her material payoff. Thus, the unique prediction of this model would be $c = 0$, reflecting the problematic feature of pure intentionalist models.

We proceed with the public good game with the punishment option:

Proposition 12 *In SRE of the public good game with punishment played by players having utility function (8):*

1. *Agent who contributes equal or more than* $\frac{y}{2}$ *and does not punish is never punished.*
2. *Agent* i *punishes* j *if* $p_{ji} > a(g_j - \frac{y}{2}) - \frac{c}{Y_{ij}}$.
3. *If, for all* i, $\sum_{j\neq i} Y_{ij} > \frac{2(1-a)}{a^2 y}$, $g_i = y$ *and no punishment for each* i *is an equilibrium.*
4. *No common contribution such that* $g_i = \overline{g} \in \left(\frac{y}{2}, y\right)$ *for each* $i \in N$ *can be an equilibrium, unless for all* $i \in N$, $\sum_{j\neq i} Y_{ij} = \frac{1-a}{a^2(\overline{g}-\frac{y}{2})}$.
5. *A common contribution such that* $g_i = \overline{g} \in (0, \frac{y}{2}]$ *for each* $i \in N$ *cannot be an equilibrium.*

[17] Note that rejection is completely inefficient. This implies that the equitable payoff of Proposer is $1 - c$. Hence, Responder can only be neutral accepting or mean rejecting.

6. *No contribution without punishment can be an equilibrium if each i has Y_{ij} low enough for each opponent j.*

The predictions of this model in public good game with punishment contrast with the observations in the laboratory. Experimental subjects, who deviate downwards from a common contribution, use to be punished independently of the level of common contribution that is, even if they still contribute more than $\frac{y}{2}$. The second part predicts that the less j contributes, the more seriously she is punished by others, in harmony with observations. The third part shows that if everybody is enough concerned with reciprocity, full contribution is an equilibrium. However, the cooperative equilibrium is sustained by high social concern, rather than the threat of being punished, as observed in the lab. Note that, since this prediction is independent of the punishment stage, it extends into the public good game without the punishment option. This illustrates even better why the prediction of this model regarding the cooperative outcome does not match the stylized facts. Moreover, since too small deviations are not worth of punishment, only the selfish part of deviator's utility is affected by this small deviation, which increases. Therefore, common contribution scenarios are mostly not sustained in equilibrium.

To summarize, the predictions of Rabin's and Dufwenberg and Kirchsteiger's model are in some sense in contrast with the experimental evidence. The main reason is that they deal with pure intentions. In the next section, we discuss models that combine distributional and intention-based models.

3.3 Combinations

In Falk *et al.* (2003), subjects play four mini-ultimatum games from Figure 3. In all games, Proposer chooses which option to offer to Responder who either accepts and the proposed distribution realizes, or rejects and both get nothing. Falk *et al.* concentrates on rejections of the (8,2) offer, which appears in all games and allocates 8 monetary units to Proposer and 2 units to Responder. Let us call each game according to the alternative distribution. There are two possible hypothesis. First, purely distributional models predict that the rejection rates of (8,2) offer do not change across the four games, because the payoff consequences are always the same. In particular, acceptance leads to distribution of (8,2) and rejection to (0,0) in all four left-hand parts of the four games. Distributional models, in other words, predict that the unchosen alternatives play no role. Intentionalist models, on the other hand, suggest decreasing rejection rates of this order: the highest rejection of (8,2) distribution in 5/5 game, following with 2/8 game, 8/2 and the lowest in 10/0 game.[18] Since Proposer is forced to offer (8,2) distribution in 8/2 game, the rejection rate measures pure distributional motives. Figure 4. plots the rejection rates of the (8,2) offers of

[18] Choosing (8,2) over (5,5) in 5/5 game shows Proposer's bad intentions, since a reasonable, completely equitable allocation exists. In 2/8, a big sacrifice has to be made to offer Responder better alternative. In 8/2 game, intentions play no role and in 10/0, (8,2) distribution signals good intentions.

the four games. The ranking of rejection rates is exactly as intention-based models predict. Moreover, observe that there is a almost 20% of subjects rejecting the (8,2) allocation, even in the absence of any intention, and around 10% still rejecting (8,2) allocation, even in the presence of only good intentions, benefiting so the arguments of distributional models. This intelligent test shows that the combination of both distributional and intention-based models is necessary to explain experimental evidence. This chapter reviews models that attempt to combine both types of concepts.

Put Figure 3. around here.

Put Figure 4. around here.

Levine (1998) proposes the following utility function:[19]

$$U_i = \pi_i + \sum_{j \neq i} \frac{\pi_j (a_i + \lambda a_j)}{1 + \lambda} \tag{9}$$

where $\lambda \in [0, 1]$ and for all $i \in N$, $a_i \in (-1, 1)$, is the altruism parameter of player i. Observe that the utility function depends on the opponents' altruism. Referring to λ, two cases can occur: $\lambda = 0$ and $\lambda > 0$. In the first case, i's utility reduces to $U_i = \pi_i + a_i \sum_{j \neq i} \pi_j$. Then, if $a_i > 0$ ($a_i < 0$) agent i is altruist (spiteful) and we have a purely distributional model. This case cannot explain while a person behaves altruistically in some situations and spitefully in others. The other case is $\lambda > 0$. If such, i behaves more altruistically toward altruistic players and viceversa. Moreover, if $0 \leq a_i < -\lambda a_j$ an altruistic player can behave spitefully. This logic is parallel to modelling reciprocity.

Levine differs in a an important way from the rest of models discussed in this survey. Since Levine's individual utility function depends on the coefficient of altruism of opponents which, in practice, characterizes j's utilities, the utility of agents depend on opponents utilities, rather than payoffs or actions. This approach is broader that simple payoff interdependence. It seems reasonable to believe that individuals treats better good people and meanly mean ones. The problem of this model is that the parameters of other agents' utilities are not observable. They are deduced from actions of others what, on the other hand, serves as an argument in favor of models of the previous and this section, which model the perception of intentions using payoffs and actions. Levine is aware of that and argues that players may reveal information about their altruism coefficient through their play (p.598).

Let us illustrate the properties of this model on ultimatum game predictions:

[19] Sethi and Somanathan (2003) argue that Levine's (1998) specification of preferences does not survive the evolutionary arguments and propose an alternative

$$U_i = x_i + \sum_{j \neq i} \frac{x_j (a_i + \lambda (a_i - a_j))}{1 + \lambda}.$$

Such a utility function does not have the evolutionary problems of Levine (1998).

Proposition 13 *Consider ultimatum-game Proposer and Responder with utility (9). In equilibrium, Proposer offers*

$$c \begin{cases} = 0 & \text{if} \quad a_R + \lambda_R a_P \geq 0 \text{ and } a_P < 1 \text{ and/or } a_R < 1 \\ = -\frac{a_R + \lambda_P a_P}{1 + \lambda_R - a_R - \lambda_R a_P} & \text{if} \quad a_R + \lambda_R a_P < 0 \text{ and } a_P < 1 \text{ and/or } a_R < 1 \\ \in [0,1] & \text{if } a_P = a_R = 1 \end{cases}$$

and Responder accepts.

The first case shows the weakness of this model. In this case, both Proposer and Responder are positively concerned with each other's payoff, but this feature allows Proposer to offer nothing. In experiments, however, offering nothing is considered unfair and widely rejected. This reflects why it is important to get rid of this type of equilibria and the existence of such equilibria is a drawback of Levine's model. The second case - $a_R + \lambda_R a_P < 0$ - seems to match the best the experimental evidence: it predicts a positive offers depending on the altruism of the Responder. Hence, the signaling of types through past play is very relevant in this model. Levine calibrates the possible types of opponents using a particular distribution of types, but does not provide any particular way, how to deduct the altruism parameters of opponents from the past play. The prediction for the public good game with punishment leads to a similar problem:

Proposition 14 *In equilibrium of public good game with punishment played by players with utility (9):*

1. *Second-stage punishment is completely independent of contributions made in the first stage of the game and agent i punishes only agents j with the altruism parameter low enough.*
2. *If for each i, $\sum_{j \neq i} a_j > \left[\frac{(1-a)(1+\lambda_i)}{a} - (n-1)a_i \right] \frac{1}{\lambda_i}$, all players contribute fully.*
3. *If for each i, $\sum_{j \neq i} a_j < \left[\frac{(1-a)(1+\lambda_i)}{a} - (n-1)a_i \right] \frac{1}{\lambda_i}$, nobody contributes.*
4. *For i such that $\sum_{j \neq i} a_j = \left[\frac{(1-a)(1+\lambda_i)}{a} - (n-1)a_i \right] \frac{1}{\lambda_i}$, any $g_i \in [0, y]$ is equilibrium.*
5. *There can be equilibria with some players contributing fully, some contributing nothing and some contributing any $g_i \in [0, y]$.*

Again, similar problems as in ultimatum game appear here. The punishment stage is completely independent on contribution in stage 1, in contrast with laboratory experiment where the first stare contributions signal the altruism of players.

Charness and Rabin (2002), apart from the reciprocity-free version from Section 3.1, propose a model that combines intentions and distribution of payoffs.

Since Charness and Rabin observe no positive reciprocity in their experiment, they develop a model capturing what they call *concern withdrawal*[20] and combine it with the purely distributional model. Charness and Rabin extend their preferences with a demerit profile, $d = (d_1, ..., d_n)$ such that $d_i \in [0, 1]$, measuring how much a player deserves from the point of view of other players. The smaller is d_i the more the others concern i's payoff. Charness and Rabin define a utility function:

$$U_i(\pi, d) = (1-\lambda)\pi_i + \lambda \left[\begin{array}{c} \delta \min \left\{ \pi_i, \min\limits_{j \neq i} \{\pi_j + b d_j\} \right\} + \\ (1-\delta)(\pi_i + \sum\limits_{j \neq i} \max\{1 - k d_j, 0\} \pi_j) - f \sum\limits_{j \neq i} d_j \pi_j \end{array} \right] \tag{10}$$

where b, k and f are non-negative model parameters.

The utility function (10) shares purely distributional features of (2) combined with that the less i believes j deserves, the less he is concerned about j's payoff. If $b = k = f = 0$ (10) coincides with (2). With f large, i even wants to hurt players with high demerit parameter.

Even if Charness and Rabin do not count for positive reciprocity, allowing d to lay between -1 and 1 can capture cases where agents reward kind behavior with kind actions.[21]

Whether a player has misbehaved is determined on basis of a certain parameter λ^* such that if a player has $\lambda < \lambda^*$ he invokes a negative reaction from his opponents. This manner of modelling reciprocity is parallel to Levine (1998). In both cases, reciprocity lies upon the observebility of individual preferences and in both models, reciprocity works as a reinforcement device for purely distributional models. The problem is that neither Charness and Rabin (2002) nor Levine (1998) offer a formal definition of reciprocity, contrary to Rabin (1993) and Dufwenberg and Kirchsteiger (2004), what reduces the generality of the model. Nevertheless, we can combine Levine (1998) and Charness and Rabin (2002) with the definition in Dufwenberg and Kirchsteiger (2004).[22] Suppose, for instance, $\lambda = \lambda(c)$ in Ultimatum game with $\lambda'(c) > 0$ and $\lambda(\frac{1}{2}) = 0$. In such a case, high (low) c reinforces (lowers) kindness of Responder toward Proposer.[23]

Charness and Rabin's (2002) specification provides interesting insights, but there are too many free parameters in their model. This makes the model practically impossible to test in the laboratory.

A very sophisticated model of combination of distributional and intention-based model is Falk and Fischbacher (2004). They integrate the ideas of reciprocity and inequity aversion into one model, using the framework of psycholog-

[20] Charness and Rabin (2002) define concern withdrawal as a situation, in which people "withdraw their willingness to sacrifice to allocate the fair share toward somebody who himself is unwilling to sacrifice for the sake of fairness"(p. 820).

[21] An alternative approach would be not to restrict parameters b, k and f to be non-negative. In such a case, d_j would take value asigning the level of reciprocity j deserves and the sign of individual parameters would determine whether the reciprocity is positive or negative.

[22] Falk and Fischbacher (2004), below, offers a different definition of reciprocal behavior, which can also be applied here.

[23] In public good game with punishment, the solution would be $\lambda = \lambda_{ij}(g_j)$.

ical game theory. Qualitatively, Falk and Fischbacher provide similar insights as Dufwenberg and Kirchsteiger (2004). Nevertheless, Falk and Fischbacher redefine the reference standard upon which agents base their perception of (un)kindness. They base it on the idea of inequity aversion of Fehr and Schmidt (1999) and propose equity-based self-centered reference point for this evaluation.

Due to the extreme complexity of the model, we relegate the technical details of this model into Appendix B and, in the main text, we only illustrate the logic of their model, using examples.

In this model, when an individual earns less than another person, he feels treated unkindly. Consider game 10/0 in Figure 3. It is reasonable suppose that if Proposer behaves kindly and chooses the (8,2)-sideshoot, any responder will accept this offer. In contrast, this model predicts that any disadvantageous distribution of payoffs is considered as undesirable. Nevertheless, the probability to reject (8,2) distribution should be lower that in case of 5/5 or 2/8 game. The experimental observation of Falk *et al.* (2003) shows that 10% of subjects, in fact, reject and that this fraction is lower that in the other cases. This empirical finding goes hand in hand with this model: People exhibit some inequity aversion even when treated as kindly as possible.

As a next step, we state the predictions of the model in both ultimatum game and public good game with punishment:[24]

Proposition 15 *The unique reciprocity equilibrium of the ultimatum game played by players with utility function (14) is:*

1. *Responder accepts with probability* $p = \begin{cases} \min\left\{\frac{c}{\rho_R(1-2c)(1-c)}, 1\right\} & \text{if } c < 0.5 \\ 1 & \text{if } c \geq 0.5 \end{cases}$.

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

Proof. See Falk and Fischbacher [15], Proposition 1. ■

In harmony with experimental evidence, more than half amount is never rejected and very low offer are rejected with very high probability. Moreover, the closer the offer to one half the more likely it is accepted. If $\frac{1+3\rho_R-\sqrt{1+6\rho_R+\rho_R^2}}{4\rho_R} > \frac{1}{2}(1-\frac{1}{\rho_P})$ the Responder's social concerns are crucial and in this case, the Dictator's offer is lower than in ultimatum game. On the other case, Proposers with higher regards may offer the same in ultimatum and dictator game. Thus, this model reproduces all the stylized facts of these 2 games.

Proposition 16 *In the equilibrium of Public good game with punishment played by n players with utility function (14):*

1. *i never punishes j whose* $g_j \geq g_i$ *and* $p_{ji} = 0$.

[24] Falk and Fischbacher (2004) define the reciprocity equilibrium as the analogue of the sequential reciprocity equilibrium of Dufwenberg and Kirchsteiger (2004).

2. *If player i punishes, her optimal punishment is* $p_{ij} = p_{ji} + (1-c)^{-1}[g_i - g_j - \frac{c}{\rho_i \vartheta(.)}]$.

3. *Any common contribution, such that* $g_i = \overline{g} \in (0, y]$, *is not equilibrium.*

4. *No contribution is an equilibrium.*

The first two parts of Proposition 4 show the conditions for punishing and the optimal punishment level. The lower is the contribution of an opponent the more i punishes her. It is consistent with experimental results. The weakness of this model is that it cannot sustain high contribution levels in equilibrium, in spite of that they are widely observed in experiments. As in the model of Dufwenberg and Kirchsteiger (2004), the reason is that any player can deviate such a small amount downwards that her payoff advantage is not worth being punished. Despite the common features of this model and Fehr and Schmidt (1999), the type of equilibria from Proposition 9 does not survive here using the same argument.

4 Discussion

This paper surveys the models of social preferences. As shown, no existing model can explain all the observational regularities in the two games we analyze. Consequently, even though it is clear that a theory combining the properties of distributional and intention-based models is required, the analysis of this paper suggests that the existing literaure should serve as a cornerstone for further empirical and theoretical modeling of human prosocial behavior.

It is worth mentioning that the aim of this survey is not to say that selfish behavior does not matter. On the contrary, Andreoni and Miller (2002), for instance, report that almost 23% of their subjects behave as completely self-interested; too significant fraction to simply reject the self-interest hypotheses. On the other hand, more than 77% of subjects seem to maximize different utilities. This fraction is also high enough in order not to stuck in simple payoff maximization, as applied economic theory standartly assumes. Hence, any suffciently realistic model should contain the self-interested individual among its particular cases. This actually occurs in case of all the models discussed in this paper.

One direction of research reports that the population is rather a mixture of preference types. Most of the results here allows for heteregneity with respect to the parameters of the models and, since self-interest is always a special case, it allows for the coexistence of selfish and prosocial individuals. Nevertheless, we do not analyze the consequences of a richer coexistence of behavioral types. Therefore, one path for research could be enriching economics and game theory by models with a more complex strucutre of population. This approach could lead to exploration of other empirical findings that cannot be explained by today's models.

In this respect, as well as reciprocity, heterogeneous population can also predict why the same people can in one situation behave as selfish and in another one as altruists. The behavior should clearly differ if treating with completely selfish individual from the treatment with an altruist. Since we suppose people to play Nash equilibria, the extended utility functions can predict that an optimal behavior differ, depending on the opponent an individual faces. It is, for instance, known that the presence of very few selfish agents drives markets toward very "unfair" competitive outcomes in market-like experiments.

Other direction of research is to develop a theoretical model general enough to encompasses great part of the existing models of social preferences. A big step in this direction is Segal and Sobel (2007), who provide a set of axioms that generate the type of models discussed in this survey. Even though they discuss the connection of their model with psychological game theory, a formal merge of the ideas of both approaches could provide a sufficiently general setup, containing both Segal and Sobel's utility functions and psychological game theory as special cases.

References

[1] Andreoni, J. and J. Miller (2002): "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism," *Econometrica* **70**(2), 737-753.

[2] Battigalli, and M. Dufwenberg (2005): *Dynamic Psychological Games*, Working paper Bocconi University.

[3] Beker, G. (1974): "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology", *JET* **14**(3), 817-826.

[4] Berg, J., J. Dickhaut and K. McCabe (1995), "Trust, Reciprocity and Social History", *Games and Economic Behavior* **10**, 122-42.

[5] Blount, S. (1995): "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences", *Organizational Behavior and Human Decision Processes* **63**(2), 131-144.

[6] Bolton, G.E. (1991): "A Comparative Model of Bargaining: Theory and Evidence", *American Economic Review* **81**, 1096-1136.

[7] Bolton, G.E. and A. Ockenfels (2000): "A Theory of Equity, Reciprocity and Cooperation", *American Economic Review* **100**, 166-193.

[8] Camerer, C. (2003): *Behavioral Game Theory*, Princeton University Press.

[9] Carpenter, J. (2007): "Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods," *Games and Economic Behavior* **60**, 31-52.

[10] Charness, G. and M. Rabin (2002): "Understanding Social Preferences with Simple Tests", *Quarterly Journal Economics* **117**, 817-869.

[11] Dufwenberg, M. and U. Gneezy (2000): "Measuring Beliefs in an Experimental Lost Wallet Game", *Games and Economic Behavior* **30**, 163-182.

[12] Dufwenberg, M. and G. Kirchsteiger (1998): *A Theory of Sequential Reciprocity*, mimeo, Tilburg University.

[13] Dufwenberg, M. and G. Kirchsteiger (2004): "A Theory of Sequential Reciprocity", *Games and Economic Behavior* 47, 269-298.

[14] Engelmann, D. and M. Strobel (2004): "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments", *American Economic Review* **94**, 857-869.

[15] Falk, A. and U. Fischbacher (2004): "A Theory of Reciprocity," *Games and Economic Behavior* **54**, 293-315.

[16] Falk, A., E. Fehr and U. Fischbacher (2003): "On the Nature of Fair Behavior", *Economic Inquiry* **41**(1), 20-26.

[17] Fehr, E., and S. Gächter (2000): "Cooperation and Punishment in Public Goods Experiments", *American Economic Review* **90**(4), 980-994.

[18] Fehr, E., Gächter, S., Kirchsteiger, G. (1997): "Reciprocity as a contract enforcement device – experimental evidence," *Econometrica* **65**, 833–860.

[19] Fehr, E., G. Kirchgeister and A. Riedl (1993): "Does Fairness Prevent Market Clearing? An Experimental Investigation," *Quarterly Journal of Economics* **108**, 437-460.

[20] Fehr, E., G. Kirchgeister and A. Riedl (1998): "Gift Exchange and Reciprocity in Competitive Experimental Markets," *European Economic Review* **42**, 1-34.

[21] Fehr, E. and K. Schmidt (1999): "A Theory of Fairness, Competition and Cooperation", *Quarterly Journal of Economics* **114**, 817-868.

[22] Fehr, E. and K. Schmidt (2003): "A Theory of Fairness, Competition and Cooperation: Evidence and Economic Applications", in: *Advances in Economic Theory, Eighth World Congress of the Econometric Society*, S.T.M. Dewatripont, L.P. Hansen (eds.), Cambridge University Press.

[23] Fischbacher, U., S. Gachter and E. Fehr (2001): "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment", *Economic Letters* **71**, 397-404.

[24] Geanakoplos, J. D. Pearce and E. Stacchetti (1989): "Psychological Games and Sequential Rationality", *Games and Economic Behavior* **1**, 60-79.

[25] Gűth, W., R. Schmittberger and B. Schwarze (1982): "An Experimental Analysis of Ultimatum Bargaining", *Journal of Economic Behavior and Organization* **3**, 367-88.

[26] Harsanyi, J. (1955): "Cardinal Welfare, Individualistic Ethics and Interpersonal Comparison of Utility", *Journal of Political Economy* **63**, 309–321.

[27] Isaac, M. and J. Walker (1988): "Group Size Effects in Public Good Provision: The Voluntary Contribution MEchanism," *Quarterly Journal of Economics* **103**, 179-199.

[28] Kagel, J.H., and A.E. Roth (1995): *The Handbook of Experimental Economics*, Princeton University Press.

[29] Kirchsteiger, G. (1994): "The Role of Envy in Ultimatum Games"; *Journal of Economic Behavior and Organization* **25**(3), 373-389.

[30] Loewenstein, G., L. Thompson and M. Bazerman (1989): "Social utility and decision making in interpersonal contexts", *Journal of Personality and Social Psychology* **57**, 426-441.

[31] Levine, D.K. (1998): "Modeling Altruism and Spitefulness in Experiments", *Review of Economic Dynamics* **1**, 593-622.

[32] Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics", *The American Economic Review* **83**(5), 1281-1302.

[33] Rawls, J. (1971): *A Theory of Justice*. Harvard University Press, Cambridge (Mass.).

[34] Segal, U. and J. Sobel (2007): "Tit for Tat: Foundation of Preferences for Reciprocity in Strategic Settings," *Journal of Economic Theory* **136**, 197-216.

[35] Sen, A.K. (1977): "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory ", *Philosophy and Public Affairs* **6**, 317-344.

[36] Sethi, R. and E. Somanathan (2003): "Understanding reciprocity", *Journal of Economic Behavior and Organization* **50**, 1-27.

[37] Simon, C.P. and Blume, L. (1994): *Mathematics for Economists*, W.W. Norton, NY.

[38] Slonim, R. and A. E. Roth (1998) "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic," *Econometrica* **66**(3), 569–96.

[39] Smith, A. (1976): *The Theory of Moral Sentiments*. Clarendon Press, Oxford.

[40] Sobel, J. (2005): "Interdependent Preferences and Reciprocity," *Journal of Economic Literature* **43**, 392-436.

[41] Veblen, T. (1922): *The Theory of the Leisure Class - An Economic Study of Institutions*, George Allen Unwin, London, UK.

[42] Vega Redondo, F. (2003): *Economics and the Theory of Games*, Cambridge Univ. Press.

5 Appendix A

Proof of Proposition 1. In the second stage, self-interested Responder derives utility $U_R(c, A) = c$ if she accepts and $U_R(c, R) = 0$, otherwise. Then $U_R(c, A) \geq U_R(c, R)$ for any $c \geq 0$ and she always accepts. Proposer knows it and offers the smallest c, that is zero. ■

Proof of Proposition 2. Since $\frac{\partial \pi_i}{\partial p_{ij}} = -c \;\; \forall i, j \in N$, $p_{ij} = 0 \quad \forall i, j \in N$. In the first stage, players anticipate this. Then, given that $\frac{\partial \pi_i}{\partial g_i} = -1 + a < 0 \; \forall i$ and having $a < 1$, $g_i = 0 \; \forall i$. ■

Proof of Proposition 3. (i) *Second stage.* Responder accepts if $U_R(c, A) \geq U_R(c, R)$.

(a) If $c \leq 1 - c$, he accepts if $(1 - \lambda_R)c + \lambda_R [\delta_R c + (1 - \delta_R)1] \geq 0$ what is satisfied for any c.

(b) If $c > 1 - c$, he accepts if $(1 - \lambda_R)c + \lambda_R [\delta_R(1 - c) + (1 - \delta_R)1] \geq 0$. Again, he accepts for any $c \in \left[\frac{1}{2}, 1\right]$.

(ii) *First stage.* $U_P(\frac{1}{2}, A) \geq U_P(\frac{1}{2} + \varepsilon, A)$ for each $\varepsilon \in \left(\frac{1}{2}, 1\right]$. Then, Proposer never offers $c > \frac{1}{2}$. For $c \leq \frac{1}{2}$, $\lambda_P < \frac{1}{1+\delta_P} \implies \frac{\partial U_P}{\partial c} < 0 \implies c = 0$. If $\lambda_P > \frac{1}{1+\delta_P}$, $\frac{\partial U_P}{\partial c} > 0$ what implies $c = \frac{1}{2}$, and $\lambda_P = \frac{1}{1+\delta_P}$ with $U_P(c, A)$ being the same for any $c \in \left[0, \frac{1}{2}\right]$. ■

Proof of Proposition 4. $\frac{\partial U_i}{\partial p_{ij}} < 0$ for each $i, j \in N \implies p_{ij} = 0$ for each $i, j \in N$. (i) If $g_i \notin \max_{j \in N} \{g_j\}$, $U_i(0, g_{-i}) > U_i(g_i > 0, g_{-i})$ when $\lambda_i < \lambda^1 = \frac{1-a}{(n-1)a(1-\delta_i)+\delta_i}$. If this holds for each i, $g_i = 0 \; \forall i$. (ii) For $g_i = \max\{g_j\}$, $\frac{\partial U_i}{\partial g_i} = (1 - \lambda_i)(-1 + a) + \lambda_i [\delta_i(-1 + a) + (1 - \delta_i)(-1 + na)] > 0$ when $\lambda_i > \lambda^2 = \frac{1-a}{(n-1)a(1-\delta_i)}$. Again, having this for all agents, $g_i = y$ for each i. (iii) For $\lambda_i \in (\frac{1-a}{(n-1)a(1-\delta)+\delta_i}, \frac{1-a}{(n-1)a(1-\delta_i)})$, i's utility is decreasing with respect to g_i if she is the worst off player and increasing if he is not. Then, $g_i = \max\{g_{-i}\}$. (iv) If $\lambda_i = \lambda^1$, i contributes such that she is not the worst of player, i.e. $g_i \in [0, \max\{g_{-i}\})$. (v) If $\lambda_i = \lambda^2$, i contributes such that $g_i = \max\{g_j\}$.

Observe that a mixture of population such that there exist at least one j with $\lambda_j > \lambda^2$ and least one $k \neq j$ with $\lambda_k < \lambda^1$ and $\nexists i$ with $\lambda_i = \lambda^1$ or λ^2, the equilibrium outcome is so that a part of population contributes fully and the others do not contribute. ■

Proof of Proposition 5. Since $U_R(\frac{1}{2}, \frac{1}{2}) \geq U_R(0, \frac{1}{2})$ by assumption, hence $c = \frac{1}{2}$ is never rejected. For any $c > \frac{1}{2}$, $U_P(\frac{1}{2}, \frac{1}{2}) > U_P(1 - c, 1 - c) \implies c > \frac{1}{2}$ is not offered in equilibrium. By the definiton of $s_R(.)$, Responder accepts any $c \in \left[s_R(.), \frac{1}{2}\right]$. Proposer maximizes his utility in $1 - c = r_P(.) \in \left[\frac{1}{2}, 1\right]$. If $1 - r_P(.) \geq s_R(.)$, $r_P(.)$ is accepted. Otherwise, the concavity implies that Proposer offers the closest share that is accepted, i.e. $s_R(.)$. The strict concavity of the utility function guarantees the unicity of this equilibria. ■

Proof of Proposition 6. In the punishment stage of the game, player i punishes j if $\frac{\partial U_i}{\partial p_{ij}} = \frac{\partial U_i}{\partial \pi_i} . \frac{\partial \pi_i}{\partial p_{ij}} + \frac{\partial U_i}{\partial \varpi_i} . \frac{\partial \varpi_i}{\partial p_{ij}} = -c \frac{\partial U_i}{\partial \pi_i} + \frac{\partial U_i}{\partial \varpi_i} \frac{\pi_i - c \sum_{j \neq i} \pi_j}{(\sum_j \pi_j)^2} > 0$. Consider a common contribution scenario with $g_i = \overline{g} \in [0, y] \; \forall i$. In the second stage,

$\frac{\partial U_i}{\partial \varpi_i} = 0 \Longrightarrow \frac{\partial U_i}{\partial p_{ij}} \leq 0$. Thus, $p_{ij} = 0 \ \forall i, j \in N$ in any common contribution scenario. Since $\frac{\partial \pi_i}{\partial g_i} < 0$ and $\frac{\partial \varpi_i}{\partial g_i} \leq 0$ (due to $\varpi_i = 1/n$), no i deviates from the common contribution upwards. This proves the first part of the proposition. Consider any downward deviation. The concavity of U_i with respect to π_i ensures that any i, such that $r_i > \frac{1}{n}$, can find it optimal to contribute less until $\varpi_i = r_i$. She will do so costlessly if she is not punished. This does not happen if for $\forall k \neq i \ \frac{\partial U_k}{\partial p_{ki}} \leq 0$. We can ensure this for $c \geq \pi_k / \sum_{j \neq k} \pi_j$. Given the deviation of i, $\pi_k / \sum_{j \neq k} \pi_j < 1/(n-1)$. This proves the second part of the proposition. ■

Proof of Proposition 8.

1. Responder can only accept. $U_D(\frac{1}{2}, \frac{1}{2}) = \frac{1}{2} > U_D(1-c, c) \ \forall c > \frac{1}{2}$. Then, Dictator never offers more then $\frac{1}{2}$. Given that $c \leq \frac{1}{2}$ and $\pi_D = 1 - \pi_R$, $\frac{\partial U_D(1-c,c)}{\partial c} = 0 \Longleftrightarrow \pi_D = \frac{1+4\beta_D^2}{8\beta_D^2}$. Clearly, $\frac{1+4\beta_D^2}{8\beta_D^2} < 1 \Longleftrightarrow \beta_D > \frac{1}{2}$. In such a case, Dictator offers positive amounts. Furthermore, $\frac{\partial}{\partial \beta_D}\left[\frac{1+4\beta_D^2}{8\beta_D^2}\right] < 0 \Longrightarrow \pi_D$ is minimal in $\beta_D = 1$ where $\pi_D = \frac{5}{8}$.

2. $\frac{\partial U_R}{\partial c} > 0$ for any $c \geq \frac{1}{2}$. For $c < \frac{1}{2}$, $U_R(c, A) \geq U_R(c, R) = 0 \Longleftrightarrow c \in \left[\frac{1+4\alpha_R^2 - \sqrt{1+8\alpha_R^2}}{8\alpha_R^2}, \frac{1}{2}\right]$. Proposer's optimal behavior is identical with Dictator, but she has to take into account that too low offer will be rejected. Then, if $\frac{4\beta_P^2-1}{8\beta_P^2} \geq \frac{1+4\alpha_R^2-\sqrt{1+8\alpha_R^2}}{8\alpha_R^2}$, Proposer prefers $c = \frac{4\beta_P^2-1}{8\beta_P^2}$ and otherwise, she proposes the lowest acceptable c that is $\frac{1+4\alpha_R^2-\sqrt{1+8\alpha_R^2}}{8\alpha_R^2}$.

■

Proof of Proposition 11. Consider first the second stage of the game. The equity payoff of Responder is $\frac{1+0}{2}$. Denote p'' the second order belief of Responder (that is the Responder's belief about with which probability Proposer believes Responder accepts the corresponding c).[25] Responder believes Proposer gives her c with probability p'' and 0 with probability $1-p''$. Then, the kindness function $\lambda_{RPR}(.) = cp'' - \frac{1}{2}$.

Let us turn to the kindness of Responder toward Proposer. By receiving offer c, Responder can give Proposer $1-c$ by accepting or 0 by rejecting. However, note that rejection is an inefficient action. Thus, the equity payoff of proposer is $1-c$, and the kindness of Responder toward Proposer can be either 0 if she accepts or $-(1-c)$ otherwise.

Responder accepts if $U_P(accept, c, accept) > U(reject, c, reject) \Longleftrightarrow c + Y_R\frac{1-c}{2}(cp'' - \frac{1}{2}) > Y_R(-\frac{1-c}{2})(cp'' - \frac{1}{2})$. The belief consistency condition requires beliefs to match in equilibrium. That is, if Responder accepts $p'' = 1$ and the solution is $c > c^1 = \frac{0.5+0.75Y_R-0.5\sqrt{1+3Y_R+0.25Y_R^2}}{Y_R}$. If she rejects, $p'' = 0$ and

[25] Recall that, in the second stage, c is already know and fixed. Thus, p'' does not require to be a function of c.

$c < c^2 = \frac{Y_R}{2+Y_R}$. Note that $c^2 \geq (>) c^1$ for $Y_R \geq (>) 0$. Thus for $c \in [c^1, c^2]$, both accept and reject may be optimal, depending on the beliefs players hold.

Let us now turn to the behavior of Proposer. Responder can be either neutral (accepting) or mean (rejecting) to Responder. Accepting leads to kindness of 0. Thus, only material payoff matters in Proposer's utility and she offers the minimal acceptance offer. Beliefs consistency implies that, since Proposer believes Responder accepts this offer, she actually accept it. This argument is compeltely independent of the values of Y_P and Y_R.

Next step is to show that there can be an equilibrium, in which the Proposer makes an offer that is rejected if Y_P and Y_R are large enough. If this really is an equilibrium, the beliefs are correct and, by offering less than acceptable, Proposer's kindness and preceived kindess are $\kappa_{PR} = -\frac{c}{2}$ and $\lambda_{PRP} = -1 + c$, leading to $U_P(c, reject, c) = 0 + Y_P(-\frac{1}{2})(-1 + c)$ for any $c \leq c^2$. If Proposer deviates to any $c' > c^2$, his perceived kindeness does not change and he always offers the lowest such c'. He would does not deviate as long as: $U_P(c', accept, c) < U_P(c, reject, c)$ that is if $Y_P > \frac{1-c'}{c'(1-c)}$. On the other hand, there has to exist a c that is rejected with probability 1 and it exist for any $Y_R > 0$. This concludes the proof. ■

Proof of Proposition 12. .

(i) How i feels treated by j.

$$\pi_i^{ei}((b_{ji})_{i\neq j}) = \frac{1}{2}\left[\begin{array}{c} (y - g_i + a\sum_{j\neq k} g_k + ay - \sum_{\substack{i\neq k\\ j\neq k}} p'_{ki} - c\sum_{\substack{i\neq k\\ j\neq k}} p''_{ik}) \\ +(y - g_i + a\sum_{j\neq k} g_k + 0 - \sum_{\substack{i\neq k\\ j\neq k}} p'_{ki} - c\sum_{\substack{i\neq k\\ j\neq k}} p''_{ik}) \end{array}\right]$$

Then, kindness perceived by i:

$$\begin{aligned} \kappa_{ij}(a_i(h), (b_{ij}(h))_{j\neq i}) &= \\ &\left[\begin{array}{c} (y - g_i + a\sum g_k - \sum_{i\neq k} p'_{ki} - c\sum_{i\neq k} p''_{ik}) \\ -(y - g_i + a\sum_{j\neq k} g_k + \frac{ay}{2} - \sum_{\substack{i\neq k\\ j\neq k}} p'_{ki} - c\sum_{\substack{i\neq k\\ j\neq k}} p''_{ik}) \end{array}\right] \\ &= ag_j - \frac{ay}{2} - p'_{ji} \end{aligned}$$

(ii) How i treats j.

$$\pi_j^{ei}((c_j)_{i\neq j}) = \frac{1}{2}\left[\begin{array}{c} (y - g_j + a\sum_{i\neq k} g_k + ay - \sum_{\substack{i\neq k\\ j\neq k}} p'_{ki} - c\sum_{\substack{i\neq k\\ j\neq k}} p''_{ik}) \\ +(y - g_j + a\sum_{i\neq k} g_k + 0 - \sum_{\substack{i\neq k\\ j\neq k}} p'_{ki} - c\sum_{\substack{i\neq k\\ j\neq k}} p''_{ik}) \end{array}\right]$$

$$\lambda_{iji}(b_{ij}(h),(c_{ijk}(h))_{k\neq j}) = \begin{bmatrix} (y-g_j+a\sum g_k-p_{ij}-\sum_{\substack{j\neq k\\ i\neq k}} p'_{kj}-c\sum_{j\neq k} p''_{jk}) \\ -(y-g_j+a\sum_{i\neq k} g_k+\frac{ay}{2}-\sum_{\substack{i\neq k\\ j\neq k}} p'_{kj}-c\sum_{\substack{i\neq k\\ j\neq k}} p''_{jk}) \end{bmatrix}$$
$$= ag_i-\frac{ay}{2}-p_{ij}$$

The resulting utility function is:

$$U_i(.\mid g)=y-g_i+a\sum_j g_j-c\sum_{i\neq j} p_{ij}-\sum_{i\neq j} p_{ji}+\sum_{i\neq j} Y_{ij}(ag_j-\frac{ay}{2}-p'_{ji})(ag_i-\frac{ay}{2}-p_{ij})$$

with $\frac{\partial U_i(.|g)}{\partial p_{ij}} = -c-Y_{ij}(ag_j-ay/2-p'_{ji})$. Given that belief are consistent in equilibrium so that $p'_{ji}=p_{ji}$, i punishes j if $\frac{\partial U_i(.|g)}{\partial p_{ij}} > 0 \iff p_{ji} > a(g_j-\frac{y}{2})-\frac{c}{Y_{ij}}$. If $g_j \geq y/2$ and $p_{ji}=0\ \forall i$, $p_{ij}=0$.

Consider full contribution scenario: $U(y)=any+\sum_{j\neq i} Y_{ij}\frac{ay}{2}\frac{ay}{2}$ and $U(y-\varepsilon)=\varepsilon+any-a\varepsilon+\sum_{j\neq i} Y_{ij}\frac{ay}{2}(\frac{ay}{2}-a\varepsilon)$. Then, if $\sum_{j\neq i} Y_{ij} > \frac{2(1-a)}{a^2y}$, $U(y) > U(y-\varepsilon)$.

Consider a common contribution $g_i=\overline{g}\in\left(\frac{y}{2},y\right)$ for each $i\in N$. In such a case, no punishment occurs till nobody deviates below $\frac{y}{2}$. $U(\overline{g})=y-\overline{g}+an\overline{g}+\sum_{j\neq i} Y_{ij}(a\overline{g}-\frac{ay}{2})(a\overline{g}-\frac{ay}{2})$. If i deviates downwards (resp. upwards) by an ε small, $U(\overline{g}-\varepsilon)=y-\overline{g}+\varepsilon+an\overline{g}-a\varepsilon+\sum_{j\neq i} Y_{ij}(a\overline{g}-\frac{ay}{2})(a\overline{g}-a\varepsilon-\frac{ay}{2})$ (resp. $U(\overline{g}+\varepsilon)=y-\overline{g}-\varepsilon+an\overline{g}+a\varepsilon+\sum_{j\neq i} Y_{ij}(a\overline{g}-\frac{ay}{2})(a\overline{g}+a\varepsilon-\frac{ay}{2})$). It is profitable to deviate unilaterally for i if:

- $U(\overline{g}-\varepsilon)>U(\overline{g}) \iff \sum_{j\neq i} Y_{ij} < \frac{1-a}{a^2(\overline{g}-\frac{y}{2})}$
- $U(\overline{g}+\varepsilon)>U(\overline{g}) \iff \sum_{j\neq i} Y_{ij} > \frac{1-a}{a^2(\overline{g}-\frac{y}{2})}$

These conditions are exclusive, so - unless for all $i\in N$ $\sum_{j\neq i} Y_{ij}=\frac{1-a}{a^2(\overline{g}-\frac{y}{2})}$ - no such common contribution is optimal.

To prove that $g_i=\frac{y}{2}$ for each i, let everybody but i play $g_j=\frac{y}{2}$. Consider $g_i=\frac{y}{2}-\varepsilon$ with $\varepsilon\in(0,\frac{c}{a\max\{Y_{ji}\}})$ and i does not punish. The utility of generic j is: $U_j(.)=\frac{y}{2}+an\frac{y}{2}-a\varepsilon-cp_{ji}+a\varepsilon Y_{ji}.p_{ji}$. She does not punish if $\frac{\partial U_j}{\partial p_{ji}}<0 \iff \frac{c}{aY_{ji}}>\varepsilon$ what is so for all i by assumption on ε. Given, i is not punished, her psychological utility does not change and she earns $(1-a)\varepsilon>0$ in the selfish part. Then, he prefers to deviate downwards.

Suppose a common contribution scenario with $g_i=\widehat{g}\in\left(0,\frac{y}{2}\right)$ for each $i\in N$. In such a case, $U(\widehat{g})=y-\widehat{g}+an\widehat{g}-c\sum_{i\neq j} p_{ij}-\sum_{i\neq j} p_{ji}+\sum_{i\neq j} Y_{ij}(a\widehat{g}-\frac{ay}{2}-p_{ij})(a\widehat{g}-\frac{ay}{2}-p_{ji})$. Any agent want to contribute less if $U(\widehat{g}-\varepsilon)>U(\widehat{g})$, what happens

if $\frac{1-a}{a^2} > \sum_{i\neq j} Y_{ij}(\widehat{g} - \frac{y}{2} - p_{ji})$. Observe that $\frac{1-a}{a^2} > 0 \geq \sum_{i\neq j} Y_{ij}(\widehat{g} - \frac{y}{2} - p_{ji})$ for all considered $\widehat{g}$ and whatever p_{ji}.

With no contribution $U(0) = y - c\sum_{i\neq j} p_{ij} - \sum_{i\neq j} p_{ji} + \sum_{i\neq j} Y_{ij}(-\frac{ay}{2} - p_{ji})(-\frac{ay}{2} - p_{ij})$. $\frac{\partial U(0)}{\partial p_{ij}} < 0 \iff Y_{ij} < c(\frac{ay}{2} + p_{ji})^{-1}$. If noone punishes and $Y_{ij} < \frac{2c}{ay}$ $\forall i, j$ s.t $j \neq i$, no contribution can be optimal. ■

Proof of Proposition 13. 2^{nd} stage. $U_R(c, A) = c + (1-c)\frac{a_R + \lambda_R a_P}{1+\lambda_R} \geq U_R(c, R) = 0 \iff c \geq -\frac{a_R + \lambda_R a_P}{1+\lambda_P - a_R - \lambda_R a_P}$, Thus,

$$\text{Responder accepts} \begin{cases} \text{any } c & \text{if} \quad a_R + \lambda_R a_P \geq 0 \\ \text{any } c \geq -\frac{a_R + \lambda_P a_P}{1+\lambda_R - a_R - \lambda_R a_P} & \text{if} \quad a_R + \lambda_R a_P < 0. \end{cases}$$

If $a_P = 1$ and $a_R = 1$, the utility is the same for any c.

1^{st} stage. $U_P(c, A) = (1-c) + c\frac{a_P + \lambda_P a_R}{1+\lambda_P}$, $U_R(c, R) = 0$ and $\frac{\partial U(c,A)}{\partial c} = -1 + \frac{a_P + \lambda_P a_R}{1+\lambda_P}$. Since $a_P + \lambda_P a_R \leq 1 + \lambda_P$, $\frac{\partial U(c,A)}{\partial c} < 0$ if $a_P + \lambda_P a_R < 1 + \lambda_P$. Otherwise, $a_P = a_R = 1$, $U_P(c, A) = 1 > 0 = U_R(c, R)$ for any c. ■

Proof of Proposition 14. 2^{nd} stage.

$$\begin{aligned} U_i(g, p) \quad = \quad & y - g_i + a\sum\nolimits_j g_j - c\sum\nolimits_{j\neq i} p_{ij} - \sum\nolimits_{j\neq i} p_{ji} \\ & + \sum\nolimits_{j\neq i} (\frac{a_i + \lambda_i a_j}{1+\lambda_i})(y - g_j + a\sum\nolimits_k g_k - c\sum\nolimits_{j\neq k} p_{jk} - \sum\nolimits_{j\neq k} p_{kj}). \end{aligned}$$

Since $\frac{\partial U_i}{\partial p_{ij}} = -c - (\frac{a_i + \lambda_i a_j}{1+\lambda_i})$, first stage contributions have no effect.

Since the first stage have no influence on the second, we can separate the game into two independent parts: public good game without punishment and a punishment game.

In the second stage, i punishes if $\frac{\partial U_i}{\partial p_{ij}} > 0$ what happens if $a_j < -\frac{a_i + c(1+\lambda_i)}{\lambda_i}$.

In Public good game without punishment, i (does not) contributes if her altruistic profit from contributing over weights the negative effect on his material payoff, that is if

$$\sum_{j\neq i} (\frac{a_i + \lambda_i a_j}{1+\lambda_i}) > (<) 1 - a \Longrightarrow \sum_{j\neq i} a_j > (<) \left[\frac{(1-a)(1+\lambda_i)}{a} - (n-1)a_i\right] \frac{1}{\lambda_i}. \tag{11}$$

Agent i, then, contributes fully (nothing). If (11) is satisfied with equality, i is indifferent how much she contributes. Therefore, the heterogeneity of players allow for three possible equilibrium outcomes: full contribution of all players, zero contribution and polymorphic equilibrium. ■

Proof of Proposition 16. The utility function of a player in the second stage

when the vector of contributions g is given is

$$U_i = y - g_i + a\sum_j g_j - c\sum_{j\neq i} p_{ij} - \sum_{j\neq i} p_{ji} + \rho_i \sum_{j\neq i} \vartheta(.)\left[g_j - g_i - p'_{ji} - cp''_{ij} + p''_{ij} + cp'_{ji}\right]\left[-p_{ij} + p''_{ij}\right]. \quad (12)$$

Then,

$$\frac{\partial U_i}{\partial p_{ij}} = 0 \Longleftrightarrow p_{ij} = p_{ji} + (1-c)^{-1}[g_i - g_j - \frac{c}{\rho_i \vartheta(.)}]. \quad (13a)$$

If $g_j \geq g_i$ and $p_{ji} = 0$, no punishment takes place.

To prove the second part of Proposition, consider a full contribution scenario without any punishment. If it is an equilibrium, beliefs are consistent and for each i, $U(y) = any$. Take an agent i who deviates by an amount $\varepsilon < \min_{j\neq i}\left\{\frac{c}{\rho_j}\right\}$.[26] The second part proves that in this case, i is not punished in the second stage and her psychological part of her utility is not affected. Since she wins in material terms $\varepsilon(1-a) > 0$, she will deviate. Then, full contribution cannot be an equilibrium. The same argument holds for any common contribution, except zero contribution.

Consider, now, no contribution scenario with no punishment. Let i deviate some ε. If $\varepsilon \leq \frac{c}{\rho_i}$, her disadvantage is not high enough to punish and she only looses in material terms. Then, only $\varepsilon > \frac{c}{\rho_i}$ can be considered. Nevertheless, even this is not optimal for player i. Consider the second stage. If i does not deviate, $U_i(0) = y$. If he does, the relevant utility is (12). Then, take into account that each $j \neq i$ anticipates from (13a) the punishment and i knows it, i.e. $p_{ij} = p''_{ij}$. Since i cannot alter j's payoff, the psychological part of (12) is zero and only material payoff matter. It decreases by this deviation, i, therefore, prefers not to deviate. ■

Proof of Corollary 17. The last part of Proposition 4 shows that it is not optimal to deviate upwards from common contribution level in the case of null contribution scenario. The same argument can be used for any common contribution scenario.

Let us now check any possible unilateral deviation downwards. Consider an agent i who deviates by an amount ε. Since there is always at least one player who punishes him, the utilities without and with deviation are $U_i(\overline{g}) = y - \overline{g} + an\overline{g}$ and $U_i(\overline{g} - \varepsilon) = y - \overline{g} + \varepsilon + an\overline{g} - a\varepsilon - \sum_{j\in K} \frac{\varepsilon}{k-c}$, respectively. $U_i(\overline{g}) > U_i(\overline{g} - \varepsilon) \iff k > (1-a)(k-c)$ what always holds. Now consider an agent $j \in K$ who punishes i and let prove that she does not prefer to deviate from the punishment strategy. If she punishes, $U_j(., p_{ji} = 0) = y - \overline{g} + an\overline{g} - a\varepsilon - c\frac{\varepsilon}{k-c} + \rho_j(-\varepsilon - p'_{ji} - cp''_{ij} + p''_{ij} + cp'_{ji})(-p_{ij} + p''_{ij})$. We have $p_{ji} = p'_{ji} = 0$ and $p''_{ij} = p_{ij} = \frac{\varepsilon}{k-c}$. Let verify that the punisher does not like to deviate in his punishment strategy downward by a small ϵ. If so, he earns ϵc in material

[26] This deviation is clearly fully intentional.

payoff and looses $\rho_j(c\epsilon - \epsilon)(\epsilon) < 0$ in psychological utility. Then, he does not deviate if $\rho_j(-c\epsilon + \epsilon)(\epsilon) > \epsilon c \Leftrightarrow \rho_j > \frac{c}{\epsilon}(1-c)^{-1}$. ■

6 Appendix B

In this section, we formally introduce the two-player version of the model of Falk and Fischbacher (2004).[27]

Let D_i be the the set of nodes where i moves with d being a node of this player, A_d the set of actions in node d and F the set of end nodes of the game. Being $P(A_d)$ the probability distribution over actions in node d, $A_i = \prod_{d \in D_i} P(A_d)$ is player i's set of behavior strategies. Last, $\pi_i(d, a_i, a_j)$ denotes the expected payoff conditional on node d.

Again, let $b_{ij} \in A_j$ and $c_{iji} \in A_i$ be first and second-order beliefs of player i, respectively.

The utility function of Falk and Fischbacher can be divided into 2 parts: selfish and psychological. The latter part can be further divided into the *kindness* and *reciprocation terms.* The kindness term is a product of the *outcome term*, which determines whether the individual feels treated kindly or not, and the *intentional factor* defining the intention of action on basis of all alternatives. Reciprocation term reflects the response of a player to (un)kindness

First, define the outcome term as $\triangle(d) = \pi_i(d, c_{iji}, b_{ij}) - \pi_j(d, b_{ij}, c_{iji})$. Whether an action is considered kind depends on the sign of the outcome term. If the outcome term, $\triangle$, is positive (negative) the individual is in (dis)advantageous position. Observe that $\pi_j(d, b_{ij}, c_{iji})$ measures i's belief about how much j want to keep for herself. The outcome term reflects the crucial difference between this model and that of Dufwenberg and Kirchsteiger (2004). In this case, the reference point is not the equitable payoff. Its role picks the opponent's payoff. Then, i feels treated unkindly by j if she earns less than j does.

Note that the outcome term does not take into account (un)kindness of the path to d. This feature is included in the intentional factor, ϑ. It depends on the set of all alternatives. Traditionally, let S_j be the set of pure strategies of j and $\Pi_i = \{(\pi_i(c_{iji}, s_j), \pi_j(n, c_{iji}, s_j)) \mid s_j \in S_j\}$ the set of all alternatives i beliefs j can offer her. Whether payoff distribution (π_i^0, π_j^0), given that j had an alternative (π_i, π_j), is intentional is given by function Ω :

$$\Omega(\pi_i, \pi_j, \pi_i^0, \pi_j^0) = \begin{cases} 1 & \text{if } \pi_i^0 \geq \pi_j^0 \text{ and } \pi_i < \pi_i^0 \\ \varepsilon_i & \text{if } \pi_i^0 \geq \pi_j^0 \text{ and } \pi_i \geq \pi_i^0 \\ 1 & \text{if } \pi_i^0 < \pi_j^0,\ \pi_i > \pi_i^0 \text{ and } \pi_i \leq \pi_j \\ \max(1 - \frac{\pi_i - \pi_j}{\pi_i^0 - \pi_j^0}, \varepsilon_i) & \text{if } \pi_i^0 < \pi_j^0,\ \pi_i > \pi_i^0 \text{ and } \pi_i > \pi_j \\ \varepsilon_i & \text{if } \pi_i^0 < \pi_j^0 \text{ and } \pi_i \leq \pi_i^0 \end{cases}$$

where $\varepsilon_i \in [0, 1]$ is an individual pure outcome concern parameter. It measures player's pure concern for equity. If, for instance, it equals one we get back to pure distributional inequity aversion of Fehr and Schmidt (1999).

[27] The Appendix of Falk and Fischbacher (2004) provide the multiperson version.

The intentional factor is formally defined as follows:

$$\vartheta(d) = \max\left\{\Omega(\pi_i, \pi_j, \pi_i(d, c_{iji}, b_{ij}), \pi_j(d, c_{iji}, b_{ij})) \mid (\pi_i, \pi_j) \in \Pi_i\right\}.$$

Then, $\vartheta \in [0, 1]$ such that $\vartheta = 1$ means that $\triangle$ is induced by fully intentional play. The formal definition seems extremely difficult, so it deserves more attention. The intuition full intentionality is that an agent's action is considered intentionally kind, if he had the opportunity to be less kind, and it is intentionally unkind if more generous and still reasonable alternative exists; reasonable meaning that the opponent would still be in payoff advantage. Figure 5 illustrates graphically the logic of the latter case. From the point of view of a player, the black point represents a fully intentional, unkind choice, because her opponent has a possibility to give her a higher payoff, remaining still in better position

Put Figure 5. around here

As mentioned above, the kindness term, $\varphi(d)$, is the product of distributional term and intentional factor and express in monetary terms how the player feels treated by his opponent.

Last, to define the reciprocation term, let $\upsilon(d, f)$ denote unique node that directly follows d on the way to an end node f. Set $\sigma(d, f) = \pi_j(\upsilon(d, f), c_{iji}, b_{ij}) - \pi_j(d, c_{iji}, b_{ij})$. It measures how much i alters j's payoff by his move in d.

With all the terms defined above, the utility function has the following form:[28]

$$U_i(f) = \pi_i(f) + \rho_i \sum_{\substack{d \to f \\ d \in D_i}} \varphi(d)\sigma(d, f) \tag{14}$$

The parameter ρ_i measures individual concern for reciprocity.

[28] In the utility function, $d \to f$ states for nodes f that follows d (both directly or undirectly).

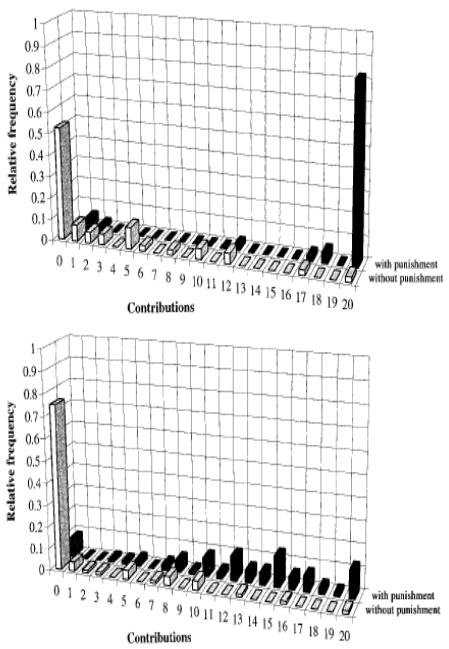

Figure 1: The contrast between Public good game with/without punishment.
Above: Partner treatment. *Below*: Stranger treatment
(Fehr and Gächter (2000)).

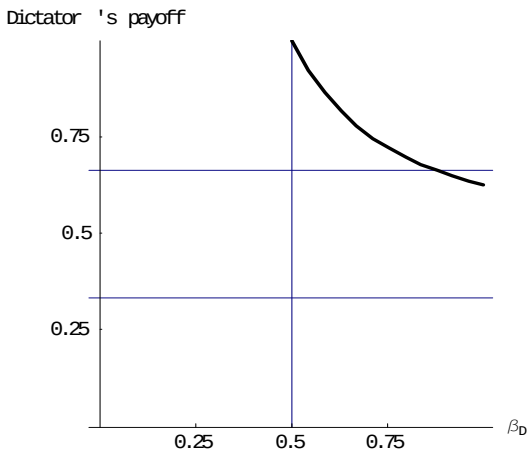

Figure 2. Equilibrium offer of a Dictator
with non-linear version of Fehr and Schmidt (1999)

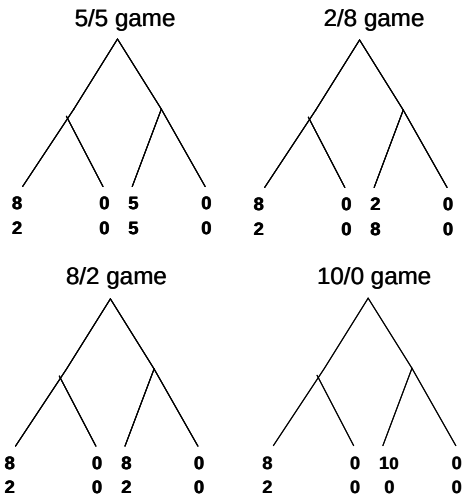

Figure 3. Mini-ultimatum games (Falk *et al.* (2003)).

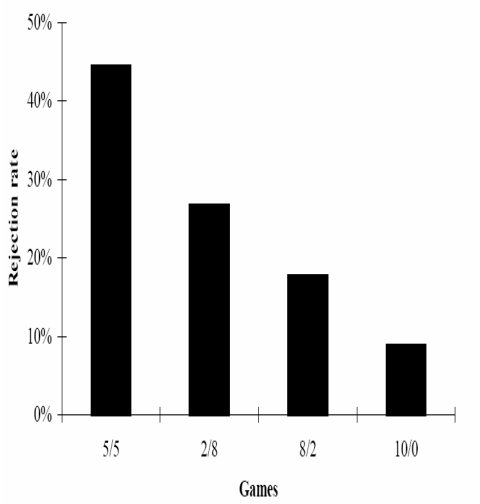

Figure 4. Experimental results of Falk *et al.* (2003).

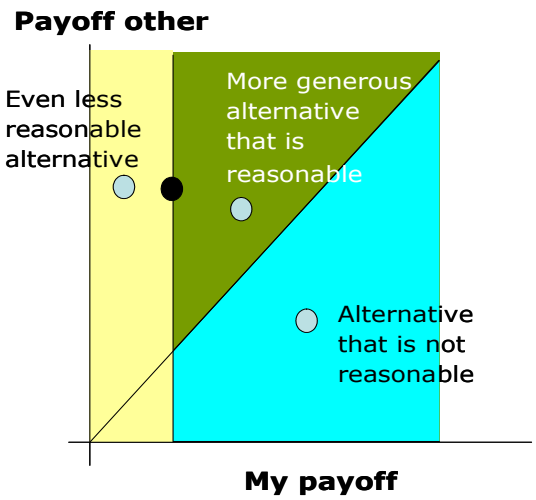

Figure 5. Example of fully intentional choice.[29]

[29] This graph is taken from a persentation of Simon Gachter in Summer School in Trento (Italy), July 2005.

	DG	UG						PGG w/Punishment			
	positive giving	positive offers	offers of 40% of stake	offers > than in DG	no offers > than 50%	low offers rejected	high offers accepted	positive punish.	punish. decreases in contr.	equilib. w/positive contrib.	contrib. stand on punish.
selfish	No	No	No	No	Yes	No	Yes	No	-	No	-
distributional mod.:											
Charness & Rabin (2002)	Yes	Yes	No	No	Yes	No	Yes	No	-	Yes	No
Bolton & Ockenfels (2000)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	-	-
Fehr & Schmidt (1999)	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
reciprocity:											
Dufwenberg & Kirch. (2004)	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	-
combinations:											
Levine (1998)	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	No
Falk & Fischbacher (2004)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	-

Note: "Yes" ("No") appears if the model in the corresponding row is (not) capable to explain the stylized fact in the corresponding column.

Table 1. The capacity of models of social preferences to explain experimental observations.