The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

CESifo Working Papers
www.cesifo.org/wp

Robust Mechanism Design and Social Preferences

Felix Bierbrauer
Axel Ockenfels
Andreas Pollak
Désirée Rückert

CESIFO WORKING PAPER NO. 4713
CATEGORY 12: EMPIRICAL AND THEORETICAL METHODS
MARCH 2014

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 4713*

Robust Mechanism Design and Social Preferences

Abstract

We study a classic mechanism design problem: How to organize trade between two privately informed parties. We characterize an optimal mechanism under selfish preferences and present experimental evidence that, under such a mechanism, a non-negligible fraction of individuals deviates from the intended behavior. We show that this can be explained by models of social preferences and introduce the notion of a social-preference-robust mechanism. We characterize an optimal mechanism in this class and present experimental evidence that it successfully controls behavior. We finally show that this mechanism is more profitable only if deviations from selfish behavior are sufficiently frequent.

JEL-Code: C920, D030, D820.

Keywords: robust mechanism design, social preferences, bilateral trade.

Felix Bierbrauer
University of Cologne
Cologne / Germany
bierbrauer@wiso.uni-koeln.de

Axel Ockenfels
University of Cologne
Cologne / Germany
ockenfels@uni-koeln.de

Andreas Pollak
University of Cologne
Cologne / Germany
pollak@wiso.uni-koeln.de

Désirée Rückert
University of Cologne
Cologne / Germany
rueckert@wiso.uni-koeln.de

February 19, 2014
We benefited from comments by Carlos Alós-Ferrer, Dirk Engelmann, Alia Gizatulina, Jacob Goeree, Hans-Peter Grüner, Johannes Münster, Nick Netzer and Bettina Rockenbach. Financial support of the German Science Foundation through the Leibniz-program and through the research unit "Design & Behavior" (FOR 1371) is gratefully acknowledged.

1 Introduction

Inspired by Wilson (1987), Bergemann and Morris (2005) have provided a formalization of mechanisms that are robust in the sense that they do not rely on a common prior distribution of material payoffs. We add another dimension in which we seek robustness. A mechanism that works well under selfish preferences might fail under social preferences. Some agents might be selfish and others might be motivated by social concerns, differing with respect to the nature and intensity of their social preferences. We want a mechanism to work for a large set of selfish and social preferences, including altruism, inequity-aversion, and intentionality. So, we introduce the notion of social-preference-robust mechanism: a mechanism must not depend on specific assumptions about the nature of preferences. The following quote of Wilson (1987), which can also be found in Bergemann and Morris (2005), suggests that this is a "natural" next step: *"Game theory has a great advantage in explicitly analyzing the consequences of trading rules that presumably are really common knowledge; it is deficient to the extent it assumes other features to be common knowledge, such as one player's probability assessment about another's* ***preferences*** *or* ***information*** *(Emphasis added). I foresee the progress of game theory as depending on successive reductions in the base of common knowledge required to conduct useful analyses of practical problems. Only by repeated weakening of common knowledge assumptions will the theory approximate reality."*

While Bergemann and Morris (2005) have focused on common knowledge assumptions regarding the information structure, we seek robustness with respect to common knowledge assumptions on the content of preferences. To this end, we study one of the classic applications of mechanism design theory, the bilateral-trade problem due to Myerson and Satterthwaite (1983). We argue that solutions to this problem which are derived under the assumption of selfish preferences are not robust to the possibility that individuals are motivated by social preferences. We then introduce the notion of a social-preference-robust mechanism, and derive a mechanism that is optimal in this class. Finally, we use a laboratory experiment to compare the optimal mechanism under selfish preferences and the optimal social-preference-robust mechanism.

We focus on the the bilateral-trade problem because it is a simple, and stylized setup that facilitates a clear exposition. Moreover, it admits interpretations that are of interest in public economics, environmental economics, or contract theory. The basics are as follows: A buyer either has a high or low valuation of a good produced by a seller. The seller either has a high or a low cost of producing the good. An economic outcome specifies, for each possible combination of the buyer's valuation and the seller's cost, the quantity to be exchanged, the price paid by the buyer and the revenue received by the seller. Both the buyer and the seller have private information. Thus, an allocation mechanism has to ensure that the buyer does not understate his valuation so as to get a desired quantity at a lower price. Analogously, the seller has to be incentivized so that she does not exaggerate her cost in order to receive a larger compensation. This environment can be reinterpreted as a problem of voluntary public-goods provision in which one party benefits from larger provision levels, relative to some status quo outcome, and the other party is harmed. By how much the first party benefits and the second party loses is private information. The allocation problem then is to determine the public-goods provision level and how the provision costs should be divided between the two parties. It can also be reinterpreted

as a problem to control externalities. One party can invest so as to avoid emissions which harm the other party. The cost of the investment to one party and the benefit of reduced emissions to the other party are private information. In a principal-agent-framework, we may think of one party as benefiting from effort that is exerted by the other party. The size of the benefit and the disutility of effort are, respectively, private information of the principal and the agent.

Our analysis proceeds as follows: We first characterize an optimal direct mechanism for the bilateral trade problem under the standard assumption of selfish preferences, i.e., both the buyer and the seller are assumed to maximize their own payoff, respectively, and this is common knowledge. We solve for the mechanism that maximizes the seller's expected profits subject to incentive constraints, participation constraints, and a resource constraint. We work with ex post incentive and participation constraints, i.e. we insist that after the outcome of the mechanism and the other party's private information have become known, no party regrets to have participated and to have revealed its own information.

Our reason for imposing these constraints is twofold: First, as has been shown by Bergemann and Morris (2005), they imply that a mechanism is robust in the sense that its outcome does not depend on the individual's probabilistic beliefs about the other party's private information. Second, we use the arguments in Bergemann and Morris (2005) for our experimental testing strategy. In their characterization of robust mechanisms *complete information environments* play a key role. In such an environment, the buyer knows the seller's cost and the seller knows the buyer's valuation, and, moreover, this is commonly known among them. The mechanism designer, however, lacks this information and therefore still has to provide incentives for a revelation of privately held information. Bergemann and Morris provide conditions so that the requirement of robustness is equivalent to the requirement that a mechanism generates the intended outcome in every complete information environment, which in turn is equivalent to the requirement that incentive and participation constraints hold in an ex post sense.

In our experimental approach, we investigate the performance of an optimally designed mechanism for selfish agents in all complete information environments. This approach is useful, because it allows us to isolate the role of social preferences in a highly controlled setting, which eliminates complications that relate to decision-making under uncertainty. For instance, it is well-known that, even in one-person decision tasks, people often do not maximize expected utility, and that moreover, in social contexts, social and risk preferences may interact in non-trivial ways (see, e.g., Camerer (2003), Bolton and Ockenfels (2010), and the references therein). The complete information environments in our study avoid such complicating factors.

The robust mechanism which maximizes the seller's expected profits under selfish preferences has the following properties: (i) The trading surplus is allocated in an asymmetric way, i.e. the seller gets a larger fraction than the buyer, (ii) whenever the buyer's valuation is low, his participation constraint binds, so that he does not realize any gains from trade, (iii) whenever the buyer's valuation is high, his incentive constraint binds, so that he is indifferent between revealing his valuation and understating it. Experimentally, we find that under this mechanism, a non-negligible fraction of high valuation buyers understates their valuation. In all other situations, deviations - if they occur at all - are significantly less frequent.

We argue that this pattern of deviation from truth-telling is consistent with models of social preferences such as Fehr and Schmidt (1999), and Falk and Fischbacher (2006), among others. The reason is that a buyer with a high valuation can understate his valuation at a very small personal cost since the relevant incentive constraint binds. The benefit of this strategy is that this reduces the seller's payoff and therefore brings the seller's payoff closer to his own, thereby reducing inequality. In fact, as we will demonstrate later, many social preference models would predict this behavior.

We then introduce a class of direct mechanisms that "work" if the possibility of social preferences is acknowledged. Specifically, we introduce the notion of a direct mechanism that is externality-free. Under such a mechanism, the buyer's equilibrium payoff does not depend on the seller's type and vice versa; i.e. if, say, the buyer reveals his valuation, his payoff no longer depends on whether the seller communicates a high or a low cost to the mechanism designer. Hence, the seller cannot influence the buyer's payoff.

Almost all widely-used models of social preferences satisfy a property of selfishness in the absence of externalities, i.e. if a player considers a choice between two actions a and b, and moreover, if everybody else is unaffected by this choice, then the player will choose a over b if her own payoff under a is higher than her own payoff under b. Now, suppose that a direct mechanism is ex post incentive-compatible and externality-free. Then truth-telling will be an equilibrium for any social preference model in which individuals are selfish in the absence of externalities.

We impose externality-freeness as an additional constraint on our problem of robust mechanism design, i.e. we have to design the mechanism so that it has the following property: Suppose that the traded quantity goes up because we move from a state of the world in which the seller's cost is high to a state in which the seller's cost is low. Then, there has to be an accompanying change in the price the buyer has to pay. This change needs to be calibrated in such a way that the buyer's trading surplus remains unaffected.[1] We then characterize the optimal robust and externality-free mechanism and investigate its performance in an experiment. We find that there are no longer deviations from truth-telling. We interpret this finding as providing evidence for the relevance of social preferences in mechanism design: If there are externalities, a significant fraction of individuals deviates from truth-telling. If those externalities are shut down, individuals behave truthfully.

Externality-freeness is an additional constraint. While it makes sure that individuals behave in a predictable way, it reduces expected profits relative to the theoretical benchmark of a model with selfish preferences. This raises the following question: Does the seller make more money if she uses an externality-free mechanism? We answer this question both theoretically and empirically: The externality-free mechanism makes more money if the number of individuals whose behavior is motivated by social preferences exceeds a threshold. In our experimental

[1] In settings different from the bilateral trade problem, externality-freeness may arise naturally. E.g. price-taking behavior in markets with a large number of participants gives rise to externality-freeness. If a single individual changes her demand, this leaves prices unaffected and so remain the options available to all other agents. Market behavior is therefore unaffected by social preferences, see Dufwenberg et al. (2011). Bierbrauer (2011) studies a problem of redistributive income taxation and provides conditions under which the optimal solution is externality-free, i.e. such that the taxes paid by any one individual depend only on the own income, and not on the income earned by other individuals.

context, however, this number was below the threshold, so that the "conventional" mechanism made more money than the externality-free mechanism.

The next section discusses related literature. Section 3 describes the economic environment. Section 4 contains a detailed description of the mechanism design problems that we study. In addition, we elaborate on why models of social preferences are consistent with the observation that individuals deviate from truth-telling under a mechanism that would be optimal if all individuals were selfish, and with the observation that they do not deviate under a mechanism that is externality-free. Section 5 describes our experimental findings. In Section 6, we clarify the conditions under which an optimal externality-free mechanism outperforms an optimal mechanism for selfish agents and relate them to our experimental data. The last section contains concluding remarks.

2 Related literature

Our work is related to different strands of the literature. For one, we draw on the model of Myerson and Satterthwaite (1983) which establishes an impossibility result for efficient trade in a setting with two privately informed parties.[2] We embed this problem into a model of robust mechanism design, see Bergemann and Morris (2005). Other contributions to the literature on robust mechanism design include Ledyard (1978), Gershkov et al. (2013) and Börgers (2013).

There is a large experimental economics literature testing mechanisms. Most laboratory studies deal with mechanisms to overcome free-riding in public goods environments (Chen (2008)), provides a survey), auction design (e.g., Ariely et al. (2005), Kagel et al. (2010)), and the effectiveness of various matching markets (e.g., Kagel and Roth (2000), Chen and Sönmez (2006)). Roth (2012) provides a survey. Some studies take into account social preferences when engineering mechanisms. For instance, it has been shown that feedback about others' behavior or outcomes, which would be irrelevant if agents were selfish, can strongly affect social comparison processes and reciprocal interaction, and thus the effectiveness of mechanisms to promote efficiency and resolve conflicts (e.g., Chen et al. (2010), Bolton et al. (2013), Ockenfels et al. (2013), Bolton and Ockenfels (2012) provide a survey). Social preferences are also important in bilateral bargaining with complete information, most notably in ultimatum bargaining (Güth et al. (1982); Güth and Kocher (2013) provide a survey). In fact, this literature has been a starting point for various social preference models that we are considering in this paper - yet the observed patterns of behavior have generally not been related to the mechanism design literature. This is different with laboratory studies of bilateral trade with incomplete information, such as Radner and Schotter (1989), Valley et al. (2002) and Kittsteiner et al. (2012). One major finding in this literature is, for instance, that cheap talk communication among bargainers can significantly improve efficiency. These findings are generally not related to social preference models, though.

Our study contributes to the literature by linking prominent models of social preferences with the mechanism design literature mentioned above, in the important context of bilateral trade. Bierbrauer and Netzer (2012) explore the implications of a specific model of social preferences, namely the one by Rabin (1993), for a Bayesian mechanism design problem - as opposed

[2] Related impossibility results hold for problems of public-goods provision, see Güth and Hellwig (1986) and Mailath and Postlewaite (1990).

to a problem of robust mechanism design. They show that, to any mechanism that is incentive compatible, one can construct an "essentially" equivalent version which is externality-free and therefore should generate the intended behavior even if individuals have social preferences. Bartling and Netzer (2013) use this observation to construct an externality-free version of the second-price auction. They show experimentally that there is significant overbidding in a standard second-price auction. Overbidding disappears with the externality-free version. The work of Bartling and Netzer is related to this paper in that it also makes use of externality-freeness. There is, however, an important difference. Since we work with ex post – as opposed to Bayesian – incentive and participation constraints, the equivalence result in Bierbrauer and Netzer (2012) no longer holds, i.e. externality-freeness becomes a substantive constraint. A contribution of this paper is the characterization of a mechanism that is optimal in the set of those which are externality-free and ex post incentive-compatible.

3 The economic environment

There are two agents, referred to as the buyer and the seller. An economic outcome is a triple (q, p_s, p_b), where $q \in \mathbb{R}_+$ is the quantity that is traded, $p_b \in \mathbb{R}$ is a payment made by the buyer, and $p_s \in \mathbb{R}$ is a payment received by the seller. Monetary payoffs are $\pi_b = \theta_b q - p_b$, for the buyer and $\pi_s = -\theta_s k(q) + p_s$, for the seller where k is an increasing and convex cost function. The buyer's valuation θ_b either takes a high or a low value, $\theta_b \in \Theta_b = \{\underline{\theta}_b, \bar{\theta}_b\}$. Similarly, the seller's cost parameter θ_s can take a high or a low value so that $\theta_s \in \Theta_s = \{\underline{\theta}_s, \bar{\theta}_s\}$. A pair $(\theta_b, \theta_s) \in \Theta_b \times \Theta_s$ is referred to as a state of the economy. A social choice function or direct mechanism $f : \Theta_b \times \Theta_s \to \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}$ specifies an economic outcome for each state of the economy. Occasionally, we write $f = (q^f, p_b^f, p_s^f)$ to distinguish the different components of f.

We denote by

$$\pi_b(\theta_b, f(\theta_b', \theta_s')) := \theta_b q^f(\theta_b', \theta_s') - p_b^f(\theta_b', \theta_s')$$

the payoff that is realized by a buyer with type θ_b if he announces a type θ_b' and the seller announces a type θ_s' under direct mechanism f. The expression $\pi_s(\theta_s, f(\theta_b', \theta_s'))$ is defined analogously.

We assume that the buyer has private information on whether his valuation θ_b is high or low. Analogously, the seller privately observes whether θ_s takes a high or a low value. Hence, a direct mechanism induces a game of incomplete information. Our analysis in the following focuses on a very specific and artificial class of incomplete information environments, namely the ones in which the types are commonly known among the players but unknown to the mechanism designer. In total there are four such complete information environments, one for each state of the economy.[3] It has been shown by Bergemann and Morris (2005) that the implementability of a social choice function in all such complete information environments is not only necessary but also sufficient for the robust implementability of a social choice function, i.e. for its implementability

[3] "Complete information" refers to a situation in which the players' monetary payoffs are commonly known. Information may still be incomplete in other dimensions, e.g. regarding the weight of fairness considerations in the other player's utility function.

in all conceivable incomplete information environments. Thus, our focus on complete information environments is not only useful to cleanly isolate the effect of social preferences from uncontrolled behavior under risk, but also justified by the robustness criterion.

Suppose that individuals are only interested in their own payoff. Then truth-telling is an equilibrium in all complete information environments if and only if the following ex-post incentive compatibility constraints are satisfied: For all $(\theta_b, \theta_s) \in \Theta_b \times \Theta_s$,

$$\pi_b(\theta_b, f(\theta_b, \theta_s)) \geq \pi_b(\theta_b, f(\theta_b', \theta_s)) \quad \text{for all} \quad \theta_b' \in \Theta_b \ , \tag{1}$$

and

$$\pi_s(\theta_s, f(\theta_b, \theta_s)) \geq \pi_s(\theta_s, f(\theta_b, \theta_s')) \quad \text{for all} \quad \theta_s' \in \Theta_s \ . \tag{2}$$

Moreover, individuals prefer to play the mechanism over a status quo outcome with no trade if and only if the following ex-post participation constraints are satisfied: For all $(\theta_b, \theta_s) \in \Theta_b \times \Theta_s$,

$$\pi_b(\theta_b, f(\theta_b, \theta_s)) \geq \bar{\pi}_b \quad \text{and} \quad \pi_s(\theta_s, f(\theta_b, \theta_s)) \geq \bar{\pi}_s \ , \tag{3}$$

where $\bar{\pi}_b$ and $\bar{\pi}_s$ are, respectively, the buyer's and the seller's payoffs in the absence of trade.

Throughout, we limit attention to direct mechanisms and to truth-telling equilibria. For models with selfish individuals, or more generally, for models with outcome-based preferences – which possibly include a concern for an equitable distribution of payoffs – this is without loss of generality by the revelation principle. For models with intention-based social preferences, such as Rabin (1993) or Dufwenberg and Kirchsteiger (2004), the revelation principle does not generally hold, see Bierbrauer and Netzer (2012) for a proof. Still, it is a sufficient condition for the implementability of a social choice function that it can be implemented as the truth-telling equilibrium of a direct mechanism. We focus on this sufficient condition, and note that it is also necessary if preferences are outcome-based.

Another property of interest to us is the externality-freeness of a social choice function f. This property holds if, for all $\theta_b \in \Theta_b$,

$$\pi_b(\theta_b, f(\theta_b, \underline{\theta}_s)) = \pi_b(\theta_b, f(\theta_b, \bar{\theta}_s)), \tag{4}$$

and if, for all $\theta_s \in \Theta_s$,

$$\pi_s(\theta_s, f(\underline{\theta}_b, \theta_s)) = \pi_s(\theta_s, f(\bar{\theta}_b, \theta_s)). \tag{5}$$

If these properties hold, then the buyer, say, cannot influence the seller's payoff, provided that the latter tells the truth. I.e. the buyer's report does not come with an externality on the seller. As we will argue later in more detail, many models of social preferences give rise to the prediction that externality-freeness in conjunction with ex post incentive compatibility is a sufficient condition for the implementability of a social choice function.

4 Mechanism design with and without social preferences

This section contains theoretical results which relate mechanism design theory to models of social preferences. We begin with the benchmark of optimal mechanism design under the assumption that individuals are purely selfish. We then show that many models of social preferences give rise to the prediction that such mechanisms will not generate truthful behavior. However, while there is only one way to maximize expected payoffs, there are many ways to behave socially. In fact, one of the most robust insights from behavioral economics and psychology is the large variance of social behaviors across individuals (e.g., Camerer (2003)). As a result, there is now a plethora of social preference models, and almost all models permit individual heterogeneity by allowing different parameter values for different individuals (e.g., Cooper and Kagel (2009)). This poses a problem for mechanism design, because optimal mechanisms depend on the agents' preferences. Our approach to deal with this problem is neither to just select one of those models, nor are we even attempting to identify the best model. We will also not assume that idiosyncratic social preferences are commonly known. All these assumptions about preferences would violate the spirit of robust mechanism design and the Wilson doctrine. This is why we restrict our attention to a property of social preferences which is shared by almost all, widely-used social preference models and which is independent of the exact parameter values: individuals are selfish if there is no possibility to affect the payoffs of others. As we will show, this general property of social behavior is already sufficient to construct "externality-free" mechanisms which generate truthful behavior, regardless of what is known about the specific type and parameters of the agents' social preferences.

Our approach comes at a cost. While we will be able to better control behavior than when we assume selfish preferences, not knowing the exact details of preferences will impair the profitability of the mechanism. As we show in Section 6, the optimal robust and externality-free mechanism outperforms the optimal mechanism for selfish agents only if the probability of behavior that is motivated by social preferences is sufficiently high.

4.1 Optimal mechanism design under selfish preferences

We consider a problem of optimal robust mechanism design for selfish agents: There is an ex ante stage. At this stage, a mechanism designer wishes to come up with a mechanism for trade. The designer acts in the interest of one of the parties, here the seller. The designer does not know what information the buyer and the seller have about each other at the moment where trade takes place. Hence, he seeks robustness with respect to the information structure and employs ex post incentive and participation constraints. The designer assumes that individuals are selfish so that these constraints are sufficient to ensure that individuals are willing to play the corresponding direct mechanism and to reveal their types. Finally, he requires budget balance only in an average sense. (Possibly, the mechanism is executed frequently, so that the designer expects to break even if budget balance holds on average.)

Formally, we assume that a social choice function f is chosen with the objective to maximize expected seller profits, $\sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s)\pi_s(\theta, f(\theta_b, \theta_s))$, where g is a probability mass function that gives the mechanism designer's subjective beliefs on the likelihood of the different states of

the economy. The incentive and participation constraints in (1), (2) and (3) have to be respected. In addition, the following resource constraint must hold

$$\sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) p_b^f(\theta_b, \theta_s) \geq \sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) p_s^f(\theta_b, \theta_s) \ . \tag{6}$$

To solve this *full problem*, we first study a *relaxed problem* which leaves out the incentive and participation constraints for the seller. Proposition 1 characterizes the solution to the relaxed problem.

Proposition 1. *A social choice function f solves the relaxed problem of robust mechanism design if and only if it has the following properties:*

(a) For any one $\theta_s \in \Theta_s$, the participation constraint of a low type-buyer is binding:

$$\pi_b(\underline{\theta}_b, f(\underline{\theta}_b, \theta_s)) = \bar{\pi}_b \ .$$

(b) For any one $\theta_s \in \Theta_s$, the incentive constraint of a high type-buyer is binding:

$$\pi_b(\overline{\theta}_b, f(\overline{\theta}_b, \theta_s)) = \pi_b(\overline{\theta}_b, f(\underline{\theta}_b, \theta_s)) \ .$$

(c) The trading rule is such that, for any one $\theta_s \in \Theta_s$, there is a downward distortion at the bottom

$$q^f(\underline{\theta}_b, \theta_s) \quad \in \quad argmax_q \left(\underline{\theta}_b - \frac{g(\overline{\theta}_b, \theta_s)}{g(\underline{\theta}_b, \theta_s)} (\overline{\theta}_b - \underline{\theta}_b) \right) q - \theta_s k(q) \ ,$$

and no distortion at the top

$$q^f(\overline{\theta}_b, \theta_s) \quad \in \quad argmax_q \, \overline{\theta}_b q - \theta_s k(q) \ .$$

(d) The payment rule for the buyer is such that, for any one θ_s,

$$p_b^f(\underline{\theta}_b, \theta_s) = \underline{\theta}_b q^f(\underline{\theta}_b, \theta_s) \ ,$$

and

$$p_b^f(\overline{\theta}_b, \theta_s) = \overline{\theta}_b q^f(\overline{\theta}_b, \theta_s) - (\overline{\theta}_b - \underline{\theta}_b) q^f(\underline{\theta}_b, \theta_s) \ .$$

(e) The revenue for the seller is such that

$$\sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) p_b^f(\theta_b, \theta_s) = \sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) p_s^f(\theta_b, \theta_s) \ .$$

We omit a formal proof of Proposition 1, but provide a sketch of the main argument: Since we leave out the seller's incentive constraint, we can treat the seller's cost parameter as a known quantity. Hence, we think of the relaxed problem as consisting of two separate profit-maximization problems, one for a high-cost seller and one for a low-cost seller, which are linked only via the resource constraint. In each of these problems, however, the buyer's incentive and

participation constraints remain relevant. Hence, we have two profit-maximization problems. The formal structure of any one of those problems is the same as the structure of a non-linear pricing problem with two buyer types. This problem is well-known so that standard arguments can be used to derive properties (a)-(e) above.[4]

The solution to the relaxed problem leaves degrees of freedom for the specification of the payments to the seller. Consequently, any specification of the seller's revenues so that the expected revenue is equal to the buyer's expected payment is part of a solution to the relaxed problem. If there is one such specification that satisfies the seller's ex post incentive and participation constraints, then this solution to the relaxed problem is also a solution to the full problem.

Example 1: An optimal social choice function. Suppose that $\underline{\theta}_b = 1$, $\overline{\theta}_b = 1.3$, $\underline{\theta}_s = 0.2$, and $\overline{\theta}_s = 0.65$. Also assume that the seller has a quadratic cost function $k(q) = \frac{1}{2}q^2$. Finally, assume that the reservation utility levels of both the buyer and the seller are given by $\bar{\pi}_b = \bar{\pi}_s = 2.68$. For these parameters, the traded quantities of the optimal social choice function f are given by

$$q^f(\underline{\theta}_b, \underline{\theta}_s) = 3.5,\ q^f(\underline{\theta}_b, \bar{\theta}_s) = 1.08,\ q^f(\bar{\theta}_b, \underline{\theta}_s) = 6.5 \quad \text{and} \quad q^f(\bar{\theta}_b, \bar{\theta}_s) = 2\ .$$

The buyer's payments are

$$p_b^f(\underline{\theta}_b, \underline{\theta}_s) = 3.5,\ p_b^f(\underline{\theta}_b, \bar{\theta}_s) = 1.08,\ p_b^f(\bar{\theta}_b, \underline{\theta}_s) = 7.4 \quad \text{and} \quad p_b^f(\bar{\theta}_b, \bar{\theta}_s) = 2.28\ .$$

Finally, the seller's revenues are

$$p_s^f(\underline{\theta}_b, \underline{\theta}_s) = 3.5,\ p_s^f(\underline{\theta}_b, \bar{\theta}_s) = 1.08,\ p_s^f(\bar{\theta}_b, \underline{\theta}_s) = 7.4 \quad \text{and} \quad p_s^f(\bar{\theta}_b, \bar{\theta}_s) = 2.28\ .$$

By construction, f is ex post incentive compatible and satisfies the ex post participation constraints. However, it is not externality-free. These properties can be verified by looking at the games which are induced by this social choice function on the various complete information environments. For instance, the following matrix represents the normal form game that is induced by f in a complete information environment so that the buyer has a low valuation and the seller has a low cost.

Table 1: The game induced by f for $(\theta_b, \theta_s) = (\underline{\theta}_b, \underline{\theta}_s)$.

(π_b^f, π_s^f)	$\underline{\theta}_s$	$\overline{\theta}_s$
$\underline{\theta}_b$	$(2.68, 5.52)$	$(2.68, 3.88)$
$\overline{\theta}_b$	$(1.56, 6.65)$	$(2.33, 5.03)$

The first entry in each cell is the buyer's payoff, the second entry in the cell is the seller's payoff. If both individuals truthfully reveal their types, the payoffs in the upper left corner are realized. Note that under truth-telling both payoffs are weakly larger than the reservation utility of 2.68 so that the relevant ex post participation constraints are satisfied. Also note that the seller does not benefit from an exaggeration of her cost, if the buyer communicates his low valuation truthfully. Likewise, the buyer does not benefit

[4] A classical reference is Mussa and Rosen (1978), see Bolton and Dewatripont (2005) for a textbook treatment.

from am exaggeration of his willingness to pay, given that the seller communicates her low cost truthfully. Hence, the relevant ex post incentive constraints are satisfied. Finally, note that externality-freeness is violated: If the seller behaves truthfully, her payoff is higher if the buyer communicates a high willingness to pay.

For later reference, we also describe the normal form games that are induced in the remaining complete information environments.

Table 2: The game induced by f for $(\theta_b, \theta_s) = (\underline{\theta}_b, \overline{\theta}_s)$.

(π_b, π_s)	$\underline{\theta}_s$	$\overline{\theta}_s$
$\underline{\theta}_b$	$(2.68, 2.08)$	$(2.68, 3.56)$
$\overline{\theta}_b$	$(1.56, 5.23)$	$(2.33, 3.90)$

Table 3: The game induced by f for $(\theta_b, \theta_s) = (\overline{\theta}_b, \underline{\theta}_s)$.

(π_b, π_s)	$\underline{\theta}_s$	$\overline{\theta}_s$
$\underline{\theta}_b$	$(3.97, 5.52)$	$(3.06, 3.88)$
$\overline{\theta}_b$	$(3.99, 6.65)$	$(3.08, 5.03)$

Table 4: The game induce by f for $(\theta_b, \theta_s) = (\overline{\theta}_b, \overline{\theta}_s)$.

(π_b, π_s)	$\underline{\theta}_s$	$\overline{\theta}_s$
$\underline{\theta}_b$	$(3.97, 2.08)$	$(3.06, 3.56)$
$\overline{\theta}_b$	$(3.99, -5.23)$	$(3.08, 3.90)$

An inspection of Tables 1 and 4 reveals the following properties of f: (i) Under truth-telling the seller's payoff exceeds the buyer's payoff in all states of the economy, (ii) if the buyer's type is low (Tables 1 and 2), then his payoff under truth-telling is equal to his reservation utility level of 2.68, i.e. the participation constraint of a low type buyer binds, (iii) if the buyer's type is high (Tables 3 and 4), then the buyer's incentive constraint is binding in the sense that understating comes at a very small personal cost (the payoff drops from 3.99 to of 3.97).

4.2 An observation on models of social preferences

We now show that the social choice function in Proposition 1 is not robust, because it provokes deviations from truth-telling if individuals are motivated by social preferences. Consider the game induced by a direct mechanism f on some complete information environment. To formalize a possibility of social preferences, we assume that any one individual $i \in \{s, b\}$ has a utility function $U_i(\theta_i, r_i, r_i^b, r_i^{bb})$ which depends in a parametric way on the individual's true type θ_i and, in addition, on the following three arguments: the individual's own report r_i, the individual's (first-order) belief about the other player's report, r_i^b, and the individual's (second-order) belief about the other player's first-order belief, r_i^{bb}. Different models of social preferences make different assumptions about these utility functions. Second-order beliefs play a role in models with

intention-based social preferences such as Rabin (1993), Dufwenberg and Kirchsteiger (2004) or Falk and Fischbacher (2006). In these models, the utility function takes the following form

$$U_i(\theta_i, r_i, r_i^b, r_i^{bb}) = \pi_i(\theta_i, f(r_i, r_i^b)) + y_i\, \kappa_i(r_i, r_i^b, r_i^{bb})\, \kappa_j(r_i^b, r_i^{bb})\ . \tag{7}$$

The interpretation is that the players' interaction gives rise to sensations of kindness or unkindness, as captured by $y_i\, \kappa_i(r_i, r_i^b, r_i^{bb})\, \kappa_j(r_i^b, r_i^{bb})$. In this expression, $y_i \geq 0$ is an exogenous parameter, interpreted as the weight that agent i places on kindness considerations. The term $\kappa_i(r_i, r_i^b, r_i^{bb})$ is a measure of how kindly i intends to treat the other agent j. While the models of Rabin (1993), Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) differ in some respects, they all make the following assumption: Given r_i^b, and r_i^{bb} for any two reports r_i' and r_i'', $\pi_j(\theta_j, f(r_i', r_i^b)) \geq \pi_j(\theta_j, f(r_i'', r_i^b))$ implies that $\kappa_i(r_i', r_i^b, r_i^{bb}) \geq \kappa_i(r_i'', r_i^b, r_i^{bb})$, i.e. the kindness intended by i is larger if her report yields a larger payoff for j. Second-order beliefs are relevant here if player i expresses kindness by increasing j's payoff relative to the payoff that, according to the beliefs of i, j expects to be realizing. The latter payoff depends on the beliefs of i about the beliefs of j about i's behavior.

Whether or not i's utility is increasing in κ_i depends on i's belief about the kindness that is intended by player j and which is denoted by κ_j. If $\kappa_j > 0$, then i believes that j is kind and her utility increases, ceteris paribus, if j's payoff goes up. By contrast, if $\kappa_j < 0$, then i believes that j is unkind and her utility goes up if j is made worse off. Rabin (1993), Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) all assume that the function κ_j is such that, for given second-order beliefs r_i^{bb}, $\kappa_j(r_i^{b'}, r_i^{bb}) \geq \kappa_j(r_i^{b''}, r_i^{bb})$ whenever $\pi_i(\theta_i, f(r_i^{b'}, r_i^{bb})) \geq \pi_i(\theta_i, f(r_i^{b''}, r_i^{bb}))$. Second-order beliefs play a role here because, in order to assess the kindness that is intended by j, i has to form a belief about j's belief about i's report.

In models with outcome-based social preferences such as Fehr and Schmidt (1999), Bolton and Ockenfels (2000), or Charness and Rabin (2002) second order beliefs play no role, yet individuals are assumed to care about the distribution of payoffs among the players. For instance, with Fehr-Schmidt-preferences, the utility function of individual i reads as

$$\begin{aligned} U_i(\theta_i, r_i, r_i^b, r_i^{bb}) \quad = \quad & \pi_i(\theta_i, f(r_i, r_i^b)) \quad -\alpha_i \max\{\pi_j(\theta_j, f(r_i, r_i^b)) - \pi_i(\theta_i, f(r_i, r_i^b)), 0\} \\ & \qquad\qquad\quad -\beta_i \max\{\pi_i(\theta_i, f(r_i, r_i^b)) - \pi_j(\theta_j, f(r_i, r_i^b)), 0\}\ , \end{aligned} \tag{8}$$

where it is assumed that $\alpha_i \geq \beta_i$ and that $0 \leq \beta_i < 1$.

Many models of social preferences give rise to the prediction that a social choice function that would be optimal if individuals were selfish will trigger deviations from truth-telling. Specifically, for our bilateral trade problem, high valuation buyers will understate their valuation. Models of outcome-based and intention-based social preferences provide different explanations for this: The social choice function f is a profit-maximizing one and hence allocates the gains from trade in a way that favors the seller. With outcome-based social preferences, the buyer may wish to harm the seller so as to make their expected payoffs more equal. The reasoning for intention-based models would be different. Consider the game in Table 4. With intention-based social preferences as in Rabin (1993), the buyer would reason in the following way: My expected payoff would be higher if the seller deviated from truth-telling and communicated a low cost. Since the

seller does not make use of this opportunity to increase my payoff, he is unkind. I therefore wish to reduce his expected payoff. Whatever the source of the desire to reduce the seller's payoff, a high valuation buyer can reduce the seller's payoff by understating his valuation. Since the relevant incentive constraint binds, such an understatement is costless for the buyer, i.e. he does not have to sacrifice own payoff if he wishes to reduce the seller's payoff.

Observation 1 states this more formally for the case of Fehr-Schmidt-preferences. In Appendix A, we present analogous results for other models of social preferences.

Observation 1. *Consider a complete information types space for state* (θ_b, θ_s) *and suppose that* $\theta_b = \bar{\theta}_b$. *Suppose that* f *is such that*

$$\pi_s(\theta_s, f(\bar{\theta}_b, \theta_s)) > \pi_s(\theta_s, f(\underline{\theta}_b, \theta_s)) > \pi_b(\bar{\theta}_b, f(\underline{\theta}_b, \theta_s)) = \pi_b(\bar{\theta}_b, f(\bar{\theta}_b, \theta_s)) \tag{9}$$

Suppose that the seller behaves truthfully. Also suppose that the buyer has Fehr-Schmidt-preferences as in (8) with $\alpha_b \neq 0$. *Then the buyer's best response is to understate his valuation.*

The social choice function in Example 1 fulfills Condition (9). Consider Tables 3 and 4. The buyer's incentive constraint binds. Moreover, if the buyer understates his valuation this harms the seller. The harm is, however, limited in the sense that the seller's reduced payoff still exceeds the buyer's payoff. For such a situation the Fehr-Schmidt model of social preferences predicts that the buyer will deviate from truth-telling, *for any* pair of parameters (α_b, β_b) so that $\alpha_b \neq 0$. Put differently, truth-telling is a best response for the buyer only if $\alpha_b = 0$, i.e. only if the buyer is selfish.

4.3 Social-preference-robust mechanisms

The models of social preferences mentioned so far differ in many respects. They are, however, all consistent with the following assumption of *selfishness in the absence of externalities.*

Assumption 1. *Given* r_i^b *and* r_i^{bb}, *if* r_i' *and* r_i'' *are such that* $\pi_j(\theta_j, f(r_i', r_i^b)) = \pi_j(\theta_j, f(r_i'', r_i^b))$ *and* $\pi_i(\theta_i, f(r_i', r_i^b)) > \pi_i(\theta_i, f(r_i'', r_i^b))$, *then* $U_i(\theta_i, r_i', r_i^b, r_i^{bb}) \geq U_i(\theta_i, r_i'', r_i^b, r_i^{bb})$.

Assumption 1 holds provided that individuals prefer to choose strategies that increase their own payoff, whenever they can do so without affecting others. This does not preclude a willingness to sacrifice own payoff so as to either increase or reduce the payoff of others. It is a ceteris paribus assumption: In the set of strategies that have the same implications for player j, player i weakly prefers the ones that yield a higher payoff for herself. Assumption 1 has the following implication: In situations where players do not have the possibility to affect the payoffs of others, social preferences will be behaviorally irrelevant, and the players act as if they were selfish payoff maximizers.

The following observation illustrates that the utility function underlying the Fehr and Schmidt (1999)-model of social preferences satisfies Assumption 1 for all possible parametrization of the model. Appendix A.2 confirms this observation for other models of social preferences.[5]

[5] Assumption 1 is also satisfied in models of pure altruism, see Becker (1974). The only exception among prominent social preference models that we encountered is the model by Bolton and Ockenfels (2000), which is more general. While parameterized versions of their model are consistent with Assumption 1, their general model does not rule out the possibility of preferences which violate Assumption 1.

Observation 2. *Suppose the buyer and the seller have preferences as in (8) with parameters* (α_b, β_b) *and* (α_s, β_s)*, respectively. The utility functions* U_b *and* U_s *satisfy Assumption 1, for all* (α_b, β_b) *so that* $\alpha_b \geq \beta_b$ *and* $0 \leq \beta_b < 1$ *and for all* (α_s, β_s) *so that* $\alpha_s \geq \beta_s$ *and* $0 \leq \beta_s < 1$*.*

We now define a mechanism that is robust in the following sense: For any individual i, given correct first-and second-order beliefs, a truthful report maximizes U_i, for all utility functions satisfying Assumption 1.

Definition 1. *A direct mechanism for social choice function* f *is said to be a social-preference-robust if it satisfies the following property: On any complete information environment, given correct first and second-order beliefs, truth-telling by any player* $i \in \{b, s\}$ *is a best response to truth-telling by player* $j \neq i$*, for all utility functions* U_i *satisfying Assumption 1.*

Social-preference-robustness of a mechanism is an attractive property. It is robust against widely varying beliefs of the mechanism designer about what is the appropriate specification and intensity of social preferences across individuals. As long as preferences satisfy Assumption 1, we can be assured that individuals behave truthfully under such a mechanism.

Proposition 2 justifies our interest in externality-free mechanisms. If we add externality-freeness to the requirement of incentive compatibility, we arrive at a social-preference-robust mechanism.

Proposition 2. *Suppose that* f *is ex post incentive-compatible and externality-free. Then* f *is social-preference-robust.*

Proof. Consider a complete information environment for types (θ_i, θ_j). Suppose that player i beliefs that player j acts truthfully so that $r_i^b = \theta_j$ and that he beliefs that player j beliefs that he acts truthfully so that $r_i^{bb} = \theta_i$. By ex post-incentive compatibility, $\pi_i(\theta_i, f(r_i, r_i^b))$ is maximized by choosing $r_i = \theta_i$. By externality-freeness, $\pi_j(\theta_j, f(r_i', r_i^b)) = \pi_j(\theta_j, f(r_i'', r_i^b))$ for any pair $r_i', r_i'' \in \Theta_i$. Hence, by Assumption 1, $r_i = \theta_i$ solves $\max_{r_i \in \Theta_i} U_i(\theta_i, r_i, r_i^b, r_i^{bb})$. □

4.4 Optimal robust and externality-free mechanism design

We now add the requirement of externality-freeness to our mechanism design problem. To characterize the solution it is instructive to begin, again, with a relaxed problem in which only a subset of all constraints is taken into account. Specifically, the relevant constraints are: the resource constraint in (6), the participation constraints for a low valuation buyer,

$$\pi_b(\underline{\theta}_b, f'(\underline{\theta}_b, \theta_s)) \geq \bar{\pi}_b, \quad \text{for all} \quad \theta_s \in \Theta_s \ , \tag{10}$$

the incentive constraint for a high type buyer who faces a low cost seller,

$$\pi_b(\bar{\theta}_b, f'(\bar{\theta}_b, \underline{\theta}_s)) \geq \ \pi_b(\bar{\theta}_b, f'(\underline{\theta}_b, \underline{\theta}_s)) \ , \tag{11}$$

and, finally, the externality-freeness condition for a high valuation buyer

$$\pi_b(\bar{\theta}_b, f'(\bar{\theta}_b, \underline{\theta}_s)) = \pi_b(\bar{\theta}_b, f'(\bar{\theta}_b, \bar{\theta}_s)) \ . \tag{12}$$

We will later provide conditions under which the solution of the relaxed problem is also a solution to the full problem.

Proposition 3. *A social choice function f' solves the relaxed problem of robust and externality-free mechanism design if and only if it has the following properties:*

(a)' For any one $\theta_s \in \Theta_s$, the participation constraint of a low type-buyer is binding:

$$\pi_b(\underline{\theta}_b, f'(\underline{\theta}_b, \theta_s)) = \bar{\pi}_b \ .$$

(b)' For $\theta_s = \underline{\theta}_s$, the incentive constraint of a high type-buyer is binding.

(c)' The trading rule is such that there is a downward distortion only for state $(\underline{\theta}_b, \underline{\theta}_s)$

$$q^{f'}(\underline{\theta}_b, \underline{\theta}_s) \quad \in \quad argmax_q \left(\underline{\theta}_b - \frac{g^m(\overline{\theta}_b)}{g(\underline{\theta}_b, \underline{\theta}_s)} (\overline{\theta}_b - \underline{\theta}_b) \right) q - \theta_s k(q) \ ,$$

where $g^m(\overline{\theta}_b) := g(\overline{\theta}_b, \underline{\theta}_s) + g(\overline{\theta}_b, \overline{\theta}_s)$. Otherwise, there is no distortion.

(d)' The payment rule for the buyer is such that, for any one θ_s,

$$p_b^{f'}(\underline{\theta}_b, \theta_s) = \underline{\theta}_b q^{f'}(\underline{\theta}_b, \theta_s) \ .$$

In addition

$$p_b^{f'}(\overline{\theta}_b, \underline{\theta}_s) = \overline{\theta}_b q^{f'}(\overline{\theta}_b, \underline{\theta}_s) - (\overline{\theta}_b - \underline{\theta}_b) q^{f'}(\underline{\theta}_b, \underline{\theta}_s) \ ,$$

and

$$p_b^{f'}(\overline{\theta}_b, \overline{\theta}_s) = \overline{\theta}_b q^{f'}(\overline{\theta}_b, \overline{\theta}_s) - (\overline{\theta}_b - \underline{\theta}_b) q^{f'}(\underline{\theta}_b, \underline{\theta}_s) \ ,$$

(e)' The revenue for the seller is such that

$$\sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) p_b^{f'}(\theta_b, \theta_s) = \sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) p_s^{f'}(\theta_b, \theta_s) \ .$$

We provide a sketch of the proof: The first step is to show that all inequality constraints of the relaxed problem have to be binding. Otherwise, it would be possible to implement the given trading rule q^f with higher payments of the buyer. This establishes $(a)'$ and $(b)'$. Second, we solve explicitly for the payments of the buyer as a function of the trading rule $q^{f'}$ – this yields $(d)'$ – and substitute the resulting expressions into the objective function. This resulting unconstrained optimization problem has first order conditions which characterize the optimal trading rule, see the optimality conditions in $(c)'$.

After having obtained the solution to the relaxed problem, we need to make sure that it is also a solution to the full problem. For the buyer, it can be shown that the neglected participation, incentive and externality-freeness constraints are satisfied provided that the solution to the relaxed problem is such that the traded quantity increases in the buyer's valuation and decreases in the seller's cost. If there is a solution to the relaxed problem that satisfies the seller's incentive, participation and externality-freeness constraints, then this solution to the relaxed problem is

also a solution to the full problem. The social choice function f' in Example 2 below has all these properties.

The substantive difference between the optimal mechanism for selfish agents in Proposition 1 and the optimal externality-free mechanism in Proposition 3 is in the pattern of distortions. The mechanism in Proposition 1 has downward distortions whenever the buyer has a low valuation. The mechanism in Proposition 3 has a downward distortion in only one state, namely the state in which the buyer's valuation is low and the seller's cost is low. This distortion, however, is more severe than the distortion that arises for this state with the mechanism that would be optimal if all agents were selfish.

Example 2: An optimal externality-free social choice function. Suppose the parameters of the model are as in Example 1. The social choice function f' solves the problem of optimal robust and externality-free mechanism design formally defined in the previous paragraph: The traded quantities are given by

$$q^{f'}(\underline{\theta}_b, \underline{\theta}_s) = 2,\ q^{f'}(\underline{\theta}_b, \bar{\theta}_s) = 1.54,\ q^{f'}(\bar{\theta}_b, \underline{\theta}_s) = 6.5 \quad \text{and} \quad q^{f'}(\bar{\theta}_b, \bar{\theta}_s) = 2\ .$$

The buyer's payments are

$$p_b^{f'}(\underline{\theta}_b, \underline{\theta}_s) = 2,\ p_b^{f'}(\underline{\theta}_b, \bar{\theta}_s) = 1.54,\ p_b^{f'}(\bar{\theta}_b, \underline{\theta}_s) = 7.85 \quad \text{and} \quad p_b^{f'}(\bar{\theta}_b, \bar{\theta}_s) = 2\ .$$

Finally, the seller's revenues are

$$p_s^{f'}(\underline{\theta}_b, \underline{\theta}_s) = 2.52,\ p_s^{f'}(\underline{\theta}_b, \bar{\theta}_s) = 1.99,\ p_s^{f'}(\bar{\theta}_b, \underline{\theta}_s) = 6.35 \quad \text{and} \quad p_s^{f'}(\bar{\theta}_b, \bar{\theta}_s) = 2.52\ .$$

To illustrate the property of externality-freeness, we consider, once more, the various complete information games which are associated with this social choice function.

Table 1': The game induced by f' for $(\theta_b, \theta_s) = (\underline{\theta}_b, \underline{\theta}_s)$.

(π_b, π_s)	$\underline{\theta}_s$	$\overline{\theta}_s$
$\underline{\theta}_b$	$(2.68, 5.33)$	$(2.68, 4.86)$
$\overline{\theta}_b$	$(0.97, 5.33)$	$(2.66, 5.31)$

Along the same lines as for Table 1, one may verify that the relevant ex post incentive and participation constraints are satisfied. In addition, externality-freeness holds: If the seller communicates her low cost truthfully, then she gets a payoff of 5.33 irrespectively of whether the buyer communicates a high or a low valuation. Also, if the buyer reveals his low valuation, he gets 2.68 irrespectively of whether the seller communicates a high or a low cost.

Again, we also describe the normal form games that are induced by f' in the remaining complete information environments.

Table 2': The game induced by f' for $(\theta_b, \theta_s) = (\underline{\theta}_b, \overline{\theta}_s)$.

(π_b, π_s)	$\underline{\theta}_s$	$\overline{\theta}_s$
$\underline{\theta}_b$	$(2.68, 4.19)$	$(2.68, 4.21)$
$\overline{\theta}_b$	$(0.97, -6.57)$	$(2.66, 4.21)$

Table 3': The game induced by f' for $(\theta_b, \theta_s) = (\overline{\theta}_b, \underline{\theta}_s)$.

(π_b, π_s)	$\underline{\theta}_s$	$\overline{\theta}_s$
$\underline{\theta}_b$	$(3.41, 5.33)$	$(3.24, 4.86)$
$\overline{\theta}_b$	$(3.43, 5.33)$	$(3.43, 5.31)$

Table 4': The game induced by f' for $(\theta_b, \theta_s) = (\overline{\theta}_b, \overline{\theta}_s)$.

(π_b, π_s)	$\underline{\theta}_s$	$\overline{\theta}_s$
$\underline{\theta}_b$	$(3.41, 4.19)$	$(3.24, 4.21)$
$\overline{\theta}_b$	$(3.43, -6.57)$	$(3.43, 4.21)$

On top of externality-freeness, the social choice function f' in Tables 1' to 4' has the following properties: (i) The seller's payoff under truth-telling is higher than the buyer's payoff under truth-telling, (ii) a low type buyer realizes his reservation utility (see Tables 1' and 2'), and (iii) the buyer's incentive constraint binds if the seller's cost is low, but not if the seller's cost is high (see Tables 3' and 4').

5 A laboratory experiment

We conducted a laboratory-experiment with two different treatments, one based on the optimal mechanism f under selfish preferences in Example 1 (T1), and one based on the optimal externality-free mechanism f' under social preferences in Example 2 (T2). Each treatment consisted of two phases: A learning phase and a decision phase, both based on the complete information games of the respective social choice functions.

The experiment was conducted in the *Cologne Laboratory for Economic Research* at the University of Cologne and had been programmed with *z-Tree* developed by Fischbacher (2007). Participants were recruited via e-mail from a subject pool with about 5,000 registered subjects by using the online recruitment system *ORSEE* developed by Greiner (2004). We held in total eight sessions, four for each treatment. Each session consisted of 32 participants, with the exception of one session, were two subjects did not show up. Registered subjects were students from all faculties. Of the 254 participating subjects 151 were female and 103 were male. The average age was 24.4 years. Each subject was allowed to participate in one session and in one treatment only (between subject design). Average payments to subjects, including the show-up fee, was 10.99 Euro for about 45-60 minutes in the laboratory.

At the beginning of each session, subjects were randomly assigned to computer-terminals. Prior to starting the experiment, all subjects received identical instructions which informed them about all rules and procedures of the experiment. The instructions were the same for all treatments and roles (treatment- and role-specific information was given on the computer-screen), and written in neutral terms. Specifically, player-roles were labeled *Participant A (B)* and strategies were labeled *Top (Left)* and *Bottom (Right)* respectively.[6]

[6]However, in the following we refer to the specific roles within the experiment as buyers and sellers. This is done to make this section consistent with previous ones. A translated version of the instructions can be found in the Appendix B.

Before the payoff-relevant decision phase started, subjects went through a learning phase, with no interaction among subjects and no decision-dependent payments. This was done to familiarize participants with the decision situation. Within the learning phase, subjects had to choose actions for the buyer and the seller in each information game and then to state the resulting payoffs for the corresponding strategy combination. Subjects had to give the right answer before proceeding to the decision phase. By using this procedure we assured, that all subjects were able to read the payoff tables correctly, without giving them defaults which might create anchoring or experimenter demand effects.

After all subjects completed the learning phase, the decision phase began by informing subjects about their role in their group. The matching into groups and roles was anonymous, random and held constant over the course of experiment. Within the decision phase subjects had to choose one strategy for each of the four complete information games of their specific treatment. Only after all subjects submitted their choices, feedback was given to each subject on all choices and resulting outcomes in their group. Finally, one of the games was randomly determined for payments in addition to the show-up fee.

Table 5: DATA

		Buyer		*Seller*	
	Game induced by	$\underline{\theta}_b$	$\overline{\theta}_b$	$\underline{\theta}_s$	$\overline{\theta}_s$
T1:	f for $(\underline{\theta}_b, \underline{\theta}_s)$	63	0	63	0
	f for $(\underline{\theta}_b, \overline{\theta}_s)$	63	0	0	63
	f for $(\overline{\theta}_b, \underline{\theta}_s)$	8	55	63	0
	f for $(\overline{\theta}_b, \overline{\theta}_s)$	10	53	1	62
T2:	f' for $(\underline{\theta}_b, \underline{\theta}_s)$	64	0	62	2
	f' for $(\underline{\theta}_b, \overline{\theta}_s)$	64	0	0	64
	f' for $(\overline{\theta}_b, \underline{\theta}_s)$	1	63	64	0
	f' for $(\overline{\theta}_b, \overline{\theta}_s)$	2	62	0	64

Results. Table 5 summarizes the decisions made in the experiment. Overall, behavior is well much in line with selfish preferences. 992 out of 1016 (97.6%) reports are truthful. There is just one notable exception, and this is exactly where social preference models would predict the deviation: a non-negligible share of high valuation buyers (14.3%) does not truthfully report their valuation. We can reject the null-hypothesis that deviations from truth-telling are distributed evenly across all eight games (chi-square test of goodness of fit, $p < 0.0001$). More specifically, Table 6 shows the p-value for comparisons of truthful reports across games. For buyers with high valuation the differences between treatments is significant, while in all other cases we do not find statistical evidence for differences in behavior.

Table 6: Statistical comparison between treatments

Games induced by	*Buyer*	*Seller*
$(\underline{\theta}_b, \underline{\theta}_s)$	—	$p = 0.496$
$(\underline{\theta}_b, \overline{\theta}_s)$	—	—
$(\overline{\theta}_b, \underline{\theta}_s)$	$p = 0.017$	—
$(\overline{\theta}_b, \overline{\theta}_s)$	$p = 0.016$	$p = 0.496$

All p-values stated above refer to the two-sided Fisher exact test. The null hypothesis is that the fraction of deviations from truth-telling is equal in both treatments. If deviations from truth-telling were observed in neither treatments, no p-value is stated.

6 Which mechanism is more profitable?

We now turn to the question which of the two mechanisms the designer would prefer. We first clarify the conditions under which the optimal mechanism for selfish agents outperforms the optimal externality-free mechanism in the sense that it yields a higher value of the designer's objective, here, maximal expected profits for the seller. We then check whether these conditions are satisfied in our experimental data.

Based on our experimental results, we introduce a distinction between different *behavioral types* of buyers. There is the "truthful type" and the "understatement type".[7] The former communicates his valuation truthfully in all the complete information games induced by the optimal robust mechanism f. The latter communicates a low valuation in all such games. We assume throughout that the seller always behaves truthfully, which is also what we observed in the experiment. We denote the probability that a buyer is of the "truthful type" by σ. We denote by $\Pi^f(\sigma)$ the expected profits that are realized under f. We denote by $\Pi^{f'}$ the expected profits that are realized under the optimal externality-free social choice function f', under the assumption that the buyer and the seller behave truthfully in all complete information games.

Proposition 4. *Suppose that* $\Pi^f(0) < \Pi^{f'}$. *Then there is a critical value* $\hat{\sigma}$ *so that* $\Pi^f(\sigma) \geq \Pi^{f'}$ *if and only if* $\sigma \geq \hat{\sigma}$.

Proof. We first note that

$$\begin{aligned}\Pi^f(\sigma) &= \sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) \Big\{ \sigma(p_b^f(\theta_b, \theta_s) - \theta_s k(q^f(\theta_b, \theta_s))) \\ &\qquad + (1-\sigma)(p_b^f(\underline{\theta}_b, \theta_s) - \theta_s k(q^f(\underline{\theta}_b, \theta_s))) \Big\} \\ &= \sigma \Pi^f(1) + (1-\sigma)\Pi^f(0)\,.\end{aligned}$$

[7] We refer to behavioral types because we wish to remain agnostic with respect to the social preference model that generates this behavior. Truthful behavior, for instance, can be rationalized both by selfish preferences and by preferences that include a concern for welfare. In the latter case, understatement is not attractive because it is Pareto-damaging.

We also note that $\Pi^f(1) > \Pi^{f'}$ since $\Pi^f(1)$ gives expected profits if there are only truthful buyer types, which is the situation in which f is the optimal mechanism. The term $\sigma\Pi^f(1) + (1-\sigma)\Pi^f(0)$ is a continuous function of σ. It exceeds $\Pi^{f'}$ for σ close to one. If $\Pi^f(0) < \Pi^{f'}$, it falls short of $\Pi^{f'}$ for σ close to zero. Hence, there is $\hat{\sigma} \in (0,1)$ so that $\Pi^f(\sigma) = \sigma\Pi^f(1) + (1-\sigma)\Pi^f(0)$ exceeds $\Pi^{f'}$ if and only if σ exceeds $\hat{\sigma}$. □

Examples 1 and 2 and our experimental data revisited. For the Examples 1 and 2 on which our experiments were based, the premise of Proposition 4 that $\Pi^f(0) < \Pi^{f'}$ is fulfilled. Specifically,

$$\Pi^f(0) = 4.54\,, \quad \Pi^f(1) = 4.91\,, \quad \Pi^f(\sigma) = 4.54 - 0.37\sigma\,, \quad \Pi^{f'} = 4.77 \quad \text{and} \quad \hat{\sigma} = 0.622$$

Thus, the fraction of deviating buyers must rise above 0.38 if the optimal externality-free mechanism is to outperform the optimal mechanism for selfish agents. In our experimental data, however, the fraction of deviating buyer types was with 0.14 smaller. Hence, in our experimental context the optimal mechanism for selfish agents is superior.

One might have expected more deviations from truth-telling. For instance, the social preference model by Fehr and Schmidt (1999) is consistent with truthful buyers only for one special case, namely the case in which buyers are completely selfish so that $\alpha_b = 0$, and Fehr and Schmidt estimate that often roughly 50% of subjects behave in a fair manner. This would have been more than enough to make the externality-free mechanism more profitable. However, the degree of selfishness may vary with the framing of the context, size of payments, etc., and moreover not all social preferences predict deviations. For instance, according to Charness and Rabin's (2002) model, individuals have a concern for welfare, so that a Pareto-damaging action such as communicating a low valuation instead of high valuation seems even less attractive. It is exactly this uncontrolled uncertainty about the mix of preferences among negotiators in a specific context, which justifies our approach to not further specify (beliefs about) social preferences. That said, however, an important insight is that the ability to control behavior is not the same as the ability to reach a given objective, here, maximal seller profits. Under an externality-free mechanism deviations from truth-telling are no longer tempting, i.e. this mechanism successfully controls behavior. One may, however, still prefer to lose control and use a mechanism under which some agents deviate if the complimentary set of agents who do not deviate is sufficiently large.

7 Concluding remarks

This paper shows how social preferences can be taken into account in robust mechanism design. We have first characterized an optimal mechanism for bilateral trade under the assumption that social preferences are irrelevant. We have argued theoretically that such a mechanism will not generate the desired behavior if individuals have social preferences, and we have illustrated experimentally that deviations from the intended behavior indeed occur. This has motivates us to introduce an additional constraint on mechanism design, which we termed externality-freeness. We have shown theoretically that such a mechanism does generate the intended behavior if individuals are motivated by social preferences, without a need to specify (beliefs about) the nature

and intensity of social preferences. Experiments confirmed that an externality-free mechanism does indeed generate the intended behavior.

These observations raise the question whether externality-freeness is a desirable property. We have assumed that mechanisms were designed with an objective of profit-maximization. In our experimental data, profits were higher with the mechanism that was derived on the assumption that individuals are selfish.[8] Hence, the ability to make money is not necessarily the same as the ability to predict behavior.

This observation reflects that externality-freeness is a sufficient but not a necessary condition for the ability to control behavior. Its advantage is that it successfully controls the underlying motivations across a wide variety of social preferences discussed in the literature, as well as the frequently observed large heterogeneity in parameter values across individuals. A fully fledged mechanism design exercise would require to elicit not only the monetary payoffs of individuals but also the precise functional form of their social preferences. We conjecture that with such a more fine-tuned mechanism design approach, there would no longer be a tension between the ability to predict behavior and the ability to reach a given objective. However, a need to specify the details of the nature and intensity of social preferences, which typically differ across individuals and contexts, would work against our goal to develop robust mechanisms in the spirit of the Wilson doctrine. We leave the question what can and what cannot be reached with a fine-tuned approach to future research.

Another important extension of our research would study social-preference-robustness with incomplete information about monetary payoffs. This would require a good understanding of how social preferences interact with uncertainty, which currently is an active research field. In fact, some recent evidence and theory suggest that social motivations are less significant for behavior when there is uncertainty about the comparison standard such as the opponent's final payoff (see Ockenfels et al. (2013), and the references cited therein), and that some patterns of risk taking in social context are not easily explained by either standard models of decision making under uncertainty nor standard models of social preferences (e.g., Bohnet et al. (2008), Saito (2013)). The implications of such findings for robust mechanism design need further attention.

References

Ariely, D., Ockenfels, A., and Roth, A. (2005). An experimental analysis of ending rules in internet auctions. *RAND Journal of Economics*, 36:890–907.

Bartling, B. and Netzer, N. (2013). An externality-robust auction: Theory and experimental evidence. Mimeo.

Becker, G. (1974). A theory of social interactions. *Journal of Political Economy*, 82:1064–1093.

Bergemann, D. and Morris, S. (2005). Robust mechanism design. *Econometrica*, 73:1771–1813.

Bierbrauer, F. (2011). On the optimality of optimal income taxation. *Journal of Economic Theory*, 146:2105–2116.

[8] The externality-free mechanism would have been more profitable only if more individuals had deviated from truth-telling.

Bierbrauer, F. and Netzer, N. (2012). Mechanism design and intentions. University of Zurich, Department of Economics, Working Paper No. 66.

Bohnet, I., Hermann, G., and Zeckhauser, R. (2008). Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review*, 98:249–310.

Bolton, G., Greiner, B., and Ockenfels, A. (2013). Engineering trust - reciprocity in the production of reputation information. *Management Science*, 59:265–285.

Bolton, G. and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90:166–193.

Bolton, G. and Ockenfels, A. (2010). Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States: Comment. *American Economic Review*, 100:628–633.

Bolton, G. and Ockenfels, A. (2012). Behavioral economic engineering. *Journal of Economic Psychology*, 33:665–676.

Bolton, P. and Dewatripont, M. (2005). *Contract Theory*. Cambridge, MA, MIT Press.

Börgers, T. (2013). An introduction to the theory of mechanism design. Mimeo, University of Michigan.

Camerer, C. (2003). *Behavioral Game Theory. Experiments in Strategic Interaction*. Princton Universtiy Press.

Charness, A. and Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117:817–869.

Chen, Y. (2008). Incentive-compatible mechanisms for pure public goods: A survey of experimental literature.

Chen, Y., Harper, M., Konstan, J., and Li, S. (2010). Social comparisons and contributions to online communities. *American Economic Review*, 100:1358–1398.

Chen, Y. and Sönmez, T. (2006). Social choice: An experimental study. *Journal of Economic Theory*, 127:202–231.

Cooper, D. and Kagel, J. (2009). Other-regarding preferences: A selective survey of experimental results.

Dufwenberg, M., Heidhues, P., Kirchsteiger, G., Riedel, F., and Sobel, J. (2011). Other-regarding preferences in general equilibrium. *Review of Economic Studies*, 78:613–639.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47:268–298.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54:293–315.

Fehr, E. and Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114:817–868.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10:171–178.

Gershkov, A., Goeree, J., Kushnir, A., Moldovanu, B., and Shi, X. (2013). On the equivalence of Bayesian and dominant strategy implementation. *Econometrica*, 81:197–220.

Greiner, B. (2004). An online recruitment system for economic experiments. *Forschung und wissenschaftliches Rechnen*, 63:79–93.

Güth, W. and Hellwig, M. (1986). The private supply of a public good. *Journal of Economics*, Supplement 5:121–159.

Güth, W. and Kocher, M. (2013). More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. Jena Economic Research Papers.

Güth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3:367–388.

Kagel, J., Lien, Y., and Milgrom, P. (2010). Ascending prices and package bidding: A theoretical and experimental analysis. *American Economic Journal: Microeconomics*, 2:160–185.

Kagel, J. and Roth, A. (2000). The dynamics of reorganization in matching markets: A laboratory experiment motivated by a natural experiment. *Quaterly Journal of Economics*, 115:201–235.

Kittsteiner, T., Ockenfels, A., and Trhal, N. (2012). Heterogeneity and partnership dissolution mechanisms: Theory and lab evidence. *Economics Letters*, 117:394–396.

Ledyard, J. (1978). Incentive compatibility and incomplete information. *Journal of Economic Theory*, 18:171–189.

Mailath, G. and Postlewaite, A. (1990). Asymmetric bargaining procedures with many agents. *Review of Economic Studies*, 57:351–367.

Mussa, M. and Rosen, S. (1978). Monopoly and product quality. *Journal of Economic Theory*, 18:301–317.

Myerson, R. and Satterthwaite, M. (1983). Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, 28:265–281.

Ockenfels, A., Sliwka, D., and Werner, P. (2013). Bonus payments and reference point violations. University of Cologne Working Paper.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 83:1281–1302.

Radner, R. and Schotter, A. (1989). The sealed-bid mechanism: An experimental study. *Journal of Economic Theory*, 48:179–220.

Roth, A. (2012). Experiments in market design. Mimeo.

Saito, K. (2013). Social preferences under risk: Equality of opportunity vs. equality of outcome. *American Economic Review*, 7:3084–3101.

Valley, K., Thompson, L., Gibbons, R., and Bazerman, M. (2002). How communication improves efficiency in bargaining games. *Games and Economic Behavior*, 38:127–155.

Wilson, R. (1987). Game-theoretic analyses of trading processes. In Bewley, T., editor, *Advances in Economic Theory: Fifth World Congress*, chapter 2, pages 33–70. Cambridge University Press, Cambridge, U.K.

A Other models of social preferences

In the body of the text, we have shown that the model of Fehr and Schmidt (1999) predicts deviations from truth-telling in certain situations (see *Observation 1*). Below, we present analogous findings for two other models of social preferences, Rabin (1993) and Falk and Fischbacher (2006). The Rabin (1993)-model is an example of intention-based social preferences, as opposed to the outcome-based model of Fehr and Schmidt (1999). The model by Falk and Fischbacher (2006) is a hybrid that combines considerations that are outcome-based with considerations that are intention-based. We show that these models also satisfy *Assumption 1*, i.e. selfishness in the absence of externalities.

Similar exercises could be undertaken for other models, such as Charness and Rabin (2002), and Dufwenberg and Kirchsteiger (2004). These models are consistent with our predictions under both, selfish and social preferences: Whether or not these models would predict deviations from truth-telling under the optimal mechanism for selfish agents depends on the values of specific parameters in these models. To avoid a lengthy exposition, we do not present these details here. The preferences in Charness and Rabin (2002), and Dufwenberg and Kirchsteiger (2004) do also satisfy the assumption of selfishness in the absence of externalities (*Assumption 1*).

Rabin (1993). The utility function of any one player i utility takes the form in (7). Rabin models the kindness terms in this expression in a particular way. Kindness intended by i towards j is the difference between j's actual material payoff and an equitable reference payoff,

$$\kappa_i(r_i, r_i^b, r_i^{bb}) = \pi_j(r_i, r_i^b) - \pi_j^{e_i}(r_i^b). \tag{13}$$

The equitable payoff $\pi_j^{e_i}(r_i^b)$ is to be interpreted as a norm, or a payoff that j deserves from i's perspective. According to Rabin (1993), this reference point is the average of the best and the worst player i could do to player j, i.e.

$$\pi_j^{e_i}(r_i^b) = \frac{1}{2}\left(max_{r_i \in E_{ij}(r_i)}\pi_j(\theta_j, f(r_i, r_i^b)) + min_{r_i \in E_{ij}(r_i)}\pi_j(\theta_j, f(r_i, r_i^b))\right), \tag{14}$$

where $E_{ij}(r_i)$ is the set of Pareto-efficient reports: A report r_i belongs to $E_{ij}(r_i)$ if and only if there is no alternative report r_i' so that $\pi_i(r_i', r_i^b) \geq \pi_i(r_i, r_i^b)$ and $\pi_j(r_i', r_i^b) \geq \pi_j(r_i, r_i^b)$, with at least one inequality being strict. Rabin models the beliefs of player i about the kindness intended by j in a symmetric way. Thus,

$$\kappa_j(r_i^b, r_i^{bb}) = \pi_i(r_i^b, r_i^{bb}) - \pi_i^{e_j}(r_i^{bb}). \tag{15}$$

Observation 3. *Let f be a social choice function that solves a problem of optimal robust mechanism design as defined in Section 4.1. Consider a complete information types space for state $(\overline{\theta}_b, \underline{\theta}_s)$ and suppose that $\theta_b = \overline{\theta}_b$. Suppose that f is such that*

$$\pi_b(\overline{\theta}_b, f(\overline{\theta}_b, \underline{\theta}_s)) = \pi_b(\overline{\theta}_b, f(\underline{\theta}_b, \underline{\theta}_s)) > \pi_b(\overline{\theta}_b, f(\overline{\theta}_b, \overline{\theta}_s)) = \pi_b(\overline{\theta}_b, f(\underline{\theta}_b, \overline{\theta}_s)) \ . \tag{16}$$

Suppose that the buyer's and the seller's first and second order beliefs are as in a truth-telling

equilibrium. Also suppose that the buyer has Rabin (1993)-preferences with $y_b \neq 0$. Then the buyer's best response is to truthfully reveal his valuation.

The social choice function in Example 1 fulfills Condition (16). Consider Table 3. The buyer's incentive constraint binds. Moreover, if the buyer understates his valuation this harms the seller. Since the seller's intention, when truthfully reporting his type, is perceived as kind, the buyer maximizes utility by rewarding the seller. By (9), the buyer will therefore announce his type truth-fully *for all* y_b.

Observation 4. *Let f be a social choice function that solves a problem of optimal robust mechanism design as defined in Section 4.1. Consider a complete information types space for state $(\overline{\theta}_b, \overline{\theta}_s)$ and suppose that $\theta_b = \bar{\theta}_b$. Suppose that f is such that (16) holds. Suppose that the buyer's and the seller's first and second order beliefs are as in a truth-telling equilibrium. Also suppose that the buyer has Rabin (1993)-preferences with $y_b \neq 0$. Then the buyer's best response is to understate his valuation.*

The social choice function in Example 1 fulfills Condition (16). Consider Table 4. We hypothesize that truth-telling is an equilibrium and show that this leads to a contradiction unless the buyer is selfish: The buyer's incentive constraint binds. Moreover, if the buyer understates his valuation this harms the seller. Since the seller's intention, when truthfully reporting his type, is perceived as unkind, the buyer maximizes utility by punishing the seller. By (9), the buyer will therefore understate his type *for all* $y_b \neq 0$. Hence, the Rabin model predicts that the buyer will deviate from truth-telling, *for all* $y_b \neq 0$. Put differently, truth-telling is a best response for the buyer only if $y_b = 0$, i.e. only if the buyer is selfish.

Finally, we note that the utility function in the Rabin (1993)-model satisfies Assumption 1 for all possible parameterizations of the model. The reason is that two actions which have the same implications for the other player generate the same kindness. The one that is better for the own payoff is thus weakly preferred.

Observation 5. *Suppose the buyer and the seller have preferences as in (7) with parameters y_b and y_s, respectively. The utility functions U_b and U_s satisfy Assumption 1, for all $y_b \neq 0$ and for all $y_s \neq 0$,*

Falk and Fischbacher (2006). We present a version of the Falk-Fischbacher model that is adapted to the two player simultaneous move games that we study. The utility function takes again the general form in (7). The kindness intended by player i is now given as

$$\kappa_i(r_i, r_i^b, r_i^{bb}) = \pi_j(r_i, r_i^b,) - \pi_j(r_i^b, r_i^{bb}) \,, \tag{17}$$

Moreover, $\kappa_j(r_i^b, r_i^{bb})$ is modeled by Falk-and Fischbacher in such a way that

$$\kappa_j(r_i^b, r_i^{bb}) \leq 0 \,, \tag{18}$$

whenever $\pi_i(r_i^b, r_i^{bb}) - \pi_j(r_i^b, r_i^{bb}) \leq 0$. More specifically, the following assumptions are imposed:

(a) If $\pi_i(r_i^b, r_i^{bb}) - \pi_j(r_i^b, r_i^{bb}) = 0$, then $\kappa_j(r_i^b, r_i^{bb}) = 0$.

(b) The inequality in (18) is strict whenever $\pi_i(r_i^b, r_i^{bb}) - \pi_j(r_i^b, r_i^{bb}) < 0$ and there exists r_j so that $\pi_i(r_j, r_i^{bb}) > \pi_i(r_i^b, r_i^{bb})$.

(c) If $\pi_i(r_i^b, r_i^{bb}) - \pi_j(r_i^b, r_i^{bb}) < 0$ and there is no r_j so that $\pi_i(r_j, r_i^{bb}) > \pi_i(r_i^b, r_i^{bb})$, then $\kappa_j(r_i^b, r_i^{bb})$ may be zero or positive.

The case distinction in (c) is decisive for the predictions of the Falk-Fischbacher model. If $\kappa_j(r_i^b, r_i^{bb}) > 0$, then Observation 1 for the Fehr-Schmidt-model also holds for the Falk-Fischbacher model. If, by contrast, $\kappa_j(r_i^b, r_i^{bb}) = 0$, then Observations 3 and 4 for the Rabin-model also hold for the Falk-Fischbacher model. In any case, the Falk-Fischbacher satisfies Assumption 1, the assumption of selfishness in the absence of externalities.

Observation 6. *Suppose the buyer and the seller have preferences as in the model of Falk and Fischbacher (2006) with parameters y_b and y_s, respectively. The utility functions U_b and U_s satisfy Assumption 1, for all $y_b \neq 0$ and for all $y_s \neq 0$.*

This follows since $\pi_j(r_i, r_i^b) = \pi_j(r_i', r_i^b)$ implies that $\kappa_i(r_i, r_i^b, r_i^{bb}) = \kappa_i(r_i', r_i^b, r_i^{bb})$. Consequently, two actions that yield the same payoff for the other player generate the same value of $\kappa_i(r_i, r_i^b, r_i^{bb})\kappa_j(r_i^b, r_i^{bb})$.

B Instructions

The instructions are a translation of the German instructions used in the experiment, and are identical for all participants. The original instructions are available upon request.

Instructions — General Part

Welcome to the experiment!

You can earn money in this experiment. How much you will earn, depends on your decisions and the decisions of another anonymous participant, who is matched with you. Independent of the decisions made during the experiment you will receive 7.00 € as a lump sum payment. At the end of the experiment, positive and negative amounts earned will be added to or subtracted from these 7.00 €. The resulting total will be paid out in cash at the end of the experiment. All payments will be treated confidentially.

All decisions made during the experiment are anonymous.

From now on, please do not communicate with other participants. If you have any questions now or during the experiment, please raise your hand. We will then come to you and answer your question.

Please switch off your mobile phone during the experiment. Documents (such as books, lecture notes etc.) that do not deal with the experiment are not allowed. In case of violation of these

rules you can be excluded from the experiment and all payments.

On the following page you will find the instructions concerning the course of the experiment. After reading these, we ask you to wait at your seat until the experiment starts.

First Part — Presentation of decision settings, reading of payoffs

The purpose of this part of the experiment is to familiarize all participants with the decision settings. This ensures that every participant understands the presentation of the decision settings and can correctly infer the resulting payoffs of specific decision combinations. None of the choices in the first part are payoff-relevant.

In the course of this part, eight different decision settings will be presented to you. In all of them two participants have to make a decision without knowing the decision made by the other participant. The combination of the decisions determines the payoffs of both participants. *[These eight decision settings refer to the four complete information games of the respective social choice function of their specific treatment. Each game was presented twice: First in the original form and then in a strategically identical form where the payoffs of Participant A and B were switched. This explanation is, of course, not part of the original instructions.]*

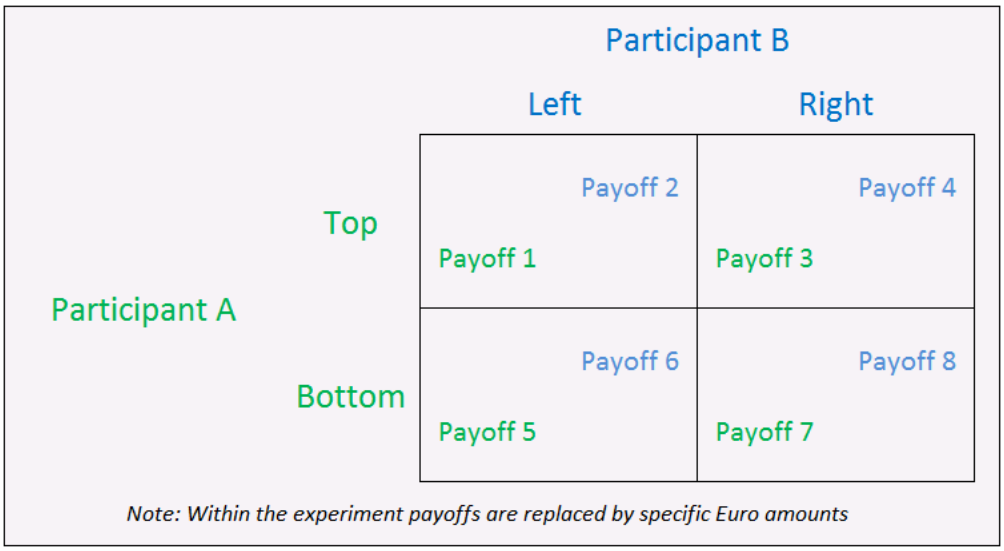

Figure 1: Exemplary Decision Setting

Participant A, highlighted in green, can decide between *Top* or *Bottom*. Participant B, highlighted in blue, can decide between *Left* and *Right*. The decision of Participant A determines whether the payment results from the upper or lower row in the table. Accordingly, the decision of Participant B determines whether the payment results from the left or right column. Both decisions combined unambiguously determine the cell of the payoff pair.

Each cell contains a payoff pair for both participants. Which payoff is relevant for which participant, is highlighted through their respective color. The green value, which can be found in the lower left corner of every cell, shows the payoff for Participant A. The blue value, which can be found in the upper right corner of every cell, shows the payoff for Participant B.
Please familiarize yourself with the payoff table. Put yourself in the position of both participants and consider possible decisions each participant would make. After a short time for consideration, you can enter a choice combination. The entry can be modified and different constellations can be tried. After choosing two decisions, please enter the payoffs which would result from this constellation. Your entry will then be verified. If your entry is wrong, you will be notified and asked to correct it.

Second Part — Decision Making

At the beginning of the second part you will be assigned to a role which remains constant over the course of the experiment. It will be the role of either Participant A or Participant B. Which role you are assigned to, will be clearly marked on your screen. Please note that the assignment is random, both roles are equally likely. It will be assured that half of the participants are assigned to the role of Participant A and the other half to the role of Participant B.
Simultaneously to the assignment of roles, you are matched with a participant of a different role. This matching is also random. In the course of the remaining experiment you will interact with this participant.

The second part of the experiment consists of four decisions settings. Exactly one decision setting is payoff relevant for you and the other participant matched with you. Which decision setting that is, is determined by chance: Every decision setting has the same chance of being chosen. Hence, please bear in mind that each of the following decision settings can be payoff-relevant.

All decision settings are presented similarly to those of the first part. The difference with respect to the first part is, that you can only make one decision, namely that for your role. Thus, you do not know the decision of the participant matched with you.

Only after you have made a decision for each of the four settings, you will learn which decision setting is relevant for your payoff and the payoff of the participant assigned to you. In addition you will learn the decisions of the other participant in all decisions settings.

After the resulting payoffs are displayed, the experiment ends. A short questionnaire will appear on your screen while the experimenters prepare the payments. Please fill out this questionnaire and wait at your seat until your number is called.

If you have any questions, please raise your hand.

Thank you for participating in this experiment!