I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16200

Social Preferences: Fundamental Characteristics and Economic Consequences

Ernst Fehr
Gary Charness

MAY 2023 (THIS VERSION: MARCH 2024)

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16200

Social Preferences: Fundamental Characteristics and Economic Consequences

Ernst Fehr
Zurich University and IZA

Gary Charness
University of California and IZA

MAY 2023 (THIS VERSION: MARCH 2024)

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Social Preferences: Fundamental Characteristics and Economic Consequences*

We review the vast literature on social preferences by assessing what is known about their fundamental properties, their distribution in the broader population, and their consequences for important economic and political behaviors. We provide, in particular, an overview of the empirical characteristics of distributional preferences and how they are affected by merit, luck, and concerns for equality of opportunity. In addition, we identify what is known about belief-dependent social preferences such as reciprocity and guilt aversion. Furthermore, we discuss and assess the empirical relevance of self image and social image concerns in prosocial behaviors. The overall evidence indicates that a large majority of individuals have some sort of social preference, while purely self-interested subjects are a minority. We also document the converging insights from lab and field evidence on the role of social preferences for a deeper understanding of important phenomena such as the consequences of wage inequality on work morale, employees' resistance to wage cuts, individuals' self-selection into occupations that are more or less prone to morally problematic behaviors, as well as issues of distributive politics. However, although much has been learned in recent decades, there are still many important, unresolved, yet exciting, questions waiting to be tackled.

JEL Classification: D0, D2, D9, H0, J0, P0

Keywords: social preferences, altruism, inequality aversion, preference heterogeneity, demand for redistribution

Corresponding author:
Ernst Fehr
Department of Economics
University of Zurich
Blümlisalpstrasse 10
8006 Zurich
Switzerland

E-mail: ernst.fehr@econ.uzh.ch

* We thank the editor (David Romer) and four excellent reviewers for their extremely useful and detailed comments on the paper. In addition, we thank Peter Andre, Alain Cohn, Martin Dufwenberg, Tore Ellingsen, David Levine, Johanna Möllerström, Larry Samuelson, Bertil Tungodden, and Joel van der Weele for constructive and useful comments on earlier versions of this paper.

1. Introduction

The topic of social preferences has received a great deal of attention in economics and in the broader social science literature in the past three decades. A large, and still growing, literature has documented the existence of social preferences, their social and economic implications, and the various conditions under which they influence the equilibria and outcomes of human interactions. The key characteristic of social preferences is that individuals are willing to sacrifice money or other material resources to help or hurt other people in order to establish fairness and justice or to increase a group's joint payoff. In other words, people with social preferences do not maximize their own material payoff but are other-regarding – which has potentially far-reaching consequences for both theory and practical applications.[1] In this paper, we will summarize what has been learned in this line of research and we will identify what needs to be known and done to make further progress.

The role of social preferences can, in principle, manifest itself in myriads of human interactions. Behaviors associated with social preferences range from helping friends and neighbors to support for co-workers, from charitable donations to participation in collective actions against oppressive dictatorships, from individuals' support for sustainable investments to workers' responses to wage inequality in companies, from affluent individuals' support for taxing the rich to workers' collective stance against wage cuts. However, a multitude of factors other than social preferences also play a role in all these cases. For example, expectation of future material benefits or reputational concerns can influence individuals' behavior in these situations. Therefore, considerable part of the research on social preferences is based on incentivized laboratory experiments where one can rule out these concerns with certainty by ensuring that the experiments are one-shot and that the parties interact anonymously with each other.[2] More recently, however, we also experienced a healthy transition to more field evidence – evidence we will also cover in this review.

[1] More formally, let s_i (π_i) and s_j (π_j) be the strategies (material payoffs) of players i and j, respectively. Then, the material payoffs $\pi_i(s_i, s_j)$ and $\pi_j(s_i, s_j)$ do not describe the players preferences. Instead, their (social) preferences are given by by $U_i(\pi_i, \pi_j)$ and $U_j(\pi_i, \pi_j)$.

[2] The implementation of incentivized anonymous one-shot experiments is thus a frequently employed conservative strategy that makes sense, but it is also important to keep in mind that important aspects of social relations, and their influence on social preferences, may be lost in these experiments (Frohlich, Oppenheimer and Kurki 2004). It is, however, possible to introduce many real-life features (e.g., lack of anonymity, reputation formation, the role of merit and entitlements, etc.) into the lab and to study them in a controlled manner.

Due to the extensive literature that has accumulated over the last three decades, it is useful to provide some guidance about the structure of this review (see Box 1 below). We will describe the evidence for the existence of social preferences gathered in canonical ("paradigmatic") experiments in Section 2. However, readers who are already familiar with these findings may directly move on to one of the other sections of the paper, as we have tried to keep the individual sections relatively self-contained. For example, readers who are mainly interested in the economic and political consequences of social preferences, may proceed to Section 6 which discusses rich evidence from the field and the lab on these issues.

Box 1: Table of Contents

The paradigmatic experiments discussed in Section 2, while indicating the existence of social preferences and capturing essential components of important real-life interactions, are a poor tool for the precise identification of the properties of social preferences. Moreover, these experiments led to a fundamental puzzle because the same individuals sometimes displayed very selfish behavior, while they behaved very prosocially at other times. Section 3 discusses the theories of distributional preferences that helped resolve this puzzle by providing a

unifying explanation of these seemingly disparate facts, which then also led to a rich empirical literature on the distribution of distributional preferences in the broader population.

Section 3 takes the recent literature on this topic into account and discusses what we know about the key motivational drivers underlying distributional preferences and the relative frequencies of the different distributional preference types. This section also documents one of the key findings of the recent literature that a considerable majority of individuals in the broader population has other-regarding distributional preferences in the form of inequality aversion and altruism while predominantly selfish individuals constitute a minority. In addition, Section 3 explores how notions of entitlement, merit, and luck affect distributional preferences and how psychologists and economists examined and modelled these factors. This section also indicates that considerable progress has been made, including interesting insights into the large cultural differences in how luck affects individuals' fairness views and how cognitive biases shape individuals' perception of merit.

Section 4 discusses evidence that theories of distributional preferences can hardly explain; this evidence led to the development of belief-dependent social preferences theories such as reciprocity theory and guilt aversion theory. A key characteristic of these approaches is that individuals' first and second order beliefs affect their utility and behavior. This poses challenges in identifying these preferences because the beliefs are endogenous. This section will also provide a critical discussion on the extent to which these challenges have been met and on the empirical relevance of the notion of reciprocity in the lab and the field.

People may not only care directly about others' payoff or others' deservingness and kindness, but they may also care about their own identity and how others perceive them – which leads us to theories of self-image and social image. We examine these theories, the extent to which the evidence supports them, and some of their economic implications in Section 5 of the paper.

Section 6 documents the widespread and far-reaching economic and political consequences of social preferences. In this context, we discuss their important role in human cooperation; their effects on labor relations including the potential consequences on macroeconomic phenomena; their role in affecting incentives, contractual arrangements, and institutions; as well as their effects on politics. Finally, we discuss important open questions – such as the determinants of social preferences or the relationship between social norms and social preferences – that offer exciting research opportunities in Section 7.

2. Social Preferences and Self-Interest in Paradigmatic Economic Experiments

There are several paradigmatic experiments that – when played one-shot and under anonymity between the players involved – document the widespread existence of social preferences. We call these experiments "paradigmatic" because they all capture an essential feature of an important economic or social situation. The experiments are described in more detail in Box 2 on "Paradigmatic Economic Experiments" (see below). All of these experiments have been replicated dozens of times (or more) and show robust, replicable behavioral patterns (see Camerer (2003) for a review). The critical player in these experiments (i.e., the person for whom one can measure social preferences) always has a simple and transparent money maximizing choice, while systematic (i.e., non-random) deviations from that choice indicate some form of social preference.

For example, many responders in the *ultimatum experiment*[3] from societies across the globe indicate a social preference by rejecting uneven offers and earning nothing as a consequence (Guth, Schmittberger and Schwarze 1982; Henrich et al. 2001) . The proposer often anticipates this responder behavior, inducing him or her to make relatively fair offers so that the responder appropriates on average roughly 40% of the pie even though the self-interest model predicts that he or she should get 0%.

In the *dictator experiment*, many subjects in the role of dictators make positive unilateral transfers – on average typically around 20% of the available money – to the recipient (Forsythe et al. 1994; Camerer 2003). Subjects in the role of workers in the *gift exchange experiment* (Fehr, Kirchsteiger and Riedl 1993) respond to higher wages with higher costly effort levels, although the minimal effort would always be the money-maximizing choice. In the *trust experiment* (Berg, Dickhaut and McCabe 1995), the trustees respond to trustors' positive transfers with positive back-transfers, although the money maximizing choice would be to transfer back nothing. In *social dilemma experiments* (Dawes, Mctavish and Shaklee 1977; Dawes 1980) and *public good experiments* with (complete) defection as a dominant strategy (Andreoni 1988), many subjects nevertheless cooperate and make positive contributions to the public good.

[3] Note that the experiments described in Box 2 are not games in the game theoretic sense because the definition of a game includes a full description of the players' preferences but the (known) material payoffs in the game typically do not describe the subjects' (unknown) social preferences.

Box 2: Paradigmatic Economic Experiments

In a ***dictator experiment*** one player – the dictator – is given a sum of money that she can unilaterally allocate between herself and a passive recipient who cannot make a decision. Self-interest predicts zero transfers to the recipient. In an ***ultimatum experiment***, a proposer can make a single proposal of how to distribute a given a sum of money between herself and a responder. The responder can accept or reject the proposed allocation. If she rejects, both players receive nothing. If she accepts, the proposed allocation is implemented. Self-interest predicts that the proposer makes the lowest possible money offer, which the responder will then accept. In a ***third party punishment experiment,*** two players, the dictator A and the recipient B, participate in a dictator game. A third player, the potential punisher C, observes how much A gives to B; then C can spend a proportion of his endowment on punishing A. A third party with a punishment option can, in principle, be added to any constituent base game such as, for example, the prisoners' dilemma game. Selfish third parties will never punish.

In a ***trust experiment***, two players, A and B, each have an identical initial endowment. First, A decides whether to keep his endowment or to send some or all of it to B. Then B observes A's action and decides whether to keep the amount she receives or share some if it with A. The experimenter increases A's transfer by some proportion, so that both players are better off collectively if A transfers money and B sends back a sufficient amount. Self-interest predicts that B sends back zero money and, therefore, A also sends no money. In a ***gift exchange experiment***, a subject in the role of an employer can offer a fixed wage to a subject in the role of a worker. After observing the wage, the worker chooses a costly effort level that raises the overall surplus that can be distributed among the parties. Self-interest predicts the lowest possible wage offer and the lowest possible effort level. In a generic **linear *public goods experiment***, players have a token endowment they can invest in any proportion in a private project in their own favor or a public project to be shared equally among the group, where the experimenter increases any donation to the public project. Investment into the public project therefore maximizes the group's aggregate earnings, but each individual has a dominant money-maximizing strategy to invest the whole endowment into the private rather than the public project. All games described in this paragraph are examples of **social dilemma *experiments*** where non-selfish cooperation causes pareto-superior outcomes, but selfish behavior prevents this from happening.

In the ***market experiment with responder competition***, the proposer decides how to split a given sum of money between herself and one of the competing responders. All n>1 responders have to decide simultaneously whether to accept or reject the proposal. If all reject, all parties receive zero payoff; if some responders accept, one of them is randomly chosen to receive the proposed amount, and all other responders receive zero. In the ***market experiment with proposer competition***, all proposers simultaneously make a proposal of how to split a given sum of money with a single responder who can accept one of the proposals or reject all of them. The accepted proposal is implemented, while all other proposers earn zero. If all proposals are rejected, all players earn zero payoff.

The predictions for the experimental games mentioned above assume that the games are played one-shot and that interactions are anonymous.

And third parties in *third party punishment experiments* (Fehr and Fischbacher 2004) who observe that other individuals have been treated unfairly punish the perpetrators at a cost to themselves. Interestingly, many of these other-regarding behaviors have also been observed at very high stake levels of up to $10'000 (Slonim and Roth 1998; Cameron 1999; Fehr, Tougareva and Fischbacher 2014; Larney, Rotella and Barclay 2019; Dwyer et al. 2023).

The robust documentation of other-regarding behaviors in experiments has led many observers to assume that in addition to self-interest, people also care for other people's payoffs (Thibaut and Kelley 1959; Messick and Mcclintock 1968). However, one fundamental question that has remained unanswered for a long time is how one can reconcile the existence of social preferences with evidence from other paradigmatic experiments that appear to indicate that people are largely selfish. How can we explain this without arbitrarily assuming that individuals' preferences are different across experiments? For example, in a *market experiment with responder competition* (see Box 2), there is not just one responder as in the bilateral ultimatum experiment but several, so that the responders compete with each other for the share of the surplus the proposer offers. In this experiment, the parties behave in a much more self-interested manner – compared to the bilateral ultimatum experiment: the proposers make much more unequal offers, and the rejection of these uneven offers decreases dramatically, allowing the proposer to appropriate the lion's share of the surplus (Fischbacher, Fong and Fehr 2009).

In a *market experiment with proposer competition*, one adds several competing proposers to a bilateral ultimatum experiment. Here again, the players behave much more in line with the self-interest prediction because the competition among the proposers drives up their offers such that the responder appropriates almost the whole surplus, i.e., responders accept a distribution of payoffs (Roth et al. 1991; Fischbacher, Fong and Fehr 2009) that appears very unfair. These facts are also consistent with the observations from competitive double auctions and competitive posted offer markets that indicate that the observed prices and quantities traded in these markets tend to quickly converge to the competitive equilibrium derived from selfish preferences (Smith 1982). Similarly, cooperation rates are often extremely low in the final period of finitely repeated public goods experiments, seemingly indicating that social preferences may play no role when sufficient learning has been possible.

How is the claim that many people have social preferences compatible with these facts? Does a little bit of learning or competition wipe out social preferences, i.e., do they simply vanish under competitive pressure; even worse, is behavior in these paradigmatic experiments that document social preferences just a strange behavioral aberration or can models of social preferences also explain the conditions under which other-regarding individuals behave as if they were completely self-interested? Section 3 below will discuss how these puzzles could be solved by distributional theories of social preferences.

3. Fundamental Properties of Distributional Preferences

3.1. Social Preferences over payoff distributions

3.1.1. Models of distributional preferences

A number of social preference models (see Box 3) are based on the assumption that people care about the distribution of payoffs between themselves and a set of relevant reference agents (Fehr and Schmidt 1999; Bolton and Ockenfels 2000; Andreoni and Miller 2002; Charness and Rabin 2002; Fisman, Kariv and Markovits 2007). These models are a natural starting point for the modelling of social preferences because the behavioral patterns observed in the paradigmatic experiments discussed above strongly suggest that some sort of interdependent preferences are at play. Another main motivation for the construction of some of these models (Fehr and Schmidt 1999; Bolton and Ockenfels 2000) was the existence of important facts from (i) market experiments with proposer and responder competition and final periods of public good experiments that appeared to contradict the existence – or at least the widespread relevance – of social preferences, and (ii) the observation that the same people are willing to engage in seemingly contradictory behaviors by increasing another's payoff at a cost to themselves in some situations, while decreasing the other's payoff in other situations. For example, positive transfers in the dictator experiment indicate that the dictators value the recipients' payoff positively, while rejections in the ultimatum experiment reduce the proposer's payoff and thus indicate a negative evaluation of the proposer's payoff. Likewise, a rise in the effort level in response to a higher wage leads to an increase in the employer's payoff in the gift exchange experiment, while the third party's sanctions in the third-party punishment experiment reduce the dictator's payoff in that experiment.

Thus, the question is whether social preference models can account for these seemingly contradictory facts. Obviously, a simple model of altruism that assumes that other-regarding

individuals value others' payoffs positively cannot account for these facts nor can a model of spiteful/envious preferences, where individuals value others' payoff negatively, do so. Since simple models of altruism and spite cannot explain these facts, the concept of a social value orientation (SVO), that social psychologists developed (Liebrand 1984; Liebrand and Mcclintock 1988; Van Lange et al. 1997) is not capable of doing so because the SVO concept views an other-regarding individual as *either* altruistic/cooperative *or* as envious.

Surprisingly, however, relatively simple models of other-regarding preferences over payoff-distributions – such as those by Fehr and Schmidt (1999) and Bolton-Ockenfels (2000) – go a long way towards reconciling these facts. While this does not mean that these models are necessarily empirically correct – because this requires further empirical testing (see below) – it means that widespread seemingly selfish *behaviors* in certain strategic experiments can be perfectly consistent with the existence of widespread *social preferences*.

In the two-player case, a simple, linearized, version of player i's distributional social preferences $U_i\,(\pi_i, \pi_j)$ can be written as

$$U_i = (1-\alpha)\pi_i + \alpha\pi_j \qquad if\ \pi_i < \pi_j \tag{1a}$$

$$U_i = (1-\beta)\pi_i + \beta\pi_j \qquad if\ \pi_i \geq \pi_j, \tag{1b}$$

where both α and β are from the interval (-1, +1); π_i and π_j represent the material payoffs of players i and j, respectively; α is the weight on player j's payoff when *i* is behind ($\pi_i < \pi_j$), while β is the weight when *i* is ahead ($\pi_i > \pi_j$). For simplicity, we drop the subscript *i* for the weights that player *i* assigns to the other player's payoff.

Depending on the parameters α and β, this model captures a number of distributional social preference motives discussed in the literature. If $\alpha = \beta = 0$, players are selfish and their indifference curves in the (π_i, π_j) space are vertical (see Figure 1a). If both α and β are negative, a player has *competitive or spiteful* social preferences because he or she always values the other player's payoff negatively.

The indifference curves of *spiteful* players in (π_i, π_j) space are illustrated in Figure 1b. Equations (1a) and (1b) can also capture the *inequality averse preferences* described in Fehr and Schmidt (1999), where subjects dislike both being ahead and being behind. This preference implies that subjects value the other player's payoff positively if ahead but negatively when behind, which can be captured in terms of the parameters in equations (1a)

and (1b) by $\alpha < 0$ and $\beta > 0$.[4] The indifference curves of inequality averse players are illustrated in Figure 1c below. The key property of these indifference curves is that individuals are willing to sacrifice resources to reduce the other player's payoff in order to diminish disadvantageous inequality. In the two-player case, the social preferences Bolton-Ockenfels (2000) assume are quite similar to those of Fehr and Schmidt, i.e., the players basically dislike being either behind or ahead.[5] Therefore, if a player is behind, she values the other player's preferences negatively, but she values them positively when ahead.

It is easy to see how social preferences like those in Fehr and Schmidt (1999) and Bolton-Ockenfels (2000) can explain that people sometimes increase and sometimes decrease other agents' payoffs. In the Fehr-Schmidt model, the criterion for when individuals switch from being benevolent to malevolent is whether they have more or less than the other's material payoff. As Fehr and Schmidt (1999) point out, and as discussed in depth in in Section 3.2, this criterion may not always be the empirically relevant one – because equity is not always identical to equality – but it may nevertheless be a useful criterion in many lab and field environments. In the Bolton-Ockenfels model, the criterion for switching from benevolent to malevolent is whether the player receives an equal share of the overall surplus. $\bar{\alpha}_i$ $\bar{\beta}_i$

How can these theories explain why other-regarding individuals behave very selfishly in competitive market experiments? A simple example may suffice to illustrate this. Consider a population of players consisting only of people with a relatively strong aversion against unequal payoffs in the sense of Fehr-Schmidt (e.g., assume they reject every offer in a *bilateral* ultimatum experiment that is below 40% of the available pie). Now put these people into the responder position of a market experiment with *two competing responders* who face a very unfair offer of, say, 3% of the pie. The rules of the game are such that if one responder accepts and the other rejects, the accepting responder gets the 3% and the proposer gets 97%. If both accept, the responder who receives the 3% will be randomly drawn and if both reject, all three players receive nothing.

[4] To show this more explicitly: In the two-player case inequality aversion is defined by Fehr and Schmidt as $U_i = \pi_i - \bar{\alpha}_i(\pi_j - \pi_i)$ $if\ \pi_j > \pi_i$ and $U_i = \pi_i - \bar{\beta}_i(\pi_i - \pi_j)$ $if\ \pi_i > \pi_j$ with both $\bar{\alpha}_i > 0$ and $0 < \bar{\beta}_i < 1$, implying that players dislike inequality. These utilities can be rewritten as $U_i = (1 + \bar{\alpha}_i)\pi_i - \bar{\alpha}_i\pi_j$ if $\pi_j > \pi_i$ and $U_i = (1 - \bar{\beta}_i)\pi_i + \bar{\beta}_i\pi_j$ $if\ \pi_i > \pi_j$. Define $\alpha = -\bar{\alpha}_i$ and $\beta \equiv \bar{\beta}_i$ to arrive at equations (1a) and (1b).

[5] In the two-player case, a version of the Bolton and Ockenfels preferences can be written as $U_i = \pi_i + f(\sigma)$ where $\sigma = \pi_i/(\pi_i + \pi_j)$ measures A's relative payoff. $f(\sigma)$ reflects the other-regarding part of an individual's utility function; it is increasing in σ for $\sigma <$ ½ (i.e., $\pi_i < \pi_j$) and decreasing for $\sigma >$ ½ (i.e., $\pi_i > \pi_j$). Note that the preferences in Figure 1c represent a piece-wise linear approximation of Bolton-Ockenfels preferences.

Box 3: Models of Distributional Preferences

Let π_i and π_j represent the material payoffs of players i and j, respectively. In the **Fehr-Schmidt (1999)** model, individuals are assumed to derive disutility from inequitable outcomes. Inequity is formalized as inequality in the experimental games under consideration. This led to the following utility function:

$$U_i = \pi_i - \frac{\bar{\alpha}_i}{n-1}\sum_{j\neq i}\max(\pi_j - \pi_i, 0) - \frac{\bar{\beta}_i}{n-1}\sum_{j\neq i}\max(\pi_i - \pi_j, 0), \quad (2)$$

where $\bar{\alpha}_i > 0$ measures the disutility from the average disadvantageous inequality while $\bar{\beta}_i > 0$ measures the disutility from the average advantageous inequality. The Fehr-Schmidt model implies that players care positively for others' payoffs if they are better off than the other player (advantageous inequality) and negatively if they are worse off than the other player (disadvantageous inequality).

The model by **Bolton-Ockenfels (2000)** stipulates a utility function

$$U_i = f_i(\pi_i, \sigma_i), \quad (3)$$

i. e., a player's utility is a function f_i that depends positively on the player's own material payoff and on the relative share σ_i that the player receives in the game. If the total payoff in the game, $\Pi = \sum_{j=1}^{n}\pi_j$, is positive, the relative share is given by $\sigma_i = \pi_i/\Pi$, but if the total payoff is zero, σ_i is assumed to be equal to 1/n. The model assumes that individuals derive additional utility from a higher σ_i if $\sigma_i < 1/n$ while if $\sigma_i > 1/n$, a higher σ_i is utility-decreasing. In the Bolton-Ockenfels model, players do not care about the payoffs of specific other players as long as their "own" relative share σ_i is unaffected. In the two-player case, individuals care positively for the other's payoff if they are better off and negatively if they are worse off than the other player.

The distributional preferences in **Charness and Rabin (2002)** are assumed to be given by

$$\begin{aligned} U_i &= (1-\lambda_i)\pi_i + \lambda_i\, W(\pi_1, \pi_2, \ldots, \pi_n) \\ &= (1-\lambda_i)\pi_i + \lambda_i[\delta_i \min(\pi_1, \pi_2, \ldots, \pi_n) + (1-\delta_i)\,\Pi] \end{aligned} \quad (4)$$

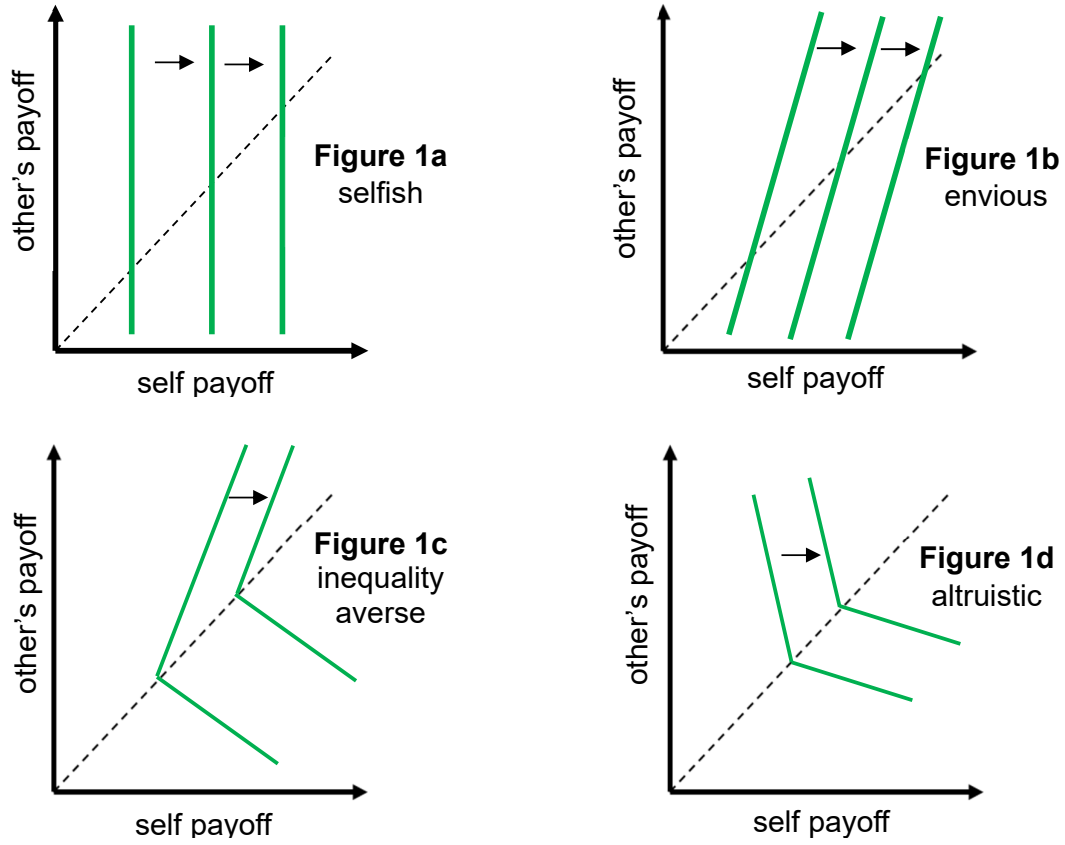
Here, subjects value a *social welfare function* $W(\pi_1, \pi_2, \ldots, \pi_n)$ positively with weight $\lambda_i \in [0,1]$, while a subject's own payoff π_i has weight (1-λ_i). $\lambda_i > 0$ indicates "other-regardingness", i.e., that player *i* cares in some way for others' payoffs. The social welfare function has two arguments: the payoff of the least well-off player enters with weight $\delta_i \in (0,1)$, and the sum of the payoffs Π enters W with weight (1-δ_i). Thus, if player *i* cares for others' welfare ($\lambda_i > 0$) and player *i* cares for the total payoff ($0 < \delta_i < 1$), player i values the payoff of *each* other player positively *regardless of whether player i has a higher or a lower payoff than the other player*. The Charness-Rabin model therefore predicts altruistic behavior if player *i* cares for the total payoff, and this tendency is particularly pronounced towards the player with the smallest payoff.

Altruistic CES preferences, as in Andreoni and Miller (2002) and Fisman, Kariv and Markovits (2007), take the form

$$U_i(\pi_i, \pi_j) = \left[(1-\alpha')\pi_i^{\rho} + \alpha'\pi_j^{\rho}\right]^{1/\rho} \quad (5)$$

where $\alpha' \in [0,1]$ represents the weight on the other player's payoff, while ρ, which obeys $-\infty < \rho \leq 1$, captures the trade-off between equity and efficiency. Altruistic CES preferences and Charness-Rabin type distributional preferences share the property that the payoff of other individuals is never valued negatively.

Figure 1: Indifference curves for different types of distributional preferences



Note: The figure shows the indifference curves for different types of distributional preferences captured by equations (1a) and (1b) above. **a**. selfish preferences ($\alpha = \beta = 0$). **b**. spiteful preferences ($-1 < \alpha < 0, -1 < \beta < 0$). **c**. inequality averse preferences ($-1 < \alpha < 0, 0 < \beta < 1$). **d**. altruistic preferences ($0 < \alpha < 1, 0 < \beta < 1$). The slope of the indifference curve in the domain of disadvantageous payoffs is given by $((\alpha - 1)/\alpha)$ while the slope in the domain of advantageous payoffs is given by $((\beta - 1)/\beta)$.

If our fair-minded responders believe that the other responder will accept, they both believe that they will be unable to prevent a very unequal distribution of income. Given these beliefs, they essentially face two very unattractive options in which the proposer always gets 97% of the pie and one of the responders gets 3%. Therefore, even strongly inequality averse responders will accept the 3% offer (Proposition 3 in Fehr-Schmidt 1999). Fischbacher, Fong and Fehr (2009) provide empirical evidence that these belief mechanisms are a key driver of responder behavior in experiments with responder competition. Dufwenberg et al (2011) show more generally that social preferences do not affect individuals' demand behavior in a

perfectly competitive economy if the utility function is separable between an individual's own consumption and the consumption vectors and budget sets of others.

The example above illustrates that the mere belief that other players *behave* selfishly might sometimes induce fair-minded players to do so as well. More generally, perhaps the most important message from papers such as those of Fehr and Schmidt and Bolton and Ockenfels is that the heterogeneity of social preferences is key because – depending on the game played - the existence of selfish players can induce fair-minded players to behave as if they were selfish and the existence of fair-minded players can induce selfish players to behave as if they were fair-minded.[6]

The simple linear social preference model described in equations (1) can also capture the distributional preferences assumed in Charness and Rabin (2002) that are described in more detail in Box 3. In the two-player case, the weights given to the other player's payoff in the disadvantageous and the advantageous domain, respectively, are given by[7]

$$\alpha = \frac{\lambda(1-\delta)}{1+\lambda(1-\delta)} \tag{6a}$$

$$\beta = \frac{\lambda}{1+\lambda(1-\delta)} \tag{6b}$$

Thus, if *i* cares about welfare ($\lambda > 0$) *and* about the total payoff ($0 < \delta < 1$), she values the other's payoff positively even in the disadvantageous payoff domain, and the valuation is even higher in the advantageous domain. Note, however, that the only reason why a Charness-Rabin player cares for the other player's payoff in the disadvantageous domain is a concern for the total payoff. Figure 1d illustrates the indifference curves this model implies.

From a behavioral viewpoint, the Charness-Rabin preferences are qualitatively quite similar *in the two-person case* to non-linear other-regarding preferences as captured by a CES utility function (see Box 3). Andreoni and Miller (2002) and Fisman, Kariv and Markovits (2007) have assumed CES preferences. However, while the CES utility function is capable of

[6] Selfish proposers in the ultimatum experiment have an incentive to make relatively fair offers if fair-minded responders reject low offers. Selfish "employers" in the gift-exchange experiment have a reason to make fair wage offers if fair-minded workers respond to fair offers with higher effort. Selfish players in public goods experiments with punishment have an incentive to cooperate if cooperative individuals punish defectors. In all cases, the existence of players with social preferences can induce fundamental changes in the selfish players' incentives.

[7] In the case where *i* earns more than *j*, utility is given by $U_i = (1-\lambda)\pi_i + \lambda[\delta\pi_j + (1-\delta)(\pi_i + \pi_j)]$ while $U_i = (1-\lambda)\pi_i + \lambda[\delta\pi_i + (1-\delta)(\pi_i + \pi_j)]$ represents the utility in the case where *j* earns more than *i*. Reformulating the first function in terms of the weight α on the other player's payoff and the second function in terms of the weight β on the other's payoff yields the expressions in the text.

capturing very similar behaviors, the Charness-Rabin model provides a psychological foundation for altruistic behaviors in terms of a motive to increase the total surplus and to help the worst-off player. Knowing that these motives may underlie other-regarding behaviors can provide a deeper understanding of these behaviors.

As in Charness and Rabin (2002), subjects with these CES preferences always value others' payoffs positively and preferences are linear and given by $U_i\left(\pi_i, \pi_j\right) = (1 - \alpha)\pi_i + \alpha\pi_j$ in the special case of $\rho = 1$. It is well known that preferences are Cobb-Douglas for the case of $\rho = 0$, while social preferences take the Rawlsian form $U_i\left(\pi_i, \pi_j\right) = \min(\pi_i, \pi_j)$ for $\rho \to -\infty$. In the latter case, indifference curves in the $\left(\pi_i, \pi_j\right)$-space become horizontal for $\pi_i > \pi_j$ and vertical for $\pi_i < \pi_j$ with a kink at the 45° line, implying that subjects' social preferences are strictly egalitarian. These egalitarian preferences arise in Charness and Rabin for the two player case if subjects completely disregard their own payoffs (i.e., $\lambda = 1$) and the sum of payoffs (i.e., $\delta = 1$), but care only for the payoff of the least well-off player.

As in Charness and Rabin, CES preferences can thus capture a high willingness to pay to increase the other's payoff when ahead and a very low willingness to pay when behind. However, similar to Charness and Rabin's distributional preferences, CES preferences cannot capture subjects' willingness to reduce others' payoffs at a cost to themselves because the other player's payoff is never valued negatively.

3.1.2. Empirical Frequency of Different Distributional Preferences

In the following, we will discuss the empirical properties of distributional preferences that emerged from more systematic attempts to identify the relevant parameters. Fehr-Schmidt (1999) and Bolton-Ockenfels (2000) were inspired by the evidence from the paradigmatic experiments. Their papers provided a unifying account for important, and seemingly contradictory, empirical regularities but they did not involve a systematic attempt to identify the model parameters empirically. To achieve identification of the parameters of distributional preferences like those in equations (1), one needs information about subjects' behavior on many budget lines with different slopes in the $\left(\pi_i, \pi_j\right)$ space. In other words, one needs a large set of generalized dictator experiments in which one subject can unilaterally determine the distribution of payoffs between herself and another subject at various costs of redistributing payoffs. By systematically varying the slope of negatively sloped and

positively sloped budget lines, it is possible to identify individuals' indifference curves in the domain of advantageous and disadvantageous inequality and to determine which of the fundamentally different distributional preference types illustrated in Figure 1 describes the behavior of (sub)groups or individuals best.[8] In addition, this approach also enables a clean separation of truly selfish individuals from other-regarding individuals because it allows for deviations from selfishness when the costs of other-regarding acts are low.[9]

Concern for equality or concern for the total payoff?

Charness and Rabin (2002) took an important step towards answering this question in a study with roughly 220 undergraduate students in 29 distinct experimental games or allocation tasks. In one of their tasks 69% of the participants chose (Own, Self) payoffs of (400, 750) rather than (400, 400). This result, and similar other results (Charness and Grosskopf 2001) imply that roughly two thirds of these student populations are unwilling to reduce others' income for the sake of equality, implying that $\alpha \geq 0$. However, these results do not imply that subjects have a positive concern for the total payoff because it costs nothing to choose the allocation that maximizes the total payoff. Therefore, choosing the allocation that maximizes the total payoff in this task is compatible with selfishness (i.e., with $\alpha = 0$).

A positive concern for the total payoff is, however, indicated in an allocation task from Charness and Rabin (2002) where 50% of the subjects choose (375, 750) rather than (400, 400), implying that half of the students are willing to pay \$1 to increase the other's payoff by \$14. Based on these results, Charness and Rabin conclude that "inequality reduction is not a good explanation of Pareto-damaging behavior" (p. 819).

[8] There is a considerable literature on social value orientation (SVO) in psychology that uses the so-called ring measure of SVO (Liebrand 1984; Liebrand and Mcclintock 1988) and/or the triple-dominance measure of SVO (Van Lange et al. 1997; Van Lange 1999) or the slider task (Murphy, Ackermann and Handgraaf 2011; Murphy and Ackermann 2014). These measures are based on generalized dictator experiments, but they do not lend themselves easily to the estimation of utility functions. For example, they do not allow us to differentiate between the four qualitatively different types of indifference curves displayed in Figure 1. Instead, they use measures such as the ratio between the total payoff given to the other player and the total payoff assigned to "self" (across all dictator experiments) to assign individuals to predefined SVO types such as "cooperative" (= desire to maximize joint gains), "altruistic" (= desire to maximize the other player's payoff), and "competitive" (= desire to maximize the payoff difference). In particular, the SVO measures cannot identify inequality aversion as defined in Fehr and Schmidt (1999) or Bolton-Ockenfels (2000). For this reason, we do not use them in the following review of distributional preferences.

[9] Some people may be inclined to discount situations in which the cost of other-regarding acts is low, but social life is in fact pervaded by situations in which low-cost favors, and low-cost punishments, can be given to other people. When a colleague in the workplace asks for help, when a stranger in a city asks for directions, or when a colleague punishes you for your performance with a dismissive smile, the costs for the helper or punisher are often very low, while the benefits and costs for the receiving party are high.

Engelmann and Strobel (2004) conducted experiments with roughly 380 undergraduate economics and business administration students. They studied choices in which the chooser would receive the same material payoff in all available choices. Participants had to choose between different three-person allocations that enabled the authors to disentangle the motive (i) to maximize the groups' joint payoff, (ii) to reduce overall inequality as predicted by the Fehr-Schmidt or the Bolton-Ockenfels model, or (iii) to maximize the payoff of the worst-off individual in the three-person group (max-min motive). Their results support the Charness-Rabin model because motive (i) and (iii) play a substantial role, while motive (ii) appears to play no significant role in their three-person dictator experiments.

However, the generality of this conclusion has been challenged by studies that show that non-economists put a substantially higher weight on equality – compared to the maximization of the group's overall payoff – than economists do (Fehr, Naef and Schmidt 2006). While a majority of economists and business administration students preferred total payoff maximization over equality, various groups of non-economists, ranging from students of various other disciplines to low-level employees of banks and financial institutions, showed the opposite pattern.

Many other studies have, in the meantime, estimated the structural parameters α and β of model (1) above based on observations from strategic and non-strategic experiments. A recent meta-analysis by Nunnari and Pozzi (2022) uses data from 41 articles, comprising student and non-student populations, with a total of 297 estimates of the different populations' *average* α and β values. The authors estimate the overall distribution of parameters across populations and find weighted mean values for $\alpha = -0.434$ and $\beta = 0.337$, both significant at $p < 0.001$. This indicates that across all the (strategic and nonstrategic) experiments in the meta-analysis the observed behaviors are consistent with inequality aversion for the representative agent. A similar qualitative conclusion is reached if one restricts attention to the dictator experiments, which arguably offer the cleanest preference interpretation of the estimated parameters: the mean values of α and β are then given by $\alpha = -0.266$ and $\beta = 0.387$ (both significant at $p < 0.001$), again indicating that subjects display inequality aversion *on average*. In contrast, in strategic games – where players are more likely to view each other as opponents – disadvantageous inequality aversion is much higher ($\alpha = -0.727$) while advantageous inequality aversion is significantly lower ($\beta = 0.214$).

Preference estimates of the representative agent typically hides important heterogeneity at the individual level. A systematic analysis of individual heterogeneity in distributional preferences has been conducted in the seminal papers by Andreoni and Miller (2002) and Fisman, Kariv and Markovits (2007), who estimated individuals' CES utility functions (see Box 3). Their results indeed show enormous heterogeneity in subjects' preferences. In addition, the Fisman-Jakiela-Kariv-Markovits group shows striking differences between students and the general population's distributional preferences in a series of papers: the general population is much more other-regarding (i. e., has a much higher α' in equation (5)) and puts a much higher weight on equality (i.e., displays $\rho < 0$ in equation (5)) compared to students. This also holds if one controls for age.[10] Because of space limitations and because the CES approach neglects important classes of distributional preferences –inequality aversion and envy – we describe the results of these studies in more detail in Online Appendix 1.

How prevalent are inequality aversion and spite?

Kerschbamer (2015) developed a systematic approach – the Equality Equivalence Test (EET) – that enables the identification of all four distributional preference types displayed in Figure 1. For this purpose, subjects are presented with choice lists in the domain of disadvantageous inequality (DA-lists) and the domain of advantageous inequality (A-lists). In any given list, the subjects face a series of binary choices where the equal payoff distribution E is always paired with an alternative allocation (See Figure A2 in Online Appendix 2). A list essentially confronts subjects with a series of positively and negatively sloped budget lines in the "self-payoff / other payoff" space with two discrete options on each budget line.

Table 1 shows the results of a relatively large number of EETs in student samples (Kerschbamer (2015); Balafoutas, Kerschbamer & Sutter (2012); Paetzel, Sausgruber & Traub (2014) ; Balafoutas et al. (2014); Krawczyk and Lee (2021)) and in broad, demographically diverse, population samples (Chapman et al. 2018; Kerschbamer and Muller 2020; Hedegaard et al. 2021). Several striking facts emerge from these studies. First, the share of selfish individuals is much larger in the student samples, where it varies between 29% and 58%, while it varies only between 5% and 20% in the broad population samples.

[10] The large differences between student samples and the broader population are consistent with research reported in Snowberg and Yariv (2021) and Cappelen et al. (2015).

Table 1 Empirical Frequency of Different Distributional Preference Types

			Preference types in the Equality Equivalence Test			
Study		**Subject Pool**	**Altruistic**	**Inequality Averse**	**Envy/ Spite**	**Selfish**
Kerschbamer (2015)	**Student Samples**	N = 92 Austria	33.7%	11.9 %	3.2%	48.9%
Krawczyk & Lee (2021)		N = 101 Poland	48.5%	11.9%	9.8%	28.7%
Balafoutas et al. (2012)		N = 132 Austria	28%	8.3%	5.3%	58.3%
Balafoutas et al. (2014)		N = 195 Austria	49.7%	6.7%	7.2%	33.8%
Paetzel et al. (2014)		N = 280 Germany	29.3%	9.3%	1.8%	58.3%
Hedegard et al. (2021)	**Broad population samples**	N = 885 Denmark	47.1%	23.2%	8.6%	20.0%
Kerschbamer & Müller (2020)		N = 2794 Germany	13.4 %	64.8%	14.0%	5.0%
Chapman et al. (2018)		N = 1000 USA	27.5%	41.9%	8.3%	16.6%
2nd Wave German Internet Panel[11]		N = 2583 Germany	11.5%	67.8%	11.0%	7.9%
			Preference types in endogenously emerging preference clusters			
			Altruistic	**Inequality Averse**	**Predominantly Selfish**	
Epper et al. (2020)	**Broad population samples**	N = 3691 Denmark	30.2%	37.3%	32.5%	
Fehr®Epper® Senn (2021)		N = 816 Switzerland 2017	34.4%	50.8%	14.8%	
Henkel®Fehr ®Senn®Epper (2024)		N = 916 Switzerland 2020	30.5 %	45.5%	24%	

Note: All studies mentioned in the table are based on incentivized one-shot experiments with anonymous partners. An individual is classified as altruistic in the EET (in terms of the parameters of equation 1a and 1b) if her choices imply $\alpha \geq 0$ and $\beta \geq 0$ with at least one strictly positive. To be classified as envious/spiteful, either $\alpha < 0$ *and* $\beta \leq 0$ or $\alpha \leq 0$ *and* $\beta < 0$ holds. Individuals are classified as inequality averse in the EET if $\alpha < 0$ *and* $\beta > 0$ holds. The derivation of endogenous preference clusters in the three final studies is based on the application of a nonparametric Bayesian clustering algorithm (Dirichlet Process Means).

Second, the much larger share of other-regarding subjects in the broad population samples is primarily due to a much larger share of inequality averse individuals in these samples. While the share of inequality averse individuals in student samples varies between 7% and 12%, this

[11] The results in the Kerschbamer and Müller (2020) paper are based on a first application of the EET to the German Internet Panel (GIP) in 2016. In the meantime, the EET was performed with this sample a second time in 2018. We therefore also include the results of this second wave of data collection to illustrate the stability of the preference classification over time in Table 1.

share is between 23% and 68% in the broad population samples. Third, the share of envious subjects is also slightly higher in the broad population samples, where it varies between 8% and 14%, while in the student samples it varies between 2% and 10%.

Finally, the share of individuals with altruistic preferences is considerably larger in student samples than in broad population samples. The average share of altruistic subjects over samples consisting only of students amounts to 37%, while the average share of altruists is only 19% in the broad samples. Taken together, the general population seems much more other-regarding and this higher share of other-regarding individuals reflects a much larger share of inequality averse subjects and a somewhat larger share of envious subjects. Interestingly, however, altruism is more prevalent in student samples.[12]

The existence of large differences between the student samples and the general population samples are corroborated by two earlier studies by Bellemare, Kröger and van Soest (2008; 2011). Using a representative sample of the Dutch population that played ultimatum and dictator experiments, they estimated a non-linear version of the Fehr-Schmidt model and find that the young and educated individuals in their sample display a substantially lower aversion against disadvantageous inequality than the rest of the sample.

The main findings from the EETs are further corroborated by studies that identify preference clusters without committing to a pre-specified number of preference types (Fehr®Epper®Senn (2023), Epper et al. (2020), Henkel et al.(2024))[13]. In these studies, a non-parametric Bayesian clustering algorithm ("Dirichlet Process Means, DP-Means) is applied to two large Swiss and one Danish data set (see Table 1) where subjects faced many positively and negatively sloped budget lines that allow for the identification of distributional preferences. The algorithm determines the number and the behavioral properties of the preference clusters endogenously and assigns each individual to one of the clusters.

The clustering algorithm yields the same three (qualitative) preference clusters in all three broad population samples: an altruistic cluster, an inequality averse cluster, and a cluster of individuals who behave predominantly selfishly.[14] Moreover, the inequality averse cluster represents the majority group and the altruistic and inequality averse cluster together

[12] When comparing student samples and broad population samples one may worry about stake size effects because in the general population income variation is much larger. However, the prevailing evidence (e.g., in Epper®Senn®Fehr (2024)) indicates that income levels do not affect the estimates of α and β.
[13] The symbol ® indicates that the author order has been randomly determined.
[14] The frequency of envious/spiteful subjects is generally small and their location in preference space is scattered; they thus do not constitute a distinct cluster but are subsumed either under the selfish or the inequality averse cluster.

comprise between two-thirds and 85% of the sample across the different data sets. Thus, the endogenous clustering approach is also in line with the conclusion that inequality aversion and altruism are the dominant patterns of distributional preferences in the general population, while predominantly selfish subjects constitute a smaller, yet significant, share of the population.

3.2. The Role of Merit and Luck in Distributional Preferences[15]

In the standard dictator experiment and other pure distribution experiments, the experimenter exogenously provides the income that can be divided among the parties. Moreover, the role of the powerful player, who can make a unilateral allocation decision, is assigned randomly, and the subjects typically interact anonymously with each other and have no information about the other players' income, wealth, and social background. It seems natural that no player in such an environment has a priori a greater normative claim on the available resources compared to the other players, i.e., equality is a natural reference point for judging the fairness and equitability of outcomes. It is, however, also clear that social background, the saliency and status of particular individuals, the social proximity among individuals, as well as their effort and contribution to the available resources, including the risks and conditions under which they had to produce these resources, can play a role in what is considered an equitable distributional claim. The potential importance of these factors is widely acknowledged and may often drive a wedge between equality and equity (fairness) and offers ample opportunities for examining these factors in a controlled way.[16]

In this context, the notions of merit, luck, and effort play a key role. In many modern societies. meritocracy seems a widely held normative ideal. A prominent theory of equality of opportunity (Roemer and Trannoy 2015) argues that individuals should be held responsible for factors under their own control, while they should not be held responsible for factors beyond their control. Based on this approach, there is no normative reason to redistribute

[15] Due to space constraints, we have moved the discussion of the role of risk in social preferences to Online Appendix 4.

[16] The very fact that equity and equality may diverge induced Fehr and Schmidt to first call the preferences they examine as "*inequity* aversion", i.e., "a general dislike for outcomes that are perceived as inequitable". However, under the special circumstances in the laboratory experiments they examined, where subjects entered the laboratory as equals, knew nothing about each other's background, and were randomly allocated to different roles, they assumed that equality may be a reasonable reference point.

resources that are entirely generated by an individual's effort, while earnings that accrue at least partly from non-controllable external factors ("luck") should be subject to redistribution.

3.2.1. Equity and Entitlements

Psychologists and sociologists (Homans 1961; Adams 1963; Adams 1965) have developed positive theories of justice that incorporate a widely applicable merit principle. According to this view, individuals perceive inequity if their outcome/input ratios diverge from the ratio of their relevant comparison agent. The outcomes and inputs that enter the equity calculus have typically been interpreted very broadly, i.e., any kind of reward that the individuals experience can represent an outcome and any kind of perceived contribution to the outcome can represent an input.

Equity theory thus allows for many subjective influences that may make its predictions rather malleable and imprecise. However, it makes sharper predictions if restricted to more easily measurable outcomes (e.g., received material payoffs) and inputs (e.g., time spent on a task or material payoff contributed to the group payoff). Mikula (1973) conducted, for example, an experiment with recruits from the Austrian Army to examine whether subjects follow the equity principle or the equality principle. Two matched parties produced a joint monetary payoff, and then one of the individuals could unilaterally allocate this payoff to the two parties. He found that the parties with poorer performance assigned themselves more than the proportional monetary payoff equity theory predicted but less than the equal monetary payoff. Likewise, the parties with higher performance requested *on average* more than the equal split but less than the reward equity theory predicted. For example, when the performance ratio between the poorer and better performers was between 62.5% and 37.5%, the higher performers requested on average 54.5% of the joint payoff, while the lower performers requested 42.7%. More generally, the data confirm the qualitative equity theory prediction that better performers request (if they have the power to decide) and are conceded (if the partner has the power to decide) a higher share of the joint payoff. Likewise, poorer performers allocate themselves and are allocated a lower share of the joint payoff (Leventhal and Michaels 1969; Leventhal and Anderson 1970; Leventhal and Lane 1970; Lane and Messe 1971; Leventhal and Michaels 1971). This literature thus shows that subjects behave as if they feel entitled to a larger (smaller) share if they contribute more (less) to the joint payoff than their experimental counterpart.

One drawback in these experiments was that participants were typically deceived in various ways. Often, there was no actual working partner who contributed to a joint surplus or the performance ratios to which the parties were randomly allocated did not reflect the subjects' actual performances, or individuals had to make hypothetical choices without real economic consequences. These practices may have generated doubts about the credibility of the implemented procedures or might have affected subjects' behavior in other ways. For this reason, it makes sense to ask whether the notion of "earned entitlements" or "earned property rights" implied by equity theory is indeed robust to the methods used in experimental economics which rule out deception while implementing designs with transparent economic consequences for the involved parties.

The experimental economics literature (e.g., Hoffman et al. (1994), Ruffle (1998), Fahr and Irlenbusch (2000) Konow (2000); Cherry, Frykblom and Shogren (2002); Frohlich, Oppenheimer and Kurk (2004); Cappelen et al. (2007); Krawczyk (2010); Lefgren et al. (2016)) also strongly suggests that "earned entitlement" effects exist. For example, Hoffman et al. (1994) and Cherry et al. (2002) show that the dictators take a higher share of the pie when they earned the role of the dictator in a quiz or when they generated the pie that can be distributed.[17] While Hoffman et al. (1994) and Cherry et al. (2002) show that the dictators behave more selfishly if they acquired an earned entitlement, Konow (2000) and Ruffle (1998) show that the dictators also respect the recipients' earned entitlement, i.e., the recipients' earned entitlements also constrain the dictators' selfish behavior. Subjects in Konow's experiment *jointly* produce the pie to be distributed in the subsequent dictator experiment in a (letter production) task that enables the exact measurement of the matched subjects' relative contribution (i.e., the number of letters). He hypothesizes that fairness considerations will induce the dictators to tilt the allocation given to the recipients towards their relative contribution in the production task – a finding that his data nicely corroborates.

[17] Hoffman et al. (1994) also claim that if one not only introduces anonymity between the subjects but also anonymity between the subjects and the experimenter, the dictators behave more selfishly and take a higher share. However, subsequent research (Frohlich, Oppenheimer and Kurk 2004) suggests that the way Hoffman et al. implemented experimenter-subject anonymity also generated doubts among the subjects about whether the recipient was a real person, which may well induce selfish behavior. Other research on double anonymity (e.g., Bolton, Katok and Zwick 1998; Barmettler, Fehr and Zehnder 2011) indicates that lack of subject-experimenter anonymity has only minor, insignificant effects in dictator, ultimatum, and trust experiments.

3.2.2. Modelling Entitlement Effects

How should we model the motivational forces underlying the earned entitlement effect? To set the stage for this discussion, it is useful to rely on a concrete example that is taken from the seminal paper by Cappelen et al. (2007) who introduced the notion of "pluralistic fairness ideals" to the literature. In their experiment, subjects participate first in two production tasks where they can finance an investment q_i that generates an output $x_i = a_i q_i$ from an identical endowment E. Each participant was randomly assigned to a high ($a_i = 4$) or a low ($a_i = 2$) rate of return on investment.

After the two production tasks, each subject was matched twice randomly with a different partner and played a bilateral dictator experiment with each of the two partners. The pie size in each dictator experiment was given by the paired players' jointly produced total output $X = a_1 q_1 + a_2 q_2$. Note that some pairs in this experiment face a total output produced with identical returns on investment ($a_1 = a_2$) and identical investments ($q_1 = q_2$), while others can distribute an output that is produced with different rates of return ($a_1 \neq a_2$) and/or different investment levels ($q_1 \neq q_2$).

In the following, we discuss the behavioral predictions for this experiment for (i) inequality averse individuals, (ii) individuals motivated by equity theory and (iii) individuals who are motivated by one of the fairness ideals stipulated in Cappelen et al. (2007). An inequality averse dictator would implement the allocation that equalizes the two players' incomes. If we denote s_1 (s_2) as player 1's (player 2's) share of the total payoff X, income equalization Implies $E - q_1 + s_1 X = E - q_2 + s_2 X$. From this, and the fact that $s_1 + s_2 = 1$, follows that player 1's share is given by

$$s_1^{IA} = \frac{1}{2} + \frac{q_1 - q_2}{2X}, \tag{7}$$

where the superscript IA refers to the predicted share under inequality aversion. Thus, inequality averse dictators assign individuals with identical investments the same income because they view them as equally meritorious, while individuals who invest more are assigned a higher income share proportional to the investment gap between the players. Inequality aversion therefore incorporates a notion of meritocracy because the principle of equality is applied to *all* material payoffs, and effort or investment costs are certainly part of this. In contrast, according to equation (7), inequality averse individuals do not honor the luck of those participants who were randomly assigned a higher rate of return a_i.

Which allocation would a dictator who is motivated by the forces stipulated by equity theory implement in the Cappelen et al. experiment? He or she would equalize the output/input ratios between the players. While a plausible interpretation of the "output" seems to be the earnings received, there is more ambiguity when considering the inputs. This depends crucially on whether one considers the investment levels q_i or the contributions to the total earnings, $x_i = a_i q_i$, as an "input". However, if we view equity theory as embodying a merit principle, the investment levels are the appropriate input measure, implying that equity is established once $s_1 X/q_1 = s_2 X/q_2$ or $s_1/q_1 = s_2/q_2$ holds.

If we consider that $s_1 + s_2 = 1$, the payoff shares predicted by equity theory, s_i^E, are given by the players' relative investment shares:

$$s_1^E = \frac{q_1}{q_1 + q_2} \qquad \text{and} \qquad s_2^E = \frac{q_2}{q_1 + q_2}. \tag{8}$$

This coincides with the notion of fairness advocated by Konow (2000) and with the liberal egalitarianism fairness ideal proposed in Cappelen et al. (2007). Note that the notion of equity described in (8) also implies that liberal egalitarians do not honor luck – in the form of randomly assigned investment returns – while honoring merit – in the form of players' relative investment shares.

In addition to the liberal egalitarian fairness ideal, Cappelen et al. introduce two further potentially relevant fairness ideals: the strictly egalitarian ideal (s_i^{SE}) and the libertarian ideal (s_i^L). According to the strictly egalitarian ideal, the earnings should be distributed equally regardless of the players' investments and rates or return, i.e., $s_i^{SE} = 1/2$. In contrast, according to the libertarian ideal each player is entitled to the earnings he or she produced regardless of any luck or effort considerations, i.e., $s_i^L = a_i q_i / X$.

What kind of utility function rationalizes the behavior of players who are motivated by a fairness ideal? To answer this question let s_i^F be the income share of *i* under a generic fairness ideal $s_i^F \in \{s_i^E,\ s_i^{SE},\ s_i^L,\ s_i^{IA}\}$. A plausible utility function that incorporates the players' fairness ideals could look similar to the Fehr and Schmidt utility function where the reference point is no longer the other players' payoff, but the dictator's own fair share as defined by s_i^F, and with actual income denoted by y_i: $\bar{\alpha}_i\ \bar{\beta}_i\ \gamma_i$

$$U_i = y_i - \bar{\alpha}_i max[s_i^F X - y_i, 0] - \bar{\beta}_i max[y_i - s_i^F X, 0] \tag{9}$$

A dictator with this utility function would never give herself less than $s_i^F X$ and would allocate herself exactly $s_i^F X$ if $\bar{\beta}_i > 1$. In contrast, Cappelen et al. (2007) assumed that deviations from the fair payoff impose symmetric (quadradic) costs on subjects, which led them to postulate the utility function:

$$U_i = y_i - \gamma_i \frac{\left(y_i - s_i^F X\right)^2}{2X}. \quad (10)$$

The two utility functions above imply that there is a nonpecuniary disutility from receiving an inequitably large share in case of $y_i > s_i^F X$. Thus, because the players' income shares add up to one, subjects in this situation are willing to reduce their own incomes and behave generously towards the other player. Likewise, the associated nonpecuniary disutility from receiving a too low share generates a willingness to pay to reduce the other player's payoff in case of $y_i < s_i^F X$.

To what extent are the different fairness ideals capable of capturing earned entitlement effects? Liberal egalitarians and inequality averse individuals who behave according to utility functions (9) or (10) will exhibit earned entitlement effects if they contributed more effort or investments to the production process. In contrast, libertarians with a fairness ideal of $s_i^L = a_i q_i / \mathrm{X}$ will show an earned entitlement effect even if they provide higher output merely because of luck (i.e., a higher a_i).

What are the empirical results of the Cappelen et al. (2007) study? Based on the assumption that there are three types of fairness ideals – s_i^E, s_i^{SE} and s_i^L – and that the subjects behave according to utility function (9), the authors estimate a mixture model that provides the relative share of individuals who are assigned to the different fairness ideals and also an estimate of the distribution of the strength of fairness preferences as captured by γ. They find that the share of strict egalitarians is 43.5%, the share of liberal egalitarians is 38.1% and the share of libertarians is 18.4%. In another study with a broader sample of Norwegian and German students (Cappelen et al. 2013), the allocation of individuals to the different fairness ideals is different, however, as only 22.5% of individuals are strict egalitarians, while 42.5% are liberal egalitarians, and 35% are libertarians.

The introduction of heterogenous fairness ideals by Cappelen et al (2007) represents an interesting innovation; it implies that if people do not share in one particular situation, then this may not mean that they do not care about fairness (i.e., put no weight on fairness in (9) or (10)) but that they consider sharing unfair. Many interpretations of experiments overlook this

distinction and too often categorize people who do not share as selfish. The distinction between the weight put on fairness and the concrete fairness ideals also opens an avenue towards studying the impact of institutions on distributional preferences. It has been argued, for example, that studying economics makes people more selfish (Frank, Gilovich and Regan 1993). An interesting question in this context is whether studying economics indeed reduces the other-regarding component in individuals' utility function (i.e., $\bar{\alpha}_i$, $\bar{\beta}_i$, or γ_i in equations (9) and (10) or whether it changes people's fairness ideals. More generally, fairness ideals are likely to be shaped by educational and political institutions as well as by the general moral infrastructure of societies.

The Cappelen et al. study deserves credit for introducing and estimating the share of subjects assigned to *heterogenous* fairness ideals that are based on different normative recognitions of "luck" and "effort". At the same time, it is clear that the assumed existence of three exogenously given fairness ideals is a strong assumption, and that it would be desirable to develop experimental designs and econometric methods that make it possible to infer individual subjects' normative reference points from the data, instead of exogenously assuming them.

Likewise, while assuming that different reference points can capture the distinction between luck and effort, it is also possible that individuals' fairness ideals directly affect their interpretation of an individual's deservingness. For example, an individual with a meritocratic fairness ideal who faces a "lazy" individual may put a different weight on being fair towards this individual compared to a situation where she faces a hard-working individual. Thus, we believe it is an important task for future research to find ways to *simultaneously* identify normative reference points *and* the strength of fairness preferences (in terms of $\bar{\alpha}_i$ $\bar{\beta}_i$ or γ_i) at the individual level, and to determine how different environments affect these parameters. [18]

Another important unresolved question is how do empirical measures of inequality aversion, altruism, and selfishness, which are identified in experiments without obvious entitlements (like, e. g., in Table 1), relate to the different fairness ideals? For example, are subjects identified as inequality averse in Table 1 liberal egalitarians or strict egalitarians?

[18] Cabeza (2021) conducted a study that varied the deservingness of players along the "effort" and "luck" dimensions. She reports that changes in deservingness are associated with changes in the strength of fairness preferences in both the domain of advantageous and disadvantageous inequality. If subjects become "more deserving", the decision-makers' willingness to behave altruistically in the domain of advantageous inequality increases while the willingness to behave enviously decreases.

Likewise, how are subjects identified as selfish in experiments without obvious entitlements allocated to the different fairness ideals? These questions are largely unexplored, but a first study by Fehr®Epper®Senn (2023) shows that individuals who are identified as selfish in experiments without obvious entitlements also behave very selfishly in experiments with entitlements, i.e., fairness ideals do not seem to matter to them. In contrast, individuals classified as inequality averse or altruistic in experiments without obvious entitlements display considerable concern for meritocracy. 70 percent of the altruists (60 percent of the inequality averse individuals) displayed a concern for meritocracy by, e. g., conceding a higher payoff to a better performing player, while 30 percent of the altruists (40 percent of the inequality averse) chose always the strictly egalitarian allocation.

This finding confirms that meritocratic concerns only motivate other-regarding individuals, and it has implications for interpreting the relevance of impartial spectator experiments that are frequently used to identify individuals' fairness ideals. In these experiments, an impartial spectator typically allocates payoffs between two *other* players, but the allocation decision does not affect the spectator's payoff. Therefore, one cannot identify the selfish individuals for whom fairness ideals have little behavioral bite when their self-interest is at stake. The assignment of fairness ideals based on impartial spectator experiments may thus run the danger of overstating the behavioral relevance of fairness ideals because if α, β or γ in equation (9) or (10) are zero, fairness ideals are behaviorally irrelevant.

3.2.3. Cultural Differences and the Relative Importance of Fairness and Efficiency Concerns

Almost all experiments documenting entitlement effects have been conducted with Western student populations. These subjects are part of an educational and employment environment that permanently evaluates their performance and requires high effort levels to pass frequent examinations, provide satisfactory work results, or move up the career ladder. Meritocratic and libertarian ideas may well flourish in this environment, raising the question about the prevalence of difference fairness ideals in the broader population and in other cultures.

Jakiela (2015) studies entitlement effects in a younger population from rural villages in Kenya and compares them with those observed in a US student population. She implemented a luck treatment, where the roll of a die determines the size of the overall budget, and an effort treatment, where either the dictator or the recipient produced the budget in a real effort

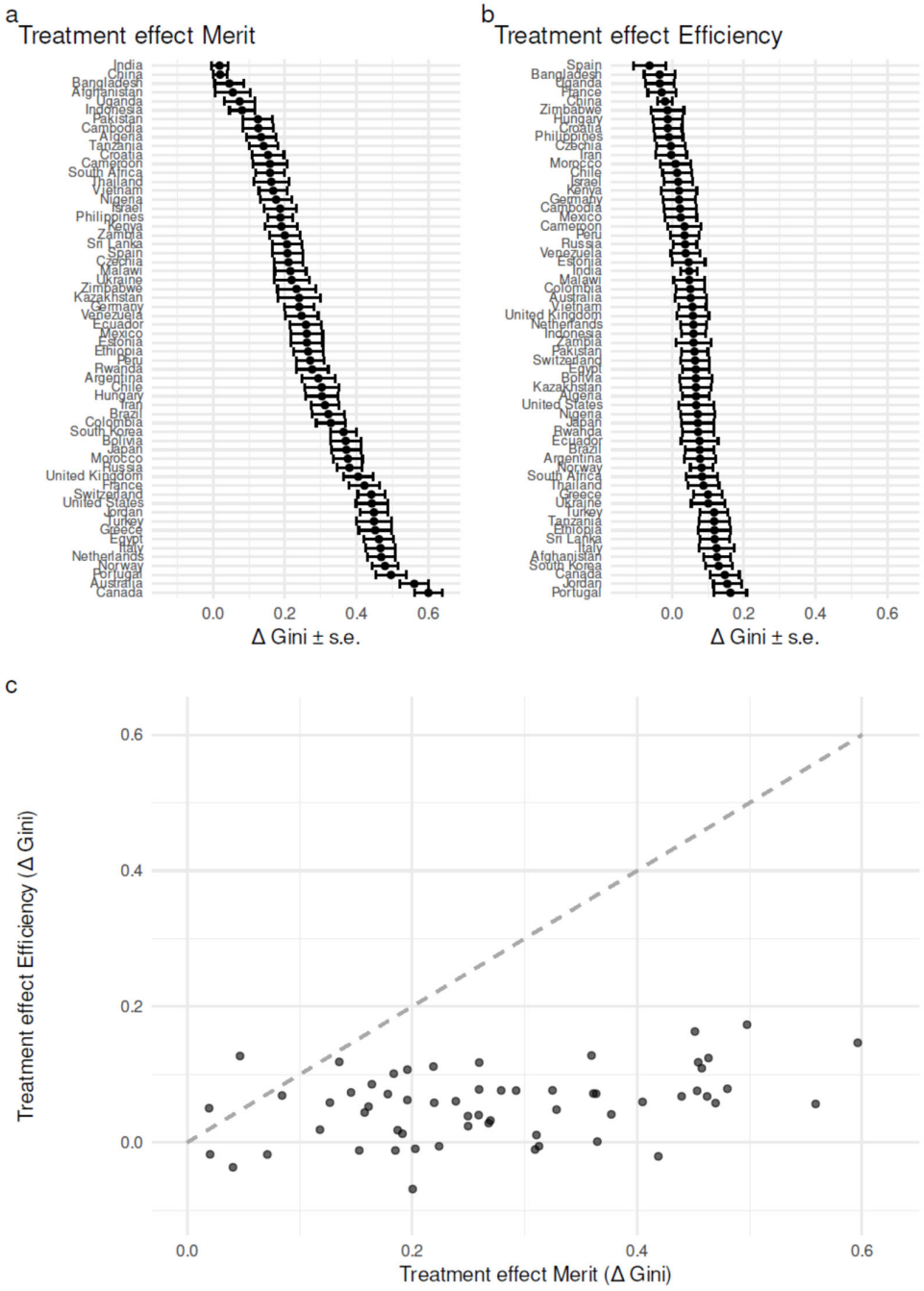
task. Jakiela's experiment is motivated by observations that these entitlement effects may not exist in poor rural communities with strong traditions of solidarity and mutual assistance (Platteau 2000).

In Jakiela's design, potential entitlement effects are strengthened with an additional feature – the so-called "Giving" and "Taking" treatment. In the Giving treatments, the budget is provisionally allocated to the dictators who also rolled the die in the luck condition and performed the real effort task in the effort condition. In the Taking treatment, the budget is provisionally allocated to the recipients who also rolled the die and performed the real effort task. Jakiela finds that the Kenyan dictators behave very similarly in the luck and the effort treatments. This contrasts sharply with the behavior of US students who showed strong behavioral differences between the Luck and Effort treatments: dictators allocated significantly higher income shares to those parties who had generated the budget through real effort rather than the role of a die.

Almas, Cappelen and Tungodden (2020) and Almas et al. (2021; 2022) were the first to study the relevance of the distinction between "luck" and "effort/merit" in *broad* population samples. In Almas et al. (2021; 2022), they conducted a large study with 65'800 participants that comprise representative country samples taken from 60 different countries around the globe. In each country, at least 1000 individuals participated in the experiment as impartial spectators who had the option to redistribute income between two subjects ("workers"). Initially, before redistribution, one of the workers had an income equivalent to US $6 while the other had US $0. In the luck treatment, a lottery draw (from an earlier experiment on Mechanical Turk in which the workers participated) generated the inequality between the subjects. In contrast, the productivity differences between the two workers in a real effort task determined the initial inequality in the merit treatment. The impartial spectator could redistribute the $6 in $1 units without any cost in both the luck and the merit treatments.

In addition to the luck and merit treatments, the authors also conducted a so-called efficiency treatment. This treatment is identical to the luck treatment except that redistribution is associated with large cost: for every $1 given to the poorer workers, the income of the worker who was initially richer is reduced by $2. With such a large redistribution cost, one would expect a substantial increase in inequality acceptance. This study enables the examination of the potential universality of the role of merit vs. luck vs. efficiency costs on preferences for redistribution in a highly controlled experimental set-up that is kept constant across all countries.

Figure 2: The relative importance of merit and luck for distributional preferences



Note: The figure is based on data from Almas et al. (2021; 2022). It shows how the introduction of performance-dependent inequality in the merit treatment and costly redistribution in the efficiency treatment increases the implemented inequality (delta Gini) relative to the luck treatment for each country. The treatment effects are estimated with a country specific regression controlling for pre-specified background characteristics. The figure shows much larger treatment effects for the merit compared to the efficiency treatment.

Several findings in Almas et al. (2021; 2022) stand out (see Figure 2). First, there are huge cultural differences in the extent to which the merit treatment causes an increase in inequality acceptance relative to the luck treatment (Figure 2a). The merit treatment causes basically no increase in inequality in countries like India and China, presumably because the implemented inequality is already very high in the luck treatment. However, in the overall sample, the Gini coefficient in the merit treatment is 26 percentage points higher than in the luck treatment, and the implemented Gini coefficient increases by 50-60 percentage points in the merit vs. the luck treatment in countries like Canada, Australia, or Portugal.

Second, the effect of the efficiency treatment – when compared to the luck treatment – is much lower than the effect of the merit treatment in almost all countries (Figure 2b and 2c). On average, the Gini coefficient in the efficiency treatment is only 5 percentage points higher than in the luck treatment, and the treatment effect is zero or even negative in many countries. The efficiency treatment increases inequality acceptance at most by (slightly less than) 20 percentage points. Almas et al. (2021, 2022) thus document the near universality of the important role of merit in redistributive preferences, including the near universality of the greater role of fairness considerations relative to efficiency considerations for redistribution.

Third, the study finds vast differences in fairness ideals across countries. While almost 75% of the subjects display a libertarian ideal in countries like India and China, this share is in the range of 10-15% in countries like Canada, Australia, or Norway. Conversely, the share of meritocratic ideals is vanishingly small in India and China, while it is close to 50% in Canada, Australia, and Norway. Interestingly, the cultural/country differences in fairness ideals primarily show up in the different shares of libertarians versus meritocrats, while the share of strict egalitarians is roughly between 20 and 30% in most countries.

3.2.4. Shallow Meritocracy?

The distinction between earnings generated through individuals' choices and earnings that accrue to individuals because of lucky circumstances is at the heart of the meritocratic fairness ideal. A key issue in this concept of fairness is, however, that choices are endogenous. In other words, individuals' choices are themselves affected by circumstances. For example, an individual who faces discrimination will generally have weaker incentives to exert high effort because the reward for effort is lower. There are myriads of external circumstances – such as gender norms, socio-economic background, ethnicity, or race – that

are associated with unequal opportunities that generate differences in individuals' choices and earnings. Moreover, when people make merit judgements based on earnings or effort information, they often do not know the external circumstances under which earnings and effort choices took place. In this context, the question then arises to what extent individuals' merit judgements take these unequal opportunities into account.

Andre (2022), Dong, Huang and Lien (2022), Preuss et al (2022) and Bhattacharya and Mollerstrom (2022) provide persuasive evidence that a majority of individuals neglect the presence of unequal opportunities in their assessment of merit. Even if individuals receive credible information about the fact that disadvantaged workers would perform much better in the absence of unequal opportunities, they largely tend to disregard this information and base their merit assessment on the *actual performance* of agents *in the disadvantaged position*. Taken together, this growing literature shows that the disregard of unequal opportunities in merit judgements is robust and arises in different settings and cultures.

4. Belief-Dependent Social Preferences

In distributional models of social preferences, the players care only about the final *material* payoff consequences for themselves and others, i.e., they are consequentialist models. The players' actions are not affected by how kind or unkind they perceive others' actions or by feelings of guilt and other emotions that may be associated with actions. But players' beliefs about others' (un)kindness, their emotions, and intentions are often relevant for players' behaviors and the utility they derive from material payoffs.[19] In view of the potential importance of such beliefs, this section presents prominent classes of belief-dependent theories of social preferences – reciprocity theories and guilt aversion theory – and discusses the evidence on the ability of these theories to explain facts beyond what purely distributional models can explain. In addition, we also evaluate the shortcomings of the evidence.

[19] A small example makes this point: Two brothers are dividing a pie. One cuts it into two pieces and offers the choice of pieces to his brother. The second brother takes the larger piece, and the first brother complains that he would have taken the smaller piece. The second brother says: "You got what you wanted. What is the problem?" Perceptions of (un)kindness may well play a role here. The second brother may, for example, infer unkind intentions from the fact that the first brother split the cake unequally.

4.1. Reciprocity

4.1.1. Models of Reciprocity

In a seminal paper, Rabin (1993) applied the tools of psychological game theory (Geanakoplos, Pierce and Stachetti 1989) to the study of intentions-based reciprocity. The key idea in this research is that people have beliefs about others' beliefs and that these higher-order beliefs can affect their utility. Subsequent research by Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) has modified and extended Rabin's model, but the basic framework remains similar. The models assume that the desire to raise or lower others' payoffs depends on the perceived fairness or unfairness of their behavior: the perception of kind intentions leads to kind responses, while the perception of unkind intentions cause unkind responses. In Rabin's model, player *i*'s utility includes both his material payoff π_i and a non-pecuniary fairness payoff that is the product of player *i*'s belief about player *j*'s kindness $\tilde{f}_j$ and player *i*'s kindness f_i as follows:

$$U_i \equiv \pi_i + \alpha \cdot \tilde{f}_j \cdot f_i. \qquad (11)$$

The fairness payoff is given by $\tilde{f}_j \cdot f_i$ and $\alpha \geq 0$ measures the weight given to the fairness payoff. $\tilde{f}_j > 0$ means that *i* believes that *j* is kind to her while $\tilde{f}_j < 0$ indicates that *i* believes that *j* is unkind. Likewise, $f_i > 0$ indicates that *i* is kind to *j* and $f_i < 0$ means that *i* is unkind to *j*. From this follows that player *i* can increase his utility by responding to perceived kindness (i.e., $\tilde{f}_j > 0$) with kindness ($f_i > 0$), and to perceived unkindness ($\tilde{f}_j < 0$) with unkindness ($f_i < 0$). The above kindness terms are affected by the players' first and second order beliefs about each others' actions, the players' actual actions, and the associated material payoff consequences.

In Rabin's model, mutual cooperation is an equilibrium in the simultaneous prisoners' dilemma because mutual cooperation is associated with mutual kindness, which triggers reciprocal fairness incentives for cooperation. Likewise, mutual defection is an equilibrium involving mutual unkindness. In addition, Rabin's model can explain rejections in the bilateral ultimatum experiment. However, one limitation of his model is that it only considers simultaneous two-player games.

Dufwenberg and Kirchsteiger (2004) develop a theory of reciprocity for n-person extensive form games. Their approach retains Rabin's idea that reciprocal fairness, and the associated (un)kindness perceptions, provide non-pecuniary incentives for the players. To

adapt this idea to extensive form games, they propose a new solution concept – sequential reciprocity equilibrium and prove an equilibrium existence result. Another advantage of the Dufwenberg and Kirchsteiger model is that it can – in contrast to Rabin's model – explain conditional cooperation of second movers in a *sequentially* played prisoners' dilemma – a behavior that has been widely observed in many experiments. Falk and Fischbacher (2006) combine the notion of inequity aversion with reciprocity concerns, which enables them to explain behavioral phenomena that inequity aversion alone cannot explain (see below).

Finally, Charness and Rabin (2002) also combine distributional preferences with negative reciprocity. In addition, they introduce the notion of "demerit" which basically means that a player deserves less merit and thus receives less weight in the decision-maker's utility function. A player deserves demerit if he misbehaves relative to a group's normative behavioral standard. Recall that individuals in their purely distributional model care with weight $(1-\lambda)$ about their own payoff π_i, with weight $\lambda\delta$ about the payoff of the least well-off player, and with weight $\lambda(1-\delta)$ about the group's total payoff. In the extended version of their model, the least well-off player's payoff as well as that of any other player receives less weight if they misbehave. In addition, a misbehaving player's payoff directly reduces the decision-maker's utility, which causes an incentive to reduce the misbehaving player's income. To capture these motivational forces, Charness and Rabin (2002) introduce three further preference parameters, in addition to λ and δ.[20]

One objection to reciprocity models and, in particular, to that proposed by Charness and Rabin (2002), is that they are considerably more complex than purely distributional preference models. One might thus wonder whether an increase in complexity is needed. However, it seems fair to say that there are many papers that provide data that distributional models cannot explain without considering reciprocity.

Consider, e. g., the following binary ultimatum experiment where the proposer can propose (8 for self, 2 for other) or (5, 5) which is taken from Falk, Fehr and Fischbacher (2003). In this experiment, the rejection rate of the (8, 2) offer is 44%, while in a slightly different experiment, where the equal-payoff alternative (5,5) is replaced by (8,2), the first mover clearly has no choice whatsoever and the rejection rate drops to 18%. Thus, depending on the alternative to the (8, 2) offer, the responder can make different inferences

[20] Other interesting models of reciprocity exist, such as those by Levine (1998) and Cox, Friedman, and Gjerstad (2007). We focus, however, our attention on intention based models of reciprocity because they have led to a considerable empirical literature.

about the proposer's intention. When (5, 5) is the alternative, a proposer's unfairness intention when offering (8, 2) becomes very visible while if (8, 2) is the alternative to itself, the proposer has no meaningful choice and thus no unkind intention can be inferred. No distributional preference model discussed in section 3, including fairness ideals models, can explain the rise in the rejection rate from 18% to 44% when (5,5) is the alternative but models such as those discussed above can. We offer several other examples like this below.

At the same time, however, reciprocity models do have drawbacks. They frequently generate multiple equilibria even in very simple games, and some of these equilibria appear unintuitive. To see this, consider again the binary ultimatum experiment with the (8, 2) and the (5,5) offer. Here, the strategy pair (8, 2) for the proposer and "accept (5, 5) but reject (8, 2)" for the responder can be a fairness equilibrium (involving mutual hostility) in Rabin's model. However, why should a proposer who knows that the responder will reject (8, 2) with certainty ever make that proposal? Likewise, why should a proposer interpret the rejection of the unfair (8, 2) offer as a hostile act rather than as an understandable response to an unfair offer?

One point is important before we discuss the empirical evidence for reciprocity. Kind or unkind intentions are assumed to play a key role in intention based models of reciprocity. One would, therefore, expect empirical researchers to have invested a lot of effort in the identification of subjects' kindness perceptions. Unfortunately, however, with a few exceptions (Offerman 2002; Dhaene and Bouckaert 2010), the empirical literature has often ignored this point. Instead of identifying the subjects' (un)kindness perceptions, the experimenters have typically been the judges of what behavior is kind or unkind. Yet, what really matters is what the participants in the experiment consider to be kind or unkind. Thus, if an experimenter assumes, for example, that a particular behavior is kind but the subjects themselves do not perceive it this way, and hence do not reciprocate, one may erroneously conclude that positive reciprocity (i.e., a positive response to a kind action) is absent when in fact the theory predicts the absence of reciprocal responses.

4.1.2. Laboratory Evidence on Reciprocity

To identify reciprocity preferences, many lab experiments used bilateral two-stage experiments that are played sequentially. In the main treatment, a human first-mover could make a supposedly kind or unkind choice, while a random device exogenously determined

the first-mover's choice in the control treatment so that the human second mover could not infer any intentions.

Charness (1996; 2004) applied this method to a bilateral gift exchange experiment where an "employer" or a random device determines a wage rate w for a "worker" to which the latter can respond with the choice of a costly effort level e that produces a positively valued output for the employer. Charness observes a strong positive relationship between w and e in both treatments. However, the effort level at low wages is lower when the employer chose the wage than when it was generated exogenously, suggesting the presence of negative reciprocity. On the other hand, there was little difference across treatments in the effort levels at high wages. Thus, intentionally receiving a low wage is likely to be seen as unkind because the worker knows that higher wages are possible, and frequently even experiences that higher wages are paid (as the worker faces offers from 10 different employers in sequence). In contrast, it appears less obvious that higher wages are kind because they may be offered for strategic reasons to elicit a higher effort level. This may help to explain why we observe that positive reciprocation is less prevalent than negative reciprocation.

Offerman (2002) also uses the random versus intentional first-mover choice approach of Blount (1995) and of Charness (1996) to study positive and negative reciprocity. He considers players' responses to an unambiguous helpful or hurtful choice. The helpful choice generates a positive payoff for the responder, while the hurtful choice causes a negative payoff. The responders could sacrifice one unit to either increase or decrease the first mover's payoff by four units. 75% of the responders reciprocate intentional helpful choices, while only 50% of the responders reciprocate randomly determined (unintentional) helpful choices. The difference of 25 percentage points is not significant but may well reflects the limited number of observations (12). In contrast, the effect of negative intentionality is quite strong: Responders reciprocate 83.3% of the intentional versus 16.7% of the unintentional hurtful choices. This difference is significant at $p < 0.01$.

To what extent are these large differences in reciprocation patterns a result of subjects' kindness perceptions? Offerman collected proxy measures of responders' (un)kindness experiences by measuring their positive and negative emotions. He finds that intentional hurtful choices generate much stronger negative emotions than unintentional hurtful choices do. In contrast, intentional helpful choices generate about the same positive emotions as unintentional helpful choices do. Moreover, negative (positive) emotions after a hurtful

(helpful) first-mover choice are significantly correlated with punishing (rewarding) responses. Taken together, these results suggest that the differences in reciprocation patterns are driven by differences in the extent to which intentional hurtful and helpful choices trigger (un)kindness perceptions. The results are thus nicely in line with reciprocity theory but also suggest that inducing kindness experiences through intentional choices can be more difficult than inducing unkindness experiences. This differential ease with which one can induce kindness versus unkindness may well have been a factor in the results reported in Charness (1996, 2004).

Several other papers (Brandts and Sola 2001; Brandts and Charness 2003; Falk, Fehr and Fischbacher 2003; Charness and Levine 2007; Falk, Fehr and Fischbacher 2008; Dhaene and Bouckaert 2010) support the conclusion that intentions matter for reciprocal responses. The evidence in these papers, as well as those discussed previously in this section, typically suggests that in addition to distributional fairness concerns, intentions are likely to play a role for deviations from self-interested behavior. While several of these papers do not explicitly measure kindness perceptions, they nevertheless implement plausible manipulations of intentions via random versus intentional first mover choice designs or via other means.

Not all laboratory evidence appears to support the relevance of intention based reciprocity. Bolton, Brandts, and Ockenfels (1998) find strong evidence for distributional preferences, no evidence for positive reciprocity, and only weak evidence for negative reciprocity. We believe, however, that their findings do not show the absence of reciprocity but are rather the result of an experimental design that very likely failed to elicit kindness and unkindness perceptions. Participants made choices in 2x6 matrix games via the strategy method. The payoff matrices used for their different treatments are shown in Appendix 4. In our view, the presentation of payoffs in this way makes it very difficult to induce any kindness or unkindness judgements for at least two reasons. First, the 2x6 game matrix is so complex that even a researcher might appreciate a guiding hand to understand the (un)kindness interpretations built into the matrix. Second, the game is a simultaneous move game and although the authors use the strategy method, the kindness judgements only emerge through a complex reasoning chain. This is rather demanding and presupposes a "theory of mind" capacity that many people are unlikely to have naturally.

4.1.3. Field Evidence on Reciprocity

There is evidence from field settings that suggest negative reciprocity is often present.

Workers have been known to engage in sabotage or increased theft rates after a pay cut or other actions perceived to be unfair (see for example Greenberg (1990), Shminki, Cropanzano, and Rupp (2002)), particularly when procedural justice in the organization is low (Skarlicki and Folger 1997). And the studies by Krueger and Mas (2004) and Mas (2006; 2008) present results suggesting retribution with real firms and workers. The case for positive reciprocity in the field is weaker; cases involving tipping when on the road or higher response rates to mailed surveys that include small gifts may instead reflect guilt aversion, discussed below.

Field evidence from uncontrolled natural environments rarely allows researchers to unambiguously pin down motives because reputational and repeated game issues may play a role. Field experiments are, however, capable of controlling these factors. Gneezy and List (2006) conducted such a field experiment that examined the effect of paying more than the participants were led to expect. In experiments in two different locations, they advertised to the public a job with a wage of $10 (or $12) per hour, which was also the wage actually paid in the control treatment. However, in the main treatment the people who showed up received to their surprise a wage of $20 per hour. The job was six hours of work, split into morning and afternoon sessions. The research question is whether this surprise overpayment leads to greater effort.

In fact, there was significantly higher productivity in the morning but no effect after lunch. Breaking production into four 90-minute segments, the average production in the gift treatment is 27%, 11%, 1% and 2% higher than in the control treatment, indicating that the treatment effect completely disappears after the first 3 hours. The authors argue that this shows that reciprocity is a weak and ephemeral phenomenon and that paying higher wages may not lead to higher net profits. While we do not dispute the behavioral findings, we note that the authors did not measure the workers' kindness perceptions. Thus, it remains unclear whether the higher productivity in the morning reflects intention-based reciprocity. Likewise, it remains unclear whether the absence of higher productivity in the afternoon occurred even though workers saw the $20 wage as kind, i.e., whether reciprocity is indeed a very short-lived response.

In this context, a key question is how the recruited subjects interpreted the surprise wage increase. For example, perhaps subjects inferred from the higher wage that the employer is much wealthier than they thought or that the employer made an advertising error

when the subjects were initially recruited for $12. If so, then the initial effort response to the higher wage may not reflect an increase in fairness perceptions but it may have resulted from short-run effects on subjects' mood. Bellemare and Shearer (2009) conducted a field experiment on the influence of wage gifts on worker productivity that circumvents this problem by providing a natural explanation for the wage increase. In addition, they increased the statistical power for identifying a wage effect with a within-subject design. Bellemare and Shearer found that the wage gift significantly increased workers' productivity.

The importance of fairness perceptions in response to a wage increase was documented in Cohn, Fehr and Goette (2015), who elicited independent measures of fairness perceptions and reciprocity preferences. They show that workers respond to a wage increase with higher effort if two conditions are met: (i) workers perceive the wage increase as an increase in fairness and (ii) workers have reciprocal preferences (that were independently identified in a lab experiment).

At the time Gneezy and List conducted their studies (2004), $10 or $12 was a fair wage for this type of work (prevailing rates were roughly $8.50), i.e., workers were already well-paid at the baseline wage. Perhaps this made it difficult to further increase the kindness perception in the group that received $20 per hour. This interpretation would be consistent with the "fair wage – effort" hypothesis put forward by Akerlof and Yellen (1990). This hypothesis stipulates that workers respond to wage levels with their effort only if they are underpaid relative to a fair reference wage. Amounts paid above this fair wage are not predicted to have positive effects on effort. This theory is also consistent with results from experiments that tested equity theory (Adams 1963). These experiments often found that overpaying subjects did not increase effort (Walster, Walster and Berscheid 1977). Thus, the surprise in the Gneezy and List study may be not that positive higher productivity vanishes over time, but that it was found in the first place.

Kube, Maréchal, and Puppe (2012) consider how the form of payment or reward affects behavior. Their experiment consisted of a 3-hour job with an advertised pay of €12. In the baseline treatment, this was done as advertised. In two other treatments, there was a surprise bonus announced; in one case this was €7, and in the other it was a thermos bottle worth €7. There were also treatments where it was clear that there was more effort put into the gift (for example, the additional money was folded into a complex origami).

Kube, Maréchal, and Puppe (2012) find that people are 25 percentage points more

productive in the bottle treatment (and in other treatments where a non-monetary gift was made) than in the baseline. In contrast, the monetary bonus had only a modest increase of 5% over the baseline. There is thus no significant positive reciprocity in response to a higher wage but a great deal of positive reciprocity when the worker is given a gift. Although this paper did not measure subjects' kindness perceptions, a plausible interpretation of this difference is that giving a gift is much more likely to be perceived as kind. These findings suggest that the "currency of reciprocity" may indeed not be surprise money *that is given without further explanation*.

Esteves-Sorenson (2018) and Della Vigna et al. (2022) also examined the impact of surprise wage increases on workers' effort in field experiments similar to those in Gneezy and List (2006). Estevez-Sorensen carefully controls for a large number of possible confounds and finds little effect of the wage increase on workers' output but – like Gneezy and List – the wage increases implied large overpayments relative to the going market wage and she did not control for workers' fairness perceptions. Della Vigna et al. estimated workers' effort cost functions and reciprocity parameters by implementing effort tasks (across experiments) where output was more or less responsive to effort. These changes enabled them to make cleaner inferences on how gifts affect workers' effort choices. They find that in settings where output is relatively inelastic to effort, such that even sizable shifts in gift-induced prosociality towards the employer would result in small productivity improvements, gifts have little effect on output. In settings where output is more elastic to effort, they find significant effects of the gift on output, but the magnitude of this effect is considerably smaller than what can be achieved with piece rates.

Kube, Maréchal, and Puppe (2013) implemented a wage cut treatment to study the role of negative reciprocity. Since internal review boards might not be happy with a researcher promising a wage and then reneging, they chose their words carefully when they advertised a job. The announcement stated: "The hourly wage is projected to be €15", leaving some room for later wage changes.[21] There were three treatments: No change in pay (€15), pay reduction (€10), and pay increase (€20). In the pay increase condition, productivity is on average very similar to the no change in pay condition, while there is an immediate and sustained decline in productivity by 21% in the pay cut condition, indicating the considerable strength of negative reciprocity.

[21] The exact German wording was "Ihr Stundenlohn beträgt voraussichtlich €15".

The limited overall impact of "wage gifts" on output is further corroborated by the field studies of De Ree et al. (2018) and Jayaraman et. al (2016). De Ree et al. document that a substantial pay raise for teachers in Indonesia neither increased their students' test scores nor their own test scores in teacher subject knowledge tests. Jayaraman et al. show that a large pay reform that led to considerable rise in fixed wages among Indian plantation workers resulted in substantial output increases during the first 3-4 post-reform months after which output converged to the pre-reform baseline.

The discussion above raises the question why there is substantial evidence for gift exchange in laboratory experiments but considerably less evidence from the field. We believe that the different role of inequality aversion in the lab compared to the field is a potentially important reason. Paying a high wage in a typical laboratory experiment implies a strong distributional advantage for the "worker" relative to the "employer". As shown in the random-wage treatment of Charness (1996 and 2004), higher distributional advantages are associated with higher effort levels, suggesting that aversion to advantageous inequality drives the gift exchange, while positive reciprocity plays little role because reciprocation is not higher in the intentional-wage treatment. In contrast, gifts in the form of higher wages in the field studies are typically not associated with a distributional advantage for the worker relative to a clearly defined individual in the role of a principal because the principal is an organization (e.g., a charity, a firm, a ministry, etc.). This means that advantageous inequality aversion is much less likely to play a role in the field compared to the lab.

This interpretation is further corroborated by the findings in Engelmann and Ortmann (2009) and Bellemare, Kröger and van Soest (2011). Engelmann and Ortmann conducted a lab experiment where the equilibrium (with self-interested parties) involves a large distributional advantage for "employer". Thus, unless the wage gift is huge, the gift still implies a distributional advantage for the employer. As a consequence, they observe little extra effort to wage gifts. Kröger et al. show that reciprocity preferences are weaker than inequality aversion in the general population. They simultaneously estimated structural preference parameters measuring inequality aversion and intention based reciprocity in a representative sample of the Dutch population that participated in ultimatum experiments with randomly determined versus intentional offers. They find that "inequity aversion tends to be more important than perceived intentions in the population as a whole". Thus, if the stronger of the two motivational forces (i.e., inequality aversion) is absent in field settings, higher wages may have little effect on effort.

4.1.4. Summary

What did we learn about reciprocity in this section? First, there are widespread and numerous examples of negative reciprocity in the lab (and in the field, although instrumental concern for the future may color the choices made). On the other hand, positive reciprocity is found in some but not all cases. Why this is the case?

A first important reason could be loss aversion, i.e., the notion that losses loom larger than gains. Recall that reciprocity theory defines (un)kindness always relative to a reference point. Kindness is naturally related to receiving more than the reference point, while unkindness means receiving less. The finding of Offerman (2002) that negative emotions after an unkind choice are much stronger than positive emotions are after a kind choice is consistent with this view. A second reason is that certain actions – like surprisingly paying higher wages without providing a plausible explanation for the wage increase – can be interpreted in different ways and may thus not be viewed as kind. This interpretation is in line with the findings in Kube, Marechal and Puppe (2012), which indicate that clearly interpretable gifts (payoff-equivalent to a wage increase) trigger strong reciprocal effort responses, while mere wage increases show insignificant effects. When employers want to induce reciprocal effort responses, they must carefully design their gifts, and not just offer more money. We typically do not give money to our spouses at Christmas but think carefully about the gifts for them.

A third potential reason for the difficulties in inducing positive reciprocity is that a certain minimal level of kindness is taken for granted in many modern societies: people usually hold the door for each other. Typically, if one expects kind or favorable treatment and receives it, there is no strong emotional jolt; on the other hand, if one expects kindness and receives unkind or hurtful treatment, the emotional response is much stronger. This means that the threshold for inducing further kindness perceptions is higher than the threshold for inducing unkindness perceptions.

A fourth potential reason is related to the fact that reciprocity is a cognitively demanding concept because it requires reading other people's intentions. This is much harder than just noticing you received less than another individual. In this context, it is interesting that there appear to be large cultural differences with regard to the extent to which intent is taken into account in moral reasoning. This has been shown by anthropologists who assessed

how people judge the badness and punish-worthiness of bad outcomes that either occur accidentally or intentionally (Barrett et al. 2016; Curtin et al. 2020). While intentions played an important role for Western populations, they played a more minor role for many other cultures. These findings suggest that it may be interesting to study the role of intentions versus outcomes in social preferences across different cultures. In addition, except for the Bellemare, Kröger, van Soest (2011) study, very little appears to be known about the distribution of reciprocity in broad population samples.

4.2. The Role of Guilt Aversion in Social Preferences

Guilt aversion is based on the idea that people prefer to avoid feeling guilty. The basic idea is that decision makers experience guilt if they believe they disappointed others who depended on them. The avoidance of guilt may thus induce prosocial behaviors. However, unlike reciprocity and inequality aversion, guilt aversion cannot explain behaviors that reduce the payoff of others. Guilt aversion leads to a concept of utility in which a player's preferences over strategies depend on her beliefs about the beliefs of others, even if there is no strategic uncertainty.

Guilt aversion has its roots in social psychology: Baumeister, Stillwell, and Heatherton (1994; 1995) advance the notion that people suffer from guilt if they inflict harm on others. One way to inflict harm is to let others down. To illustrate the intuition behind guilt, Dufwenberg (2002) introduces the following marital investment game[22]: A wife can provide support for the husband to pursue a profitable education. If she refuses to do so, each person receives one monetary payoff unit. If she agrees to support the husband, then she forgoes the chance to invest in herself, while the spouses' total payoff is doubled to 4. In case of divorce, her personal earnings are zero, but her husband's earnings are 4. After the wife's investment decision, her husband can choose to stay in the marriage and share these earnings or choose divorce, keeping the rewards. Since the husband promised not to do this, the wife will be quite upset, and the husband might therefore feel guilty.

In this game, guilt aversion naturally enters the scene as follows (Dufwenberg 2002,

[22] Huang and Wu (1994) present the first applied theoretical work in economics to incorporate guilt. Two different forms of guilt aversion have been described. What we describe above is *simple guilt aversion*, where one experiences guilt for disappointing another person's expectations. A second form of guilt aversion is *guilt-from-blame* where one experiences guilt only to the extent that one expects to be blamed for a bad outcome (Battigalli and Dufwenberg 2007; Battigalli and Dufwenberg 2009). Here, we limit our attention to simple guilt.

p. 61): "When a husband suddenly divorces his wife … the stronger the wife's belief that her husband would stay, the more *disappointed* she is. … The husband may be averse to letting a trusting wife down, and the stronger he believes that she believes that he will stay the more *guilty* he feels by forcing divorce." Dufwenberg models guilt-averse preferences here by presuming some individual sensitivity to guilt ($\gamma > 0$) for the husband, and then multiplying this sensitivity by the husband's second order beliefs τ'' concerning what the wife expects. Thus, if the husband stays his utility is 2, while his utility in case of divorce is 4 - $\gamma\tau''$: the stronger the sensitivity to guilt and the higher the second order beliefs, the greater the chance that the husband will stay in the marriage.

4.2.1. Early Experimental evidence

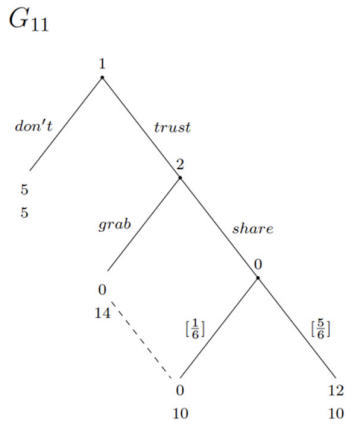
The first experiment to examine psychological game theory is Dufwenberg and Gneezy (2000).[23] They study the lost-wallet game where player 1 (the finder) finds a wallet containing a money amount *x*. At stage 1 of the game, the finder observes the money in the wallet and then decides whether to take the money ("Take") or give the wallet back to the owner ("Leave"). In case of Take, the payoffs are *x* for the finder and 0 for the owner. In case of Leave, the owner can reward the finder with an amount *y* from a given endowment of 20. For example, if the owner gives the finder $y = 5$, then the parties payoffs are (5 for the finder, 15 for the owner). The authors used *x* as a treatment variable set at *x* at 4, 7, 10, 13, 16 in the different treatments. The key issue here is whether the owner's choice of the reward *y*, after the finder has chosen Leave, is positively correlated with her expectation of the finder's expectation of *y*.

In addition to observing the parties' choices, the authors also elicited the finders' incentivized beliefs about *y* (for the known *x*). The owners were asked to make an incentivized guess about the average guess of the finders who chose to leave the money in the envelope. The results show a significant positive correlation between the reward *y* and the owner's expectation of the finder's expectation of *y*. This positive correlation is consistent with the view that guilt aversion drives the owners' reward behavior, but it is also consistent with the existence of a social norm of rewarding finders that affects both the owner's behavior and the parties' beliefs about rewarding.

[23] For a more exhaustive discussion of the evidence on guilt aversion, see Cartwright (2019).

Guilt aversion can potentially also play an important role in economic interactions under incomplete contracting. Charness and Dufwenberg (2006) show this by studying the game below where player 1 can enter a trade that involves the sequential exchange of goods. If 1 does not trust player 2 and so forgoes the trading opportunity, each party receives a payoff of 5. In case player 1 trusts by sending his goods to 2, player 2 can send her goods to player 1 ("share") or keep her goods ("grab"). If player 2 keeps her goods, player 1 receives nothing while player 2 has a payoff of 14. If player 2 sends her goods and the batch arrives, which happens with probability 5/6, the payoffs are (12, 10). However, the sent goods may not arrive with probability 1/6 in which case player 1 receives nothing but player 2 still receives 10. Note that in this game player 1 cannot distinguish whether bad luck or bad behavior by player 2 generated a payoff of zero.

Figure 3: Sequential Exchange with Moral Hazard



In this game a self-interested player 1 who expects to face a self-interested player 2 will never trust. If player 2 is guilt averse, however, then it may be rational for player 1 to trust. From the viewpoint of guilt aversion theory, the key question is whether player 2's belief about player 1's expectation about 2 choosing *share* is positively correlated with player 2's actual probability of choosing *share*. Guilt aversion theory predicts that higher beliefs by player 2 about 1's expectation will increase the probability that player 2 will actually choose *share*.

Charness and Dufwenberg conjectured that giving player 2 the option to send free-form messages to player 1 may be a particularly powerful way of strengthening player 2's second order belief because player 2 has a strong incentive to convince player 1 that 2 will share. Intuitively, if player 2 tries to persuade player 1 that he will choose to share, then

player 2 can hardly avoid the inference that in case of trust player 1 had a high belief that 2 will share. Thus, guilt aversion also provides a mechanism that can explain why communication between parties often enhances cooperative behavior.

The experimental results clearly show the effectiveness of communication. As expected, players 2 often, but not always, promised to share, and these promises were very effective: Player 1 chose *trust* 79% of the time and 2's chose *share* 79% of the time when a promise was made; this compares to 37% and 33%, respectively, when no promise was made. In addition, the authors observe a strong relationship between 2's belief about 1's belief that 2 will choose *share* and 2's actual choice of whether to *share*. These beliefs were significantly higher for 2's who chose *share* compared to 2's who chose *grab (*66.7% versus 42.7%) – which is consistent with guilt aversion theory.

Vanberg (2008) argued, however, that player 2 may keep her promise not because of guilt aversion but because she experiences an obligation to keep her promise. He developed an experimental design that distinguishes between the guilt aversion and the obligation hypothesis. His evidence suggests that feeling obliged to keep promises is indeed a driver of player 2's behavior.

4.2.2. Overcoming Problems in Identifying Guilt Aversion

A key issue in identifying guilt aversion by measuring players' second order beliefs is that these beliefs are endogenous. In the sequential exchange game in Figure 4, player 2's belief about 1's expectation could easily be a result of 2's choice instead of driving it. For example, if 2 believes that player 1 correctly anticipates her choice, then 2's second order belief coincides perfectly with 2's action. In addition, because players' second order beliefs are typically elicited after the players have made their choices, the beliefs may simply reflect an ex-post rationalization of the choices made.

Several approaches have been put forward to solve this identification problem, such as the disclosure approach and approaches that experimentally manipulate second-order beliefs. Applied to a dictator experiment, for example, the disclosure approach elicits the recipient's expectations about the dictator's transfer and informs the dictator about this expectation. Thus, there is no need to elicit the dictator's second order belief because the experimenter knows that belief directly. The disclosure approach has been implemented in several papers and led to mixed evidence for guilt aversion. Some papers found no evidence

for a positive correlation between second order beliefs and the players' prosocial choices (Ellingsen et al. 2010; Bellemare, Sebald and Suetens 2017), while others found a significant correlation (Reuben, Sapienza and Zingales 2009; Bellemare, Sebald and Strobel 2011).

The disclosure approach is not without drawbacks, however. While it circumvents the endogeneity problem associated with the direct elicitation of second order beliefs, it introduces the possibility of other confounds. For example, providing information about a partner's expectation may also signal something about the prevailing social norm, implying that guilt aversion may be confounded with social norm compliance. In addition, if players are informed about the other players' expectations, they may wonder why the experimenter does this, whether the other players know this, and if so, whether other players have biased their expectations strategically.

Is there a way to overcome the difficulties in reliably identifying guilt aversion through the exogenous variation of second-order beliefs? The papers by Ederer and Stremnitzer (2017) and Khalmetski (2016) indeed solve this problem. Due to space constraints, we focus below on the first paper. Ederer and Stremnitzer use a variation of the sequential exchange game displayed in Figure 3. As in Charness and Dufwenberg (2006), player 2 can send free form messages to player 1 before player 1 decides whether to trust.

Ederer and Stremnitzer successfully implement an exogenous variation in beliefs by introducing a chance move that generates either a reliable or an unreliable trading technology in case that player 1 plays *trust*. If the trading technology is reliable, there is a 5/6 probability that player 2 has the option to trade (i.e., the option to play *share*), while in case of an unreliable technology the option to choose *share* only materializes with probability (1/6).[24] Importantly, once the type of trading technology is determined, both players are informed about it, which will obviously affect the players' expectations. In particular, player 1 will expect a higher payoff in case of a reliable technology than in an unreliable trading technology because player 2 has the option of choosing *share* with a high probability in the former case. Moreover, player 2 will know this and will thus have higher second order expectations which, according to guilt aversion theory, should trigger a higher rate of share choices in the presence of a reliable technology.

[24] Ederer and Stremnitzer also allow for different degrees of sharing, but we abstract from this to keep the exposition simple.

Ederer and Stremnitzer (2017) verify that the reliable trading technology indeed induces higher second order beliefs regarding the probability of playing *share*. Moreover, this exogenous increase in second order beliefs holds regardless of whether player 2 made a promise, sent no message, or merely engaged in empty talk. Guilt aversion thus predicts that this exogenous increase in beliefs will be associated with a higher percentage of *share* choices under a reliable technology regardless of the message player 2 sends. Interestingly, however, the percentage of sharing choices is only higher under the reliable technology *in case of a promise*. When player 2 made no promise or simply engaged in empty talk, the exogenous shift in second order expectations was not accompanied with an increase in choosing *share*.

These results may help us understand the conditions under which guilt aversion theory applies. It seems that moving second order expectations is not always enough to move behavior in the direction of guilt aversion, but that additional psychological conditions – such as a promise – need to be present to induce guilt averse behavior. The special status of promises is also supported by evidence in Bracht and Regner (2013) who find a significant effect of promises even after controlling for second-order beliefs.

4.2.3. Summary

The theory of guilt aversion provides a simple and intuitively powerful account of prosocial behavior. There is considerable behavioral evidence that is consistent with guilt aversion and behavioral measures of individuals' guilt aversion have been shown to be correlated with psychological measures of guilt proneness (Bellemare, Sebald and Suetens 2019). However, the rigorous identification of behavioral guilt aversion is non-trivial and involves difficult identification problems. Guilt aversion also does not appear to manifest itself in all environments. The setting must give a sense of fairness, otherwise guilt sensitivity might diminish. Messages must be credible and persuasive to move beliefs, which in turn dictate choices. In addition, once beliefs change, the content of message matters for this change to become behaviorally relevant. We also feel that there is something to the commitment-based story of why promises are so effective. We do not view this idea as being in opposition to guilt aversion but rather as complementary.

5. The Role of Self-Image and Social Image Concerns

One's image can be important for a variety of reasons. Just like looking in a mirror and adjusting physical appearance, people will take actions to appear more favorably. This applies to both self-image and social image, which to some extent seem inextricably intertwined: improving one's self-image or self-esteem may well have positive spillovers socially, and one's social image can lead to a better self-image. This topic was first considered in the field of psychology, where impression management, self-concept, and self-presentation are terms that reflect the care taken to appear more favorably to self and others. The Tesser (1988) self-evaluation maintenance model proposes that people are motivated to maintain both positive self-views as well as their perception about how other individuals view them. Overall, there is considerable evidence indicating that social comparisons affect self-esteem (Gastorf and Suls 1978; Molleman, Pruyn and Van Knippenberg 1986).

Self-perception theory (Bem 1973) has provided an intellectual basis for the modeling of self and social image concerns; it stipulates that people observe their own behavior to infer what they are thinking and how they are feeling. People's behavior may therefore also serve the purpose of self-signaling and social signaling. One way of formalizing self-signaling is to assume that it is an attempt to influence the beliefs of a future self, who cannot remember the original motivation for the behavior. Bodner and Prelec (2003) were the first to develop a dual-self signaling model, an approach that was later also applied in Benabou and Tirole (2006; 2011), Grossman (2015), and Grossman van der Weele (2017).[25]

Consider, for example, a simplified version of Benabou and Tirole's (2006) model. Here, an agent has an intrinsic valuations v_a for contributing a to a charity that generates a costs $C(a)$, so the direct benefit of choosing a is $v_a a - C(a)$. Because the agent cares about her social image, there are also reputational costs and benefits R that depend on what others infer on average about the agent's intrinsic prosociality, $E(v_a|a)$, given the agent's observable prosociality a:

$$R(a) \equiv x[\gamma_a E(v_a|a)], \text{ with } \gamma_a \geq 0.$$

[25] Dufwenberg and Battigalli (2022) make a strong case that image concerns imply that the decision maker's belief about other's beliefs enters the utility function. In other words, proper modelling of image concerns requires the tools of psychological game theory. In the case of self-image concerns, beliefs about other self's beliefs, and in case of social image concerns, beliefs about other people's beliefs enter the utility function. To ease exposition, we use a simplified version of the Benabou-Tirole (2006) model.

The sign of γ_a indicates that people would like to appear prosocial and $x > 0$ is the visibility of their actions. Note that $R(a)$ could represent an instrumental or an affective value of one's social image. Letting $\mu_a \equiv x\gamma_a$, an agent with reputational concerns chooses a to satisfy:

$$\max_{a \in A}[v_a a - C(a) + \mu_a E(v_a|a)] \quad (12)$$

In a nutshell, this model assumes that individuals care for the intrinsic value of prosociality $v_a a$, the material cost involved in behaving prosocially $C(a)$, and their social image $\mu_a E(v_a|a)$. They choose a to optimally balance these components. The model can also be applied to self-image concerns by assuming that there is some chance that one might later fail to remember or (remember in a self-serving manner) the reasons for making a choice. If observable actions are easier to remember than imagined motives, signaling our type to our later self with our current actions could make sense. Perhaps the exact feelings or signals at that time are obscured with some probability proportional to x and the agent later cares about "how prosocial she is".

In the following, we discuss how concerns about one's image – whether self or social – affects *desirable* behaviors. Because *social* desirability is intrinsically tied to the prevailing social norms, a concern for one's social image generates social pressure to conform to these norms. Social image models are therefore essentially models of social norm compliance. In contrast, self-image concerns are based on an individual's *private* identity, the kind of person an individual wants to be which could differ from the prevailing norms. However, in practice, social and private desirability will often coincide.

Note also that social image concerns cannot easily explain the prosocial and the punishing behaviors observed in paradigmatic economic experiments (see Box 2) because social pressure could play little role in the anonymous environments implemented in these settings.[26] However, could it be that self-image concerns can explain some of the prosocial behavior in these experiments?

[26] These experiments implemented anonymous interactions among the subjects. Moreover, the lack of subject-experimenter anonymity has only minor, insignificant effects in dictator, ultimatum, and trust experiments (Bolton, Katok and Zwick 1998; Barmettler, Fehr and Zehnder 2011).

5.1. The Evidence for Self-Image Concerns

Self-image concerns may play a role if people cross the street to avoid passing by a poor beggar. There are other plausible explanations for this behavior, however. Perhaps I give money to charities that I know and trust, so I have no self-image problem regarding my prosociality. Nevertheless, if I pass a beggar in need and give nothing, I frustrate my empathic concern and feel bad. But I don't feel bad because I have a bad self-image, I feel bad because I empathize with the beggar. In the same way, I feel bad when I see a starving child on TV or when I learn about a famine. It may have nothing to do with my self-image. Similarly, social pressure to give to a charity that randomly shows up on one's door or on the street can be aversive *per se* and thus avoiding such situations may have nothing to do with image concerns.[27]

Laboratory experiments provide a more precise tool to identify these concerns. Dana, Weber, and Kuang (henceforth DWK) (2007) conducted a paradigmatic and influential study providing evidence consistent with self-image concerns in altruistic choice. They implemented a binary dictator experiment in a full information treatment and a hidden information treatment. In the *full information treatment*, the allocator can choose between option A, yielding allocation (6,1), and option B with allocation (5,5). In the *hidden information* treatment, the allocator has again two options (A and B) with corresponding payoffs of 6 or 5 for herself but unknown payoffs for the other party. There are two possible states of the world. In one state, the players' preferences over A and B are aligned because choosing A yields payoffs of (6,5) while choosing B leads to (5,1). There is a conflict of interest in the other state because the payoffs are the same as in the full information treatment. In the hidden information treatment, the chooser does not know which state of the world prevails, but she has the option to reveal the true state at no cost by clicking a box on the computer screen.

74% of the dictators chose the fair (5,5) option in the full information treatment, while only 37.5% of the dictators who were randomly assigned to the conflict state chose the fair option in the hidden information treatment. Thus, despite the fact that dictators in the conflict

[27] This argument applies, in our view, to the study of DellaVigna, List and Malmendier (2012) who convincingly show that some people dislike the social pressure associated in door-to-door funding campaigns. In their experiment, flyers posted on their front doors told some households that an individual will come to solicit donations for a charity at a particular time. In the control treatment, no flyers were distributed. They find that flyers significantly reduced the share of households opening the door and they also reduced overall donations.

state of the hidden information treatment faced identical cost and benefits from a fair choice compared to the full information treatment, they were much less likely to behave fairly. A reason for this is that only 50% of the dictators who were assigned to the conflict state actively chose to seek information about the true state. All of those who remained uninformed chose the selfish option, while the informed dictators chose the fair option in 75% of the cases.

To what extent can self-image concerns really explain willful ignorance, i.e., to what extent can we go beyond mere intuitions about the potential role of self-image? A key paper in this regard is Grossman and van der Weele (2017), who developed a multiple-self model where a decision-maker (DM) self-manages her image vis-à-vis an observer-self. One can illustrate the basic intuition of their paper with the help of the notation used when describing the simplified Benabou and Tirole (2006) model above where the decision-maker's utility is given by

$$v_a a - C(a) + \mu_a E(v_a|a,\sigma).$$

The *DM self* knows her preferences and derives altruistic utility $v_a a$ from behaving prosocially (a = 1), and also puts a positive weight μ_a on her self-image $E(v_a|a,\sigma)$. The *observer self* lacks introspective knowledge of the DM's altruism parameter v_a but can infer it from the DM's actions and knowledge σ about the state of the world. The state of the world σ is either unknown, the conflict state, or the aligned state. The observer's inference about the DM's altruism is given by $E(v_a|a,\sigma)$ which is the basis for the DM's self-image.

The key insight derived from this model is based on the existence of a "willful ignorance equilibrium" in which the selfish individuals (v_a = 0) neither behave prosocially (a = 0) in the full information treatment nor in the hidden information treatment. Altruistic individuals, however, can be sorted into two groups: those with a strong preference for altruism ($v_a \geq v^* > 0$) and those with a weak preference for altruism ($v_a < v^*$). The strong altruists (i) choose the prosocial option in the full information treatment and (ii) acquire information and choose the prosocial option in the conflict state of the hidden information treatment.

The third category of players, the weak altruists, are necessary to explain willful ignorance in the DWK paper. In the full information treatment, the DM and the observer know that the conflict state prevails. Choosing option A with payoffs (6,1) thus provides a

clear signal to the observer that the DM's altruism is low, which hurts the DM's self-image. In the hidden information treatment, however, neither the DM nor the observer know the state of the world. Therefore, maintaining the uncertainty and choosing A has a much lower signaling value regarding the DM's altruism because there is a 50% probability that choosing A was not an unfair choice. But the *critical assumption* here is that the *observer does not infer something negative* about the DM's altruism from the mere fact that *the DM decides to remain ignorant*. In this regard, self-image models are also self-deception models because individuals can fool themselves into believing that they are more prosocial than they in fact are. This kind of self-deception has been observed, e. g., in Carlson et al. (2020), who report that a non-negligible share of individuals recalls being more generous in the past than they actually were – an effect that occurred mainly in those individuals who violated their own fairness standards.

Thus, the essence of the model is that the hidden information treatment enables the DM to reduce the negative signaling value of behaving unfairly by obfuscating the observer's inferences about her altruism in case of an unknown state of the world. For weak altruists, the total value of an altruistic act – which consists of intrinsic altruistic utility plus the signaling value – thus declines below the cost of the altruistic act.

The previous considerations indicate that self-image concerns are consistent with the evidence in DWK (2007) but they do *not* show that individuals remain ignorant *because* of self-image concerns. In fact, a recent article by Exley and Kessler (2023) documents that self-image concerns can at most explain a minority of willful ignorance choices. These authors conducted a series of six different studies with roughly 6400 individuals. They introduced a new condition that is identical to the hidden information condition in DWK except that the allocator is now an impartial third party who allocates payoffs between *two other* parties.

The authors compare this other/other condition with the traditional self/other condition where the allocators' self-interest is at stake. Note that in the other/other condition self-image concerns about the allocator's selfishness cannot affect choices because the allocator's self-interest is not at stake. Therefore, if information avoidance in the self/other condition is driven by the desire to reduce signals of selfishness, information avoidance should completely vanish in the other/other condition. However, in contrast to this prediction, depending on the study, only between 14% and 34% of information avoidance vanishes in the other/other condition. In addition, Exley and Kessler show with additional treatments that,

rather than being driven by self-image concerns, information avoidance is largely driven by subjects' aversion to making a trade-off between two participants, their desire to avoid the bad news of being in a nonaligned state, and their inattention, confusion, or laziness.

The limits of self-image driven information avoidance under hidden information is also indicated by Grossman (2014), who examined the robustness of willful ignorance to small, seemingly innocuous changes in the experimental design. He changes the default for learning the true state in the hidden information treatment. In their "Default Non-Revealed" treatment, which is similar to DWK, subjects need to click a box to receive information; they remain uninformed otherwise. In "Active Choice" treatment, participants must click on a box to either learn the state or to not learn the state. In the "Default Revealed" treatment, the default is reversed so that participants learn the state unless they click on a box to say they don't want the information.

The data in Grossman (2014) confirm the basic result in DWK: About 45% of the participants in "Default Non-Revealed" maintained their ignorance and then almost invariably chose allocation (6, 1). Yet matters change dramatically when one must make an active choice of one of the two boxes, with the non-reveal rate dropping significantly to 25%. The most compelling evidence, though, comes from reversing the default. When one is required to affirmatively state that one does not wish to learn the true state (as in "Default Revealed"), only 3% of the participants opt not to learn it. This is a remarkable departure from the rate in "Default Non-Revealed", and it suggests that self-image-driven willful ignorance is not a very robust phenomenon.

Finally, to what extent is the reduction in altruism due to hidden information a robust phenomenon? A recent meta-analysis of a large number of studies similar to those of DWK (Vu et al. 2023), comprising roughly 6500 individuals and 56 treatment effects, shows that the role of hidden information for altruistic choice is considerably smaller than in DWK (2007): In DWK 74% of individuals chose altruistically in the full information treatment but only 37.5% did so in the conflict state of the hidden information treatment. This large decline in altruistic behavior under hidden information is due to the 50% of individuals who remain uninformed *and* behave selfishly. In contrast, in the meta-analysis 57% make the altruistic choice under full information while under hidden information this drops to 41.4%, indicating a much smaller reduction in altruism. It is, however, important to keep in mind that – given

the findings in Exley and Kessler (2023) – only a relatively small part of this smaller decline in altruism can be attributed to self-image concerns.

Carpenter and Robbett (2022) conducted another robustness test examining how moral wiggle room affect the estimates of structural distributional preference parameters. In their experiment, the subjects participated in 45 different binary dictator experiments that included many different costs of altruistic choices. In addition, subjects also faced many situations where they could reduce the other's payoff at a cost to themselves. This set-up is likely to decrease demand effects associated with standard single shot dictator experiments because the large variation across experiments provides many justifications for selfish behaviors (e.g., high costs of altruism in some situations). Carpenter and Robbett show that the existence of moral wiggle room does not lead to a significant change in the estimated distributional preference parameters. This finding is important because it indicates that moral wiggle room may play little role in the typical experiments (Fisman, Kariv and Markovits 2007; Fisman et al. 2015; Fehr®Epper®Senn 2023) used to estimate distributional preference parameters.

The literature on information avoidance in moral wiggle room experiments has undoubtedly increased our knowledge about the intricacies and driving forces underlying prosocial behavior. The overall evidence indicates, however, that the influence of self-image concerns is limited and that phenomena such as willful ignorance, initially attributed to self-image concerns, are not very robust; they can be overcome with small changes in choice architecture such as requiring active choice or changing defaults.

5.2. The Evidence for Social Image Concerns

The formal framework in Benabou and Tirole (2006) can be applied to self-image concerns and to social image concerns. In the latter case, the multiple-self interpretation can be discarded and the term $\mu_a E(v_a|a,\sigma)$ represents the decision makers' intrinsic valuation of being viewed as a prosocial actor by relevant third parties. As we will see below, there is strong evidence for impact of social image concerns on prosocial behaviors, and many authors have provided models of such concerns (Hollander 1990; Benabou and Tirole 2006; Ellingsen and Johannesson 2007; Ellingsen and Johannesson 2008).

Social image concerns are, in a sense, a preference for social approval. Rege and Telle (2004), e.g., conducted a simultaneously played one shot public goods experiment where complete free-riding is the dominant strategy. They implemented a private condition and a public condition where all group members saw how much an individual had contributed. In both conditions, each individual first privately committed to a contribution level by deciding how much of her monetary endowment to put into an envelope dedicated to the public good. Thereafter, each subject had to stand up and put her envelope into a box in front of a room. In the private condition, the envelopes remained sealed; in contrast, each subject had to open her envelope, count the money in the envelope, and write the amount contributed on a blackboard in the public condition. Obviously, the public condition generates strong social image/approval incentives such that the average contribution rate increased from 34% in the private condition to 68 percent in the public condition.

Ariely, Bracha, and Meier (2009) also exploit the distinction between a private setting (where choices remained private), and a public setting (where choices were made public to the other participants in the study) to examine social image effects on prosocial behavior. In their lab study, participants performed a real-effort task (pressing two keys sequentially) that led to donations on their behalf to the American Red Cross. In the public setting, the subjects worked much harder to generate donations to the Red Cross. In addition, the authors tested an interesting prediction that follows from several social signaling models: that monetary incentives crowd out prosocial behaviors driven by social image concerns because public observers will attribute at least a part of subjects' effort to the monetary incentive rather than to their prosociality. Ariely, Bracha and Meier confirm this prediction neatly.

Andreoni and Bernheim (2009) consider social signaling in a $20 dictator experiment with a stochastic element to the decision. Specifically, there is some probability p that Nature transfers a fixed value x_0 to the recipient, i.e., Nature overrides the dictator's decision. To heighten social image concerns, participants were informed that all participants and outcomes would be publicly identified at the end of the session. Note that only the *outcomes* for the two players but *not* the dictators' *choices* are made public. This means that selfish dictators can hide their choices in those dictator experiments in which Nature chooses a low transfer with positive probability, and thus avoid damaging their social image by making the same transfer that Nature would have made.

Andreoni and Bernheim elicit dictators' choices for four values of p (0, 0.25, 0.50, and 0.75) and two values (0 and 1) for Nature's fixed transfer. Their findings provide strong support for the idea that people care about their social image. Between 57% (in case of x_0 =0) and 69% (in case of x_0 = 1) of the dictators chose the equal split when p = 0. However, the frequency of equal splits dramatically declines with increasing p, and more and more dictators hide behind Nature's fixed value and choose exactly x_0. For example, for p = 0.5, and x_0 = 0, 72% of the dictators choose exactly 0 and only 28% the equal split. Likewise, for p = 0.5, and x_0 = 1, roughly 45% of the dictators choose x_0 = 1 and the frequency of equal split is slightly below 40%. Thus, taken together, the evidence from laboratory experiments suggests that social image concerns (being viewed as fair) constitute a strong motive.

In recent years, there has also been increasing evidence from the field suggesting that social image concerns can have powerful effects on important economic behaviors such as labor supply (Bursztyn, González and Yanagizawa-Drott 2020), education investments (Bursztyn and Jensen 2015), conspicuous consumption (Bursztyn et al. 2018), effort in the workplace (Mas and Moretti 2009), or voting (Dellavigna et al. 2017). In the latter paper, the authors left flyers on the doors of homes informing the households that they would return the next day to conduct a survey. Half of the households were randomly informed that the survey would ask them whether they voted in a recent election while the other half did not receive this information. Under the plausible assumption that people who did not vote want to avoid both lying to an interviewer *and* admitting that they did not vote, one would expect that among the informed households the willingness to open the door the next day would be lower compared to the uninformed households. The reason is that by not opening the door individuals can avoid both lying *and* the stigma or shame of being seen as a low civic type. Consistent with this hypothesis, the informed household were indeed 20 percentage points less likely to open the door.

5.3. Summary

The evidence suggests that social image concerns can have strong effects on behaviors that are clearly normatively desirable, while the evidence for self-image driven prosocial behaviors is more limited. In contrast to other-regarding preferences like altruism, inequity aversion, reciprocity, or guilt aversion, self and social image concerns share a more self-regarding flavor. This does not mean, however, that truly other-regarding preferences and

image concerns are mutually exclusive. In fact, it appears quite plausible that individuals simultaneously hold various other-regarding and image-based motives. In terms of quantitative importance, a recent meta-analysis finds that other-regarding motives appear to be considerably more important than self-image concerns (Vu et al. 2023). Moreover, neither self nor social image concerns appear to easily explain the rejection of low offers in ultimatum experiments or the payoff-reducing behaviors observed in allocation tasks with positively sloped budget lines because these behaviors lack a clear prosocial meaning.

Image concerns imply that situational and institutional factors that are completely irrelevant and innocuous in a world without image concerns can become highly relevant. Self-image concerns can be mobilized for prosocial behaviors by removing uncertainties about the prosocial effects of these behaviors, by reducing opportunities that enable moral wiggle room, or by making decisions to remain uninformed public knowledge. All these precautions reduce the likelihood that players can deceive themselves into believing they are more altruistic than they in fact are. Likewise, making decisions transparent and publicly known to relevant others activates social image concerns.

6. Economic and Political Consequences of Social Preferences

Like time and risk preferences, social preferences have broad implications for a wide variety of domains. In this section, we review evidence on how social preferences affect (i) cooperation and (ii) employees' responses to wage inequality and wage cuts, (iii) the extent to which this affects firms' employment decisions, (iv) how they affect the allocation of workers with varying prosociality levels to different industries, (v) how they influence incentives and contract, and (vi) how they affect the political demand for redistribution. These topics do not exhaust the consequences of social preferences, but space constraints force us to limit consideration on the above-mentioned issues.

6.1. The Role of Social Preferences in Cooperation

There is a large theoretical and empirical literature on the role of social preferences in cooperation and collusion, including several review papers (e.g., van Lange et al. (2014); Fehr and Schurtenberger (2018); Balliet, Mulder and van Lange (2011)). Theory suggests that altruistic preferences tend to facilitate cooperation, while the role of disadvantageous inequality aversion or negative reciprocity is more nuanced and depends on the possibility

and the specific features of peer punishment opportunities. If public goods experiments offer these opportunities, disadvantageous inequality aversion and negative reciprocity are motivational forces that facilitate the punishment of free-riders, which helps maintain cooperation levels (Fehr and Schmidt 1999, Fehr and Gächter (2000), Falk and Fischbacher 2006). However, if these opportunities are absent, these forces may have detrimental effects on cooperation and induce players to cease cooperating.

Recent field studies on large-scale cooperation have also established the relevance of peer pressure and peer punishment. Breza, Kaur and Krishnaswamy (2019) studied whether large groups of decentralized workers cooperate to prevent downward pressure on wages. They implemented a field experiment in 183 local labor markets in rural India and show that almost none of the agricultural workers are willing to accept jobs below the prevailing wage when other workers can observe this choice. In addition, they document that this unwillingness to accept low wages is due to workers' willingness to sanction those who accept wage cuts. However, if acceptance of low wages is not observable (and thus not subject to peer sanctioning) the willingness to undercut the going wage increases substantially. Moreover, consistent with the aggregate implications of downward rigidity, Breza et al. also show that measures of social cohesion in local markets correlate with downward wage rigidity and its employment effects across India.

The hugely successful recruitment of soldiers for the British Army at the beginning of World War I when the army relied entirely on the voluntary recruitment of soldiers provides another powerful real-world example of large-scale cooperation supported by peer punishment. Roughly 479000 volunteers were recruited between August 1914 (when Britain declared war on Germany) and September 1914, and approximately 2.5 million men had voluntarily joined the British Army by December 1915. Those who did not join faced the contempt of their community members, who attached big red patches to the free-riders' front doors at night, so that everybody could see that the person living there was a dodger (Simkins (1988)). Recently, Becker (2022) documented how young women publicly shamed young men who refused to join the army. In many towns and cities, the women handed out white feathers to men in civilian clothes, marking them out as cowards. The young women often took substantial risk when doing so because the affected men retaliated. Becker collects evidence from local newspaper articles and exploits the gradual spread of the movement to show that during the 10 days after the first mention of White Feather Girls in the news, volunteering surged by one-third.

Cooperation is not always a good for the overall society. The cooperation between companies for the purpose of maintaining high prices and the cooperation within criminal organizations are examples. Another example is vote-buying, a frequent practice in many countries with weak democratic institutions. In a fascinating study, Finan and Schechter (2012) document how social preferences for reciprocity facilitate vote-buying in municipal elections in Paraguay. Vote-buying constitutes a serious puzzle in a secret voting environment with selfish voters because they cannot ultimately be forced to vote for the candidate who tries to buy their vote. A selfish voter would just take the bribe and merely claim that he or she voted for the bribing politician. However, this commitment problem can be overcome if voters have an intrinsic preference for reciprocity.

In Paraguay, politicians hire respected community leaders in each village to interact with voters and offer them money and other forms of aid for the promise of their vote. Finan and Schechter (2012) show that these community leaders have a very good knowledge of individual voters' preferences for reciprocity and preferentially target reciprocal voters for vote-buying. A one standard deviation increase in reciprocity (measured by the change of trustees' back-transfers in a trust experiment in response to changes in trustors' investments) increases the likelihood of being targeted for vote-buying by 44% - a finding that is robust to a large set of controls including other social preferences and voters' network relationships in the village.

Another domain where social preferences matter is the role of individual heterogeneity in public goods provision. The theory of inequality aversion predicts that groups that are more heterogeneous – in terms of their wealth (endowments) or in terms of the benefits they derive from public goods – are less likely to achieve and maintain successful cooperation, a prediction that a large experimental literature supports (Chan et al. 1996; Chan et al. 1999; Anderson, Mellor and Milyo 2008) and that is consistent with field observations (Mayer 2001; Fajnzylber, Lederman and Loayza 2002).

The influence of prosocial preferences on groups' abilities to maintain high levels of cooperation has been documented in Gächter and Thöni (2005). They measured individuals' willingness to cooperate in a one-shot social dilemma experiment and subsequently formed three types of homogeneous groups: (i) groups comprising individuals with a high prosocial preference, (ii) intermediate groups and (iii) groups comprising selfish individuals. Gächter and Thöni show that aggregate group cooperation in public good experiments with a dominant free-riding strategy is close to maximal in groups of type (i), intermediate levels of

cooperation are achieved in groups of type (ii), and the lowest cooperation levels prevail in groups with predominantly selfish individuals.

These lab findings are nicely echoed in field evidence. Rustagi and Kosfeld (2015) measured prosocial and antisocial tendencies of village leaders in Ethiopia in a third-party punishment experiment. These leaders were responsible for monitoring and sanctioning of free-riders in communities that are strongly reliant on the successful management of forest commons. Rustagi and Kosfeld show that villages with prosocial leaders have significantly better forest outcomes. These results continue to hold after careful consideration of reverse causality issues and omitted variable bias. Similar results of the effects of prosocial preferences on cooperation have been reported in Rustagi, Engel and Kosfeld (2010) and Carpenter and Seki (2011).

6.2. Implications of Social Preferences for Labor Relations and Macroeconomics

6.2.1. Fairness Concerns, Wage Inequality, and Job Satisfaction

If people care for equity and reciprocity, it is likely that wage inequalities that violate their equity standards will have detrimental effects on performance and satisfaction. Many practitioners in Human Resource Management share this viewpoint. Bewley (1999) interviewed several hundred personnel managers about pay-related issues and concluded, for example: “The main function of internal structure is to ensure internal pay equity, which is critical for good morale” (p. 82). However, do *behavioral* data from laboratory experiments and field studies back the views that managers express in surveys and interviews? As we will see below, the behavioral data generally provide a strong endorsement for the role of fairness and equity concerns in the assessment of wage inequalities and indicate the conditions under which pay inequalities have detrimental effects on performance.

The notion of inequity aversion implies that employees who work under identical conditions and provide identical effort should be paid identically. If, instead, they are offered unequal wages, the prediction is that workers who dislike disadvantageous inequality and who receive a lower wage will reduce their effort. Gächter and Thöni (2010) conducted a laboratory experiment in a repeated one-shot gift exchange setting (i.e., an environment with noncontractible effort) with three players, where one experimental employer faces two

workers and makes flat wage offers, w_i and w_j, to each of the two workers. The workers are then informed about w_i and w_j after which they choose their effort levels e_i and e_j which are associated with effort costs of $c(e_i)$ and $c(e_j)$. Gächter and Thöni indeed find that disadvantageous wage discrimination for worker i (i.e., increasing w_j for a given level of w_i) reduces her effort level, while advantageous wage discrimination (i.e., decreasing w_j for a given level of w_i) leaves e_i unaffected.

In subsequent studies (Gächter, Nosenzo and Sefton 2012; Gächter, Nosenzo and Sefton 2013), the authors also showed that social preferences affect workers' effort behavior even in the absence of wage inequality. Their design is similar to that of Gächter and Thöni (2010), but after the workers observed the employer's wage offer, they chose their effort levels sequentially. Thus, the effort level of employee 1, who chose first, could affect the effort level of employee 2, who chose second. It turns out that if both workers received generous wage offers, the effort level of employee 2 is strongly positively correlated with the effort level of employee 1 – a finding that inequality averse preferences predict.

The lab-based papers discussed above implemented a situation where all parties knew wages, effort levels, and the workers' output at different effort levels. In this set-up, the involved parties have clean data that enables them to compare their outcomes. However, what happens if, for example, there are large productivity differences between the workers, but they do not know the exact differences and are only informed that their productivity differs? In this case, social comparison processes are necessarily based on less precise information which – depending on workers' beliefs about the co-worker's productivity – may strengthen or weaken the impact of differential wage payments on effort choices. For example, if workers believe that productivity differences are minor, then wage inequality may have negative effects on effort, while they may consider wage differences to be more justified if workers believe that there are large productivity differences. Charness and Kuhn (2007) implemented a three-player gift exchange experiment, where one employer faces two workers who do not know the effort-output schedules. They report that workers' effort responded positively to their own wages, which indicates the presence of social preferences, but co-workers' wages did not affect workers' effort choices. A plausible interpretation of this finding is that workers believed that there are large productivity differences that justified differential wage payments.

To what extent are the findings above about the negative effort spillovers of higher co-workers' wages generalizable to the field? Cohn et al. (2014) implemented a wage cut in a

field experiment that offered a one-time job opportunity to workers who performed the job in teams of two. The firm placed a job advertisement stipulating an hourly wage of about €10 on an online search platform. The task for both workers was identical and consisted of selling promotional cards that permitted entrance to specific nightclubs. The experiment had two phases that were spread over two subsequent weekends with 6 hours of work per weekend.

In phase one (the first weekend) both workers received the same the same hourly wage of €12 while in phase two (the second weekend) either (i) none of the workers or (ii) only one of the two workers or (iii) both of the workers received a wage cut of €3 relative to the first weekend. Note that even the workers who received a wage cut on the second weekend earned a higher total income for the overall job (6x12 + 6x9 = 126) than that which they initially could have expected with the hourly wage of about €10 initially announced (12x10 = 120).

The authors find that a unilateral wage cut leads to a 34% reduction in the performance of the worker subject to the wage cut relative to the no-wage cut group, while the performance of the worker whose wage is not cut remains unchanged.[28] Thus, similar to the lab experiments discussed above, disadvantageous wage inequality is associated with a large negative effect on effort, while advantageous wage inequality leaves effort unchanged. One might therefore expect that a multilateral wage cut would lead to a smaller effort reduction compared to a unilateral wage cut, which is indeed what Cohn et al. observed. A multilateral wage cut to both workers reduced their performance by "only" 15% relative to the no-wage cut group, and this difference between the unilateral and the multilateral cut is highly significant. These findings are in line with the predictions of a model of inequality aversion.

Breza, Kaur and Shamdasani, henceforth BKR, (2018) implemented a month-long field experiment with Indian manufacturing workers who worked for a daily wage. Their study provides a fascinating and rich collection of facts regarding the relevance of wage inequality for workers' daily labor supply (i.e., showing up for work), their effort during work, and the impact of wage disparities on the subsequent ability to cooperate in other tasks.

BKR assembled production units consisting of three workers who sit together in a separate physical space during work and lunch breaks. The workers in a production unit thus form a natural reference group. They randomized workers into (i) a pay disparity condition

[28] To interpret this finding, it is useful to understand that while the two workers in a "team" worked during the same shift and in the same environment (e.g., at a well frequented subway station) there was no interdependence in their task and their interactions during a shift were minimal. In particular, they had little information about their co-workers' effort during the shift.

where the daily wage reflected workers' baseline productivity that was assessed during an initial training period, and into (ii) a pay compression condition in which all three members of a unit are paid the same wage. By randomizing workers with different baseline productivities to production units, they introduced variation in the extent to which pay differences overstate productivity differences in the disparity condition. And by randomizing production units to different tasks that differ with regard to the observability of co-workers' output, the authors can examine the impact of output observability on effort during work. In addition, the authors also measured workers' attendance on the job (they are only paid when they appear on the job) and their knowledge about co-workers' wages within and across production units. BKR show that workers within a unit are well aware of their co-workers' wages, while little wage information travels across production units.

BKR find that for workers with a similar baseline productivity and a *given* absolute pay level, a worker's output declines by 0.33 standard deviations (22 percent) on average when his two co-workers receive roughly 5 percent higher wages. Moreover, if we hold the level of absolute pay constant, there is no evidence that receiving a roughly 5 percent higher wage than one's peers increases output. In fact, pay inequality in the presence of similar baseline productivities across workers appears to cause a general dissatisfaction with the job situation in the sense that both overpaid and underpaid workers reduce their attendance at the job compared to the compressed pay condition. This means that the workers in the pay disparity group give up valuable earnings – on average by 9.3% – by substantially reducing attendance.

Note that these facts about the effect of over and underpayment on effort/output nicely coincide with the lab findings of Gächter and Thöni (2010) and the field findings of Cohn et al (2014). BKR also show that paying different wages has no negative impact on workers' output in tasks where individuals' productivity differences are easily observable. This suggests that wage inequality has no negative effect when productivity differences justify pay differences – a finding that is consistent with our interpretation of the lab evidence in Charness and Kuhn (2007).

BKR also report a remarkable finding about the detrimental impact of unjustified pay disparities on the subsequent ability of workers to cooperate even when it is in their self-interest to cooperate. On the last day of their job, the workers participated in two cooperative games in each of which they could earn money on the basis of group piece rates for performance. The outcome of these games did not affect the firm's payoff. In the first game, the members of a unit had to build towers of raw materials; the higher the tower, the more

each worker in the unit earned. In pay disparity units with little or no baseline productivity differences, the workers built towers that were 17% smaller on average compared to the compressed pay units. In contrast, when pay differences were justified – based on baseline productivity or task observability – pay disparity units and compressed pay units performed equally well.

Finally, their endline survey reveals that workers from pay disparity units show a significantly lower social cohesion – in terms of their willingness to borrow or lend, or to seek or give advice, or visit one another's homes – compared to workers in the compressed pay group.

More recently, Cullen and Perez-Truglia (2022) conducted a field experiment on the effects of salary comparisons with a sample of 2060 employees from a large corporation in Southeast Asia. They document substantial misperceptions of managers' and co-workers' wages and identify the causal impact of changes in salary perceptions with the help of an information provision experiment. They find that a higher perceived peer salary has a large negative effect on employee's own effort. A 10% increase in employees' perception of peer salaries significantly reduces the number of hours they work by 9.4%, the number of emails they send by 4.3%, and their sales performance by 7.3%. The authors also collected survey evidence that is consistent with the view that social preferences are the mechanism underlying these peer comparison effects, as they also find that higher perceived peer salaries have negative effects on pay and job satisfaction.

These negative effects of peer salaries on job satisfaction are consistent with the findings of Card et al. (2012), who conducted a field experiment with University of California employees by providing them with easy access to information about peer salaries. Employees who had salaries below the peer median subsequently displayed a reduced job satisfaction and an increased intention to switch jobs, while those with salaries above the median showed no changes in job satisfaction and no intention to switch. These findings are consistent with the view that employees have a considerable aversion against disadvantageous inequality.

D'Ambrosio, Clark and Bazzaretta (2018) provide further evidence of the role of fair reference wages on quitting behavior using data from the German Socio-Economic Panel (SOEP). The SOEP contains panel data on which income people consider as fair for their current job. They show that individuals' fair income gap (the difference between what they

earn and what they consider fair for their current job) is not only significantly associated with individuals' life and job satisfaction, but it also influences workers' emotional states such as the frequency of feeling happy or feeling sad and a strong influence on the frequency of experiencing anger. Finally, consistent with these findings on subjective assessments of well-being, the fair income gap also predicts the probability of quitting within the next year.

The evidence provided by Dube, Giuliano and Leonhard (2019) further provides strong support for the view that fairness concerns involving comparisons with (higher) peer wages have a substantial impact on workers' quitting behavior. The authors estimated the own-wage and the peer-wage elasticities of employees' job quitting behavior at a large US retailer with hundreds of stores nationwide. They exploited a regression discontinuity that resulted from the firm's response to the federal minimum wage increases in 1996 and 1997 to identify the causal effect of own and peer wages on quitting behavior. The results show that job separations are extremely sensitive to rising peer wages with peer wage elasticities of 20, 9 and 3 for three, six and nine months after the raise. This result contrasts sharply with the rather low own-wage elasticities they found. Their estimates suggest that, holding the gap between own and peer wage constant, a uniform raise in wages has no impact on quitting behavior. Thus, the overall effect of wages on separations is mostly driven by peer comparisons.

Finally, Dube et al. show that the peer wage effects are asymmetric because they are only driven by comparisons with higher paid peers, which again suggests that aversion against disadvantageous wage inequality is an important driver of labor market behavior. The overall findings from surveys, lab experiments, and field evidence suggest that aversion against disadvantageous wage inequality that cannot be justified by effort or productivity differences generates strong behavioral effects in terms of a reduced willingness to perform, an increased willingness to quit, a lower job satisfaction, a reduced social cohesion and lower willingness to cooperate.

6.2.2. Fairness Concerns and Resistance to Wage Cuts

There is a considerable literature indicating the importance of fairness concerns for the presence of downward wage rigidity. Surveys conducted by Kaufmann (1984), Blinder and Choi (1990), Agell and Lundborg (1995; 2003) and Bewley (1995; 1998; 2002) all point in the direction that workers strongly resist (nominal) wage cuts for fairness reasons even in

recessions, and that personnel managers are keenly aware of this resistance. While these surveys do not pin down the concrete social preferences underlying workers' fairness concerns, a plausible interpretation of the evidence suggests that preferences such as negative reciprocity or inequality aversion (with suitable reference points) provide the motivational raw material for these concerns.

Because fairness concerns induce workers to resist wage cuts and personnel managers anticipate this resistance, the survey evidence suggests that firms will be very reluctant to cut wages, which in turn mitigates labor market adjustments to exogenous shocks. There is indeed substantial laboratory and field evidence of downward wage rigidity and the data often point towards the existence of fairness concerns as the underlying mechanism.

Regarding the field evidence related to whole labor markets, Dickens et al. (2007) document wage rigidity in many countries. Fehr and Goette (2005) show downward nominal wage rigidity for Switzerland, while Grigsby, Hurst & Yildirmaz (2021) document it for the US. In addition, Fehr and Goette (2005) show that downward wage rigidity is negatively related to employment and Kaur (2019) also finds that it is associated with employment distortions. Kaur studies downwards rigidity in the context of village labor markets in India and documents strong rigidity. In addition, she reports that Indian village workers consider nominal wage cuts to be very unfair, suggesting that fairness related resistance to wage cuts is driving rigidity. Interestingly, workers do not consider *real* wage cuts that arise from avoiding nominal pay rises in response to inflation to be unfair – a finding that is consistent with data reported in Kahneman, Knetsch and Thaler (1986).

Firm-level evidence is provided by Greenberg (1990), who reports that workers responded to a temporary wage cut triggered by a negative demand shock, with an increase in employee theft during the period for which pay was cut. Krueger and Mas (2004) document that workers at Bridgestone/Firestone's Decatur, Illinois, plant responded to the firm's attempt to cut wages and hire replacement workers with the provision of lower quality tires. Their monthly data show that defective tires were produced primarily during those months in which the firm demanded wage reductions and incumbents worked side by side with replacement workers.

Coviello, Deserranno and Persico (2022) report evidence on workers' responses to a wage cut in a sales call center in the US. This company paid its sales representatives on the basis of two performance indicators: commissions based on net sales (gross sales minus

refunds due to dissatisfied customers) and conversion rates (percentage of calls resulting in positive gross sales). When the company raised the *required* conversion rates, which was associated with a 13% earnings reduction at a given performance, many sales representatives responded by keeping gross sales constant but increasing customer refunds by intentionally selling suboptimal items to the customers. Note that this behavior not only hurt the company but also the workers themselves, indicating that workers were willing to take costly actions to punish the firm for cutting their wages.

Labor relations are often long-term. Therefore, if workers respond to wage cuts with reduced effort, sabotage, or higher theft rates, one may interpret this as a rational punishment for employers in a repeated game, i.e., the workers' responses may not necessarily result from their social preferences. For this reason, it is useful to study responses to wage cuts in more short-term employment situations in lab and field experiments where there is no prospect for future employment.

The laboratory evidence on downward wage rigidity comes from experimental labor markets (e.g., Fehr, Kirchsteiger and Riedl (1993); Fehr and Falk (1999); Charness (2004); Charness and Brandts (2004); Brown, Falk and Fehr (2004)) that are designed in such a way that, in the absence of social preferences, employers have an incentive to pay competitive wage levels that are rather low. However, if effort is non-contractible and workers respond to the low competitive wage levels with lower effort (due to social preferences such as reciprocity or inequality aversion), even selfish employers have a pecuniary incentive to pay high, non-competitive wages. The evidence from these experiments indeed indicates that employers are reluctant to cut wages to competitive levels because they will then receive low effort levels from their workers. The experiments also show that even in the presence of a large excess supply of workers, the experimental employers shy away from cutting wages to low, competitive levels because of anticipated detrimental effects on workers' performance.

Do the negative effort responses triggered by wage cuts in the lab generalize to field experimental settings that credibly rule out repeated game effects? Several studies indicate that the answer is "yes" (Kube, Marechal and Puppe 2013; Cohn et al. 2014). We already discussed the negative productivity effects of wage cuts in Cohn et. al. in the previous section. Likewise, the evidence in Kube, Marechal and Puppe (2013) indicates large negative productivity effect of 20% from a wage cut relative to a no-wage-cut treatment.

Another interesting study documenting the employment effects of downwards wage rigidity is Breza, Kaur and Shamdasani, BKR, (2021). These authors implemented hiring shocks in local Indian labor markets by giving jobs in *external* jobsites to an average of 24 percent of the labor force of casual male workers for two to four weeks – a shock that substantially reduced how many workers remained in the local economy.

Their approach exploits the strong seasonality in labor demand in these local labor markets. The hiring shock led to immediate and strong raises in wages and a fall in employment in the local markets during the peak season, when demand for casual labor is generally high. However, the hiring shock had basically no impact on employment and wages in the local market in the lean season, when the demand for casual labor is generally low. This is a remarkable finding, since there were apparently enough unemployed workers to fill the gap generated by removing 24 percent of the available labor force, indicating severe rationing of labor supply – a finding that could not have happened in a competitive labor market with flexible wages. If the local labor market during the lean season had been cleared before the hiring shock, then the shock also should have led to large wage increases and a reduction in employment. However, there were apparently enough unemployed workers before the hiring shock who were willing to work at the going wage but who could not find employment at that wage. This made it possible for employers to find enough workers without having to raise wages despite the considerable reduction in the local labor force. Breza, Kaur and Shamdasani also provide evidence indicating that moral hazard or nutrition efficiency wage models cannot explain their data, while a model that relies on workers' resistance to wage cuts can.

Quach (2020) provides evidence for downward wage rigidity by exploiting the following natural experiment from the US. In May 2016, the federal Department of Labor announced that starting December 1, 2016, salaried workers earning less than $913 per week would be entitled to overtime compensation if they work more than 40 hours in a week. In response to this announcement, many employers promised raises to their employees in anticipation of the new rule. However, one week before the rule became effective, a federal court ordered an injunction on the new policy, implying that the employers would not face any legal obstacles if they wanted to refrain from the promised pay raises.

Quach (2020) shows that employers nevertheless increased wages. For the median worker, for example, wages rose by 5.8%, suggesting the employers shied away from cutting nominal wages to pre-announcement levels The pay raises took the form of bunching many

employees at $913 per week and reclassifying some workers from salaried pay to hourly pay. Quach also shows that workers who received pay raises through bunching experienced the same wage growth as workers slightly above the bunching threshold, suggesting that firms did not lower the future wage growth of workers whose wages exhibited rigidity. Moreover, the paper shows that even a year after the proposed overtime policy was nullified, the employers continued to bunch the salaries of new hires at the $913 threshold, indicating that wage nominal wage rigidity also affected the new hires.

The evidence from Quach (2020) suggests that even the mere promise of a pay rise based on a *temporary* legal requirement makes it hard for employers to subsequently lower wages. This finding is also consistent with the laboratory evidence documented in Falk, Fehr and Zehnder (2006), where the temporary implementation of a legal minimum wage led to lasting effects on wages that prevailed even long after the legal minimum wage was removed.

6.2.3. Screening and Selection based on Social Preferences

If social preferences are a relatively stable individual attribute, employers may want to attract workers with particular social preferences and avoid workers with others. Workers with altruistic preferences, e.g., may be valuable for employers because they generate positive spillover effects on other workers in interdependent production processes. Conversely, employers may shy away from workers with envious or spiteful social preferences because they may have detrimental effects on cooperation among employees and between the envious employee and the employer. Likewise, employers might avoid workers who are negatively reciprocal because they may have a strong tendency to engage in counterproductive activities when they are aggrieved.[29]

With regard to self-selection of employees, there is a relatively large literature suggesting that people with more prosocial inclinations tend to self-select themselves to a higher degree into the public sector in countries with a high degree of trust into the public sector (e.g., Dur & Zoutenbier (2014)). Most of these studies are based on self-reported data about motivation or self-reported prosocial actions. However, evidence based on revealed preference data also exists. Buurman et al. (2012) show that early career public sector

[29] Selection and sorting only make sense if individuals' social preferences or their assignment to a particular social preference type exhibits a reasonable degree of stability over time. In Online Appendix 5, we discuss evidence suggesting that this is the case.

workers are more likely to donate to a charity compared to observationally equivalent private sector workers. Gregg et al. (2011) study British Household Panel Data and show that workers who are more prosocial – in terms of providing unpaid overtime work – are more likely to sort into the non-profit sector. They also find that this effect is strongest for industries with "caring characteristics" such as health, education, and social care.

Prosocial individuals may not only prefer working in companies and sectors with caring or helping characteristics, but they may also shy away from sectors or companies involved in immoral business practices such as the intentional sale of toxic financial assets, the marketing of tobacco products to underage smokers, or the aggressive marketing of opioids by the pharmaceutical industry. Schneider, Brun and Weber (2020) used administrative, laboratory, and survey data to study the hypothesis that the least prosocial (i.e., most immoral) people are most likely to work in jobs perceived to involve (or actually involving) immoral activities. Moreover, if working for a company/industry that is perceived to be involved in immoral activities is emotionally aversive, standard economic theory would predict compensating wage differentials. In other words, labor market competition would induce companies/industries perceived to be more immoral to pay, ceteris paribus, higher wages.

To examine the compensating wage differentials hypothesis, they collected survey data from the Swiss population on the perceived morality/immorality of different industries in Switzerland. Then they regressed the gross hourly wages across the industries on the industries' perceived immorality, controlling for observable industry and workers' characteristics. The results indicate a strong positive correlation between the perceived immorality and the gross hourly wages, with industries such as tobacco and weapons manufacturing paying the highest wages and construction and sports facilities being among the lowest paying industries.

Because the correlational evidence from administrative data is, of course, not yet fully convincing, they exogenously varied the characteristics of competitive experimental labor markets. In the immoral work treatment, the subjects were competing for jobs that required them to give wrong advice to another individual that reduced that individual's earnings *and* the charitable donations to UNICEF. In the neutral treatment, the job involved giving advice that increased another individual's earnings and donations to UNICEF.

The striking result of this experiment is that the reservation wages, and thus the competitive equilibrium wages, for the immoral job are much higher than for the neutral job.

Moreover, subjects with weaker prosocial preferences have a much higher frequency of employment in the market involving the immoral task. Thus, the lab experiments provide causal evidence for compensating wage differentials for immoral jobs and for selective sorting of more immoral individuals into these jobs. Finally, the authors also show with the help of survey evidence that subjects who are less prosocial are more willing to work for industries perceived to be more immoral.

Dohmen et al. (2009) provide further evidence on the sorting/selection hypothesis of social preferences. Their results suggest that workers' attitudes towards positive and negative reciprocity can have quite far-reaching effects on their workplace behavior and their earnings. They exploit an interesting survey measure of reciprocity (Perugini et al. 2003) that was included in the German Socio-Economic Panel (SOEP) in 2005, which enabled them to estimate the association between workers' willingness to voluntarily provide effort (in the form of overtime work) and their positive and negative reciprocity.

Controlling for a large number of individual characteristics, Dohmen et al. show that positive reciprocity measured in 2005 is significantly associated with workers' actual overtime work in the years 2005 as well as in the years 2006 and 2007. Moreover, the coefficient on positive reciprocity is almost twice as large for workers who perceived their current wage as fair, while if workers perceive their current wage as unfair, the association between overtime work and positive reciprocity is zero. These results are consistent with the view – derived from theories of inequity aversion and reciprocity – that wages perceived as fair induce reciprocal workers to increase their work effort.

Dohmen et al. also find that negatively reciprocal workers are less willing to perform overtime work. In addition, they show that positively reciprocal workers are less likely to be absent from the workplace, while negatively reciprocal workers tend to be absent more often. Likewise, positively reciprocal workers "consume" fewer days for paid sick leave, while negatively reciprocal workers are on paid sick leave for more days.

Based on these results, one would expect that positively reciprocal workers are more valuable employees, i.e., that the labor market will reward them with higher wages, while negatively reciprocal workers are less valuable. Dohmen et al. estimate Mincer-type wage equations and indeed find that positively reciprocal workers earn higher monthly and annual labor incomes. They do not find a negative impact of negative reciprocity on wages but

instead they show that negatively reciprocal workers have a higher probability of being unemployed, while positive reciprocity reduces the probability of being unemployed.

Barr and Serneels (2009) conducted trust experiments with several hundred employees from 20 manufacturing companies in Ghana; 164 of them were in the role of the second mover, which provides a (noisy) measure of the workers' willingness to display reciprocal behavior. Note that the first mover in this experiment reaps a positive rate of return if she gets back more than what she transferred. Barr and Serneels categorize a worker as highly reciprocal if his back-transfer yields a rate of return of more than 50% for the first mover.[30] The authors show that the output per worker across companies is strongly positively correlated with the share of high reciprocators among employees. This correlation persists when controlling for capital inputs and sector fixed effects. Moreover, a Mincer-type earnings regression that includes a dummy for highly reciprocal workers indicates that these workers earn a wage premium.

Although the papers by Dohmen et al (2009) and Barr and Serneels (2009) do not establish a causal relationship between workers' willingness to reciprocate and their workplace behaviors and earnings, they nevertheless constitute suggestive correlations that deserve further scrutiny. Their findings are consistent with what one would theoretically expect based on knowledge about the behavioral properties of the involved social preferences and they are also consistent with the literature on the impact of early childhood characteristics on later life outcomes. Verdunst et al. (2019) show, for example, that teachers' ratings of kindergarten boys' prosociality are positively associated with the boys' earnings in adulthood after controlling for a large set of covariates.

The findings in Dohmen et al. and Barr and Serneels are also in line with causal laboratory evidence of Bartling, Fehr and Schmidt (2012) on the role of employer's screening in experimental labor markets. In these experiments, employers in some treatments can condition their job offers – in terms of wages, rent-sharing, and employees' opportunities for effort discretion – on information about employees' past performance levels in other firms. The experimental employers make ample use of this information and offer completely different compensation packages to the workers depending on their past performance. Workers with high past effort levels – generally based on their willingness to reciprocate to

[30] The parameters of the trust experiment are such that at a 100% return for the first mover, the payoffs between the two parties are equal, while the second mover reaps all the surplus generated from the first mover's transfer at a zero rate of return.

generous job offers – receive generous current job packages with high wages, a high share of the overall surplus, and broad opportunities for effort discretion. In contrast, workers with low past performance received mediocre job packages with low wages, no rent-sharing, and tightly controlled effort opportunities. These findings suggest that, under the realistic assumption that employers can acquire information about their employees' effort attitudes, positively reciprocally motivated workers are rewarded with better job packages.

6.3. The Role of Social Preferences for Incentives, Contracts, and Institutions

6.3.1. The Effects on Contract Enforcement and Financial Incentives

Over the past 20 years, several authors have illustrated the advantages and disadvantages of different incentive schemes in the light of fairness concerns. For example, it has been shown theoretically (e.g. Sliwka (2007)) and experimentally (e.g. Fehr and Gächter (2002); Fehr and List (2004); Falk and Kosfeld (2006)) that explicit incentive contracts may undermine voluntary cooperation based on social preferences. As a consequence, explicit incentive contracts may be less efficient than implicit alternatives based on trust, informal bonuses, or informal sanctions.

Fehr and Gächter (1997; 1998), for example, tested the impact of trust and reciprocity on contract enforcement in a standard one-shot gift-exchange experiment where principals commit to pay a wage and state a desired effort level in stage 1, and workers respond to the offer with an effort choice in stage 2. In an additional treatment they added a third stage in which the experimental firms can pay to reward or punish the worker for her effort choice. There was a positive wage-effort relation in both treatments, but the average effort level was much higher in the three-stage treatment than in the two-stage treatment. Thus, workers apparently anticipated that firms reward high effort choices and punish low ones, indicating that opportunities to informally reward or sanction workers, which in reality almost always exist, can have powerful incentive effects even in one-shot interactions.

Fehr, Klein and Schmidt (2007) provided evidence suggesting that social preferences may induce principals to prefer informal bonus contracts over more formal contracts with explicit incentives. They consider the following three types of contracts: a trust contract in a two-stage gift exchange environment like the one described above, an incentive contract that introduced an explicit incentive into the trust contract, and a bonus contract that introduced

the option of informally rewarding agents in a third stage. The incentive contract is based on a verification technology that enables the principal to fine the agent in case of verified shirking. The verification technology is imperfect and the fine is limited, implying that the highest effort level that can be implemented is positive but falls short of the efficient effort level. The informal bonus contract contains no explicit incentives but gives the principal the opportunity to reward agents ex-post, i. e., after effort is observed. The bonus contract does not rely on effort verification and enforcement by third parties. Instead, the principal promises a nonbinding, voluntary bonus payment if the agent's effort is satisfactory. This bonus contract is an implicit contract because third parties do not enforce the principal's promise.

If all actors were completely selfish, the incentive contract is the only viable contract, and the trust and bonus contracts would be equally bad. However, the incentive contract in fact dominates the trust contract, but the bonus contract turns out to be much more efficient than the incentive contract. How is it possible that social preferences are not strong enough to render the trust contract more efficient than the incentive contract, but strong enough to make the bonus contract the most efficient one? Fehr, Klein and Schmidt (2007) show that inequality aversion preferences can explain this puzzle.

6.3.2. Social Preferences as a Behavioral Foundation for Employment Contracts

The existence of simple employment contracts that pay a state-independent fixed wage and give employers the right to tell the employee what to do (i.e., to exert authority) is a long-standing puzzle in economics. Why should the trading parties ever agree to such a seemingly inefficient contractual arrangement that may prevent trade (i.e., employment) in certain states of the world? Why do they not continuously renegotiate the contract terms, allowing them to respond to changing conditions and achieve ex-post efficient outcomes, as Alchian and Demsetz (1972) suggest? And if continuous and efficient ex-post renegotiation is always possible, isn't the characterization of the employment contract as an authority relation thoroughly misguided? A negotiation, after all, means that task assignments are subject to *both* parties' agreement.

Hart and Moore (2008) tackle these and related questions by dropping the assumption that "*ex post* trade is perfectly contractible" and that "renegotiation always leads to *ex post*

efficiency" (p. 3).[31] In the absence of perfect ex post contractibility, we are in the world of incomplete contracts with gift exchanges and informal relationships where social preferences typically play a key role. Moreover, social preferences and their interactions with contractual arrangements deeply affect ex post inefficiencies (Fehr, Gachter and Kirchsteiger 1997; Fehr, Klein and Schmidt 2007), and thus also determine which contracts are most efficient. This raises the question whether social preferences could also render employment contracts with rigid wages more efficient compared to contracts that allow for the flexible adjustment of wages to the prevailing state of the world.

Fehr, Hart and Zehnder (2011) indeed show experimentally that rigid contracts can be superior to flexible contracts. In the experiment, there is ex ante uncertainty whether a good state of the world (e.g., high output prices) or a bad state of the world prevails. In addition, the parties' values and costs in the different states of the world are not verifiable so that state-contingent contracts cannot be written. The advantage of a flexible contract, that fixes only a wage *range* but not the wage *level*, is that it allows adjusting the wage w such that trade between an employer and an employe is also possible when output prices are low. In contrast, a contract with wages that are rigidly fixed ex ante (i.e., before the state of the world is known) may prevent trade in this situation. The flexible contract may, however, also be disadvantageous because it provides scope for diverging expectations regarding the wage that will be paid ex post. In other words, while the rigid contract pins down wage expectations ex ante and thus avoids ex post disappointments, workers under flexible wages may feel entitled to higher ex post wages in a good state of the world which provides scope for ex post disappointments. In the presence of (i) non-contractible effort levels and (ii) fairness concerns (social preferences) workers may thus shirk more under flexible contracts than under rigid contracts. Moreover, the lower effort levels under flexible contracts may even render that contract less profitable than the rigid one.

The experimental results confirm the conjectures above. The drawback of the rigid contract is that it prevents trade in the bad state of the world, but this is often over-compensated by the fact that rigid contracts elicit considerably higher effort levels in the good state of the world. To prevent disappointments and low effort levels in the good state of the world, employers pay much higher ex post wages under flexible contracts but still observe a non-negligible amount of shirking. In contrast, the wage is competitively fixed at

[31] Hart and Moore (2008, p. 3) ask the following fundamental question: If the relevant parties can always sit down together ex post (i.e., after the state of the world is revealed) and bargain to an efficient outcome, why should "authority, hierarchy, delegation, or indeed anything apart from asset ownership matter"?

very low levels by market competition under rigid contracts [32], and low effort levels rarely occur despite these low wages. Overall, this renders fixed wage contracts more profitable than flexible contracts. Note that this result could not occur with selfish actors because the flexible contract would always dominate the rigid one.

6.3.3. Social Preferences, Contractual Incompleteness, and Property Rights

Many investments are relationship-specific, meaning that they are valuable only within a particular relationship. In the presence of incomplete contracts, these investments bear the risk of being exploited ex-post – the so-called hold-up problem. Rational parties anticipate being held up ex post, and thus underinvest in relation-specific assets. The property rights literature shows that the appropriate allocation of asset ownership can mitigate the underinvestment incentive. Incomplete contracting and the associated hold-up problem have thus provided an important economic rationale for the allocation of asset ownership to those parties who are the most vulnerable to exploitation (Grossman and Hart 1986; Hart and Moore 1990).

The key assumption behind the property rights approach is that contracts are incomplete, which is justified by assuming that payoff-relevant information is observable to the involved parties but not verifiable by a third-party enforcer. There are many other applications of the "observable but not verifiable assumption" in economics and the assumption has therefore become one of the most important cornerstones of modern institutional economics.[33] However, all these applications of the incomplete contracting approach are subject to a fundamental criticism that Maskin and Tirole (1999) have raised. They show that if parties commonly observe payoff-relevant information, one can construct an extensive form mechanism that leads to truthful revelation of the relevant information in the *unique* subgame perfect equilibrium of the game the mechanism implies.

Thus, one could in principle design contracts that embody this extensive form mechanism, and if the mechanism works as theory predicts, all commonly observable information could be turned into truthfully reported verifiable information. This means that the *second-best* institutional arrangements derived under incomplete contracting would become superfluous because a superior contractual arrangement exists. The question,

[32] Under flexible contracts, competition only determines the lower bound on wages.
[33] The assumption has, for example, been used to understand property rights and firm boundaries, the optimal scope of governments, problems of privatization, the control of insiders by outsiders through voting rights, financial contracts, and patterns of international trade and technology adoption.

however, is whether the extensive form mechanism mentioned above, which is based on the work of Moore and Repullo (1988), indeed works as predicted.

Fehr, Powell and Wilkening (2021) examine this question experimentally, and show that negative reciprocity thoroughly undermines the functioning of Maskin-Tirole-type mechanisms. Most parties are unwilling to enter a contract that incorporates the mechanism, and if they enter them, these contracts typically perform worse than contracts without the mechanism. Intuitively, a key reason for the failure of the mechanisms is that they are based on large fines for the trading parties if they "misbehave", but the threat and execution of large fines is also likely to induce extreme hostility (i.e., negative reciprocity) between the parties. Adding the mechanism to a usual hold-up problem is like handing out guns at a fist fight. The guns are unlikely to make the fight more peaceful.

Thus, social preferences in the form of negative reciprocity undermine the criticism of the theoretical foundations of incomplete contracting models. Ironically, to the extent to which social preferences are a force that contributes to contractual incompleteness, they help sustain their own behavioral importance. Why? Because incomplete contracts provide the terrain – gift exchanges, informal sanctions and rewards, informal agreements – under which social preferences can play an important role.

Overall, social preferences may contribute to the prevalence of incomplete contracting in two ways. First, they may render institutions like the mechanisms discussed above, that render contracts more complete, dysfunctional. Second, they may mitigate the contracting problems that arise under incomplete contracts. One example of this is the relatively high efficiency of incomplete bonus contracts in Fehr, Klein and Schmidt (2007), which we discussed above. Another example is provided by the large experimental literature on behavior under the hold-up problem (see Yang (2021) for a review). This literature shows that the underinvestment problem is typically considerably less severe than the self-interest model predicts (see, e. g., Gantner, Güth and Königstein (2001); Ellingsen and Johannesson (2004a; 2004b); Dufwenberg, Smith and van Essen (2013)). A key reason for this is that fairness concerns induce the parties to take their ex-ante investments in the ex-post bargaining process partially into account. This means that the investing parties experience less exploitation than predicted under self-interest which weakens underinvestment. Negative reciprocity and disadvantageous inequality aversion are also potentially important forces in this context because they provide a preference-based commitment to credibly reject very unfair offers. Parties who can hold-up their counterparts thus often face the threat of complete

disagreement (like in the simple ultimatum experiment), and therefore they shy away from fully exploiting their ex-post bargaining power.

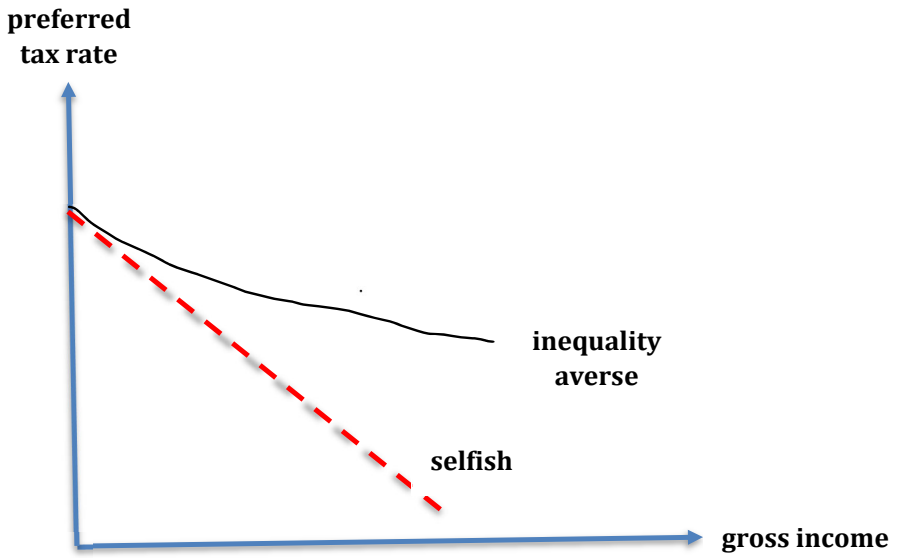
6.4. The Role of Social Preferences in Politics

Inequality aversion, quasi-maximin preferences, and altruism imply that individuals care about payoff distributions, which should affect their willingness to vote for redistributive policy proposals. Suppose, for example, an economic environment like that in the famous model of Meltzer and Richard (1981), but assume that people are inequality averse. In this environment, individuals face the decision to vote on a proportional tax rate τ to be levied on all individuals in the population and redistributed equally as a lump sum. An individual with gross income y_i who receives a lump sum transfer T will thus have a consumption level of $c_i = (1-\tau)y_i + T$. Assuming that there are quadratic resource costs of taxation of $(\frac{1}{2})\tau^2$ per tax dollar, the government budget is balanced if $T = \left(\tau - (\frac{1}{2})\tau^2\right)\bar{y}$, where $\bar{y}$ represents the average gross income in the population. If one assumes that individuals have Fehr-Schmidt preferences, the preferred tax rate τ^* (for an interior solution) is given by

$$\tau_i^* = 1 - \frac{1}{\bar{y}}\left(y_i - \bar{\alpha}_i \frac{1}{n-1}\sum_{j \neq i} \max(y_j - y_i, 0) - \bar{\beta}_i \frac{1}{n-1}\sum_{j \neq i} \max(y_i - y_j, 0)\right) \qquad (13)$$

where $\bar{\alpha}_i$ captures the aversion against disadvantageous inequality ("envy") and $\bar{\beta}_i$ the aversion against advantageous inequality ("empathy"). The above solution for τ_i^* suggests that the preferred tax rates for a selfish and an inequality averse individual may look like those in Figure 4 below.

Figure 4: Preferred tax rate as a function of gross income and social preferences



More precisely, the equation (13) yields the following predictions and implications for empirical research: (i) Low-income individuals obviously have a selfish reason for choosing redistributive taxation, i.e., even in case of $\bar{\alpha}_i = \bar{\beta}_i = 0$ they favor a high tax rate. (ii) Inequality averse individuals ($\bar{\alpha}_i > 0,\ \bar{\beta}_i > 0$) generally demand more redistribution, i.e., have a higher preferred tax rate than selfish individuals ($\bar{\alpha}_i = \bar{\beta}_i = 0$). However, this non-pecuniary driver of redistribution may be difficult to identify empirically at low incomes because selfish motivations already demand a high tax rate. (iii) A higher gross income y_i will generally lower the demand for redistribution, but this effect will be mitigated for inequality averse individuals. In fact, the term that multiplies ($1/\bar{y}$) may become close to zero for very inequality averse individuals, implying that their demand for redistribution does not decline much with gross income. (iv) For individuals with a relatively low income, disadvantageous inequality aversion $\bar{\alpha}_i$ is the main non-pecuniary driver of the demand for redistribution because the social comparisons involve many individuals who earn more than the low-income individual. (v) For individuals with a relatively high income, $\bar{\beta}_i$ is the main non-pecuniary driver of the demand for redistribution because the social comparisons involve many individuals who earn less than the high-income individual.

Several lab studies on political redistribution – such as those of Sausgruber and Tyran (2006) or Durante, Putterman and van der Weele (2014) – indicate that social preferences play a role in voting decisions. Durante et al., e. g., assembled groups of 21 subjects whose pre-tax incomes were calibrated to proportionally reproduce the actual US pre-tax income distribution. Each subject made a choice regarding the preferred proportional tax rate $t \in \{0.1, 0.2, \ldots, 0.9, 1\}$ that generated tax revenue that was equally redistributed as a lump sum to all 21 members of the group. The final group outcome was not determined by voting but by the decision of one randomly chosen subject from the group. This has the advantage that every subject's decision had the same probability of being decisive, i.e., incentive compatibility also held for subjects with extreme preferences. The study shows that most subjects are willing to pay to reduce income inequality among others, but they also take the direct costs of taxation and deadweight losses into account when voting on tax rates. Durante et al. also estimate the parameters of a Charness and Rabin model and find that the weight given to increasing the income of the worst-off player in the group is about three times higher than the weight given to aggregate earnings.

To what extent are individuals' social preferences predictive of their political behaviors outside the laboratory, i.e., their preferences for left versus right wing parties and their

demand for redistribution in the broader society? Kerschbamer and Müller (2020) measured the social preferences of a large representative sample of the German population (the German Internet Panel, GIP, see Table 1) that also contained various questions indicating individuals' views on redistribution spread over several survey waves. These are questions like "Should the government mitigate income differences?" or "Should people, who work more and consequently earn more, pay more or less taxes than they currently do?". Kerschbamer and Müller show that, compared to selfish subjects, inequality averse and altruistic subjects have (i) a higher propensity to vote for left-wing parties, (ii) self-report that they are more left-leaning, and (iii) are more in favor of redistribution as measured by the first principal component of the bundle of redistribution questions in the GIP. These results also hold when the authors control for age, gender, income, education, risk aversion, patience, and political preferences.

Yet there could be many other potential reasons why people might be for or against redistribution such as their expected future income, their history of misfortunes (i.e., unemployment or negative health shocks), their (false) beliefs about the prevailing inequality, their beliefs about their relative incomes, or their beliefs about the role of luck and effort for economic success in life. All of these reasons have been intensely discussed and examined in the political economy literature on redistribution, which raises the question whether distributional preferences are also predictive of people's demand for redistribution if one controls for these motives. In addition, there is the question of the extent to which answers to non-incentivized survey questions such as whether the government should mitigate income differences validly capture the demand for redistribution.

Fehr®Epper®Senn (2021) tackle these problems by measuring social preferences in a broad sample of the Swiss population for whom they also elicit measures of the motives for redistribution mentioned in the previous paragraph that allows them to control for these motives. They asked people for the intensity of their support of several strongly redistributive referenda that were put to vote under the rules of Swiss direct democracy during the last 10-12 years. These survey results were then validated with the actual voting results by comparing the geographic and sociodemographic distribution of votes with the distribution of survey answers. In addition, they validated the survey results with people's actual donations to organizations that support or oppose redistributive proposals.

Fehr®Epper®Senn document that differences in the support for redistribution across the different social preference groups (selfish, inequality averse, and altruistic) is very small

at low incomes but rather large at higher incomes. Both altruistic and inequality averse individuals with above-median incomes display much more support for redistribution than selfish individuals. For example, the support for redistribution among inequality averse individuals with incomes above the median is 0.57 standard deviations higher than the support of selfish individuals. These results follow from the fact that the support for redistribution declines sharply with increasing income for selfish individuals but, as depicted in Figure 4, social preferences strongly mitigate this decline in support for redistribution.

The above results do, however, not yet exhaust the role of social preferences in the demand for redistribution. The reason for this is that both the Kerschbamer & Müller paper and the Fehr®Epper®Senn paper used a distributional measure of social preferences that does not account for people's concern for meritocracy. However, as discussed in section 3.2, people with meritocratic concerns care about other people's incomes and have, therefore, social preferences, while selfish individuals show no concern for meritocracy (see equations (9) and (10) in section 3.2).

Starting with Fong (2001), there is a sizeable literature that shows that people's beliefs about the role of effort and luck in economic success is a key factor in their demand for redistribution (Alesina and La Ferrara 2005; Alesina and Giuliano 2011). This literature also comprises studies that examine how beliefs in intergenerational mobility and equality of opportunity affect the demand for redistribution (e.g., Alesina, Stantcheva and Teso (2018)). Individuals who believe that equality of opportunity already prevails, i.e., that effort (and not luck) is a primary driver for success in life often attribute low income to a lack of effort, i.e., people with low income are considered responsible for their situation and do not deserve help through redistributive legislation. The combination of beliefs about the important role of effort with meritocratic concerns thus reduces the demand for redistribution. Moreover, because selfish individuals have no meritocratic concerns, beliefs about the role of effort should affect other-regarding individuals' demand for redistribution while selfish individuals should remain unresponsive to these beliefs. This is exactly what Fehr®Epper®Senn (2021) found. Thus, taking the meritocratic dimension into account leads to a more nuanced view about the role of social preferences in redistributive politics. The widespread existence of meritocratic other-regarding preferences has been documented not only in the laboratory but also in several survey experiments conducted with general population samples (Almas, Cappelen and Tungodden 2020; Cappelen et al. 2022), and can, in particular, also explain the popularity of workfare programs (Fong, Bowles and Gintis (2005); Drenik & Perez-Truglia

(2018)). Survey experiments have also documented how social preferences and demand for redistribution in general, and meritocratic preferences in particular, are sensitive to the (mis)perceptions a person holds about society, and her own place (e.g., rank) in it (Cruces, Perez-Truglia and Tetaz 2013; Karadja, Mollerstrom and Seim 2017; Fehr, Mollerstrom and Perez-Truglia 2022).

7. Outlook

Over the previous two to three decades a lot has been learned about the properties, the prevalence and the consequences of social preferences but there is also still a lot that needs to be learned. In the following, we outline several open questions that offer exciting research opportunities.

Perhaps, the most fundamental question is related to the determinants of social preferences. There is already initial evidence suggesting that they are formed in childhood through different role models or different early childhood education practices (Van Lange et al. 1997; Cappelen et al. 2020; Kosse et al. 2020) but the set of societal determinants is probably much larger. For example, is it possible for companies and other organizations to shape the social preferences of their employees by structuring rewards, incentives, and the overall company culture in different ways? What is the effect of detrimental health and income shocks on social preference? How does the break-up of marriages or other events that disrupt or improve the relation between family members or members of a community affect social preferences? A recent paper (Cassar et al. 2022) suggests, for example, that allomaternal care increases prosocial preferences in a community, and Rao (2019) shows that having poor classmates makes rich students more prosocial, generous, and egalitarian; and less likely to discriminate against poor students.

We also know very little about how a society's governance institutions shape individuals' social preferences although Rustagi (2024) recently has made some advances in understanding how a history of self-government and democratic interactions tends to favor preferences for cooperation. He exploits a natural experiment in Switzerland, where during the middle-ages, the absence of an heir resulted in the extinction of a prominent noble dynasty, which enabled some Swiss municipalities to become self-governing whereas others remained under feudalism for another 600 years. Rustagi shows that individuals from self-governing communities display stronger preferences for conditional cooperation (measured in

a behavioral experiment) as well as higher voter turnout and higher charitable donations.[34] These findings are also consistent with those of Guiso, Sapienza and Zingales (2016) who show that Northern Italian cities that experienced a period of independence in the Middle Ages have significantly higher prosocial behaviors in terms of organ/blood donations, the frequency of cheating in national exams taken by children in each Italian town and the number of non-profit organizations.

When discussing the potential determinants of social preferences, the relationship between intrinsic social preferences and social norms may also become important. We define a social norm as a commonly known standard of behavior that is based on a widely shared view how individual group members *ought* to behave in a given situation (Fehr and Schurtenberger 2018). Thus, in contrast to preferences, which are a property of individuals, social norms are a property of whole groups of people. They constitute an external normative constraint on individuals' behavior that arises from the fact that the normative standard is widely shared and deviations from the standard are met with disapproval, ridicule, and other forms of sanctioning. However, over time external normative constraints may be internalized which turns them into preferences but very little is known conceptually and empirically about these internalization processes, the factors that shape them, and ways to model them (although see Enke (2019); Schulz et al. (2019); Ellingsen and Mohlin (2022)). In addition, there is very little empirical research that simultaneously elicits and measures social norm driven and social preference driven behaviors (for an exception see Carpenter and Robbett (2022)).

Another important unresolved question concerns the determinants of individuals reference points for their fairness and equity judgements. In the absence of reliable empirical knowledge, models like those of Fehr and Schmidt (1999) have pragmatically assumed that (at least in experiments) equality between the involved parties is a good first-order approximation of individuals actual reference point. As the section on the role of merit, luck and risk in social preferences has made clear, however, there are many situations in which equality may be the wrong reference point for many individuals. It is therefore important to develop methods that enable the reliable empirical identification of individuals' reference

[34] Because Switzerland tracks every family's place of origin in registration data, Rustagi can identify the "cultural origin" of individuals and document the persistence of cultural transmission at the individual level in a context of historically low migration rates.

agents and reference outcomes. An interesting step in this direction has recently been undertaken by Hvidberg, Thustrup-Kreiner and Stantcheva (2023) and Xu et al. (2023).[35]

Finally, it would be desirable to study the deeper implications of heterogenous social preferences for normative (public) economics. If individuals display altruistic or inequality averse distributional preferences, it does not make much sense to compute optimal policies on the basis of social welfare functions that assume that every individual only cares for his or her own consumption. Isn't economics, after all, built on a deep commitment to respect individuals' preferences? Likewise, if people care also for equality of opportunity, it appears of paramount importance to incorporate that notion into modern welfare economics rather than computing optimal policies on the basis of a standard utilitarian welfare function that assumes that individuals only care for their own consumption. In a recent AEA Distinguished Lecture, Emmanuel Saez (2021) echoed this view by stressing the importance of concerns about inequality, poverty and relative position for positive and normative public economics.

Some steps in this direction have been undertaken by Aronsson and Johansson-Stenman (2020a) who have studied optimal income taxation in the presence of externalities and inequality averse individuals, and derived optimal second-best taxation conditions when individuals have social preferences (Aronsson and Johansson-Stenman 2020b). And more recently, Eden and Piacquadio (2023) discussed the normative content of other-regarding preferences. The generalized social marginal welfare weights approach to optimal taxation (Saez and Stantcheva 2016) also opens up ample opportunities to link insights from social preference research with (normative) public economics. This could, for example be done by relating subjects' experimentally measured social preferences to measures of social marginal welfare weights that can be used for deriving optimal tax schedules. However, to answer questions like this, it is important to move beyond measuring social preferences in bilateral settings and consider how these preferences change when there are many recipients. An interesting step in this direction has recently been made by Charite, Fisman and Kuziemko (2021) who show that people exclusively care of the very poor (positively), the very rich (negatively) and their local "income neighbors" directly above them (negatively). Overall, the role of social preference research for normative and positive public economics may be substantial and, perhaps, change what economists recommend to policy makers.

[35] There exists also an older literature in labor economics that discussed reference points such as one's own past wages, peer wages in the company, the company's ability to pay, workers' perceived contributions, etc. as potential reference points (e.g., Levine (1993)). It is, however, probably fair to say that no firm conclusions have been reached by this literature.

References

Adams, J. S. "Inequity in Social-Exchange." *Advances in Experimental Social Psychology* 2, no. 4 (1965): 267-99.

———. "Toward an Understanding of Inequity." *Journal of Abnormal Psychology* 67, no. 5 (1963): 422-&.

Agell, J., and P. Lundborg. "Survey Evidence on Wage Rigidity and Unemployment: Sweden in the 1990s." *Scandinavian Journal of Economics* 105, no. 1 (2003): 15-29.

———. "Theories of Pay and Unemployment - Survey Evidence from Swedish Manufacturing Firms." *Scandinavian Journal of Economics* 97, no. 2 (1995): 295-307.

Akerlof, G. A., and Y. L. Yellen. "The Fair Wage-Effort Hypothesis and Unemployment." *Quarterly Journal of Economics* 105, no. 2 (1990): 255-83.

Alchian, A. A., and H. Demsetz. "Production, Information Costs, and Economic Organization." *American Economic Review* 62, no. 5 (1972): 777-95.

Alesina, A., and P. Giuliano. "Preferences for Redistribution." In *Handbook of Social Economics*, edited by J. Benhabib, A. Bisin and M.O. Kackson, 93-131: Elsevier, 2011.

Alesina, A., and E. La Ferrara. "Preferences for Redistribution in the Land of Opportunities." *Journal of Public Economics* 89, no. 5-6 (2005): 897-931.

Alesina, A., S. Stantcheva, and E. Teso. "Intergenerational Mobility and Preferences for Redistribution." *American Economic Review* 108, no. 2 (2018): 521-54.

Almas, I., A. W. Cappelen, E. O. Sorensen, and B. Tungodden. "Attitudes to Inequality: Preferences and Beliefs." In *IFS Deaton Review of Inequalities*: Institute for Fiscal Studies, 2022.

———. "Fairness across the World: Preferences and Beliefs." Federal Reserve Bank of New York, 2021.

Almas, I., A. W. Cappelen, and B. Tungodden. "Cutthroat Capitalism Versus Cuddly Socialism: Are Americans More Meritocratic and Efficiency-Seeking Than Scandinavians?" *Journal of Political Economy* 128, no. 5 (2020): 1753-88.

Anderson, L.R., J.M. Mellor, and J. Milyo. "Inequality and Public Good Provision: An Experimental Analysis." *The Journal of Socio-Economics* 37, no. 3 (2008): 1010-28.

Andre, P. "Shallow Meritocracy." University of Bonn, 2022.

Andreoni, J. "Why Free Ride - Strategies and Learning in Public-Goods Experiments." *Journal of Public Economics* 37, no. 3 (1988): 291-304.

Andreoni, J., and B. D. Bernheim. "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects." *Econometrica* 77, no. 5 (2009): 1607-36.

Andreoni, J., and J. Miller. "Giving According to Garp: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica* 70, no. 2 (2002): 737-53.

Ariely, D., A. Bracha, and S. Meier. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review* 99, no. 1 (2009): 544-55.

Aronsson, T., and O. Johansson-Stenman. "Inequality Aversion, Externalities, and Pareto-Efficient Income Taxation." University of Gothenburg, 2020a.

———. "Optimal Second-Best Taxation When Individuals Have Social Preferences." Umea School of Business: Department of Economics, 2020b.

Balafoutas, L., R. Kerschbamer, M. Kocher, and M. Sutter. "Revealed Distributional Preferences: Individuals Vs. Team." *Journal of Economic Behavior & Organization* 108 (2014): 319-30.

Balafoutas, L., R. Kerschbamer, and M. Sutter. "Distributional Preferences and Competitive Behavior." *Journal of Economic Behavior & Organization* 83, no. 1 (2012): 125-35.

Balliet, D., L. B. Mulder, and P. A. M. Van Lange. "Reward, Punishment, and Cooperation: A Meta-Analysis." *Psychological Bulletin* 137, no. 4 (2011): 594-615.
Barr, A., and P. Serneels. "Reciprocity in the Workplace." *Experimental Economics* 12 (2009): 99-112.
Barrett, H. C., A. Bolyanatz, A. N. Crittenden, D. M. T. Fessler, S. Fitzpatrick, M. Gurven, J. Henrich, M. Kanovsky, G. Kushnick, A. Pisor, B. A. Scelza, S. Stich, C. von Rueden, W. Y. Zhao, and S. Laurence. "Small-Scale Societies Exhibit Fundamental Variation in the Role of Intentions in Moral Judgment." *Proceedings of the National Academy of Sciences of the United States of America* 113, no. 17 (2016): 4688-93.
Bartling, B., E. Fehr, and K. M. Schmidt. "Screening, Competition, and Job Design: Economic Origins of Good Jobs." *American Economic Review* 102, no. 2 (2012): 834-64.
Battigalli, P., and M. Dufwenberg. "Belief-Dependent Motivations and Psychological Game Theory." *Journal of Economic Literature* 60, no. 3 (2022): 833-82.
———. "Dynamic Psychological Games." *Journal of Economic Theory* 144, no. 1 (2009): 1-35.
———. "Guilt in Games." *American Economic Review* 97, no. 2 (2007): 170-76.
Baumeister, R. F., A. M. Stillwell, and T. F. Heatherton. "Guilt - an Interpersonal Approach." *Psychological Bulletin* 115, no. 2 (1994): 243-67.
———. "Personal Narratives About Guilt - Role in Action Control and Interpersonal Relationships." *Basic and Applied Social Psychology* 17, no. 1-2 (1995): 173-98.
Becker, A. "Shamed to Death: Social Image Concerns and War Participation." University College London, 2022.
Bellemare, C., S. Kroeger, and A. Van Soest. "Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities." *Econometrica* 76, no. 4 (2008): 815-39.
Bellemare, C., S. Kroger, and A. van Soest. "Preferences, Intentions, and Expectation Violations: A Large-Scale Experiment with a Representative Subject Pool." *Journal of Economic Behavior & Organization* 78, no. 3 (2011): 349-65.
Bellemare, C., A. Sebald, and M. Strobel. "Measuring the Willingness to Pay to Avoid Guilt: Estimation Using Equilibrium and Stated Belief Models." *Journal of Applied Econometrics* 26, no. 3 (2011): 437-53.
Bellemare, C., A. Sebald, and S. Suetens. "Guilt Aversion in Economics and Psychology." *Journal of Economic Psychology* 73 (2019): 52-59.
———. "A Note on Testing Guilt Aversion." *Games and Economic Behavior* 102 (2017): 233-39.
Bellemare, C., and B. Shearer. "Gift Giving and Worker Productivity: Evidence from a Firm-Level Experiment." *Games and Economic Behavior* 67, no. 1 (2009): 233-44.
Bem, D. "Self-Perception Theory." *Advances in Experimental Social Psychology* 6, no. 1 (1973): 1-62.
Benabou, R., and J. Tirole. "Identity, Morals, and Taboos: Beliefs as Assets." *Quarterly Journal of Economics* 126, no. 2 (2011): 805-55.
———. "Incentives and Prosocial Behavior." *American Economic Review* 96, no. 5 (2006): 1652-78.
Berg, J, J Dickhaut, and K McCabe. "Trust, Reciprocity and Social History." *Games and Economic Behavior* 10, no. 1 (1995): 122-42.
Bewley, T. F. "A Depressed Labor-Market as Explained by Participants." *American Economic Review* 85, no. 2 (1995): 250-54.
———. "Fairness, Reciprocity, and Wage Rigidity." SSRN, 2002.
———. "Why Not Cut Pay?" *European Economic Review* 42, no. 3-5 (1998): 459-90.
Bewley, T.F. . *Why Wages Don't Fall During a Recession*. Cambridge MA: Harvard University Press, 1999.

Bhattacharya, P., and J. Mollerstrom. "Lucky to Work." *Unpublished Working Paper* (2022).
Blinder, A. S., and D. H. Choi. "A Shred of Evidence on Theories of Wage Stickiness." *Quarterly Journal of Economics* 105, no. 4 (1990): 1003-15.
Blount, S. "When Social Outcomes Arent Fair - the Effect of Causal Attributions on Preferences." *Organizational Behavior and Human Decision Processes* 63, no. 2 (1995): 131-44.
Bolton, G. E., and A. Ockenfels. "Erc: A Theory of Equity, Reciprocity, and Competition." *American Economic Review* 90, no. 1 (2000): 166-93.
Bolton, G.E, J. Brandts, and A. Ockenfels. "Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game." *Experimental Economics* 1 (1998): 207-19.
Bracht, J., and T. Regner. "Moral Emotions and Partnership." *Journal of Economic Psychology* 39, no. C (2013): 313-26.
Brandts, J., and G. Charness. "Do Labour Market Conditions Affect Gift Exchange? Some Experimental Evidence." *Economic Journal* 114, no. 497 (2004): 684-708.
———. "Truth or Consequences: An Experiment." *Management Science* 49, no. 1 (2003): 116-30.
Brandts, J., and C. Sola. "Reference Points and Negative Reciprocity in Simple Sequential Games." *Games and Economic Behavior* 36, no. 2 (2001): 138-57.
Breza, E., S. Kaur, and N. Krishnaswamy. "Scabs: The Social Suppression of Labor Supply." *NBER Working Paper 25880* (2019).
Breza, E., S. Kaur, and Y. Shamdasani. "Labor Rationing." *American Economic Review* 111, no. 10 (2021): 3184-224.
———. "The Morale Effects of Pay Inequality." *Quarterly Journal of Economics* 133, no. 2 (2018): 611-63.
Brown, M., A. Falk, and E. Fehr. "Relational Contracts and the Nature of Market Interactions." *Econometrica* 72, no. 3 (2004): 747-80.
Bursztyn, L., B. Ferman, S. Fiorin, M. Kanz, and G. Rao. "Status Goods: Experimental Evidence from Platinum Credit Cards." *Quarterly Journal of Economics* 133, no. 3 (2018): 1561-95.
Bursztyn, L., A. L. González, and D. Yanagizawa-Drott. "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia." *American Economic Review* 110, no. 10 (2020): 2997-3029.
Bursztyn, L., and R. Jensen. "How Does Peer Pressure Affect Educational Investments?" *Quarterly Journal of Economics* 130, no. 3 (2015): 1329-67.
Buurman, M., J. Delfgaauw, R. Dur, and S. Van den Bossche. "Public Sector Employees: Risk Averse and Altruistic?" *Journal of Economic Behavior & Organization* 83, no. 3 (2012): 279-91.
Cabeza, B. "Desert Concerns and Distributional Preferences." 2021.
Camerer, Colin F. *Behavioral Game Theory - Experiments in Strategic Interaction*. Princeton, New Jersey: Princeton University Press, 2003.
Cameron, L. A. "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia." *Economic Inquiry* 37, no. 1 (1999): 47-59.
Cappelen, A., J. List, A. Samek, and B. Tungodden. "The Effect of Early-Childhood Education on Social Preferences." *Journal of Political Economy* 128, no. 7 (2020): 2739-58.
Cappelen, A. W., A. D. Hole, E. O. Sorensen, and B. Tungodden. "The Pluralism of Fairness Ideals: An Experimental Approach." *American Economic Review* 97, no. 3 (2007): 818-27.
Cappelen, A. W., K. O. Moene, E. O. Sorensen, and B. Tungodden. "Needs Versus Entitlementsan International Fairness Experiment." *Journal of the European Economic Association* 11, no. 3 (2013): 574-98.

Cappelen, A. W., J. Mollerstrom, B. A. Reme, and B. Tungodden. "A Meritocratic Origin of Egalitarian Behaviour." *Economic Journal* 132, no. 646 (2022): 2101-17.
Cappelen, A. W., K. Nygaard, E. O. Sorensen, and B. Tungodden. "Social Preferences in the Lab: A Comparison of Students and a Representative Population." *Scandinavian Journal of Economics* 117, no. 4 (2015): 1306-26.
Card, D., A. Mas, E. Moretti, and E. Saez. "Inequality at Work: The Effect of Peer Salaries on Job Satisfaction." *American Economic Review* 102, no. 6 (2012): 2981-3003.
Carlson, R. W., M. A. Marechal, B. Oud, E. Fehr, and M. J. Crockett. "Motivated Misremembering of Selfish Decisions." *Nature Communications* 11, no. 1 (2020).
Carpenter, J., and A Robbett. "Measuring Socially Appropriate Social Preferences." *IZA Working Paper No. 15590* (2022).
Carpenter, J., and E. Seki. "Do Social Preferences Increase Productivity? Field Experimental Evidence from Fishermen in Toyama Bay." *Economic Inquiry* 49, no. 2 (2011): 612-30.
Cartwright, E. "A Survey of Belief-Based Guilt Aversion in Trust and Dictator Games." *Journal of Economic Behavior & Organization* 167 (2019): 430-44.
Cassar, A., A. Cristia, P. Grosjean, and S. Walker. "It Makes a Village: Allomaternal Care and Prosociality." *Working Paper, Department of Economics, UNSW Sidney* (2022).
Chan, K. S., S. Mestelman, R. Moir, and R. A. Muller. "Heterogeneity and the Voluntary Provision of Public Goods." *Experimental Economics* 2, no. 1 (1999): 5-30.
———. "The Voluntary Provision of Public Goods under Varying Income Distributions." *Canadian Journal of Economics-Revue Canadienne D Economique* 29, no. 1 (1996): 54-69.
Chapman, J., M. Dean, Ortoleva P., E. Snowberg, and C. F. Camerer. "Econographics." *NBER Working Paper 24931* (2018).
Charness, G. "Attribution and Reciprocity in a Simulated Labor Market: An Experimental Investigation." Barcelona: Universitat Pompeu Fabra, 1996.
———. "Attribution and Reciprocity in an Experimental Labor Market." *Journal of Labor Economics* 22, no. 3 (2004): 665-88.
Charness, G., and M. Dufwenberg. "Promises and Partnership." *Econometrica* 74, no. 6 (2006): 1579-601.
Charness, G., and B. Grosskopf. "Relative Payoffs and Happiness: An Experimental Study." *Journal of Economic Behavior & Organization* 45, no. 3 (2001): 301-28.
Charness, G., and P. Kuhn. "Does Pay Inequality Affect Worker Effort? Experimental Evidence." *Journal of Labor Economics* 25, no. 4 (2007): 693-723.
Charness, G., and D. I. Levine. "Intention and Stochastic Outcomes: An Experimental Study." *Economic Journal* 117, no. 522 (2007): 1051-72.
Charness, G., and M. Rabin. "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics* 117, no. 3 (2002): 817-69.
Cherry, T. L., P. Frykblom, and J. F. Shogren. "Hardnose the Dictator." *American Economic Review* 92, no. 4 (2002): 1218-21.
Cohn, A., E. Fehr, and L. Goette. "Fair Wages and Effort Provision: Combining Evidence from a Choice Experiment and a Field Experiment." *Management Science* 61, no. 8 (2015): 1777-94.
Cohn, A., E. Fehr, B. Herrmann, and F. Schneider. "Social Comparison and Effort Provision: Evidence from a Field Experiment." *Journal of the European Economic Association* 12, no. 4 (2014): 877-98.
Coviello, D., E. Deserranno, and N. Persico. "Counterproductive Worker Behavior after a Pay Cut." *Journal of the European Economic Association* 20, no. 1 (2022): 222-63.
Cox, J. C., D. Friedman, and S. Gjerstad. "A Tractable Model of Reciprocity and Fairness." *Games and Economic Behavior* 59, no. 1 (2007): 17-45.

Cruces, G., R. Perez-Truglia, and M. Tetaz. "Biased Perceptions of Income Distribution and Preferences for Redistribution: Evidence from a Survey Experiment." *Journal of Public Economics* 98 (2013): 100-12.
Cullen, Z., and R. Perez-Truglia. "How Much Does Your Boss Make? The Effects of Salary Comparisons." *Journal of Political Economy* (2022).
Curtin, C. M., H. C. Barrett, A. Bolyanatz, A. N. Crittenden, D. M. T. Fessler, S. Fitzpatrick, M. Gurven, M. Kanovsky, G. Kushnick, S. Laurence, A. Pisor, B. Scelza, S. Stich, C. von Rueden, and J. Henrich. "Kinship Intensity and the Use of Mental States in Moral Judgment across Societies." *Evolution and Human Behavior* 41, no. 5 (2020): 415-29.
D'Ambrosio, C., A. E. Clark, and M. Barazzetta. "Unfairness at Work: Well-Being and Quits." *Labour Economics* 51 (2018): 307-16.
Dana, J., R. A. Weber, and J. X. Kuang. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory* 33, no. 1 (2007): 67-80.
Dawes, R. M. "Social Dilemmas." *Annual Review of Psychology* 31 (1980): 169-93.
Dawes, R. M., J. Mctavish, and H. Shaklee. "Behavior, Communication, and Assumptions About Other Peoples Behavior in a Commons Dilemma Situation." *Journal of Personality and Social Psychology* 35, no. 1 (1977): 1-11.
de Ree, J., K. Muralidharan, M. Pradhan, and H. Rogers. "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia." *Quarterly Journal of Economics* 133, no. 2 (2018): 993-1039.
DellaVigna, S., J. A. List, and U. Malmendier. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics* 127, no. 1 (2012): 1-56.
DellaVigna, S., J. A. List, U. Malmendier, and G. Rao. "Estimating Social Preferences and Gift Exchange at Work." *American Economic Review* 112, no. 3 (2022): 1038-74.
———. "Voting to Tell Others." *Review of Economic Studies* 84, no. 1 (2017): 143-81.
Dhaene, G., and J. Bouckaert. "Sequential Reciprocity in Two-Player, Two-Stage Games: An Experimental Analysis." *Games and Economic Behavior* 70, no. 2 (2010): 289-303.
Dickens, W. T., L. Goette, E. L. Groshen, S. Holden, J. Messina, M. E. Schweitzer, J. Turunen, and M. E. Ward. "How Wages Change: Micro Evidence from the International Wage Flexibility Project." *Journal of Economic Perspectives* 21, no. 2 (2007): 195-214.
Dohmen, T., A. Falk, D. Huffman, and U. Sunde. "Homo Reciprocans: Survey Evidence on Behavioural Outcomes." *Economic Journal* 119, no. 536 (2009): 592-612.
Dong, L., L. Huang, and J. W. Lien. "They Never Had a Chance: Unequal Opportunities and Fair Redistribution." *Unpublished Working Paper* (2022).
Drenik, A., and R. Perez-Truglia. "Sympathy for the Diligent and the Demand for Workfare." *Journal of Economic Behavior & Organization* 153 (2018): 77-102.
Dube, A., L. Giuliano, and J. Leonard. "Fairness and Frictions: The Impact of Unequal Raises on Quit Behavior." *American Economic Review* 109, no. 2 (2019): 620-63.
Dufwenberg, M. "Marital Investments, Time Consistency and Emotions." *Journal of Economic Behavior & Organization* 48, no. 1 (2002): 57-69.
Dufwenberg, M., and U. Gneezy. "Measuring Beliefs in an Experimental Lost Wallet Game." *Games and Economic Behavior* 30, no. 2 (2000): 163-82.
Dufwenberg, M., P. Heidhues, G. Kirchsteiger, F. Riedel, and J. Sobel. "Other-Regarding Preferences in General Equilibrium." *Review of Economic Studies* 78, no. 2 (2011): 613-39.
Dufwenberg, M., and G. Kirchsteiger. "A Theory of Sequential Reciprocity." *Games and Economic Behavior* 47, no. 2 (2004): 268-98.
Dufwenberg, M., A. Smith, and M. Van Essen. "Hold-Up: With a Vengeance." *Economic Inquiry* 51, no. 1 (2013): 896-908.

Dur, R., and R. Zoutenbier. "Working for a Good Cause." *Public Administration Review* 74, no. 2 (2014): 144-55.
Durante, R., L. Putterman, and J. van der Weele. "Preferences for Redistribution and Perception of Fairness: An Experimental Study." *Journal of the European Economic Association* 12, no. 4 (2014): 1059-86.
Dwyer, R. J., W. J. Brady, C. Anderson, and E. W. Dunn. "Are People Generous When the Financial Stakes Are High?" *Psychological Science* 34, no. 9 (2023): 999-1006.
Eden, M., and P. G. Piacquadio. "The Normative Content of Other-Regarding Preferences." *Working Paper, Department of Economics, Brandeis University* (2023).
Ederer, F., and A. Stremitzer. "Promises and Expectations." *Games and Economic Behavior* 106 (2017): 161-78.
Ellingsen, T., and M. Johannesson. "Is There a Hold-up Problem?" *Scandinavian Journal of Economics* 106, no. 3 (2004a): 475-94.
———. "Paying Respect." *Journal of Economic Perspectives* 21, no. 4 (2007): 135-49.
———. "Pride and Prejudice: The Human Side of Incentive Theory." *American Economic Review* 98, no. 3 (2008): 990-1008.
———. "Promises, Threats and Fairness." *Economic Journal* 114, no. 495 (2004b): 397-420.
Ellingsen, T., M. Johannesson, S. Tjotta, and G. Torsvik. "Testing Guilt Aversion." *Games and Economic Behavior* 68, no. 1 (2010): 95-107.
Ellingsen, T., and E. Mohlin. "A Model of Social Duties." *Working Paper, Department of Economics, Stockholm School of Economics* (2022).
Engelmann, D., and A. Ortmann. "The Robustness of Laboratory Gift Exchange - a Reconsideration." 2009.
Engelmann, D., and M. Strobel. "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments." *American Economic Review* 94, no. 4 (2004): 857-69.
Enke, B. "Kinship, Cooperation, and the Evolution of Moral Systems." *Quarterly Journal of Economics* 134, no. 2 (2019): 953-1019.
Epper, T., E. Fehr, H. Fehr-Duda, C. T. Kreiner, D. D. Lassen, S. Leth-Petersen, and G. N. Rasmussen. "Time Discounting and Wealth Inequality." *American Economic Review* 110, no. 4 (2020): 1177-205.
Epper®Senn®Fehr. "Social Preferences across Subject Pools: Students Vs. General Population." *Working Paper, Department of Economics, University of Zurich* (2024).
Esteves-Sorenson, C. "Gift Exchange in the Workplace: Addressing the Conflicting Evidence with a Careful Test." *Management Science* 64, no. 9 (2018): 4365-88.
Exley, C. L., and J. B. Kessler. "Information Avoidance and Image Concerns." *Economic Journal* 133, no. 656 (2023): 3153-68.
Fahr, R., and B. Irlenbusch. "Fairness as a Constraint on Trust in Reciprocity: Earned Property Rights in a Reciprocal Exchange Experiment." *Economics Letters* 66, no. 3 (2000): 275-82.
Fajnzylber, P., D. Lederman, and N. Loayza. "Inequality and Violent Crime." *Journal of Law & Economics* 45, no. 1 (2002): 1-40.
Falk, A., E. Fehr, and U. Fischbacher. "On the Nature of Fair Behavior." *Economic Inquiry* 41, no. 1 (2003): 20-26.
———. "Testing Theories of Fairness - Intentions Matter." *Games and Economic Behavior* 62, no. 1 (2008): 287-303.
Falk, A., E. Fehr, and C. Zehnder. "Fairness Perceptions and Reservation Wages - the Behavioral Effects of Minimum Wage Laws." *Quarterly Journal of Economics* 121, no. 4 (2006): 1347-81.
Falk, A., and U. Fischbacher. "A Theory of Reciprocity." *Games and Economic Behavior* 54, no. 2 (2006): 293-315.

Falk, A., and M. Kosfeld. "The Hidden Costs of Control." *American Economic Review* 96, no. 5 (2006): 1611-30.
Fehr, D., J. Mollerstrom, and R. Perez-Truglia. "Your Place in the World: Relative Income and Global Inequality." *American Economic Journal-Economic Policy* 14, no. 4 (2022): 232-68.
Fehr, E., and A. Falk. "Wage Rigidity in a Competitive Incomplete Contract Market." *Journal of Political Economy* 107, no. 1 (1999): 106-34.
Fehr, E., and U. Fischbacher. "Third-Party Punishment and Social Norms." *Evolution and Human Behavior* 25, no. 2 (2004): 63-87.
Fehr, E., and S. Gachter. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90, no. 4 (2000): 980-94.
———. "Do Incentive Contracts Undermine Voluntary Cooperation7." University of Zurich, 2002.
———. "How Effective Are Trust- and Reciprocity-Based Incentives?" In *Economics, Values, and Organization*, edited by A. Ben-Ner and L. Putterman, 337-63. Cambridge: Cambridge University Press, 1998.
Fehr, E., S. Gachter, and G. Kirchsteiger. "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica* 65, no. 4 (1997): 833-60.
Fehr, E., and L. Goette. "Robustness and Real Consequences of Nominal Wage Rigidity." *Journal of Monetary Economics* 52, no. 4 (2005): 779-804.
Fehr, E., O. Hart, and C. Zehnder. "Contracts as Reference Points-Experimental Evidence." *American Economic Review* 101, no. 2 (2011): 493-525.
Fehr, E., G. Kirchsteiger, and A. Riedl. "Does Fairness Prevent Market Clearing - an Experimental Investigation." *Quarterly Journal of Economics* 108, no. 2 (1993): 437-59.
Fehr, E., A. Klein, and K. M. Schmidt. "Fairness and Contract Design." *Econometrica* 75, no. 1 (2007): 121-54.
Fehr, E., and J. A. List. "The Hidden Costs and Returns of Incentives: Trust and Trustworthiness among Ceos." *Journal of the European Economic Association* 2, no. 5 (2004).
Fehr, E., M. Naef, and K. M. Schmidt. "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Comment." *American Economic Review* 96, no. 5 (2006): 1912-17.
Fehr, E., M. Powell, and T. Wilkening. "Behavioral Constraints on the Design of Subgame-Perfect Implementation Mechanisms." *American Economic Review* 111, no. 4 (2021): 1055-91.
Fehr, E., and K. M. Schmidt. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114, no. 3 (1999): 817-68.
Fehr, E., and I. Schurtenberger. "Normative Foundations of Human Cooperation." *Nature Human Behaviour* 2, no. 7 (2018): 458-68.
Fehr, E., E. Tougareva, and U. Fischbacher. "Do High Stakes and Competition Undermine Fair Behaviour? Evidence from Russia." *Journal of Economic Behavior & Organization* 108 (2014): 354-63.
Fehr®Epper®Senn. "The Fundamental Properties, Stability and Predictive Ability of Social Preferences." *Working Paper, Department of Economics, University of Zurich* (2023).
———. "Social Preferences and Redistributive Politics." *Working Paper, Department of Economics, University of Zurich* (2021).
Finan, F., and L. Schechter. "Vote-Buying and Reciprocity." *Econometrica* 80, no. 2 (2012): 863-81.
Fischbacher, U., C. M. Fong, and E. Fehr. "Fairness, Errors and the Power of Competition." *Journal of Economic Behavior & Organization* 72, no. 1 (2009): 527-45.

Fisman, R., P. Jakiela, S. Kariv, and D. Markovits. "The Distributional Preferences of an Elite." *Science* 349, no. 6254 (2015).
Fisman, R., S. Kariv, and D. Markovits. "Individual Preferences for Giving." *American Economic Review* 97, no. 5 (2007): 1858-76.
Fisman, R., I. Kuziemko, and S. Vannutelli. "Distributional Preferences in Larger Groups: Keeping up with the Joneses and Keeping Track of the Tails." *Journal of the European Economic Association* 19, no. 2 (2021): 1407-38.
Fong, C. "Social Preferences, Self-Interest, and the Demand for Redistribution." *Journal of Public Economics* 82, no. 2 (2001): 225-46.
Fong, C.M., S. Bowles, and H. Gintis. "Behavioural Motives for Income Redistribution." *The Australian Economic Review* 38, no. 3 (2005): 285-97.
Forsythe, R., J. L. Horowitz, N. E. Savin, and M. Sefton. "Fairness in Simple Bargaining Experiments." *Games and Economic Behavior* 6, no. 3 (1994): 347-69.
Frank, R. H., T. Gilovich, and D. T. Regan. "Does Studying Economics Inhibit Cooperation." *Journal of Economic Perspectives* 7, no. 2 (1993): 159-71.
Frohlich, N., J. Oppenheimer, and A. Kurki. "Modeling Other-Regarding Preferences and an Experimental Test." *Public Choice* 119, no. 1-2 (2004): 91-117.
Gächter, S., D. Nosenzo, and M. Sefton. "The Impact of Social Comparisons on Reciprocity." *Scandinavian Journal of Economics* 114, no. 4 (2012): 1346-67.
———. "Peer Effects in Pro-Social Behavior: Social Norms or Social Preferences?" *Journal of the European Economic Association* 11, no. 3 (2013): 548-73.
Gachter, S., and C. Thoni. "Social Comparison and Performance: Experimental Evidence on the Fair Wage-Effort Hypothesis." *Journal of Economic Behavior & Organization* 76, no. 3 (2010): 531-43.
———. "Social Learning and Voluntary Cooperation among Like-Minded People." *Journal of the European Economic Association* 3, no. 2-3 (2005): 303-14.
Gantner, A., W. Guth, and M. Konigstein. "Equitable Choices in Bargaining Games with Joint Production." *Journal of Economic Behavior & Organization* 46, no. 2 (2001): 209-25.
Gastorf, J. W., and J. Suls. "Performance Evaluation Via Social-Comparison - Performance Similarity Versus Related-Attribute Similarity." *Social Psychology* 41, no. 4 (1978): 297-305.
Geanakoplos, J., D. Pierce, and E. Stachetti. "Psychological Games and Sequential Rationality." *Games and Economic Behavior* 1, no. 1 (1989): 60-79.
Gneezy, U., and J. A. List. "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments." *Econometrica* 74, no. 5 (2006): 1365-84.
Greenberg, J. "Employee Theft as a Reaction to Underpayment Inequity - the Hidden Cost of Pay Cuts." *Journal of Applied Psychology* 75, no. 5 (1990): 561-68.
Gregg, P., P. A. Grout, A. Ratcliffe, S. Smith, and F. Windmeijer. "How Important Is Pro-Social Behaviour in the Delivery of Public Services?" *Journal of Public Economics* 95, no. 7-8 (2011): 758-66.
Grigsby, J., E. Hurst, and A. Yildirmaz. "Aggregate Nominal Wage Adjustments: New Evidence from Administrative Payroll Data." *American Economic Review* 111, no. 2 (2021): 428-71.
Grossman, S. J., and O. D. Hart. "The Costs and Benefits of Ownership - a Theory of Vertical and Lateral Integration." *Journal of Political Economy* 94, no. 4 (1986): 691-719.
Grossman, Z. "Self-Signaling and Social-Signaling in Giving." *Journal of Economic Behavior & Organization* 117 (2015): 26-39.
———. "Strategic Ignorance and the Robustness of Social Preferences." *Management Science* 60, no. 11 (2014): 2659-65.

Grossman, Z., and J. J. van der Weele. "Self-Image and Willful Ignorance in Social Decisions." *Journal of the European Economic Association* 15, no. 1 (2017): 173-217.
Guiso, L., P. Sapienza, and L. Zingales. "Long-Term Persistence." *Journal of the European Economic Association* 14, no. 6 (2016): 1401-36.
Guth, W., R. Schmittberger, and B. Schwarze. "An Experimental-Analysis of Ultimatum Bargaining." *Journal of Economic Behavior & Organization* 3, no. 4 (1982): 367-88.
Hart, O., and J. Moore. "Contracts as Reference Points." *Quarterly Journal of Economics* 123, no. 1 (2008): 1-48.
———. "Property-Rights and the Nature of the Firm." *Journal of Political Economy* 98, no. 6 (1990): 1119-58.
Hedegaard, M., R. Kerschbamer, D. Muller, and J. R. Tyran. "Distributional Preferences Explain Individual Behavior across Games and Time." *Games and Economic Behavior* 128 (2021): 231-55.
Henkel®Fehr®Senn®Epper. "Beliefs About Inequality and the Nature of Support for Redistribution." *Working Paper, Department of Economics, University of Zurich* (2024).
Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies." *American Economic Review* 91, no. 2 (2001): 73-78.
Hoffman, E., K. Mccabe, K. Shachat, and V. Smith. "Preferences, Property-Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior* 7, no. 3 (1994): 346-80.
Hollander, H. "A Social-Exchange Approach to Voluntary Cooperation." *American Economic Review* 80, no. 5 (1990): 1157-67.
Homans, C. G. *Social Behavior: Its Elementary Forms*. New York: Harcourt, Brace & World, 1961.
Hvidberg, K.B., S. Stantcheva, and C. Thustrup-Kreiner. "Social Position and Fairness Views on Inequality." *Review of Economic Studies* 90, no. 6 (2023): 3083 - 118.
Jakiela, P. "How Fair Shares Compare: Experimental Evidence from Two Cultures." *Journal of Economic Behavior & Organization* 118 (2015): 40-54.
Jayaraman, R., D. Ray, and F. de Véricourt. "Anatomy of a Contract Change." *American Economic Review* 106, no. 2 (2016): 316-58.
Karadja, M., J. Mollerstrom, and D. Seim. "Richer (and Holier) Than Thou? The Effect of Relative Income Improvements on Demand for Redistribution." *Review of Economics and Statistics* 99, no. 2 (2017): 201-12.
Kaufmann, R.T. "On Wage Stickiness in Britain's Competitive Sector." *British Journal of Industrial Relations* 22 (1984): 101-12.
Kaur, S. "Nominal Wage Rigidity in Village Labor Markets." *American Economic Review* 109, no. 10 (2019): 3585-616.
Kerschbamer, R. "The Geometry of Distributional Preferences and a Non-Parametric Identification Approach: The Equality Equivalence Test." *European Economic Review* 76 (2015): 85-103.
Kerschbamer, R., and D. Muller. "Social Preferences and Political Attitudes: An Online Experiment on a Large Heterogeneous Sample." *Journal of Public Economics* 182 (2020).
Khalmetski, K. "Testing Guilt Aversion with an Exogenous Shift in Beliefs." *Games and Economic Behavior* 97 (2016): 110-19.
Konow, J. "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions." *American Economic Review* 90, no. 4 (2000): 1072-91.
Kosfeld, M., and D. Rustagi. "Leader Punishment and Cooperation in Groups: Experimental Field Evidence from Commons Management in Ethiopia." *American Economic Review* 105, no. 2 (2015): 747-83.

Kosse, F., T. Deckers, P. Pinger, H. Schildberg-Horisch, and A. Falk. "The Formation of Prosociality: Causal Evidence on the Role of Social Environment." *Journal of Political Economy* 128, no. 2 (2020): 434-67.
Krawczyk, M. "A Glimpse through the Veil of Ignorance: Equality of Opportunity and Support for Redistribution." *Journal of Public Economics* 94, no. 1-2 (2010): 131-41.
Krawczyk, M., and F. Le Lec. "How to Elicit Distributional Preferences: A Stress -Test of the Equality Equivalence Test." *Journal of Economic Behavior & Organization* 182 (2021): 13-28.
Krueger, A. B., and A. Mas. "Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires." *Journal of Political Economy* 112, no. 2 (2004): 253-89.
Kube, S., M. A. Marechal, and C. Puppe. "The Currency of Reciprocity: Gift Exchange in the Workplace." *American Economic Review* 102, no. 4 (2012): 1644-62.
———. "Do Wage Cuts Damage Work Morale? Evidence from a Natural Field Experiment." *Journal of the European Economic Association* 11, no. 4 (2013): 853-70.
Lane, I.M., and L.A. Messe. "Equity and the Distribution of Rewards." *Journal of Personality and Social Psychology* 20, no. 1 (1971): 1-17.
Larney, A., A. Rotella, and P. Barclay. "Stake Size Effects in Ultimatum Game and Dictator Game Offers: A Meta-Analysis." *Organizational Behavior and Human Decision Processes* 151 (2019): 61-72.
Lefgren, L. J., D. P. Sims, and O. B. Stoddard. "Effort, Luck, and Voting for Redistribution." *Journal of Public Economics* 143 (2016): 89-97.
Leventhal, G. S., and D. Anderson. "Self-Interest and Maintenance of Equity." *Journal of Personality and Social Psychology* 15, no. 1 (1970): 57-+.
Leventhal, G. S., and D.W. Lane. "Sex, Age, and Equity Behavior." *Journal of Personality and Social Psychology* 13, no. 4 (1970): 312-16.
Leventhal, G. S., and J.W. Michaels. "Locus of Cause and Equity Motivation as Determinants of Reward Allocation." *Journal of Personality and Social Psychology* 17, no. 3 (1971): 229-35.
Leventhal, G. S., and J.W. Michaels. "Extending the Equity Model: Perception of Inputs and Allocation of Reward as a Function of Duration and Quantity of Performance." *Journal of Personality and Social Psychology* 12, no. 3 (1969): 303-09.
Levine, D. I. "Fairness, Markets, and Ability to Pay - Evidence from Compensation Executives." *American Economic Review* 83, no. 5 (1993): 1241-59.
Levine, D. K. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics* 1, no. 3 (1998): 593-622.
Liebrand, W. B. G. "The Effect of Social Motives, Communication and Group-Size on Behavior in an N-Person Multi-Stage Mixed-Motive Game." *European Journal of Social Psychology* 14, no. 3 (1984): 239-64.
Liebrand, W. B. G., and C. G. Mcclintock. "The Ring Measure of Social Values - Computerized Procedure for Assessing Individual-Differences in Information-Processing and Social Value Orientation." *European Journal of Personality* 2, no. 3 (1988): 217-30.
Mas, A. "Labour Unrest and the Quality of Production: Evidence from the Construction Equipment Resale Market." *Review of Economic Studies* 75, no. 1 (2008): 229-58.
———. "Pay, Reference Points, and Police Performance." *Quarterly Journal of Economics* 121, no. 3 (2006): 783-821.
Mas, A., and E. Moretti. "Peers at Work." *American Economic Review* 99, no. 1 (2009): 112-45.
Maskin, E., and J. Tirole. "Unforeseen Contingencies and Incomplete Contracts." *Review of Economic Studies* 66, no. 1 (1999): 83-114.

Mayer, S. E. "How Did the Increase in Economic Inequality between 1970 and 1990 Affect Children's Educational Attainment?" *American Journal of Sociology* 107, no. 1 (2001): 1-32.
Meltzer, A. H., and S. F. Richard. "A Rational Theory of the Size of Government." *Journal of Political Economy* 89, no. 5 (1981): 914-27.
Messick, D. M., and C. G. Mcclintock. "Motivational Bases of Choice in Experimental Games." *Journal of Experimental Social Psychology* 4, no. 1 (1968): 1-25.
Mikula, G., and H. Uray. "Neglect of Individual Performances in Financial Profit with Respect to Social Situations." *Zeitschrift Fur Sozialpsychologie* 4, no. 2 (1973): 136-44.
Molleman, E., J. Pruyn, and A. Van Knippenberg. "Social Comparison Processes among Cancer Patients." *Britisch Journal of Social Psychology* 25, no. 1 (1986): 1-13.
Moore, J., and R. Repullo. "Subgame Perfect Implementation." *Econometrica* 56, no. 5 (1988): 1191-220.
Murphy, R. O., and K. A. Ackermann. "Social Value Orientation: Theoretical and Measurement Issues in the Study of Social Preferences." *Personality and Social Psychology Review* 18, no. 1 (2014): 13-41.
Murphy, R. O., K. A. Ackermann, and M. J. J. Handgraaf. "Measuring Social Value Orientation." *Judgment and Decision Making* 6, no. 8 (2011): 771-81.
Nunnari, S., and M. Pozzi. "Meta-Analysis of Social Preferences Estimates." *Working Paper, Bocconi University, Italy* (2022).
Offerman, T. "Hurting Hurts More Than Helping Helps." *European Economic Review* 46, no. 8 (2002): 1423-37.
Paetzel, F., R. Sausgruber, and S. Traub. "Social Preferences and Voting on Reform: An Experimental Study." *European Economic Review* 70 (2014): 36-55.
Perugini, M., M. Gallucci, F. Presaghi, and A. P. Ercolani. "The Personal Norm of Reciprocity." *European Journal of Personality* 17, no. 4 (2003): 251-83.
Platteau, J. P. "Does Africa Need Land Reform?" *Evolving Land Rights, Policy and Tenure in Africa* (2000): 51-73.
Prelec, D., and R. Bodner. "Self-Signaling and Self-Control." In *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*, edited by G. Loewenstein, D. Read and R. F. Baumeister, 277–98. New York: Russell Sage Foundation, 2003.
Preuss, M., G. Reyes, J. Somerville, and J. H. Wu. "Inequality of Opportunity and Income Redistribution." *Unpublsihed Working Paper* (2022).
Quach, S. "The Extent of Downward Nominal Wage Rigidity: Evidence from a Natural Experiment." Princeton, NJ: Princeton University 2020.
Rege, M., and K. Telle. "The Impact of Social Approval and Framing on Cooperation in Public Good Situations." *Journal of Public Economics* 88, no. 7-8 (2004): 1625-44.
Reuben, E., P. Sapienza, and L. Zingales. "Is Mistrust Self-Fulfilling?" *Economics Letters* 104, no. 2 (2009): 89-91.
Roemer, J., and A. Trannoy. "Equality of Opportunity." In *Handbook of Income Distribution*, edited by Atkinson. A. and F. Bourguignon. Amsterdam - New York: Elsevier, 2015.
Roth, A. E., V. Prasnikar, M. Okunofujiwara, and S. Zamir. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo - an Experimental-Study." *American Economic Review* 81, no. 5 (1991): 1068-95.
Ruffle, B. J. "More Is Better, but Fair Is Fair: Tipping in Dictator and Ultimatum Games." *Games and Economic Behavior* 23, no. 2 (1998): 247-65.
Rustagi, D. "Historical Self Governance and Norms of Cooperation." *Econometrica (forthcoming)* (2024).

Rustagi, D., S. Engel, and M. Kosfeld. "Conditional Cooperation and Costly Monitoring Explain Success in Forest Commons Management." *Science* 330, no. 6006 (2010): 961-65.
Saez, E. "Public Economics and Inequality: Uncovering Our Social Nature." *Aea Papers and Proceedings* 111 (2021): 1-26.
Saez, E., and S. Stantcheva. "Generalized Social Marginal Welfare Weights for Optimal Tax Theory." *American Economic Review* 106, no. 1 (2016): 24-45.
Schminke, M., R. Cropanzano, and D. E. Rupp. "Organization Structure and Fairness Perceptions: The Moderating Effects of Organizational Level." *Organizational Behavior and Human Decision Processes* 89, no. 1 (2002): 881-905.
Schneider, F.H., F. Brun, and R.A. Weber. "Sorting and Wage Premiums in Immoral Work." Zurich, Switzerland: University of Zurich, 2020.
Schulz, J. F., D. Bahrami-Rad, J. P. Beauchamp, and J. Henrich. "The Church, Intensive Kinship, and Global Psychological Variation." *Science* 366, no. 6466 (2019): 707-+.
Simkins, P. *Kitchener's Army: The Raising of the New Armies, 1914-1916*. Manchester: Manchester United Press, 1988.
Skarlicki, and Folger. "Retaliation in the Workplace: The Roles of Distributive, Procedural, and Interactional Justice (Vol 82, Pg 434, 1997)." *Journal of Applied Psychology* 82, no. 6 (1997): 888-88.
Sliwka, D. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes." *American Economic Review* 97, no. 3 (2007): 999-1012.
Slonim, R., and A. E. Roth. "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic." *Econometrica* 66, no. 3 (1998): 569-96.
Smith, V. L. "Microeconomic Systems as an Experimental Science." *American Economic Review* 72, no. 5 (1982): 923-55.
Snowberg, E., and L. Yariv. "Testing the Waters: Behavior across Participant Pools." *American Economic Review* 111, no. 2 (2021): 687-719.
Tesser, A. "Toward a Self-Evaluation Maintenance Model of Social Behavior." In *Advances in Experimental Social Psychology*, edited by L. Berkowitz, 181-227: Academic Press, 1988.
Thibaut, J. W., and H. H. Kelley. *The Social Psychology of Groups*. New York: Wiley, 1959.
Tyran, J. R., and R. Sausgruber. "A Little Fairness May Induce a Lot of Redistribution in Democracy." *European Economic Review* 50, no. 2 (2006): 469-85.
Van Lange, P. A. M. "The Pursuit of Joint Outcomes and Equality in Outcomes: An Integrative Model of Social Value Orientation." *Journal of Personality and Social Psychology* 77, no. 2 (1999): 337-49.
Van Lange, P. A. M., D. P. Balliet, C. D. Parks, and M. Van Vugt. *Social Dilemmas: Understanding Human Cooperation*. New York: Oxford University Press, 2014.
Van Lange, P. A. M., W. Otten, E. M. N. DeBruin, and J. A. Joireman. "Development of Prosocial, Individualistic, and Competitive Orientations: Theory and Preliminary Evidence." *Journal of Personality and Social Psychology* 73, no. 4 (1997): 733-46.
Vanberg, C. "Why Do People Keep Their Promises? An Experimental Test of Two Explanations." *Econometrica* 76, no. 6 (2008): 1467-80.
Vu, L., I. Sorraperra, M. Leib, J. van der Weele, and S. Shalvi. "Willful Ignorance: A Meta-Analytic Review." *Psychological Bulletin* forthcoming (2023).
Walster, E. , W. Walster, and E. Berscheid. *Equity: Theory and Research*. Boston: Allyn and Bacon, 1977.
Xu, X., S. Metsälampi, M. Kichler, K. Kotakorpi, P. H. Matthews, and T. Miettinen. "Which Income Comparisons Matter to People, and How? Evidence from a Large Field Experiment." *Working Paper, Hanken School of Economics, Finland* (2023).
Yang, Y. "A Survey of the Hold-up Problem in the Experimental Economics Literature." *Journal of Economic Survey* 35, no. 1 (2021): 227 - 49.

Online Appendix for

Social Preferences

Fundamental Characteristics and Economic Consequences

Ernst Fehr & Gary Charness

Table of Contents for Online Appendix

Appendix 1
Heterogeneity in Altruistic Distributional Preferences between Individuals and Subject Pools

This appendix describes the characterization of individual heterogeneity in terms of individuals' estimated CES utility functions. Andreoni and Miller (2002) appear to be the first who studied individual heterogeneity with the CES approach to social preferences. They recruited 176 student subjects who made between 8 – 11 choices in dictator games with varying prices of giving, allowing them to check for violations of the generalized axioms of revealed preferences (GARP). They find that less than 2% commit GARP violations, meaning that the choices of the remaining 98% can be represented by a quasi-concave utility function. They also classify individuals into one of three *predefined* categories: selfish subjects, egalitarian subjects who maximize $U_i\left(\pi_i, \pi_j\right)$ = min (π_i, π_j), and utilitarians who maximize $(0.5\ \pi_i + 0.5\ \pi_j)$. While 43 percent of their subjects display choices that perfectly fit these preference categories, the remaining 57 percent are allocated to these categories by minimizing the distance from the three pre-specified utility functions. Based on this procedure, they classify 47.2% of the 176 subjects as selfish, 30.4% as egalitarian and 22.4% as utilitarian.

To what extent do the 57% of "impure" subjects actually fit the three predefined preference categories? To answer this question, the authors estimate a representative CES function (2) for the "impure" individuals in each category. The results indicate that the estimated parameters deviate quite substantially from the parameters of the ideal types. For example, the average α' of the "impure" selfish subjects is 0.24, indicating a non-negligible deviation from selfishness, and the average ρ of the egalitarian types is –0.35 which is a long way from $-\infty$ which would indicate strict egalitarianism. While such deviations from the pure types are inevitable when people are classified into subgroups it is important to keep them in mind.

Two further observations related to Andreoni and Miller (2002) are worth mentioning. First, even those individuals who perfectly fit the selfish preference assumption in their choice data may not be perfectly selfish because the smallest relative price of giving was 0.25 – for every dollar given, the partner received $4. Thus, we do not know what would have

happened if the relative price had been lower.[1] Second, 34 subjects in one of their sessions also faced upwards sloping budget line in (π_i, π_j)-space that involved disadvantageous inequality. Subjects could reduce inequality in these budget lines by decreasing both players' payoffs, and 8 of the 34 subjects (23.5%) actually did so. Thus, they observed some evidence in favor of inequality aversion when behind but no strong inferences can be made here given the small sample size, and the CES utility function is not capable of capturing these preferences.

The Fisman-Jakiela-Kariv-Markovits group undertook one of the most systematic characterizations of individual heterogeneity in altruistic distributional preferences in a series of papers (Fisman, Kariv and Markovits 2007; Fisman, Jakiela and Kariv 2015; Fisman et al. 2015; Li et al. 2022). Subjects in their experiments faced many different budget constraints in the material payoff space, giving them substantial power to estimate the individual preference parameters α' and ρ of the CES utility function. In Fisman, Kariv and Markovits (2007) and Fisman, Jakiela and Kariv (2015), they report the parameter estimates of 76 and 72 Berkely undergraduates, respectively; moreover, they estimate the distributional preferences of 208 Yale Law School (YLS) students in Fisman et al. (2015) as well as of 503 US medical students in Li et al. (2017). In Figures 2a and 2b we show the cumulative distribution of the estimated α' and ρ parameter for the Berkeley and the Yale Law School students and Appendix Table A1 classifies the individuals into three categories: those close to selfishness ($\alpha' > 0.95$), intermediate altruists ($0.55 \leq \alpha' \leq 0.95$) and egalitarian altruists ($0.45 < \alpha' < 0.55$). The figures and Table A1 illustrate that between 30 and 40 percent of the students put literally a weight of zero or a weight close to zero on other individuals' payoffs ($\alpha' > 0.95$), while only between 8 and 25 percent of them are egalitarian altruists. Moreover, the student subject pools appear to be more oriented towards efficiency compared to equality because only between 30 and 37% of them reveal a $\rho < 0$.

Are these results from student samples generalizable to the general population? To answer this question, the Fisman-Jakiela-Kariv-Markovits group also conducted experiments with a large sample of roughly 1000 Adult Americans from the American Life Panel (ALP). The ALP subjects are broadly comparable with the US population in terms of demographic and socio-economic characteristics. To control for age, Fisman et al. (2015) use only the ALP subjects under age 40 for the comparison with the student sample. The figures show that the

[1] Some people may be inclined to discount situations in which the cost of altruistic acts is low, but social life is in fact pervaded by situations in which low-cost favors can be given to other people. When a colleague in the workplace asks for help, when a stranger in a city asks for directions, or when students help each other answer questions, the costs involved are often very low, while the benefits for the receiving party are high.

ALP sample under age 40 displays a much higher concern for the payoff of others (Figure 2a) and a much higher concern for equity compared to efficiency (Figure 2b) than the student sample. The following facts displayed in Table A1 are, in particular, noteworthy: (i) Among the ALP subjects under age 40, the share of individuals that are close to selfishness is only 16.2% which is much smaller than the 30-40% among the students. (ii) The share of egalitarian altruists is with 37.2% of ALP subjects under age 40 much larger than the 8-26% among the students. (iii) The share of equality-oriented individuals ($\rho < 0$) is with 47% of ALP subjects under 40 much larger than corresponding share among the students.[2]

These large differences between student samples and the broader population are consistent with research reported in Snowberg and Yariv (2021) and Cappelen et al. (2015). Snowberg and Yariv document that subjects from a representative sample of the US population transfer a much higher share of income (39%) to recipients in simple dictator games compared to the transfers given by a large sample of all Caltech undergraduate students, who gave only 14%. Likewise, Cappelen et al. (2015) report that in a representative sample of the Norwegian population the share transferred was 40.3% for men and 41.7% for women, while male students only gave 22.6% and female students gave 32.2%.

[2] The FJKM group also shows that the much higher degree of other-regardingness and the much higher equality orientation of the broad population sample does not depend on socio-economic status. In other words, the individuals with high education and income in the ALP sample (N = 152) display very similar parameters compared to the rest of the ALP sample.

Figure A1a

The estimated weight on self-payoff (α') among students and in a broad sample of the US population under age 40. High α' means a low concern for others' payoff.
(based on data from FKM 2007, FJK 2015, FJKM 2015)[3]

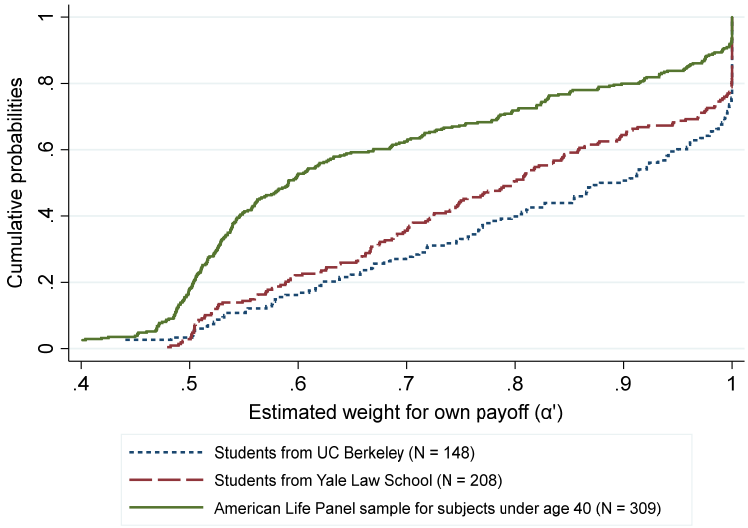
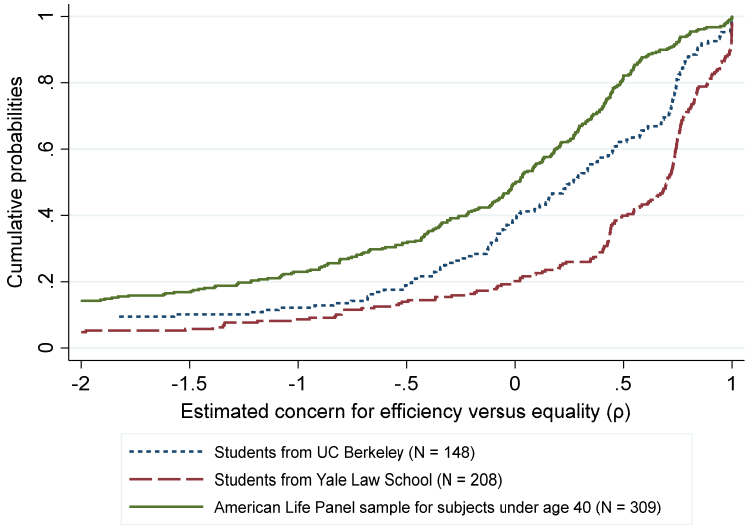


Figure A1b

The estimated weight of efficiency relative to equality concerns (ρ) among students and a broad sample of the US population under age 40. High ρ means a low concern for equality.
(based on data from FKM 2007, FJK 2015, FJKM 2015)



[3] FKM (2007) indicates Fisman, Kariv and Markovits (2007), FJK (2015) indicates Fisman, Jakiela and Kariv (2015) and FJKM indicates Fisman, Jakiela, Kariv and Markovits (2015). See also Table A1 for the type classification that follows from the estimates displayed in Figures 2a and 2b.

Table A1: Empirical Properties of Altruistic Distributional Preferences

Study	**Subject Pool**	**Egalitarian altruism $0.45 < \alpha' < 0.55$**	**Intermediate altruism $0.55 \leq \alpha' \leq 0.95$**	**Close to Selfishness $\alpha' > 0.95$**	**$\rho < 0$**
FKM 2007 & FJK 2015	N = 148 UC Berkeley students	8.1%	49.3%	39.9%	37.%
FJKM 2015	N = 208 Yale Law School Students	14.5%	53.9%	31.8%	20.3%
JDK 2017	N = 503 Students from US medical schools	25.7%	41.5%	28.2%	29.2%
FJKM 2015	N = 309 Adult Americans under 40 (ALP subjects)	37.2%	42.7%	16.2%	47.3%
	N = 693 Adult Americans over 40 (ALP subjects)	27.7%	50.5%	16.0%	57.0%
LDK 2017	N = 208 US Physicians	36.8%	42.8%	15.1%	48.3%

Note. The table shows key components of the distribution of individuals' estimated weights (α') on other persons' payoffs based on studies co-authored by D (Dow), F (Fisman), J (Jakiela), K (Kariv), L (LI) and M (Markovits). Thus. FKM (2007) indicates the paper by Fisman, Kariv and Markovits (2007). The estimates are based on the assumption that distributional preferences can be captured by a CES utility function like in equation (3) and on each subjects' distributional choices in 50 randomly chosen budget sets. The efficient frontier of the budget set (i.e., the "budget line") is always negatively sloped such that one cannot measure the willingness to pay *to reduce* others' income for the sake of equality ("inequality aversion"). However, the CES function enables the identification of individuals' preference for equality *within the class of altruistic preferences* with the parameters α' and ρ. $\alpha' = 1/2$ indicates that individuals put equal weight on others' payoff, and $\rho < 0$ implies that the income share spent on others' payoff rises as the price of giving rises, i.e., subjects are equality-oriented ($\rho < 0$) and not efficiency-oriented ($0 < \rho < 1$).

One noteworthy feature of the experimental design on which the data in Figure 1a and 1b and Table A1 are based is that the price of giving is randomly determined for every subject, i.e., different subjects see different prices. This means that some subjects may have seen a relatively large number of low prices for giving, which makes identification of purely selfish subjects very precise, while other subjects may have seen only a few low prices of giving, so that their assignment to the selfish versus intermediate category may be coarser.

Another important feature of the data collected by the Fisman-Jakiela-Kariv-Markovits group is that the subjects do not face upwards sloping budget lines in (π_a, π_b)-space. Thus, by construction, the subjects do not face a situation in which they can decrease both players' payoffs to reduce disadvantageous inequality. Given this restriction, the CES approach is a powerful tool for identifying *altruistic* distributional preferences, but it cannot capture spiteful, envious, or inequality averse preferences[4].

[4] In Fisman, Kariv, Markovits (2007), the authors had budget constraints with vertical and horizontal segments, but their student subjects never made pareto-damaging choices on these segments, which led the authors to believe that inequality aversion is not important.

Appendix 2

The Equality Equivalence Test

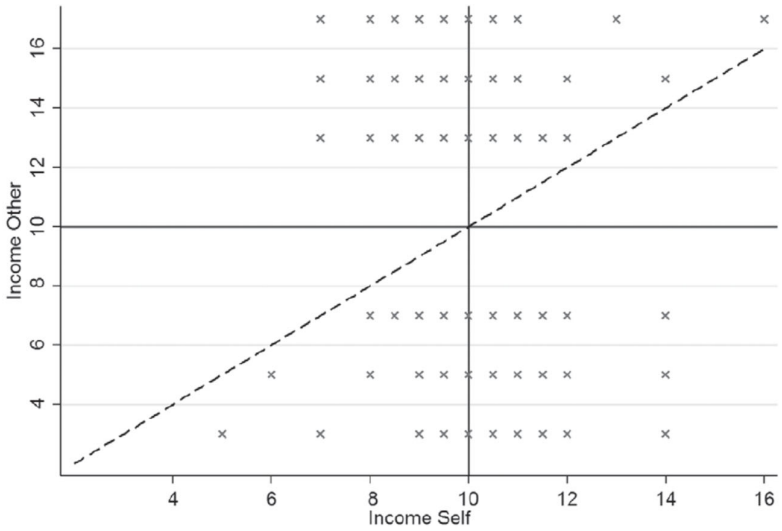
In the equality equivalence test subjects are presented choice lists in the domain of advantageous payyofs (A-lists) and the domain of disadvantageous payoffs (DA-lists). In a disadvantageous list (DA-list, see Figure A2 below), the equal payoff allocation E is always paired with a list of alternative allocations in which the other subject's payoff is kept constant at a level of $\pi_j > \pi_i$, while π_i systematically varies across alternative allocations. In an advantageous list (see Figure A2), E is always paired with a list of alternative allocations in which the other subject's payoff is kept constant at a level of $\pi_j < \pi_i$ while π_i systematically varies across alternative allocations.

Starting the binary choice list with the choice between the (π_i, π_j)-combination and E where π_i is *lowest* (and hence, below the egalitarian payoff, see Figure A1), the decision maker is more benevolent towards the other subject (i.e., willing to pay to increase the other's payoff) in the DA domain, the earlier he or she moves from E towards an alternative allocation (π_i, π_j). In the advantageous domain, the decision maker is more benevolent if he or she, starting the binary choice list with the choice between the (π_i, π_j)-combination and E where π_i is *highest* (and hence, above the egalitarian payoff), moves earlier to the equal payoff allocation E. However, the EET can also identify inequality aversion in the DA domain because some binary choice pairs essentially imply a choice on a positively sloped "budget line". Likewise, the EET can also identify positively sloped indifference curves in the A domain ("spite") because some binary choice pairs in this domain are located on positively sloped "budget lines".

A potential drawback of the EET is that the equal payoff allocation is part of every binary choice the subjects face, which may render equality very salient and thus induce a behavioral bias towards equality. However, a study by Krawczyk and Lee (2021) indicates that the results are robust to the introduction of a reference allocation that does not involve equality. In addition, the results of the EET by Kerschbamer (2015) indicates that 48.9% of his student subjects reveal selfish preferences (see Table 1 below), which is even higher than the 39.9% of selfish students in Fisman, Kariv and Markovits (2007) or the 31.8% of selfish students in Fisman et al. (2015). Likewise, Table 1 presents the data from several other studies with student samples that indicate a relatively high share of selfish subjects that approaches 60% in some student samples. Moreover, among the student subjects with other-regarding distributional preferences, those with altruistic preferences are far more prevalent

compared to inequality averse or envious preferences. The share of altruistic student subjects varies between 28 and 48%, while the share of inequality averse subjects is between 7 and 12%. Typically, envious/spiteful subjects are the least frequent across the student data with 3-10%.

Figure A2: Choice Alternatives in the Equality-Equivalence Test



The figure illustrates how the Equality Equivalence Test (EET) works by depicting the alternatives to the equal payoff allocation which is at (10, 10). The figure is taken from Kerschbamer and Müller (2021). It shows three binary choice lists in the disadvantageous domain (DA-lists) and three lists in the advantageous domain (A-lists). The DA-lists enable the identification of the slope of a subject's indifference curve in the DA domain (α), while the A-lists enable identification of the slope in the A domain (β).

Appendix 3: Distributional Preferences under Risk

When individuals care for others' payoffs, a whole new set up of questions arises if outcomes are risky. A key issue concerns the question whether people care for others' expected payoffs or for their realized payoffs. This also concerns the issue whether individuals care for equality of opportunity, i.e., have a preference for lower inequality in ex-ante expected payoffs or whether they have a preference for more equal ex-post realized payoffs. Another issue when risk is present is how individuals' own risk preferences and their beliefs about others' risk preferences affect their other-regarding behavior.

The problem of ex-ante expected payoffs versus ex-post realized payoffs comes into sharp focus in a dictator game that involves the sharing of chances to win an indivisible resource that has a value of R = 100 for both parties. The dictator chooses *x,* which determines the probability $\frac{x}{100}$ with which the recipient wins R, while the dictator wins R with probability $(1 - \frac{x}{100})$. Here, equality of opportunity implies the equalization of chances but there will always be inequality ex-post. An individual with utility function $U(\pi_a, \pi_b)$ that obeys the plausible restriction $U(R, 0) > U(0, R)$ will always choose $x = 0$. Not only inequality averse players, but players with Charness-Rabin preferences as well, may plausibly obey the restriction $U(R, 0) > U(0, R)$ and thus choose $x = 0$.

This prediction contrasts, however, with the results of experiments showing that many dictators are willing to transfer some chance of winning to the recipients (Krawczyk and Le Lec 2010; Brock, Lange and Ozbay 2013). Models that are solely based on the realized ex-post payoffs have a hard time explaining this fact, whereas models in which players also care about the ex-ante expected payoffs of others can explain it.

Now suppose that the above-described game is slightly changed so that the payoff to the two players is no longer exclusive, i.e., if the dictator transfers a chance *x*, then the dictator wins R with probability $(1 - \frac{x}{100})$ and the recipient can simultaneously also win R with $\frac{x}{100}$, i.e., there are two independent draws. Note that there may not be any ex-post inequality in this game because both players can end up with 0 or with R. Therefore, inequality averse dictators have less reason to worry about inequality, implying that they are more likely to be willing to share chances with the recipient. Krawcyk and Le Lec (2010) indeed show that dictators transfer more chances in the dictator game with independent draws compared to the game with exclusive payoffs. This result suggests that players also care about ex-post payoffs.

Further evidence for the relevance of ex-post payoffs is provided by Brock, Lange and Ozbay (2013), who designed six different dictator games where they systematically varied the risk for the dictators and the recipients across games in such a way that if players' cared only about ex-ante expected payoffs, they would behave identically across all six games. Their find treatment differences, however, that are indicative for the relevance of ex-post payoff concerns. For example, subjects in the standard dictator game without any risk (and a dictator endowment of 100) transfer a significantly higher x to the recipient compared to a dictator game where the dictator's payoff is still certain, but a transfer of x gives the recipient a payoff of 100 with probability $\frac{x}{100}$. Note that a positive transfer x in the game where the recipient faces a risky payoff implies that the dictator may end up with a lower payoff than the recipient. Inequality averse dictators, who care for ex-post inequality, will thus tend to give less in the risky dictator game.[5] Another key result documented in Brock, Lange and Ozbay (2013) is that subjects' giving in the standard dictator game is highly predictive for their willingness to equalize ex-ante expected values in dictator games involving risks.

The question whether subjects care for equality of opportunity or for equality of ex-post payoffs was also addressed in Cappelen et al. (2013). In their experiments, there was first a risk-taking phase and then a distribution phase. Subjects made 4 decisions in the risk-taking phase between the payoff y of a sure alternative ($y \in \{25, 200, 300, 400\}$ and a 50:50 chance of receiving nothing or 800 NOK. In the distribution phase, each subject was paired sequentially with 8 different subjects who participated in the risk-taking phase, and one of the four risk-taking problems was drawn randomly for each pair. Then, an "impartial" spectator, who was informed about subjects' choices and outcomes in the drawn risk-taking problem, was asked to distribute the pair's total earnings between the two subjects.

Before presenting the results, it is important to emphasize that complete equality of opportunity existed between the two paired subjects in the risk-taking phase. If spectators redistribute ex-post from the richer to the poorer subject, they thus explicitly express a preference for less ex-post inequality. Almost all of the spectators' redistributive choices

[5] Alternatively, because the certainty equivalent of a given transfer x is less valuable for risk averse recipients, dictators who care for the total payoff may give less in the risky dictator game. However, based on this logic risk averse dictators should give *more* in a dictator game in which their own payoff is risky – they receive a payoff of 100 with probability $\left(1 - \frac{x}{100}\right)$ – while the recipient receives the transfer x with certainty. The reason is that a transfer of x decreases the certainty equivalent of the dictator's payoff by less than x, i.e., giving is surplus-enhancing. The evidence strongly suggests the opposite, as dictators give much less in this game compared to the standard dictator game (Freundt and Lange 2017). Moreover, Freundt and Lange also find that the dictators who believe that recipients are risk averse do not give less to the recipients.

involved redistribution from the poorer to the richer subject.[6] If the pair consisted of two risk takers where one was lucky while the other was unlucky, the spectators strongly redistributed from the lucky to the unlucky one – they chose the equal split in more than 40% of the cases and they did not redistribute at all in only roughly 30% of the cases. In contrast, if an unlucky risk-taker was paired with an individual who chose the safe option, the unlucky risk-taker received much fewer transfers and the equal split was only chosen in roughly 15% of the cases. Spectators thus made the unlucky risk takers more responsible for their choices compared to a situation where both were unlucky. Finally, there is also a substantial amount of redistribution when a lucky risk-taker is paired with an individual who chose the safe payoff, but the lucky risk-taker was nevertheless given a higher payoff in roughly 80% of the cases.

Thus, taken together, the literature suggests that subjects on average care about both ex-ante equality of opportunity and ex-post equality of outcomes but there is strong heterogeneity in the weight that individual subjects put on the different conceptions of equality. Cappelen et al. (2013) estimate a mixture model that enables them to assign individuals to three different types – individuals who care only for ex-post equality ("ex-post egalitarians", EPs), individuals who do not care about ex-post equality ("ex-ante egalitarians", EAs), and individuals who care about ex-post equality among those who made the same choice in the risk-taking task ("choice egalitarians", CEs). Roughly 30% of their subjects (students from the Norwegian School of Economics) are EPs, 27% are CEs and 43% are EAs.

In this section we have so far mainly dealt with the question how social preferences are affected by outcome risks. However, the perceived sources of inequality may also be subject to risk and uncertainty. If individuals do not know whether a particular inequality is due to luck or differential performance, how does this affect their willingness to redistribute income? Cappelen et al (2022) study this situation, and document that this kind of uncertainty can push meritocrats towards behaving more egalitarian – with more risk averse spectators exhibiting a stronger drive towards egalitarian behavior.

Finally, we deal with the question how to combine concerns for equality of opportunity and equality of outcomes in theoretical modelling. Saito (2013) addresses this issue, providing an axiomatic foundation for "expected inequality-averse" preferences. Individuals

[6] In case that a lucky risk-taker met a subject who chose the safe payoff it would have been possible to redistribute from the poorer to the richer subject.

with such preferences put a weight of δ ($0 \leq \delta \leq 1$) on preferences for equality of opportunity and a weight (1-δ) on preferences for equal ex-post outcomes.[7]

To make things concrete, let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ denote an allocation of material payoffs to individuals. Assume that there are *m* different states of the world, each one of which is obtained with probability p_s, $s \in \{1, \ldots, m\}$, and denote the allocation obtained in state s by $\boldsymbol{x^s} = (x_1^s, x_2^s, \ldots, x_n^s)$, then the *expected* material payoff allocation is given by

$$E(\boldsymbol{x}) = \sum_{s=1}^{m} p_s \boldsymbol{x^s} = (\sum_{s=1}^{m} p_s x_1^s, \sum_{s=1}^{m} p_s x_2^s, \ldots, \sum_{s=1}^{m} p_s x_n^s),$$

where $\sum_{s=1}^{m} p_s x_i^s$ denotes the expected material payoff of individual *i* across states. Likewise, the allocation of expected utilities is given by

$$E(U(\boldsymbol{x})) = \sum_{s=1}^{m} p_s U(\boldsymbol{x^s})$$

Saito shows that if and only if a decision-maker obeys "his" axioms, the preferences of a decision-maker are represented by the following preference function V:

$$V = \delta U(E(\boldsymbol{x})) + (1-\delta)E(U(\boldsymbol{x}), \qquad (11)$$

where $U(\boldsymbol{x})$ is given by the Fehr-Schmidt Utility function. Thus, the utility of an expected inequality averse player is affected by the inequalities in the expected material payoffs with weight δ and by the inequalities in realized ex-post payoffs with weight $(1-\delta)$. It is also noteworthy that the preference function (8) also applies under further plausible assumptions if $U(\boldsymbol{x})$ is given by Charness-Rabin type preferences.

It is easy to see that an individual who puts a sufficiently high weight δ on equality of opportunity is willing to share the chances of receiving an indivisible resource in a dictator game although this creates chances for high ex-post inequality. Overall, however, the Saito model has undergone very little empirical testing. For example, it would be interesting to know to what extent the behavior of individual subjects in the six different treatment conditions of Brocks, Lange and Ozbay (2013) are consistent with the Saito model and which parameters *(α, β, δ)* explain their behaviors.[8] To our knowledge, there is no paper that jointly estimated δ and the parameters in $U(\boldsymbol{x})$. One complication in applying (8) to data is that the distributional preference models – such as Fehr-Schmidt or Charness-Rabin – assume risk neutrality, but it is well known that risk aversion also exists at the typical experimental stake

[7] Several other authors have also provided axiomatic foundations of inequality averse preferences (Neilson 2006; Rohde 2010) but none of them involves preferences for equality of opportunity.

[8] Recall that in their experiments an individual with $\delta = 1$ would behave identically across all treatments. Thus, behavioral variation across treatments may provide at least some qualitative insights with regard to the parameter constellations that may explain their data.

levels. This means that behavior in distributional problems under risk is affected by a complicated mix of risk aversion as well as by preferences for equality of opportunity and other-regarding preferences for ex-post outcomes.[9]

[9] Cettolin, Riedl and Tran (2017) and Freundt and Lange (2017) have independent measures of dictators' and recipients risk aversion and can relate them to the dictators' behavior in risk-involving dictator games. Cettolin, Riedl and Tran (2017) show that dictators' risk aversion strongly predicts lower transfers in both dictator games that render the payoff of the recipients risky and in dictator games that render the payoff of the dictators risky. Freundt and Lange (2017) also show that a rise in dictators' risk aversion is associated with a decline in generosity in games where the dictators' payoff is subject to risk.

Appendix 4: Payoff Matrices used in Bolton, Brandts and Ockenfels (1998)

Table 1. Payoff rows for the three test matrices (payoffs in Spanish pesetas).

	*c*1	*c*2	*c*3	*c*4	*c*5	*c*6
t	*C* gets 2050	*C* gets 2000	*C* gets 1950	*C* gets 1900	*C* gets 1850	*C* gets 1800
	R gets 800	*R* gets 1000	*R* gets 1200	*R* gets 1400	*R* gets 1600	*R* gets 1800
m	*C* gets 1650	*C* gets 1600	*C* gets 1550	*C* gets 1500	*C* gets 1450	*C* gets 1400
	R gets 900	*R* gets 1100	*R* gets 1300	*R* gets 1500	*R* gets 1700	*R* gets 1900
b	*C* gets 1250	*C* gets 1200	*C* gets 1150	*C* gets 1100	*C* gets 1050	*C* gets 1000
	R gets 1000	*R* gets 1200	*R* gets 1400	*R* gets 1600	*R* gets 1800	*R* gets 2000

Appendix 5 – The Stability of Social Preferences

In this appendix, we review evidence that examines the extent to which social preferences are relatively stable. Measuring the stability of social preferences over time appears straightforward as long as the measurement tools indeed deliver a preference measure and not merely a behavioral measure that is confounded by beliefs and other types of preferences (as discussed in the section on external validity), and as long as the measurement tool at different points in time is identical. In addition, the preference measure is ideally not just based on a single behavioral measure like the choice of the transfer in a standard dictator game but instead on many choice situations across which the costs and benefits of the transfer vary. Otherwise, the recovered preferences contain a lot of measurement errors and noise, which may generate spurious preference instability.

Measuring social preferences across contexts is trickier because the notion of stability is theory-dependent. To illustrate this point, consider the behavior of responders in two versions of the ultimatum game (Blount 1995). In version 1, a random mechanism determines the first-mover's offer exogenously while the first-mover herself makes the offer in version 2. Suppose that the responders are negatively reciprocal but *not* inequality averse. Then responders reject low offers in version 2 of the game but not in version 1 because a low offer does not indicate an unkind intention in version 1 but it does so in version 2 of the game. If one erroneously assumes that responders are inequality averse, one would conclude that the responders' inequality averse preferences are highly unstable because inequality averse responders should reject low offers regardless of whether they are randomly determined or volitionally chosen. However, if one correctly assumes that the responders are negatively reciprocal, their *change in behavior* across the two games is exactly what a stable preference for negative reciprocity predicts. Thus, the extent to which one can interpret changes in behavior across different contexts as changes in preferences is strongly dependent on the assumption about the underlying psychological mechanism. For this reason, care needs to be exercised when preference stability is assessed by examining behaviors across contexts.

With the above caveats in mind, what does the evidence on the stability of social preferences show? Bruhin et al. (2019) estimated the structural parameters twice for advantageous (β') and disadvantageous (α') inequality aversion in a sample of N = 196 students three months apart with the same experimental paradigm. They found that the intertemporal correlation of individuals' α' is 0.48 while the correlation for β' is 0.56.

Fehr®Epper®Senn (2022) also measured individuals' social preferences in a broad Swiss sample (N = 415) at two points in time that were three years apart (in 2017 and 2020).

The subjects faced the exact same large set of budget lines which makes it possible to study preference stability (i) at the level of choice for individual budget lines, (ii) at the level of individuals' estimated structural preference parameters and (iii) at the level of individuals' assignments to different preference types. At the choice level, roughly 55% of the choices are perfectly identical across time points and 67% of the choices are identical or coincide with the closest neighboring allocation on the budget line. At the level of individuals' structural parameters, they find an intertemporal rank correlation of 0.458 for α' and 0.428 for β'. Finally, at the level of type assignment, they find that 68% of the individuals are assigned to the same preference type (altruistic, inequality averse, selfish) across the two points in time, and that among the individuals classified as other-regarding (altruistic or inequality averse) in 2017, 89% are again classified as other-regarding in 2020. Among the individuals classified as selfish in 2017, 60% are again classified as selfish in 2020.

Moreover, two waves of the German Internet Panel implemented the same equality equivalence Test (Kerschbamer and Muller 2020). In total N = 2583 individuals participated twice in this test, 2 years apart (2016 and 2018). This permits an analysis of the stability of individuals' assignment to four pre-defined preference types (selfish, altruistic, inequality averse, envious; see Table 2). This analysis shows that 60% of individuals remain assigned to the same preference type across the two years, and that among the 76% of individuals who were classified as altruistic or inequality averse in 2016, 84.5% were again assigned to these two preference types.

Chuang and Schechter (2015) also report significantly positive intertemporal correlations between 0.21 and 0.32 involving survey measures of negative reciprocity taken in 2007, 2009 and 2010. Likewise, Carlsson, Johansson-Stenman and Nam report significantly positive intertemporal correlations of social preference related behaviors (voluntary money and labor contributions to a natural public good) at four different points in time spread across six years.

Thus, taken together, the data suggest a reasonable degree of stability in social preference when measured at the level of choices, structural parameters, or preference type assignment. However, the data also suggests a non-negligible degree of noisiness and/or measurement error. Nevertheless, the observed degree of stability appears sufficiently strong to suggest that workers with different degrees of prosociality may self-select into different sectors or to make it worthwhile for employers to screen potential employees based on certain social preference characteristics.

References

Brock, J. M., A. Lange, and E. Y. Ozbay. "Dictating the Risk: Experimental Evidence on Giving in Risky Environments." *American Economic Review* 103, no. 1 (2013): 415-37.

Bruhin, A., E. Fehr, and D. Schunk. "The Many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences." *Journal of the European Economic Association* 17, no. 4 (2019): 1025-69.

Cappelen, A. W., J. Konow, E. O. Sorensen, and B. Tungodden. "Just Luck: An Experimental Study of Risk-Taking and Fairness." *American Economic Review* 103, no. 4 (2013): 1398-413.

Cappelen, A. W., J. Mollerstrom, B. A. Reme, and B. Tungodden. "A Meritocratic Origin of Egalitarian Behaviour." *Economic Journal* 132, no. 646 (2022): 2101-17.

Cappelen, A. W., K. Nygaard, E. O. Sorensen, and B. Tungodden. "Social Preferences in the Lab: A Comparison of Students and a Representative Population." *Scandinavian Journal of Economics* 117, no. 4 (2015): 1306-26.

Cettolin, E., A. Riedl, and G. Tran. "Giving in the Face of Risk." *Journal of Risk and Uncertainty* 55, no. 2-3 (2017): 95-118.

Chuang, Y. T., and L. Schechter. "Stability of Experimental and Survey Measures of Risk, Time, and Social Preferences: A Review and Some New Results." *Journal of Development Economics* 117 (2015): 151-70.

Fisman, R., P. Jakiela, and S. Kariv. "How Did Distributional Preferences Change During the Great Recession?" *Journal of Public Economics* 128 (2015): 84-95.

Fisman, R., P. Jakiela, S. Kariv, and D. Markovits. "The Distributional Preferences of an Elite." *Science* 349, no. 6254 (2015).

Fisman, R., S. Kariv, and D. Markovits. "Individual Preferences for Giving." *American Economic Review* 97, no. 5 (2007): 1858-76.

Freundt, J., and A. Lange. "On the Determinants of Giving under Risk." *Journal of Economic Behavior & Organization* 142 (2017): 24-31.

Kerschbamer, R., and D. Muller. "Social Preferences and Political Attitudes: An Online Experiment on a Large Heterogeneous Sample." *Journal of Public Economics* 182 (2020).

Krawczyk, M., and F. Le Lec. "'Give Me a Chance!' An Experiment in Social Decision under Risk." *Experimental Economics* 13, no. 4 (2010): 500-11.

———. "How to Elicit Distributional Preferences: A Stress -Test of the Equality Equivalence Test." *Journal of Economic Behavior & Organization* 182 (2021): 13-28.

Li, J., L.P. Casalino, R. Fisman, S. Kariv, and D. Markovits. "Experimental Evidence of Physician Social Preferences." *PNAS* 119, no. 28 (2022).

Neilson, W. S. "Axiomatic Reference-Dependence in Behavior toward Others and toward Risk." *Economic Theory* 28, no. 3 (2006): 681-92.

Rohde, K. I. M. "A Preference Foundation for Fehr and Schmidt's Model of Inequity Aversion." *Social Choice and Welfare* 34, no. 4 (2010): 537-47.

Saito, K. "Social Preferences under Risk: Equality of Opportunity Versus Equality of Outcome." *American Economic Review* 103, no. 7 (2013): 3084-101.

Snowberg, E., and L. Yariv. "Testing the Waters: Behavior across Participant Pools." *American Economic Review* 111, no. 2 (2021): 687-719.