Robust Mechanism Design and Social Preferences*

Felix Bierbrauer Axel Ockenfels Andreas Pollak Désirée Rückert

University of Cologne

March 15, 2016

Abstract

One key finding of behavioral economics is that many people are motivated by social concerns. However, most of the robust mechanism design literature focuses on beliefs, and takes selfish preferences for granted. We study two classic challenges in mechanism design – bilateral trade à la Myerson and Satterthwaite (1983) and redistributive income taxation à la Mirrlees (1971) and Piketty (1993) – to show that some standard mechanism design solutions systematically fail with social preferences, while others are robust. We thus introduce the notion of a social-preference-robust mechanism which works not only for selfish but also for social preferences of different nature and intensity, and characterize the optimal mechanism in this class. We compare the performance of the optimal mechanisms for selfish agents and the optimal social-preference-robust mechanisms with the help of a series of laboratory experiments and find that behavior can indeed be better controlled with social-preference-robust mechanisms.

Keywords: Robust Mechanism Design, Social Preferences, Bilateral Trade, Income Taxation

JEL Classification: C92, D02, D03, D82, H2

*We benefited from comments by Carlos Alós-Ferrer, Tilman Börgers, Dirk Engelmann, Alia Gizatulina, Jacob Goeree, Hans-Peter Grüner, Stephen Morris, Johannes Münster, Nick Netzer and Bettina Rockenbach and participants of seminars at Bonn University, Tor Vergata in Rome, CERGE-EI in Prague, St. Gallen and the meetings of EEA-ESEM 2014, ASSA 2015 and SED 2015. Financial support of the German Science Foundation through the Leibniz program and through the DFG research unit "Design & Behavior" (FOR 1371) is gratefully acknowledged. Axel Ockenfels thanks the Economics Department at Stanford University for the generous hospitality. Felix Bierbrauer thanks the Max Planck Institute for Research on Collective Goods for its hospitality.

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. Indeed, behavioral economics has shown that many agents behave socially. One challenge is, though, that social preferences can differ with respect to their nature and intensity, leading to different kinds of social preference models, including altruism, inequity-aversion, and intentionality (Cooper and Kagel (2013)). Because we want a mechanism to work not only for selfish preferences but also for a large set of social preferences, we introduce the notion of social-preference-robust mechanism: a mechanism must not depend on specific assumptions about the nature and intensity of selfish and social preferences. The following quote of Wilson (1987), which can also be found in Bergemann and Morris (2005), suggests that our approach 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.

In this paper, we show theoretically that optimal mechanisms that are derived under the assumption of selfish preferences may not generate the intended behavior if individuals have social preferences. Second, and most importantly, we introduce the notion of a social-preference-robust mechanism and derive mechanisms that are optimal in this class. Finally, we use laboratory experiments to demonstrate that social preferences are a non-negligible factor in our context, and to compare the performance of the optimal mechanisms under selfish preferences and the optimal social-preference-robust mechanisms.

For the applications studied in this paper, the notion of robustness due to Bergemann and Morris is equivalent to the requirement that a mechanism has a dominant strategy equilibrium. Depending on the application, this may significantly restrict the set of implementable outcomes.[1] Thus, there may be the concern that adding another robustness-requirement will restrict the set of admissible mechanisms even further and is therefore problematic. In our view, comparing mechanisms that, according to theory, sacrifice performance for a more robust solution con-

[1] For instance, Hagerty and Rogerson (1987) study the bilateral trade-problem due to Myerson and Satterthwaite (1983) and show that, with incentive and participation constraints that are robust in the Bergemann and Morris (2005)-sense, the set of admissible mechanisms is restricted. For other applications, this is not a restriction at all. For a problem of redistributive income taxation, Bierbrauer (2011) shows that there is an optimal mechanism with a dominant strategy equilibrium.

cept to mechanisms that, according to theory, sacrifice robustness in return for performance, is ultimately an empirical question. Our laboratory experiments are first steps in this direction.

Throughout, we use two classic applications of mechanism design theory, a version of the bilateral-trade problem due to Myerson and Satterthwaite (1983) and versions of the optimal income tax problem due to Mirrlees (1971) and Piketty (1993) to illustrate our theoretical analysis. For our analysis of the bilateral trade problem, we focus on mechanisms that maximize the expected profits of the seller.[2] Profit-maximizing mechanisms yield an asymmetric distribution of the gains from trade between the buyer and the seller, and may therefore be particularly susceptible to provoke deviations from the "intended behavior" that are motivated by social preferences. By contrast, the theory of optimal taxation focusses on welfare-maximization. One might conjecture that, with an objective function that already incorporates social concerns, behavior that is driven by social preferences may be less of a concern.

More specifically, we use these applications to generate three important constellations: A first constellation in which the performance of the optimal mechanism under selfish preferences cannot be mitigated by the existence of social preferences (Mirrleesian income taxation); a second constellation where it is mitigated only if social preferences are sufficiently strong (bilateral trade); and finally a third constellation where it is already mitigated when social preferences play a rather minor role (income taxation á la Piketty, 1993). Our experiments confirm these predictions.

The bilateral trade problem. The bilateral-trade problem provides us with a simple, and stylized setup that facilitates a clear exposition of our approach. 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.

The essence of the bilateral trade problem is that there are two parties and that each party has private information on its benefits (or costs) from a transaction with the other party. Hence, the labels "buyer" and "seller" need not to be taken literally. 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

[2]We do not impose ex post budget balance. Instead, the budget constraint stipulates that the expected revenues of the seller must not be less than the expected payments of the buyer. Moreover, the traded quantity is continuously adjustable and not an element of $\{0, 1\}$. Therefore the set of admissible mechanism is not as restricted as in Hagerty and Rogerson (1987).

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.

As has been shown by Bergemann and Morris (2005), ex post contraints 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. Moreover, 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.[3]

In our laboratory approach, we investigate the performance of an optimally designed robust mechanism 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 are related 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 (see Camerer (1995)), and that moreover, in social contexts, social and risk preferences may interact in non-trivial ways (see, e.g., Bolton and Ockenfels (2010), and the references therein). The complete information environments in our study avoid such complicating factors.[4]

For the bilateral trade problem, the mechanism which maximizes the seller's expected profits under selfish preferences has the following properties: (i) The trading surplus is allocated in

[3]Throughout we focus on social choice functions, as opposed to social choice correspondences. Consequently, by Corollary 1 in Bergemann and Morris (2005), *ex post* implementability is both necessary and sufficient for robust implementability. Moreover, if agents are selfish, then our environment gives rise to private values so that incentive compatibility in an *ex post* sense is equivalent to the requirement that truth-telling is a dominant strategy under a direct mechanism for the given social choice function.

[4]Thus, for our experimental testing strategy, we take for granted the equivalence between implementability in all complete information environments and implementability in all incomplete information environments. We explicitly investigate the former and draw conclusions for the latter. We also take for granted the validity of the revelation principle. That is, we only check whether individuals behave truthfully under a direct mechanism for a given social choice function. We discuss the advantages and limits of this approach in our concluding section.

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 is consistent with models of social preferences such as Fehr and Schmidt (1999), and Falk and Fischbacher (2006), among others. The basic idea is the following. 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 the monetary payoffs of everybody else are 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. 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 question whether the seller makes more money if she uses an externality-free mechanism. We answer this both theoretically and empirically: The externality-free mechanism makes more money if the number of participants whose behavior is motivated by social preferences exceeds a threshold. In our laboratory context, this number was below the threshold, so that the "conventional" mechanism made more money than the externality-free mechanism.

Based on these observations, we finally engineer a mechanism that satisfies the property of externality-freeness only locally. Specifically, we impose externality-freeness for those action-

profiles where deviations from selfish behavior were frequently observed in our experiment data. We show theoretically that local externality-freeness is a constraint that can be satisfied without having to sacrifice performance: If all agents are selfish then there is an optimal mechanism that is locally externality-free. In our experiment data, however, an optimal mechanism that is locally externality-free performs significantly better than an optimal mechanism that is not externality-free. Hence, if one knows precisely which deviations from selfish behavior are tempting, one can design a mechanism that performs strictly better than both the optimal mechanism for selfish agents and the optimal globally externality-free mechanism.

Income Taxation. The bilateral trade setup is one in which externalities are at the center of the allocation problem: More consumption for the buyer can be realized only with higher costs for the seller, and additional revenue for the seller can only be realized if the buyer pays more. Hence, it seems natural that the buyer's behavior will affect the seller's payoff and vice versa. A requirement of externality-freeness which shuts down this interdependence may therefore appear demanding. In settings different from the bilateral trade problem, externality-freeness may arise naturally. For example, 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.[5] Another setting in which externality-freeness may appear natural is the design of tax systems. Here, externality-freeness requires that income taxes paid by one individual depend only on this individual's income, and not on the income earned by other individuals. Thus, when formalizing the modern approach to optimal income taxation, Mirrlees (1971) and his followers have looked exclusively at externality-free allocations.

However, as has been shown by Piketty (1993), for an economy with finitely many individuals and a commonly known cross-section distribution of types, an optimal Mirrleesian income tax system can be outperformed by one that is *not* externality-free. Specifically, Piketty shows that first-best utilitarian redistribution from high-skilled individuals to low-skilled individuals can be reached, while this is impossible with a Mirrleesian approach. A crucial feature of Piketty's approach is that types are assumed to be correlated in a particular way. For instance, if there are two individuals and it is commonly known that one of them is high-skilled and one is low-skilled, then the individuals' types are perfectly negatively correlated: If person 1 is of high ability, then person 2 is of low ability and vice versa. Piketty's construction of a mechanism that reaches the first-best utilitarian outcome heavily exploits this feature of the environment.[6]

Piketty's analysis resembles the possibility results by Crémer and McLean (1985, 1988) in auction theory.[7] Crémer and McLean have shown that, with correlated values and selfish agents, there exist Bayes-Nash equilibria that achieve first-best outcomes. These findings have then been generalized to other types of allocation problems, see e.g., Kosenok and Severinov (2008).

[5] Market behavior is therefore unaffected by social preferences, see Dufwenberg et al. (2011).

[6] If individual types are the realizations of independent random variables, then the optimal mechanism is externality-free, see Bierbrauer (2011).

[7] There are also some important differences though. Piketty uses the solution concept of a dominant strategy equilibrium which implies that his approach is robust in the sense of Bergemann and Morris (2005). The approach of Crémer and McLean is based on the solution concept of a Bayes-Nash equilibrium and strongly depends on specific properties of a common prior. It is therefore not robust in the sense of Bergemann and Morris (2005).

Importantly, the mechanisms that achieve first-best outcomes in the presence of correlated types give rise to payoff interdependencies or externalities among the players. Therefore, they raise the question whether social preferences might interfere with the possibility to achieve first-best outcomes. Piketty's treatment of the income tax problem is an example that allows us to get at this more general question in a particular context.

We run an experiment and show that Piketty's mechanism indeed provokes deviations from the intended behavior, and again, we argue that these deviations can be explained by models of social preferences. We then compare Piketty's mechanism to an optimal Mirrleesian mechanism. The latter is externality-free and we find that it successfully controls behavior; there are no longer significant deviations from truth-telling. We also find that the level of welfare that is generated by the Mirrleesian mechanism is significantly larger than the level of welfare generated by Piketty's mechanism. This last observation makes an interesting difference to our findings for the bilateral trade problem. With the income tax problem, imposing externality-freeness is also good for the performance of the mechanism. The difference is not due to different social preferences, but reflects the fact that the externally-free mechanism is preferable only if the number of socially motivated agents exceeds a threshold. This threshold is much larger for the bilateral trade problem (and too large for what we observe in the laboratory).

Outline. The next section discusses related literature. In Section 3 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. It also contains a detailed description of the bilateral trade problem that we study. In addition, Section 4 describes our laboratory findings for the bilateral trade problem, and in Section 5, we clarify the conditions under which an optimal externality-free mechanism outperforms an optimal mechanism for selfish agents and relate them to our experiment data. Section 6 looks at an engineering approach that does impose externality-freeness only locally. Section 7 contains our analysis of the income tax problem. The last section concludes.

2 Related literature

There is a rich literature on models of social preferences. Within this literature there are different subcategories, such as, for instance, the distinction between outcomes-based and intention-based models of social preferences. Well-known models of outcomes-based social preferences are Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). Models of intention-based social preferences include Rabin (1993) and Falk and Fischbacher (2006).[8] We introduce social preferences into a model of robust mechanism design. This complements the analysis of Bergemann and

[8]There is a large literature on mechanism design with interdependent valuations, see e.g. the survey in Jehiel and Moldovanu (2006). In principle, models of outcomes-based social preferences can be viewed as specific models with interdependent valuations. By contrast, models with intention-based social preferences cannot be viewed as models with interdependent valuations. In these models, preferences are menu-dependent, see Sobel (2005) for a discussion. Such a menu dependence does not arise in the literature on mechanism design with interdependent valuations.

Morris (2005) who were seeking robustness with respect to the specification of the individuals' probabilistic beliefs.[9]

For the purpose of illustration, we focus on two classic applications of mechanism design, the bilateral trade problem and the problem of redistributive income taxation. Myerson and Satterthwaite (1983) establish an impossibility result for efficient trade in a setting with two privately informed parties.[10] Our focus is different. We look at a second-best mechanism for this problem and ask how its performance is affected by social preferences. The classical reference for redistributive income taxation is Mirrlees (1971). We relate the Mirrleesian treatment to an alternative one that has been proposed by Piketty (1993).[11]

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)), 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. (2014); 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 work builds on earlier contributions by Bierbrauer and Netzer (2012) and Bartling and Netzer (2014). These papers do not seek robustness within a large class of social preference models. They focus on specific models of social preferences and trace out their implications for Bayesian mechanism design. With their approach, externality-freeness is not a substantive constraint. Bierbrauer and Netzer (2012) provide conditions under which the class of optimal mechanisms for selfish agents contains one that is externality-free. Our analysis, by contrast, gives rise to a trade-off between externality-freeness on the one hand and performance on the assumption that everyone is selfish on the other hand.

[9] Other contributions to the literature on robust mechanism design include Ledyard (1978), Gershkov et al. (2013) and Börgers (2015).

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

[11] The mechanism design approach to the problem of optimal income taxation is also discussed in Hammond (1979), Stiglitz (1982), Dierker and Haller (1990), Guesnerie (1995), and Bierbrauer (2011).

3 Mechanism design with and without social preferences

This section contains theoretical results which relate mechanism design theory to models of social preferences. Throughout, we use the bilateral trade problem to illustrate the conceptual questions that arise. 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 maximizing expected payoffs is a well-defined goal, there are many ways to be socially motivated. 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 (2013)). This poses a problem for mechanism design, because optimal mechanisms depend on the nature of 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 approaches would violate the spirit of robust mechanism design and the Wilson doctrine. Rather, 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 maximize their own payoffs, regardless of their social preferences, if there is no possibility to affect the payoffs of others. As we will show, this general property of social behavior is 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.

3.1 The bilateral trade problem

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.[12]

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 anal-

[12] Our setting differs from the one originally studied by Myerson and Satterthwaite (1983) in that we have a convex cost function for the seller and allow for quantities in $\mathbb{R}$. In the original paper, the seller's cost function is linear and quantities are in $[0, 1]$.

ogously.

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.[13] 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 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, 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.

In the body of the text, 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 (2016) 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.[14]

Another property of interest to us is the externality-freeness of a social choice function f.

[13] "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.

[14] In part B of the Appendix we derive necessary conditions for an intention-based model. There we have to allow for arbitrary non-direct mechanisms.

This property holds if, for all $\theta_b \in \Theta_b$,

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

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

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

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.

3.2 Optimal mechanism design under selfish preferences

A mechanism designer wishes to come up with a mechanism for bilateral trade. Design takes place at the *ex ante* stage, i.e. before the state of the economy is realized. 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.) The flexibility provided by the requirement of expected budget balance is important for some of the results that follow. With a requirement of *ex post* budget balance there would be less scope for adjusting the traded quantities and the corresponding payments to the privately held information of the buyer and the seller.[15]

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 has to hold

$$\sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) p_b^f(\theta_b, \theta_s) \geq \sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) p_s^f(\theta_b, \theta_s) \; . \tag{4}$$

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 its solution. This solution to the relaxed problem is also a solution to the full problem if it satisfies all constraints of the full

[15] We do not wish to argue that the requirement of expected budget balance is, for practical purposes, more relevant than the requirement of *ex post* budget balance. This will depend on the application. The mechanisms that we study in this paper are primarily meant as diagnostic tools for the relevance of social preferences in mechanism design. In this respect, the requirement of expected budget balance proved useful.

problem, as is the case for Example 1 below.

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

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

A formal proof of Proposition 1 is in part B of the Appendix. Here, we 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. Therefore, 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.[16]

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

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

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. In the following we provide a specific example in which these payments are specified in such a way that they satisfy not only these constraints, but also give rise to *ex post* budget balance, i.e. in every state (θ_b, θ_s), the price paid by the buyer equals the revenue obtained by the seller,

$$p_b^f(\theta_b, \theta_s) = p_s^f(\theta_b, \theta_s) \ . \tag{5}$$

Example 1: An optimal robust social choice function. Suppose that $\underline{\theta}_b = 1.00$, $\overline{\theta}_b = 1.30$, $\underline{\theta}_s = 0.20$, 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, an optimal robust social choice function f looks as follows: The traded quantities are given by

$$q^f(\underline{\theta}_b, \underline{\theta}_s) = 3.50, \ q^f(\underline{\theta}_b, \bar{\theta}_s) = 1.08, \ q^f(\bar{\theta}_b, \underline{\theta}_s) = 6.50 \quad \text{and} \quad q^f(\bar{\theta}_b, \bar{\theta}_s) = 2.00 \ .$$

The buyer's payments are

$$p_b^f(\underline{\theta}_b, \underline{\theta}_s) = 3.50, \ p_b^f(\underline{\theta}_b, \bar{\theta}_s) = 1.08, \ p_b^f(\bar{\theta}_b, \underline{\theta}_s) = 7.40 \quad \text{and} \quad p_b^f(\bar{\theta}_b, \bar{\theta}_s) = 2.28 \ .$$

Finally, the seller's revenues are

$$p_s^f(\underline{\theta}_b, \underline{\theta}_s) = 3.50, \ p_s^f(\underline{\theta}_b, \bar{\theta}_s) = 1.08, \ p_s^f(\bar{\theta}_b, \underline{\theta}_s) = 7.40 \quad \text{and} \quad p_s^f(\bar{\theta}_b, \bar{\theta}_s) = 2.28 \ .$$

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.[17]

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*

[17] More precisely, this and the following normal form games are generated by an approximation f^x of f which is such that, whenever an incentive constraint is binding under f, a deviation from truth-telling has a small cost of two cents under f^x. Our laboratory experiments used f^x rather than f. Thus, under f^x it is less tempting to deviate from truth-telling and we can be more confident that the deviations from truth-telling that we observe reflect social preferences, as opposed to an arbitrary selection from a set of best responses.

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 from an 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 induced by f for $(\theta_b, \theta_s) = (\overline{\theta}_b, \overline{\theta}_s)$.

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

An inspection of Tables 1 trough 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).

3.3 An observation on models of social preferences

We now show that the social choice function in Proposition 1 is not robust in the following sense: It provokes deviations from truth-telling if individuals are motivated by social preferences. To formalize a possibility of social preferences, we assume that any one individual $i \in \{b, s\}$ 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 individuals' (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.

Intention-based social preferences. 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

$$U_i(\theta_i, r_i, r_i^b, r_i^{bb}) = \pi_i(\theta_i, f(r_i, r_i^b)) + y_i\, \kappa_i(r_i, r_i^b, r_i^{bb})\, \kappa_j(r_i^b, r_i^{bb})\ . \tag{6}$$

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}, $\kappa_j(r_i^{b\prime}, r_i^{bb}) \geq \kappa_i(r_i^{b\prime\prime}, r_i^{bb})$ whenever $\pi_i(\theta_i, f(r_i^{b\prime}, r_i^{bb})) \geq \pi_i(\theta_i, f(r_i^{b\prime\prime}, r_i^{bb}))$. 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.

Outcome-based social preferences. 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 and individuals are assumed to care about their own payoff and the distribution of payoffs among the players. For instance, with Fehr-Schmidt-preferences, the utility function of individual i reads as

$$\begin{aligned} U_i(\theta_i, r_i, r_i^b, r_i^{bb}) \quad = \quad & \pi_i(\theta_i, f(r_i, r_i^b)) \quad -\alpha_i \max\{\pi_j(\theta_j, f(r_i, r_i^b)) - \pi_i(\theta_i, f(r_i, r_i^b)), 0\} \\ & \qquad\qquad\quad -\beta_i \max\{\pi_i(\theta_i, f(r_i, r_i^b)) - \pi_j(\theta_j, f(r_i, r_i^b)), 0\}\ , \end{aligned} \tag{7}$$

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

Implications for the social choice function in Proposition 1. Many models of social preferences give rise to the prediction that a social choice function that would be optimal if individual 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: 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, such as Rabin (1993), would have

a different logic. For the game in Table 4, the buyer would argue as follows: 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 reciprocally 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.

The following observation 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*

$$\pi_s(\theta_s, f(\overline{\theta}_b, \theta_s)) > \pi_s(\theta_s, f(\underline{\theta}_b, \theta_s)) > \pi_b(\overline{\theta}_b, f(\underline{\theta}_b, \theta_s)) = \pi_b(\overline{\theta}_b, f(\overline{\theta}_b, \theta_s)) \tag{8}$$

Suppose that the seller behaves truthfully. Also suppose that the buyer has Fehr-Schmidt-preferences as in (7) 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 (8). 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.

3.4 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.[18]

Observation 2. *Suppose the buyer and the seller have preferences as in (7) 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.

[18] Assumption 1 is also satisfied in models of pure altruism, see Becker (1974). All parameterized versions that Bolton and Ockenfels (2000) propose for their model are consistent with Assumption 1, too, although we note that it is theoretically possible to construct preferences that are consistent with their general assumptions and may still violate Assumption 1. Such preferences would be the only possible exception that we encountered among prominent social preference models.

Definition 1. *A direct mechanism for social choice function f is said to be 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.

The following Proposition 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 believes that player j acts truthfully so that $r_i^b = \theta_j$ and that he believes that player j believes 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 $\max_{r_i \in \Theta_i} U_i(\theta_i, r_i, r_i^b, r_i^{bb})$. □

Proposition 2 asserts that externality-freeness is a sufficient condition for social-preference-robustness. This raises the question of necessary conditions. Above we said that a condition is sufficient if it ensures implementability *for all* social preference models so that individuals are selfish in the absence of externalities. Hence, it is natural to say that a condition is necessary for social-preference-robustness if there *exists one* relevant social preference model so that it is necessary for implementability. Bierbrauer and Netzer (2016) show that, under an ancillary condition, externality-freeness is indeed necessary for the implementability of a social choice function for a version of the intention-based model of Rabin (1993) that allows for private information both on material payoffs and on the weights that reciprocity has in the players' overall utility function. However, Bierbrauer and Netzer (2016) employ Bayesian incentive compatibility constraints, as opposed to ex post incentive compatibility constraints. As we show in part B of the Appendix, this difference is of no consequence for the validity of the conclusion that externality-freeness is a necessary condition.

3.5 Optimal robust and externality-free mechanism design

We now add the requirement of externality-freeness to the bilateral trade problem. To characterize the solution of this problem 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 (4), the participation constraints for a low valuation buyer,

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

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

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

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

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

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 $(\underline{\theta}_b, \underline{\theta}_s)$;

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

where $g^m(\bar{\theta}_b) := g(\bar{\theta}_b, \underline{\theta}_s) + g(\bar{\theta}_b, \bar{\theta}_s)$. 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

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

and

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

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

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

A formal proof of the Proposition is relegated to part B of the Appendix. It proceeds as follows: 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 robust mechanism in Proposition 1 and the optimal robust and externality-free mechanism in Proposition 3 is in the pattern of distortions. The optimal robust mechanism has downward distortions whenever the buyer has a low valuation. The optimal robust and externality-free mechanism 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 optimal robust mechanism.

Example 2: An optimal robust and externality-free social choice function. Suppose the parameters of the model are as in Example 1. The social choice function f', specified in Proposition 3, solves the problem of optimal robust and externality-free mechanism design formally defined in the previous paragraph: The traded quantities are given by

$$q^{f'}(\underline{\theta}_b, \underline{\theta}_s) = 2.00,\ q^{f'}(\underline{\theta}_b, \bar{\theta}_s) = 1.54,\ q^{f'}(\bar{\theta}_b, \underline{\theta}_s) = 6.50 \quad \text{and} \quad q^{f'}(\bar{\theta}_b, \bar{\theta}_s) = 2.00\ .$$

The buyer's payments are

$$p_b^{f'}(\underline{\theta}_b, \underline{\theta}_s) = 2.00,\ p_b^{f'}(\underline{\theta}_b, \bar{\theta}_s) = 1.54,\ p_b^{f'}(\bar{\theta}_b, \underline{\theta}_s) = 7.85 \quad \text{and} \quad p_b^{f'}(\bar{\theta}_b, \bar{\theta}_s) = 2.00\ .$$

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

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').

4 A laboratory experiment

We conducted a laboratory experiment with five treatments. The first treatment is based on the optimal mechanism f under selfish preferences in Example 1 (T1), and the second treatment is based on the optimal externality-free mechanism f' under social preferences in Example 2 (T2). The three additional treatment variations (T3-5) will be described in subsequent sections. All treatments were conducted employing exactly the same laboratory procedures which are described below.

Laboratory Procedures. The experiments were conducted in the *Cologne Laboratory for Economic Research* at the University of Cologne. They had been programmed with *z-Tree* developed by Fischbacher (2007), and participants were recruited with the online recruitment system *ORSEE* developed by Greiner (2004). In total, we recruited 632 subjects who participated in twenty sessions, four for each of the five treatments. Each subject was allowed to participate in one session and in one treatment only (between-subject design). We collected at least 63 independent observations for each treatment and player role. Subjects were students from all faculties of the University of Cologne, mostly female (380 subjects), with an average

age of 24 years. A session lasted 45-60 minutes. Average payments to subjects, including the show-up fee, was 10.76 Euro.

Upon arrival, subjects were randomly assigned to computer-terminals and received identical written instructions, which informed them about all general rules and procedures of the experiment. All treatment- and role-specific information was given on the computer-screen (see Appendix C for instructions and screenshot). We used neutral terms to describe the game; e.g., player-roles were labeled Participant A (B) and strategies were labeled Top (Left) and Bottom (Right) respectively. [19]

Subjects then went through a *learning stage*, with no interaction among subjects and no decision-dependent payments. In the learning stage, subjects had to choose actions for each player role in each complete information game and then to state the resulting payoffs for the corresponding self-selected strategy combination. Subjects had to give the right answer before proceeding to the decision stage. This way we assured that all subjects were able to correctly read the payoff tables, without suggesting specific actions which might create anchoring or experimenter demand effects.

Then subjects entered the *decision stage* and were informed about their role in their matching group. The matching into groups and roles was anonymous, random and held constant over the course of the experiment. Within the decision stage, subjects had to choose one action for each of the four complete information games of their specific treatment.[20] The order of the four games was identical to the order in Table 5. 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 four games was randomly determined for being paid in addition to a 2.50 Euro show-up fee.

Results. Table 5 summarizes the decisions made in the experiment. Sellers report their true valuation in almost all cases. Buyers with a low type also make truthful reports in both treatments. This is different for high type buyers in T1, though. Here, 13% (17%) of the buyers understate their true valuation when facing a seller with a low (high) valuation.

This pattern of buyer and seller behavior is in line with models of social preferences. In particular, for T1, which is based on an optimal mechanism for selfish agents, these models imply that high type buyers cannot be expected to always make truthful reports. By contrast, for $T2$, which is based on an optimal externality-free mechanism, these models unambiguously predict truthful behavior. We observe significantly higher shares of truthful high type buyer reports in T2 in comparison to T1 (two-sided Fisher's exact test, $p = 0.017$ for the games with a low type seller and $p = 0.014$ for the games with a high type seller).

[19] In the following we refer to the specific roles within the experiment as buyers and sellers, to make this section consistent with previous ones.

[20] As mentioned before, our experimental testing strategy takes for granted the equivalence between implementability in all complete information environments and implementability in all incomplete information environments; see Section 9 for discussion.

Table 5: Choice Data T1 and T2

		Buyer		*Seller*	
	Game induced by	$\underline{\theta}_b$	$\overline{\theta}_b$	$\underline{\theta}_s$	$\overline{\theta}_s$
T1 *optimal mechanism under selfish preferences*	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 *externality-free mechanism*	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

5 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 experiment data.

Based on our experiment results, we introduce a distinction between different *behavioral types* of buyers. There is the "truthful type" and the "understatement type".[21] 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

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

[21] 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 $\hat{\sigma}$. □

Our experiment 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,

$$\Pi^f(0) = 4.54\,, \quad \Pi^f(1) = 4.91\,, \quad \Pi^f(\sigma) = 4.91 - 0.37\sigma\,, \quad \Pi^{f'} = 4.77 \quad \text{and} \quad \hat{\sigma} = 0.62$$

Thus, the fraction of deviating buyers must rise above 38% if the optimal externality-free mechanism is to outperform the optimal robust mechanism. In our experiment data, however, the fraction of deviating buyer types was only 14%. As a consequence, actual average seller profits are smaller under the externality-free mechanism (4.77) than under the optimal robust mechanism (4.82). This difference was not found to be statistically significant, though (two-sided t-test based on independent average profits, $p = 0.143$).[22]

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 way. 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 preference models predict deviations. For instance, according to the model of Charness and Rabin (2002), individuals have a concern for welfare, so that an efficiency-damaging action such as communicating a low valuation instead of high valuation seems less attractive. This uncontrolled uncertainty about the mix of preferences among negotiators in a specific context justifies our approach to not further specify (beliefs about) social preferences.

That said, 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 use a mechanism under which some agents deviate if the complementary set of agents who do not deviate is sufficiently large.

6 Finding a superior mechanism: An engineering approach

Our laboratory findings suggest that the requirement of externality-freeness is more than what is really needed to control behavior. Under the optimal mechanism for selfish agents only particular deviations from truth-telling were observed frequently: Some of the high valuation buyers understated their valuation. However, when we impose externality-freeness we also ensure that low valuation buyers do not overstate their valuation, that high cost sellers do not understate

[22] Intuitively, the sellers matched with the low valuation buyers dampen the effect of the deviant high valuation buyers on the performance measure.

their costs, and that low cost sellers do not exaggerate their costs. While such deviations could possibly be rationalized by models of social preferences, they seem empirically less likely.

In the following, we therefore consider a mechanism design problem in which the requirement of externality-freeness is imposed only locally, namely such that the buyer is unable to influence the seller's payoff. Formally, we require that, for all $\theta_s \in \Theta_s$,

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

We do not attempt to provide an axiomatic foundation for these constraints. The motivation for imposing them comes exclusively from the behavior that we observed in our laboratory tests of the mechanisms in Examples 1 and 2. This is why refer to this approach as "engineering" (see Roth (2002) and Bolton and Ockenfels (2012)).

Remember that the optimal social choice function that we characterize in Proposition 1 leaves degrees of freedom for the specification of the payments to the seller. For instance, these payments can be chosen so that, in each state, the payment of the buyer equals the seller's revenue, as stipulated by (5). Alternatively, the payments can be chosen so that the local externality-freeness condition (9) is satisfied; i.e. there is an optimal mechanism that satisfies (9). Hence, local externality-freeness can be ensured without having to sacrifice performance. This is stated formally in the following Proposition that we prove in part B of the Appendix.

Proposition 5. *There is a solution to the relaxed problem of robust mechanism design, characterized in Proposition 1, that satisfies (9).*

Proposition 5 shows that if everybody is selfish then an optimal mechanism that satisfies *ex post* budget balance (as in Example 1 above) and an optimal mechanism that satisfies local externality-freeness (as in Example 3 below) are equivalent in terms of the expected profits that they generate. However, if the locally externality-free mechanism eliminates deviations that occur under *ex post* budget balance, it will perform strictly better. To see whether this is indeed the case we ran another laboratory treatment (T3), employing the same procedures as outlined in Section 4, yet based on the following Example.

Example 3: An optimal robust and locally externality-free social choice function. We illustrate Proposition 5 in the context of our numerical example. The payoff functions, parameter values, and traded quantities are as in Example 1. We denote the optimal mechanism that is locally externality-free by f''. Under f'', payments to the seller are given by

$$p_s^{f''}(\underline{\theta}_b, \underline{\theta}_s) = 4.07,\ p_s^{f''}(\underline{\theta}_b, \bar{\theta}_s) = 1.25,\ p_s^{f''}(\bar{\theta}_b, \underline{\theta}_s) = 6.99 \quad \text{and} \quad p_s^{f''}(\bar{\theta}_b, \bar{\theta}_s) = 2.11 \ .$$

Part D of the Appendix contains a detailed description of the normal form games that f'' induces on the four different complete information type spaces. It also contains a detailed description of the experiment results. They can be summarized as follows: As predicted, all low valuation buyers communicated their types truthfully, just as in T1. For the states with high valuation buyers the locally externality-free mechanism has less deviations from truth-telling than the mechanism in Example 1. The difference is significantly different from zero for the states

with a low type seller (two-sided Fisher's exact test, $p = 0.033$). It therefore also generates higher expected seller profits ($\Pi^{f''} = 4.90$) than both the mechanism in Example 1 ($\Pi^{f} = 4.82$) and the globally externality-free mechanism in Example 2 ($\Pi^{f'} = 4.77$). Both welfare comparisons are statistically significant (two-sided t-test, $p_{T1\,vs.\,T3} = 0.037$ and $p_{T2\,vs.\,T3} < 0.001$).

7 Redistributive income taxation

We now turn to another application of mechanism design, namely redistributive income taxation. Our motivation for looking at this is twofold: First, for this application, what can be achieved with an externality-free mechanisms has a natural interpretation: Such allocations can be decentralized by means of a non-linear income tax schedule. This raises the question how these "natural" mechanisms perform relative to ones that are not externality-free and predicted to generate more welfare if all individuals are selfish. Second, this case serves as an important robustness check for our experiment results, and at the same time as a proof of concept for our notion of externality-free mechanisms. For the bilateral trade application our externality-free mechanism better controlled behavior, but failed to generate monetary gains because there were not enough socially-motivated buyers to exceed the threshold of 38 %. However, as we explain below, the fraction of socially-motivated buyers observed in the bilateral trade application (14 %) would be sufficient to make the externality-free mechanism profitable in our taxation example. That is, with our taxation example we test whether social preferences are stable across mechanism design applications (or whether a context-dependent approach is required), and whether in this specific case the externality-free mechanism outperforms the optimal mechanism under selfish preferences.

As in our analysis of the bilateral trade problem, we consider an economy with two individuals, $I = \{1, 2\}$. Individual i derives utility from private goods consumption, or after-tax-income, c_i, and dislikes productive effort. Individual i's productive effort is measured by $\frac{y_i}{\omega_i}$, where y_i denotes the individual's contribution to the economy's output, or pre-tax-income, and ω_i is a measure of the individual's productive abilities. Thus, an individual with high productive abilities can generate a given level of output with less effort than an individual with low productive abilities. We assume that individual preferences can be represented by an additively separable utility function $u(c_i) - v\left(\frac{y_i}{\omega_i}\right)$, where u is an increasing and concave function, and v is an increasing and convex function. Both functions are assumed to satisfy the usual Inada conditions. Note that the individuals' preferences satisfy the single-crossing property: For any point in a (y, c)-diagram the indifference curve of an individual with low abilities through this point is steeper than the indifference curve of an individual with high abilities.

We assume that ω_i is the realization of a random variable that is privately observed by individual i. This random variable either takes a high value, ω_h, or a low value, ω_l. A state of the economy is a pair $\omega = (\omega_1, \omega_2)$ which specifies the productive ability of individual 1 and the productive ability of individual 2. The set of states is equal to $\{\omega_l, \omega_h\}^2$. A social choice function or direct mechanism consists of functions $c_i : \{\omega_l, \omega_h\}^2 \to \mathbb{R}_+$ and $y_i : \{\omega_l, \omega_h\}^2 \to \mathbb{R}_+$ which specify, for each state of the economy, and for each individual, a consumption and an output level.

An important benchmark is the first-best utilitarian welfare optimum. This is the social choice function which is obtained by choosing, separately for each state ω, $c_1(\omega)$, $c_2(\omega)$, $y_1(\omega)$ and $y_2(\omega)$ so as to maximize the sum of utilities,

$$u(c_1(\omega)) - v\left(\frac{y_1(\omega)}{\omega_1}\right) + u(c_2(\omega)) - v\left(\frac{y_2(\omega)}{\omega_2}\right) ,$$

subject to the economy's resource constraint,

$$c_1(\omega) + c_2(\omega) \leq y_1(\omega) + y_2(\omega) .$$

For a state where one individual is high-skilled and one is low-skilled this has the following implication: Both individuals get the same consumption level because marginal consumption utilities ought to be equalized. However, the high-skilled individual has to deliver more output than the low-skilled individual because marginal costs of effort ought to be equalized as well.

It will prove useful to have specific notation which refers to the first-best utilitarian welfare maximum for an economy with one highly productive and one less productive individual. The former is assigned an income requirement of y_h^* and a consumption level of c_h^*. The latter gets a lower income requirement, denoted by y_l^*, but receives the same consumption level $c_l^* = c_h^*$.

This social choice function raises questions of incentive compatibility. Clearly, the high-skilled individual would prefer the outcome intended for the low-skilled individual since the latter has the same consumption level but a smaller workload. As we will describe in the following, whether or not the first-best utilitarian welfare optimum can be reached in the presence of private information on productive abilities depends on the economy's information structure and on whether or not we impose a condition of externality-freeness.

Information structure. We assume that it is commonly known that, with probability 1, one individual is high-skilled and one individual is low-skilled.[23] That is to say, only the states (ω_l, ω_h) and (ω_h, ω_l) have positive probability, whereas the states (ω_l, ω_l) and (ω_h, ω_h) have probability zero. This implies that any one individual knows the other individual's type: If individual 1 observes that the own productive ability is high, then she can infer that the productive ability of individual 2 is low and vice versa. Put differently, each possible state of the economy gives rise to a complete information type space, with the mechanism designer as the only uninformed party.

The Mirrleesian approach. A Mirrleesian analysis imposes externality-freeness and anonymity. Externality-freeness requires that the outcome for any one individual depends only on that individual's productive ability and not on the productive ability of the other person. Anonymity requires that these outcomes are identical across individuals, so that e.g., the outcome specified for person 1 in case that $\omega_1 = \omega_l$, equals the outcome specified for person 2 in case that $\omega_2 = \omega_l$.

[23] This setup is due to Piketty (1993). We investigate the Mirrleesian approach under the same information structure.

Consequently, a social choice function can be represented by two bundles (y_l, c_l) and (y_h, c_h) so that, for all i,

$$(y_i(\omega), c_i(\omega)) = \begin{cases} (y_l, c_l) & \text{whenever} \quad \omega_i = \omega_l \ , \\ (y_h, c_h) & \text{whenever} \quad \omega_i = \omega_h \ . \end{cases}$$

Incentive compatibility then requires that an individual with low productive ability prefers (y_l, c_l) over (y_h, c_h), and that an individual with high productive ability prefers (y_h, c_h) over (y_l, c_l).[24] Formally,

$$u(c_l) - v\left(\frac{y_l}{w_l}\right) \geq u(c_h) - v\left(\frac{y_h}{w_l}\right) \quad \text{and} \quad u(c_h) - v\left(\frac{y_h}{w_h}\right) \geq u(c_l) - v\left(\frac{y_l}{w_h}\right) \ . \tag{10}$$

Obviously, these Mirrleesian incentive constraints are violated by the first-best utilitarian welfare maximum. An optimal Mirrleesian allocation is obtained by choosing (c_l, y_l) and (c_h, y_h) so as to maximize the sum of utilities

$$u(c_l) - v\left(\frac{y_l}{\omega_l}\right) + u(c_h) - v\left(\frac{y_h}{w_h}\right) \ ,$$

subject to the incentive constraints in (10) and the resource constraint $c_l + c_h \leq y_l + y_h$.

Piketty's approach. Piketty (1993) constructs a mechanism which achieves the first-best utilitarian outcome in dominant strategies. This mechanism is anonymous, but not externality-free. The construction is illustrated in Figure 1. In this Figure, point A is the outcome for any one individual if it reports ω_l and the other individual reports ω_h. Point B is the outcome for an individual that reports ω_h if the other individual reports ω_l. Point C is the outcome for an individual that reports ω_h if the other individual also reports ω_h. Analogously, D is the outcome for an individual that reports ω_l if the other individual also reports ω_l. It can easily be verified that truth-telling is a dominant strategy for selfish individuals if (i) point C lies above point A and between the two individuals' indifference curves through A, and (ii) point D lies below point B and between the two individuals' indifference curves through B. Also note that this is incompatible with externality-freeness which would require that $A = D$ and $B = C$.

[24] According to the Taxation Principle, see Hammond (1979) and Guesnerie (1995), these incentive constraints are equivalent to the possibility to reach a social choice function by specifying a tax schedule $T : y \mapsto T(y)$ so that any one individual i chooses c_i and y_i so as to maximize utility subject to the constraint that $c_i \leq y_i - T(y_i)$.

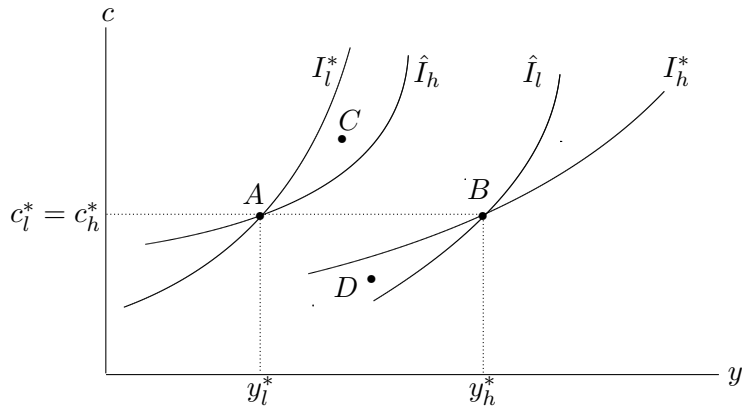


Figure 1. The Figure illustrates how the first-best utilitarian welfare maximum can be achieved with a mechanism that is not externality-free. I_l^* is the less able individual's indifference curve through $A = (y_l^*, c_l^*)$. Analogously, I_h^* is the more able individual's indifference curve trough $B = (y_h^*, c_h^*)$. The less able individual's indifference curve through B is denoted by $\hat{I}_l$, and the more able individual's indifference curve through A is denoted by $\hat{I}_h$.

Social preferences. Models of social preferences can rationalize deviations from this dominant strategy equilibrium. Consider first a model with intentions, such as Rabin (1993). The high-skilled individual might reason in the following way: The other individual could have reported a high ability type, in which case I would have gotten point C. This would have been good for me. So, the other individual is unkind since she did not make use of this possibility to increase my payoff. I am therefore willing to give up own payoff, so as to reciprocally harm the other individual. So, I should declare to be of the low ability type. In this case we both get D. This clearly harms myself and the other person. However, the point D is not that much worse for me, so the possibility to harm the other person is worth the sacrifice. Alternatively, we may consider a model with inequity aversion such as the Fehr-Schmidt-model. In this case, the same deviation could be rationalized by the observation that if both get D, their outcomes are equal, whereas they are very unequal in the dominant strategy equilibrium. Again, if point D is sufficiently close to B achieving this gain in equity is not too costly for an individual with high ability. With the Mirrleesian approach, by contrast, models of social preferences would predict truthful behavior. Since the Mirrleesian mechanism is externality-free, Proposition 2 implies that it is social-preference-robust.

An experiment. In the following we report on a laboratory experiment so as to check whether Piketty's approach does indeed provoke more deviations from truth-telling, and, if, yes, what this implies for the levels of utilitarian welfare that are generated by the two mechanisms. The experiment was based on functional form assumptions and parameter choices that are detailed in the following example.

Example 4. We impose the following functional form assumption on preferences:

$$U_i = u(c_i) - v\left(\frac{y_i}{\omega_i}\right) = \sqrt{c_i} - \frac{1}{2}\left(\frac{y_i}{w_i}\right)^2 .$$

In addition, we let $\omega_l = 4$ and $\omega_h = 6$. Under these assumptions, the optimal Mirrleesian allocation is given by $(c_l^M, y_l^M) = (3.45, 2.47)$ and $(c_h^M, y_h^M) = (6.23, 7.21)$. The normal form game that is induced by the Mirrleesian mechanism on a complete information type space so that individual 1 is of low ability and individual 2 is of high ability is summarized in the following table. The entries in the matrix are the players' utility levels under the assumption of selfish preferences.

Table 6: The game induced by the Mirrleesian mechanism for $(\omega_1, \omega_2) = (\omega_l, \omega_h)$.

(U_1, U_2)	ω_l	ω_h
ω_l	$(3.26, 3.70)$	$(3.26, 3.72)$
ω_h	$(1.99, 3.70)$	$(1.99, 3.72)$

To see that incentive compatibility holds note that first that player 1 does not benefit from claiming to be of high ability if player 2 behaves truthfully. His payoff would drop from 3.26 to 1.99. Analogously, if player 1 behaves truthfully, player 2 does not benefit from understating her ability, her payoff would drop from 3.72 to 3.70. In addition, externality-freeness holds: If player 1 communicates her low type truthfully, then she gets a payoff of 3.26 irrespectively of whether player 2 communicates a high or a low type. Also, if player 2 reveals his high type, he gets 3.72 irrespectively of whether player 1 communicates a high or a low valuation.

Piketty's mechanism is characterized by four points A, B, C and D, as illustrated in Figure 2. Points A and B coincide with the first-best utilitarian welfare maximum, so that

$$A = (c_l^*, y_l^*) = (5.53, 3.40) \quad \text{and} \quad B = (c_h^*, y_h^*) = (5.53, 7.66) \, .$$

There is a degree of freedom for the location of the points C and D. To have a completely specified example we need to determine these points in a specific way. We do this so as to capture the desire for welfare-maximizing redistribution which is the basic premise of an analysis of optimal income tax systems. In particular, suppose that there is a small probability, possibly zero, that both types have low abilities. In this case truth-telling of both individuals yields point D. Also suppose that there is an equally small probability that both types have high abilities, which would yield point C. We now allow for the possibility to redistribute resources away from the lucky state in which everybody is of high ability to the unlucky state in which everybody is of low ability. Moreover, we maximize this level of redistribution subject to the constraint of satisfying the principles of Piketty's construction. More formally, we choose point $C = (y^C, c^C)$ so that we extract a maximal tax payment subject to the constraint that C lies above point A and between the two relevant indifference curves through A.[25] We then choose point $D = (y^D, c^D)$ so as to maximize $u(c^D) - v\left(\frac{y^D}{w_l}\right)$ subject to the constraint that $c^D - y^D = y^C - c^C$ and subject to the requirement that point D lies below point B and between the two relevant indifference

[25] Formally, it is obtained as a solution to the following problem: Maximize $y^C - c^C$, s.t.

$$u(c^C) - v\left(\frac{y^C}{w_h}\right) \geq u(c^A) - v\left(\frac{y^A}{w_h}\right) \quad \text{and} \quad u(c^C) - v\left(\frac{y^C}{w_l}\right) \leq u(c^A) - v\left(\frac{y^A}{w_l}\right) \, .$$

curves through B. This construction is illustrated in Figure 3. It yields the following numerical values

$$C = (y^C, c^C) = (6.45, 7.77) \quad \text{and} \quad D = (y^D, c^D) = (4.62, 3.30)\ .$$

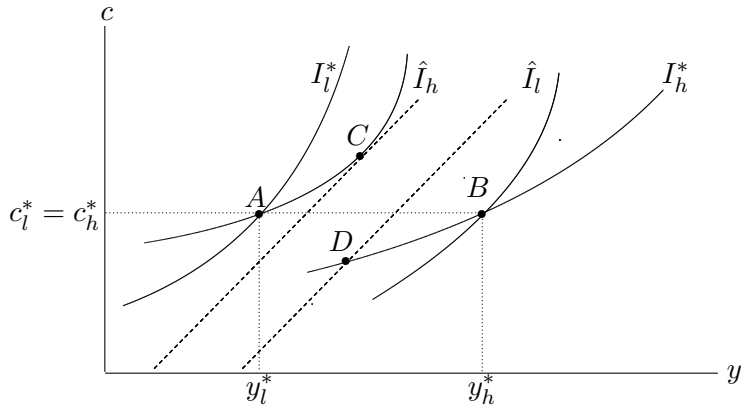


Figure 2. The Figure illustrates a specific Piketty mechanism. Point C is chosen so as to extract maximal tax revenues which yields the tangency condition that is shown in the Figure. These tax revenues are then used to make point D as attractive as possible, so that D is determined by the intersection of indifference curve I_h^* and a line with slope 1 and intercept $y^C - c^C$.

The normal form game that is induced by this version of a Piketty mechanism on a complete information type space, so that individual 1 is of low ability and individual 2 is of high ability, is summarized in Table 7. Again, the entries in the matrix are the players utility levels under the assumption of selfish preferences.

Table 7: The game induced by the Piketty mechanism in Figure 2 for $(\omega_1, \omega_2) = (\omega_l, \omega_h)$.

(U_1, U_2)	ω_l	ω_h
ω_l	$(2.32, 3.06)$	$(3.98, 3.08)$
ω_h	$(1.04, 4.38)$	$(2.94, 4.40)$

Truth-telling is a dominant strategy equilibrium under the assumption of selfish preferences. Externality-freeness is violated: If player 1 truthfully communicates a low ability type, his payoff depends on what player 2 communicates. Likewise, if player 2 communicates her high ability type truthfully, then her payoff depends on the type declared by player 1.

We conducted two laboratory treatments, one for the Mirrleesian approach and one for Piketty's approach.[26] As expected, we find hardly any deviations from truth-telling with the Mirrleesian approach: 124 of 126 low skilled individuals and 122 of 126 high skilled individuals truthfully report their ability. With Piketty's approach we also find almost no deviations from truth-telling for low skilled individuals, 121 of 126 reports are truthful. This changes with high skilled individuals in Piketty's approach where we observed 21 of 126 individuals to understate

[26] In Piketty's approach only states with a low and a high skilled individual have a positive probability in theory. Despite this, we asked subjects to report actions for all four skill combinations in order to use the same procedures as in our other treatments. The results reported in this section are based on the two states with positive probability. The full experiment data can be found in Appendix D.

their skill level. This is a significantly larger proportion of deviations than with the Mirrleesian approach (two-sided Fisher's exact test, $p<0.001$). As a result, the Mirrleesian approach reaches an average welfare level that is with 6.93 significantly larger than the average welfare level of 6.78 which results from Piketty's approach (two-sided t-test, $p = 0.014$).

At first glance, our results seem to suggest that insisting on externality-freeness is a good idea for a problem of income taxation, but a bad idea for the bilateral trade problem. Yet, in fact, social behavior is robust across applications: The fraction of individuals who deviated from selfish behavior was 14 % in our bilateral trade application and with 17 % for the income tax application not significantly different.[27] These numbers are clearly below the corresponding threshold for the profitability of the externality-free mechanism in bilateral trade (34 %), yet clearly above the threshold in taxation (5 %). We conclude that behavior is robust against our different mechanism design contexts, but the mechanisms systematically differ in their robustness towards social behavior.

8 Concluding remarks

This paper shows how social preferences can be taken into account in robust mechanism design. We have first characterized optimal mechanisms for a bilateral trade problem and a problem of redistributive income taxation under selfish preferences. We have argued theoretically that such a mechanism will not generally produce the desired behavior if individuals have social preferences, and we have illustrated in a laboratory experiment that deviations from the intended behavior indeed occur. We have then introduced 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. We have finally confirmed in a series of experiments, taking other assumptions in mechanism design for granted (see below), that an externality-free mechanism does indeed generate the intended behavior.

We also investigated under which conditions externality-freeness improves the performance of a mechanism. Our specification of the bilateral trade problem was such that, to justify externality-freeness, many deviations from selfish behavior were required. By contrast, for our income tax application, a small number of deviations was sufficient. We found that the fraction of deviating individuals was the same across applications, and moreover that this number was high enough to make externality-freeness desirable for the income tax application, but not high enough to make it desirable for the bilateral trade problem.

Externality-freeness is a sufficient but not a necessary condition for the ability to predict 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. It is not guaranteed, however, that externality-freeness also improves the performance of a mechanism. An alternative to imposing externality-freeness is a mechanism design approach that elicits not only the monetary payoffs

[27] The difference between these two fractions was not found to be statistically different from zero (two-sided Fisher's exact test, $p = 0.728$).

of individuals but also the precise functional form of their social preferences. 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.

As an alternative to such an axiomatic approach one might simply try to identify the relevant deviations from selfish behavior empirically, e.g., with a laboratory experiment, and then impose externality-freeness conditions only locally so as to eliminate the specific deviations that pose problems for the mechanism design problem at hand. This approach has the advantage that it does not impose as many additional constraints on the mechanism design problem. The disadvantage is that it does not eliminate all the deviations from selfish behavior that can be rationalized by models of social preferences. Thus, it is not as robust as an externality-free mechanism. We demonstrated the attractiveness of such an engineering approach in the context of the bilateral trade problem. Imposing externality-freeness only locally enabled us to find a mechanism that outperformed both an optimal mechanism for selfish agents and an optimal externality-free mechanism.

Adding behavioral aspects to the mechanism design literature is a promising line of research. That said, we caution that our study cannot, of course, capture all behavioral aspects that seem relevant. For instance, our experiments do not shed light on social preference robustness with incomplete information about monetary payoffs. As a first step, we rather take the theoretically predicted equivalence of implementability in all complete information environments and implementability in all incomplete information environments, as well as the revelation principle, for granted. This way, we can focus on the role of social preferences under certainty in mechanism design, abstracting away from other potential influencing behavioral factors which may arise in cognitively and socially more demanding environments. For instance, recent evidence and theory suggest 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), Bolton et al. (2015), Saito (2013), Ockenfels et al. (2014)). The implications of such findings for robust mechanism design need further attention. By the same token, our approach leaves open the question whether we can generate the behavior that is needed to implement a given social choice function also with an indirect mechanism, which may be empirically more plausible, than a direct revelation mechanism, e.g. one that simply asks individuals whether they are willing to trade at particular prices. These are fundamental questions, and their answers likely generate more important insights on how motivational and cognitive forces affect the behavioral effectiveness and efficiency of economic mechanisms. We are planning to check robustness along those lines in separate studies.

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. (2014). An externality-robust auction: Theory and experimental evidence. Available at SSRN: http://ssrn.com/abstract=2359529.

Becker, G. S. (1974). A theory of social interactions. *Journal of Political Economy*, 82(6):pp. 1063–1093.

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.

Bierbrauer, F. and Netzer, N. (2016). Mechanism design and intentions. *Journal of Economic Theory, forthcoming.*

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, G., Stauf, J., and Ockenfels, A. (2015). Social responsibility promotes conservative risk behavior. *European Economic Review*, 74:109–127.

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

Börgers, T. (2015). *An introduction to the theory of mechanism design.* Oxford University Press Inc.

Camerer, C. (1995). Individual decision making. In Kagel, J. and Roth, A., editors, *Handbook of Experimental Economics*, chapter 8, pages 587–683. Princeton University Press.

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. (2013). *Other-Regarding preferences: A selective survey of experimental results.* Princton University Press.

Crémer, J. and McLean, R. (1988). Full extraction of the surplus in bayesian and dominant strategy auctions. *Econometrica*, 56:1247–1257.

Crémer, J. and McLean, R. P. (1985). Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. *Econometrica*, 53(2):pp. 345–361.

Dierker, E. and Haller, H. (1990). Tax systems and direct mechanisms in large finite economies. *Journal of Economics*, 52:99–116.

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.

Guesnerie, R. (1995). *A Contribution to the Pure Theory of Taxation.* Cambridge University Press.

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 and Organization*, 3:367–388.

Hagerty, K. and Rogerson, W. (1987). Robust trading mechanisms. *Journal of Economic Theory*, 42:94–107.

Hammond, P. (1979). Straightforward individual incentive compatibility in large economies. *Review of Economic Studies*, 46:263–282.

Jehiel, P. and Moldovanu, B. (2006). Allocative and informational externalities in auctions and related mechanisms. In Blundell, R., Newey, W., and Persson, T., editors, *Proceedings of the 9th World Congress of the Econometric Society*.

Kagel, J., Lien, Y., and Milgrom, P. (2010). Ascending prices and package bidding: A thoretical 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 partership dissolution mechanisms: Theory and lab evidence. *Economics Letters*, 117:394–396.

Kosenok, G. and Severinov, S. (2008). Individually rational, budget-balanced mechanisms and allocation of surplus. *Journal of Economic Theory*, 140(1):126 – 161.

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.

Mirrlees, J. (1971). An exploration in the theory of optimum income taxation. *Review of Economic Studies*, 38:175–208.

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. (2014). Bonus payments and reference point violations. *Management Science*.

Piketty, T. (1993). Implementation of first-best allocations via generalized tax schedules. *Journal of Economic Theory*, 61:23–41.

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. (2002). The economist as engineer: Game theory, experimentation, and computation as tool for design economics. *Econometrica*, 70:1341–1378.

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.

Sobel, J. (2005). Interdependent preferences and reciprocity. *Journal of Economic Literature*, 43:392–436.

Stiglitz, J. (1982). Self-selection and Pareto-efficient taxation. *Journal of Public Economics*, 17:213–240.

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 *Advances in Economic Theory*, pages 33–70. Cambridge University Press. Cambridge Books Online.

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). 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, however 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 (6). 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,

$$\kappa_i(r_i, r_i^b, r_i^{bb}) = \pi_j(r_i, r_i^b) - \pi_j^{e_i}(r_i^b). \tag{11}$$

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.

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

where $E_{ij}(r_i)$ is the set of Pareto-efficient reports: A report r_i belongs to $E_{ij}(r_i^b)$ 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,

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

For later reference, it is useful to note that by equation (11), $\kappa_i(r_i, r_i^b, r_i^{bb})$ does not explicitly depend on r_i^{bb}. In the context of Rabin's model we can therefore simplify notation and write $\kappa_i(r_i, r_i^b)$ rather than $\kappa_i(r_i, r_i^b, r_i^{bb})$.

Observation 3. *Let f be a social choice function that solves a problem of optimal robust mechanism design as defined in Section 3.2. 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*

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

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 (12). 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 (8), 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 3.2. 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 (12) 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 (12). 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 (8), 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 parametrization 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 (6) 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 (6). The kindness intended by player i is now given as

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

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

$$\kappa_j(r_i^b, r_i^{bb}) \leq 0 \ , \tag{13}$$

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 (13) 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 Externality-freeness as a necessary condition.

Proposition 6 below states a condition under which externality-freeness and incentive-compatibility are not only sufficient, but also necessary for social-preference robustness. To prove Proposition 6 we focus on a specific model of social preferences, namely the one by Rabin (1993), and work with the solution concept of a *fairness equilibrium* that has been introduced in that paper. We require that a social choice function is robustly implementable as a fairness equilibrium, i.e. we require that there is a mechanism that reaches this social function on every complete information type space, and for each possible specification of the weights y_1 and y_2 that players 1 and 2 assign to kindness in their overall utility function in (6). We provide necessary conditions for robust implementability as a fairness equilibrium. Robust implementability as a fairness equilibrium is in turn a necessary condition for social-preference-robustness.

Robust implementability as a fairness equilibrium. There are two agents $I = \{1, 2\}$. We seek to implement a social choice function $f : \Theta_1 \times \Theta_2 \to X$, where X is an abstract set of economic outcomes. Thus, given a profile of preferences parameters (θ_1, θ_2), the material payoff for agent 1 is denoted by $\pi_1(\theta_1, f(\theta_1, \theta_2))$ and the material payoff for agent 2 by $\pi_2(\theta_2, f(\theta_1, \theta_2))$.

In the context of Rabins's model of social preferences, the validity of the revelation principle cannot be taken for granted, see Bierbrauer and Netzer (2016). We therefore consider the implementation of the social choice function f by means of an arbitrary allocation mechanism that consists of a set of reports R_1, with typical entry r_1, for player 1, a set of reports R_2, with typical entry r_2, for player 2 and an outcome function $g : R_1 \times R_2 \to X$ that assigns an economic outcome to each profile of reports.

The utility that individual i realizes can be written as

$$U_i(r_i \mid \theta_i, y_i, r_i^b, r_i^{bb}) = \pi_i(\theta_i, f(r_i, r_i^b)) + y_i\, \kappa_i(r_i, r_i^b)\, \kappa_j(r_i^b, r_i^{bb}) \ , \tag{14}$$

where y_i is the weight that agent i assigns to kindness sensations. This notation emphasizes that individual i chooses r_i and that θ_i, y_i, r_i^b and r_i^{bb} are parameters that enter individual i's utility function. The weights y_1 and y_2 take values in the sets $\mathbb{R}_{0+}$. In the special case with $y_1 = y_2 = 0$, both individuals are selfish.

We consider complete information type spaces where both the profile of preference parameters (θ_1, θ_2) and the kindness weights (y_1, y_2) are commonly known among the individuals. A mechanism $\mathcal{M} = [R_1, R_2, g]$ implements a social choice function f on such a type space if there exist reports $r_1^*(\theta_1, y_1)$ and $r_2^*(\theta_2, y_2)$ such that (i) the social choice function is reached, i.e.

$$g(r_1^*(\theta_1, y_1), r_2^*(\theta_2, y_2)) = f(\theta_1, \theta_2) \tag{15}$$

and (ii) given correct first and second order beliefs, $r_1^*(\theta_1, y_1)$ is the utility-maximizing report for player 1 and $r_2^*(\theta_2, y_2))$ is the utility-maximizing report for player 2. More formally, for all i and $j \neq i$,

$$r_i^*(\theta_i, y_i) \quad \in \quad \text{argmax}_{r_i \in R_i} \quad U_i(r_i \mid \theta_i, y_i, r_j^*(\theta_j, y_j)), r_i^*(\theta_i, y_i)) \ . \tag{16}$$

We say that f is robustly implementable as a fairness equilibrium if there exists a mechanism $\mathcal{M} = [R_1, R_2, g]$, and a pair of functions $r_1^* : \Theta_1 \times Y_1 \to R_1$ and $r_2^* : \Theta_2 \times Y_2 \to R_2$ so that (15) and (16) hold for all $(\theta_1, y_1) \in \Theta_1 \times \mathbb{R}_{0+}$ and all $(\theta_2, y_2) \in \Theta_2 \times \mathbb{R}_{0+}$.

A necessary condition. Consider a specific violation of externality-freeness so that player 1 has an influence on the payoff of player 2, and player 2 has a chance to lower the payoff of player 1. Part B of Proposition 6 below asserts that, if a social choice function violates externality-freeness in this specific way, then it is not robustly implementable as a fairness equilibrium. The specific violation of externality-freeness covers, in particular, environments with two types per player. With a more general structure of type spaces, a social choice function f violates externality-freeness in this way as soon as there is a type profile (θ_1, θ_2) so that, for both players, truth-telling is neither entirely selfish, nor entirely selfless, i.e. as soon as there exist (θ_1', θ_2') and (θ_1'', θ_2'') such that

$$\pi_2(\theta_2, f(\theta_1', \theta_2)) < \pi_2(\theta_2, f(\theta_1, \theta_2)) < \pi_2(\theta_2, f(\theta_1'', \theta_2))$$

and

$$\pi_1(\theta_1, f(\theta_1, \theta_2')) < \pi_1(\theta_1, f(\theta_1, \theta_2)) < \pi_1(\theta_1, f(\theta_1, \theta_2'')) \ .$$

Externality-freeness, by contrast, requires that, for all i and all θ_j,

$$\min_{\theta_i \in \Theta_i} \ \pi_j(\theta_j, f(\theta_i, \theta_j)) = \max_{\theta_i \in \Theta_i} \ \pi_j(\theta_j, f(\theta_i, \theta_j)) \ .$$

Definition 2. *We say that social choice function f violates externality-freeness in a specific way if there is a complete information type space (θ_1, θ_2) and a pair of alternative types (θ_1', θ_2') such that $\pi_2(\theta_2, f(\theta_1, \theta_2)) < \pi_2(\theta_2, f(\theta_1', \theta_2))$ and $\pi_1(\theta_1, f(\theta_1, \theta_2)) > \pi_1(\theta_1, f(\theta_1, \theta_2'))$.*

Proposition 6.

A. If f is robustly implementable as a fairness equilibrium, then f is incentive compatible.

B. If f violates externality-freeness in a specific way, then f is not robustly implementable as a fairness equilibrium.

Proof of Proposition 6. *A.* We first show that robust implementability of f as a fairness equilibrium implies that f is incentive-compatible. Let $y_i = 0$, then implementability requires that

$$r_i^*(\theta_i, 0) \quad \in \quad \text{argmax}_{r_i \in R_i} \quad \pi_i(\theta_i, g(r_i, r_j^*(\theta_j, y_j))) \; . \tag{17}$$

for all $\theta_i \in \Theta_i$ and all $(\theta_j, y_j) \in \Theta_j \times \mathbb{R}_{0+}$. In particular, this implies that for all θ_i, all (θ_i', y_i') and all (θ_j, y_j),

$$\pi_i(\theta_i, g(r_i^*(\theta_i, 0), r_j^*(\theta_j, y_j))) \geq \pi_i(\theta_i, g(r_i^*(\theta_i', y_i'), r_j^*(\theta_j, y_j))) \; .$$

Because of (15) this implies that, for all θ_i, all θ_i' and all θ_j,

$$\pi_i(\theta_i, f(\theta_i, \theta_j)) \geq \pi_i(\theta_i, f(\theta_i', \theta_j)) \; .$$

Thus, f is incentive compatible.

B. Let f be a social choice function that violates externality-freeness in a specific way. We will show that this implies that there is a threshold $\hat{y}_2$ so that conditions (15) and (16) are incompatible whenever $y_2 \geq \hat{y}_2$. To establish this claim we have to go through a number of intermediate steps.

Step 1. We show that conditions (15) and (16) imply that every type of every player behaves selfishly, i.e. that for all i, all (θ_i, y_i) and all (θ_j, y_j),

$$r_i^*(\theta_i, y_i) \quad \in \quad \text{argmax}_{r_i \in R_i} \quad \pi_i(\theta_i, g(r_i, r_j^*(\theta_j, y_j))) \; , \tag{18}$$

and that, as a consequence, equilibrium kindness is bounded from above by 0, i.e. that

$$\kappa_i(r_i^*(\theta_i, y_i), r_j^*(\theta_j, y_j)) \leq 0 \; , \tag{19}$$

for all (θ_i, y_i) and (θ_j, y_j).

If (18) was violated for some type (θ_i, y_i) of player i, then this type could reach a higher material payoff by deviating from $r_i^*(\theta_i, y_i)$ to some other report. However, by (15) the payoff consequence of choosing message $r_i^*(\theta_i, y_i)$ is the same as the payoff consequence of choosing message $r_i^*(\theta_i, 0)$. Thus, if type (θ_i, y_i) can reach a higher a higher material payoff by deviating

from $r_i^*(\theta_i, y_i)$ then also type $(\theta_i, 0)$ can reach a higher payoff by deviating from $r_i^*(\theta_i, 0)$. But this would contradict (17).

Step 2. Fix a pair (y_1, y_2) and consider a complete information type space on which externality freeness is violated in the sense of Definition 2. If player 1 behaves according to r_1^*, then player 1's equilibrium kindness is strictly negative. To see this, note that by *Step 1*, player 1 behaves selfishly. Hence, he chooses the action that minimizes player 2's payoff from the set of Pareto-efficient action profiles $E_{12}(r_2^*(\theta_2, y_2))$. Hence,

$$\pi_2(\theta_2, f(\theta_1, \theta_2)) = \pi_2(\theta_2, g(r_1^*(\theta_1, y_1), r_2^*(\theta_2, y_2))) = \min_{r_1 \in E_{12}(r_2^*(\theta_2, y_2))} \pi_2(\theta_2, g(r_1, r_2^*(\theta_2, y_2))) \, .$$

By the specific violation of externality-freeness,

$$\min_{r_1 \in E_{12}(r_2^*(\theta_2, y_2))} \pi_2(\theta_2, g(r_1, r_2^*(\theta_2, y_2))) < \min_{r_1 \in E_{12}(r_2^*(\theta_2, y_2))} \pi_2(\theta_2, g(r_1, r_2^*(\theta_2, y_2))) \, ,$$

as player 1 could increase player 2's payoff be choosing action $r_1^*(\theta_1', y_1)$ rather than action $r_1^*(\theta_1, y_1)$. Consequently $\kappa_1(r_1^*(\theta_1, y_1), r_2^*(\theta_2, y_2) < 0$, for all (y_1, y_2).

Step 3. Consider the type profile (θ_1, θ_2) for which the specific violation of externality-freeness occurs and a hypothetical fairness equilibrium in which player 1 behaves according to $r_1^*(\theta_1, y_1)$ and player 2 behaves according to $r_2^*(\theta_2, y_2)$. Given correct first- and second-order beliefs, the best response problem for player 2 looks as follows: Choose $r_2 \in R_2$, so as to maximize

$$\pi_2(\theta_2, g(r_1^*(\theta_1, y_1), r_2)) + y_2 \, \kappa_1^* \, \pi_1(\theta_1, g(r_1^*(\theta_1, y_1), r_2))$$

where we omitted some constant terms from the objective function that do not affect the solution of the optimization problem, and $\kappa_1^* < 0$ is a shorthand for $\kappa_1(r_1^*(\theta_1, y_1), r_2^*(\theta_2, y_2))$, i.e. the kindness of player 1 in the hypothetical equilibrium.

Now, if player 2 behaves according to $r_2^*(\theta_2, y_2)$, then, by *Step 1*, this yields the maximal value of $\pi_2(\theta_2, g(r_1^*(\theta_1, y_1), r_2))$ over R_2. If player 2 behaves according to $r_2^*(\theta_2', y_2)$, then because of (15), this yields a lower value of $\pi_1(\theta_1, g(r_1^*(\theta_1, y_1), r_2))$ than behaving according to $r_2^*(\theta_2, y_2)$. Moreover, if y_2 is sufficiently large, overall utility will then be larger if the action $r_2^*(\theta_2', y_2)$ is taken. But this contradicts the best response condition in (16).

□

C Proofs

Proof of Proposition 1. The relaxed problem imposes only the buyer's *ex post* participation and incentive constraints, as well as the constraint that the expected payments to the seller are equal to the expected payments of the buyer, with expectations computed using the designer's subjective beliefs. Thus, the problem is to choose, for every state $(\theta_b, \theta_s) \in \Theta_b \times \Theta_s$, $q^f(\theta_b, \theta_s)$, $p_b^f(\theta_b, \theta_s)$ and $p_s^f(\theta_b, \theta_s)$ so as to maximize

$$\sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) \left(p_s^f(\theta_b, \theta_s) - \theta_s k(q^f(\theta_b, \theta_s)) \right)$$

subject to the following constraints: (i) the resource constraint

$$\sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) p_b^f(\theta_b, \theta_s) \geq \sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) p_s^f(\theta_b, \theta_s) \ , \tag{20}$$

(ii) the incentive and participation constraints for the buyer that are relevant if the seller is of the high cost type,

$$\underline{\theta}_b \ q^f(\underline{\theta}_b, \overline{\theta}_s) - p_b^f(\underline{\theta}_b, \overline{\theta}_s) \geq 0 \ , \tag{21}$$

$$\overline{\theta}_b \ q^f(\overline{\theta}_b, \overline{\theta}_s) - p_b^f(\overline{\theta}_b, \overline{\theta}_s) \geq 0 \ , \tag{22}$$

$$\underline{\theta}_b \ q^f(\underline{\theta}_b, \overline{\theta}_s) - p_b^f(\underline{\theta}_b, \overline{\theta}_s) \geq \underline{\theta}_b \ q^f(\overline{\theta}_b, \overline{\theta}_s) - p_b^f(\overline{\theta}_b, \overline{\theta}_s) \ , \tag{23}$$

and

$$\overline{\theta}_b \ q^f(\overline{\theta}_b, \overline{\theta}_s) - p_b^f(\overline{\theta}_b, \overline{\theta}_s) \geq \overline{\theta}_b \ q^f(\underline{\theta}_b, \overline{\theta}_s) - p_b^f(\underline{\theta}_b, \overline{\theta}_s) \ , \tag{24}$$

and finally (iii) the incentive and participation constraints for the buyer that are relevant if the seller is of the low cost type. These constraints have the same structure as those in (21)-(24), except that $\overline{\theta}_s$ is everywhere replaced by $\underline{\theta}_s$.

Obviously, the resource constraint will be binding, so that the objective becomes to maximize

$$\sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) \left(p_b^f(\theta_b, \theta_s) - \theta_s k(q^f(\theta_b, \theta_s)) \right)$$

subject to the constraints in (ii) and (iii). The solution can be obtained by solving a separate optimization problem for each seller type. Thus, optimality requires that $q^f(\underline{\theta}_b, \overline{\theta}_s)$, $q^f(\overline{\theta}_b, \overline{\theta}_s)$, $p_b^f(\underline{\theta}_b, \overline{\theta}_s)$, and $p_f^f(\overline{\theta}_b, \overline{\theta}_s)$ are chosen so as to maximize

$$\sum_{\Theta_b} g(\theta_b, \overline{\theta}_s) \left(p_b^f(\theta_b, \overline{\theta}_s) - \theta_s k(q^f(\theta_b, \overline{\theta}_s)) \right)$$

subject to the constraints in (ii); likewise $q^f(\underline{\theta}_b, \underline{\theta}_s)$, $q^f(\overline{\theta}_b, \underline{\theta}_s)$, $p_b^f(\underline{\theta}_b, \underline{\theta}_s)$, and $p_f^f(\overline{\theta}_b, \underline{\theta}_s)$ are chosen so as to maximize

$$\sum_{\Theta_b} g(\theta_b, \underline{\theta}_s) \left(p_b^f(\theta_b, \underline{\theta}_s) - \theta_s k(q^f(\theta_b, \underline{\theta}_s)) \right)$$

subject to the constraints in (iii).

The solution to these problems is well-known, see e.g., Bolton and Dewatripont (2005). Thus, at a solution, the high-valuation buyer's incentive constraint and the low-valuation buyer's participation constraints bind and the other constraints are slack. For example, if $\theta_s = \overline{\theta}_s$, then (21) and (24) bind, and (22) and (23) are not binding. The optimal quantities are then obtained by substituting

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

and

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

into the objective function which yields

$$g(\underline{\theta}_b, \overline{\theta}_s) \left(\left(\underline{\theta}_b - \tfrac{g(\overline{\theta}_b, \overline{\theta}_s)}{g(\underline{\theta}_b, \overline{\theta}_s)} (\overline{\theta}_b - \underline{\theta}_b) \right) q^f(\underline{\theta}_b, \overline{\theta}_s) - \overline{\theta}_s k(q^f(\underline{\theta}_b, \overline{\theta}_s)) \right)$$

$$+ g(\overline{\theta}_b, \overline{\theta}_s) \left(\overline{\theta}_b q^f(\overline{\theta}_b, \overline{\theta}_s) - \overline{\theta}_s k(q^f(\overline{\theta}_b, \overline{\theta}_s)) \right) .$$

Choosing $q^f(\underline{\theta}_b, \overline{\theta}_s)$ and $q^f(\overline{\theta}_b, \overline{\theta}_s)$ to maximize this expression yields the optimality conditions that are stated in Proposition 1 in the body of the text. □

Proof of Proposition 3. For the relaxed problem of optimal externality-free mechanism design the objective is, again, the maximization of

$$\sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) \left(p_s^f(\theta_b, \theta_s) - \theta_s k(q^f(\theta_b, \theta_s)) \right) .$$

The resource constraint in (20) is binding at a solution to this problem, so that the objective can be equivalently written as

$$\sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) \left(p_b^f(\theta_b, \theta_s) - \theta_s k(q^f(\theta_b, \theta_s)) \right)$$

The constraints are the low valuation buyer's *ex post* participation constraints,

$$\underline{\theta}_b q^f(\underline{\theta}_b, \underline{\theta}_s) - p_b^f(\underline{\theta}_b, \underline{\theta}_s) \geq 0 , \tag{25}$$

and

$$\underline{\theta}_b q^f(\underline{\theta}_b, \overline{\theta}_s) - p_b^f(\underline{\theta}_b, \overline{\theta}_s) \geq 0 ; \tag{26}$$

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

$$\overline{\theta}_b q^f(\overline{\theta}_b, \underline{\theta}_s) - p_b^f(\overline{\theta}_b, \underline{\theta}_s) \geq \overline{\theta}_b q^f(\underline{\theta}_b, \underline{\theta}_s) - p_b^f(\underline{\theta}_b, \underline{\theta}_s) , \tag{27}$$

and the constraint, that the seller must not be able to influence the high valuation buyer's payoff,

$$\overline{\theta}_b q^f(\overline{\theta}_b, \underline{\theta}_s) - p_b^f(\overline{\theta}_b, \underline{\theta}_s) = \overline{\theta}_b q^f(\overline{\theta}_b, \overline{\theta}_s) - p_b^f(\overline{\theta}_b, \overline{\theta}_s) . \tag{28}$$

Note first that the constraint in (26) has to bind at a solution to this problem. The payment $p_b^f(\underline{\theta}_b, \overline{\theta}_s)$ enters only in this constraint. Hence, if we hypothesize a solution to the optimization problem with slack in (26), we can raise $p_b^f(\underline{\theta}_b, \overline{\theta}_s)$ without violating any constraint, thereby arriving at a contradiction to the assumption that the initial situation has been an optimum.

Second, the constraint in (25) binds as well. Suppose otherwise, then it is possible to raise

$p_b^f(\underline{\theta}_b, \underline{\theta}_s)$ by some small $\varepsilon > 0$, without violating this constraints. If at the same time, $p_b^f(\overline{\theta}_b, \underline{\theta}_s)$ and $p_b^f(\overline{\theta}_b, \overline{\theta}_s)$ are also raised by ε, then also the constraints in (27) and (28) remain satisfied. These increases of the buyer's payments raise the objective function, again contradicting the assumption that the initial situation has been optimal.

Third, the constraint in (27) has to be binding. Otherwise, it would be possible to raise $p_b^f(\overline{\theta}_b, \underline{\theta}_s)$ without violating this constraint. If at the same time, $p_b^f(\overline{\theta}_b, \overline{\theta}_s)$ is raised by ε, then also (28) remains satisfied. One more time, this contradicts the assumption that the initial situation has been optimal.

These observations enables to express the buyer's payments as functions of the traded quantities, so that

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

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

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

and

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

Substituting these payments into the objective function yields

$$g(\underline{\theta}_b, \underline{\theta}_s) \left(\left(\underline{\theta}_b - \tfrac{g(\overline{\theta}_b, \underline{\theta}_s) + g(\overline{\theta}_b, \overline{\theta}_s)}{g(\underline{\theta}_b, \underline{\theta}_s)} (\overline{\theta}_b - \underline{\theta}_b) \right) q^f(\underline{\theta}_b, \underline{\theta}_s) - \underline{\theta}_s k(q^f(\underline{\theta}_b, \underline{\theta}_s)) \right)$$

$$+ g(\overline{\theta}_b, \underline{\theta}_s)(\overline{\theta}_b q^f(\overline{\theta}_b, \underline{\theta}_s) - \underline{\theta}_s k(q^f(\overline{\theta}_b, \underline{\theta}_s)))$$

$$+ g(\underline{\theta}_b, \overline{\theta}_s)(\underline{\theta}_b q^f(\underline{\theta}_b, \overline{\theta}_s) - \overline{\theta}_s k(q^f(\underline{\theta}_b, \overline{\theta}_s)))$$

$$+ g(\overline{\theta}_b, \overline{\theta}_s)(\overline{\theta}_b q^f(\overline{\theta}_b, \overline{\theta}_s) - \overline{\theta}_s k(q^f(\overline{\theta}_b, \overline{\theta}_s))) \ .$$

Choosing $q^f(\underline{\theta}_b, \underline{\theta}_s)$, $q^f(\overline{\theta}_b, \underline{\theta}_s)$, $q^f(\underline{\theta}_b, \overline{\theta}_s)$ and $q^f(\overline{\theta}_b, \overline{\theta}_s)$ so as to maximize this expression yields the optimality conditions stated in Proposition 3. □

Proof of Proposition 5. We need to show that there is a solution to the optimization problem in Proposition 1 that satisfies

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

and

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

or, equivalently,

$$p_s^f(\underline{\theta}_b, \underline{\theta}_s) - p_s^f(\overline{\theta}_b, \underline{\theta}_s) = \underline{\theta}_s k(q^f(\underline{\theta}_b, \underline{\theta}_s)) - \underline{\theta}_s k(q^f(\overline{\theta}_b, \underline{\theta}_s)) \ , \tag{29}$$

and

$$p_s^f(\underline{\theta}_b, \overline{\theta}_s) - p_s^f(\overline{\theta}_b, \overline{\theta}_s) = \overline{\theta}_s k(q^f(\underline{\theta}_b, \overline{\theta}_s)) - \underline{\theta}_s k(q^f(\overline{\theta}_b, \overline{\theta}_s)) \ . \tag{30}$$

The right-hand-side of equations (29) and (30) is pinned down by the characterization in Proposition 1. However, this solution leaves degrees of freedom with respect to the specification of the seller's payments. It only requires that the resource constraint binds which implies that

$$\sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) p_s^f(\theta_b, \theta_s) = \sum_{\Theta_b \times \Theta_s} g(\theta_b, \theta_s) p_b^f(\theta_b, \theta_s) \ . \tag{31}$$

Again the right-hand side of this equation is is pinned down by the characterization in Proposition 1. Thus, the four payments to the seller $p_s^f(\underline{\theta}_b, \underline{\theta}_s)$, $p_s^f(\overline{\theta}_b, \underline{\theta}_s)$, $p_s^f(\underline{\theta}_b, \overline{\theta}_s)$ and $p_s^f(\overline{\theta}_b, \overline{\theta}_s)$ need to satisfy the three linear equations in (29), (30) and (31). Obviously, there will be more than one combination of payments to the seller that satisfy all of these conditions. □

D 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.

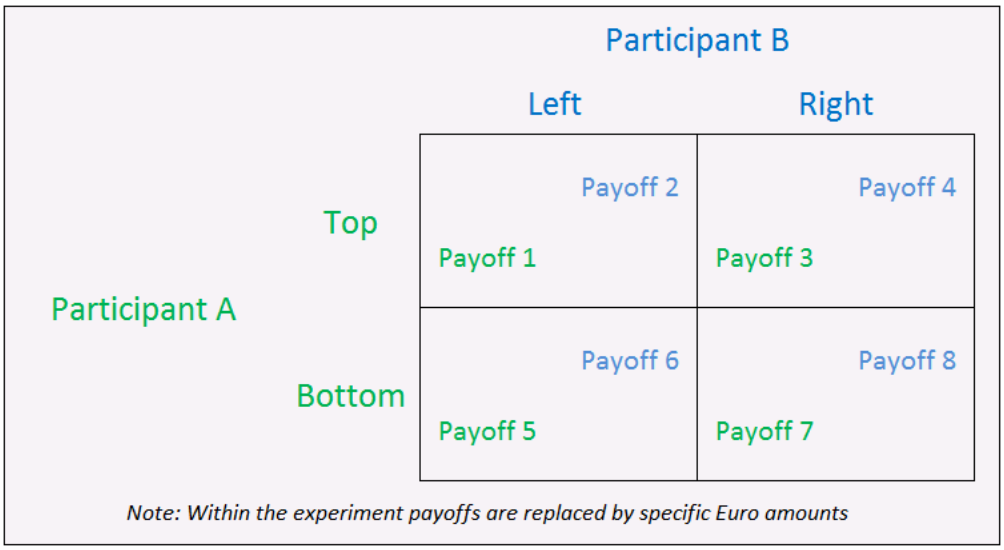
On the following page you will find the instructions concerning the course of the . After reading these, we ask you to wait at your seat until the experiment starts.

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.]



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!

E Supplementary material

E.1 The experiment reported on in Section 6

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

(π_b, π_s)	$\underline{\theta}_s$	$\overline{\theta}_s$
$\underline{\theta}_b$	$(2.68, 6.09)$	$(2.68, 4.05)$
$\overline{\theta}_b$	$(0.97, 6.09)$	$(2.66, 4.86)$

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

(π_b, π_s)	$\underline{\theta}_s$	$\overline{\theta}_s$
$\underline{\theta}_b$	$(2.68, 2.65)$	$(2.68, 3.73)$
$\overline{\theta}_b$	$(0.97, -5.79)$	$(2.66, 3.73)$

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

(π_b, π_s)	$\underline{\theta}_s$	$\overline{\theta}_s$
$\underline{\theta}_b$	$(3.41, 6.09)$	$(3.24, 4.05)$
$\overline{\theta}_b$	$(3.43, 6.09)$	$(3.43, 4.86)$

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

(π_b, π_s)	$\underline{\theta}_s$	$\overline{\theta}_s$
$\underline{\theta}_b$	$(3.41, 2.65)$	$(3.24, 3.73)$
$\overline{\theta}_b$	$(3.43, -5.79)$	$(3.43, 3.73)$

E.2 Choice data T3

		Buyer		*Seller*	
	Game induced by	$\underline{\theta}_b$	$\overline{\theta}_b$	$\underline{\theta}_s$	$\overline{\theta}_s$
T3 *locally externality-free mechanism*	f'' for $(\underline{\theta}_b, \underline{\theta}_s)$	63	0	62	1
	f'' for $(\underline{\theta}_b, \overline{\theta}_s)$	63	0	0	63
	f'' for $(\overline{\theta}_b, \underline{\theta}_s)$	1	62	63	0
	f'' for $(\overline{\theta}_b, \overline{\theta}_s)$	7	56	0	63

E.3 Normal form games which are induced by the Mirrleesian mechanism

The game induced by the Mirrleesian mechanism for $(\omega_1, \omega_2) = (\omega_l, \omega_l)$.

(U_1, U_2)	ω_l	ω_h
ω_l	$(3.26, 3.26)$	$(3.26, 1.99)$
ω_h	$(1.99, 3.26)$	$(1.99, 1.99)$

The game induced by the Mirrleesian mechanism for $(\omega_1, \omega_2) = (\omega_l, \omega_h)$.

(U_1, U_2)	ω_l	ω_h
ω_l	$(3.26, 3.70)$	$(3.26, 3.72)$
ω_h	$(1.99, 3.70)$	$(1.99, 3.72)$

The game induced by the Mirrleesian mechanism for $(\omega_1, \omega_2) = (\omega_h, \omega_l)$.

(U_1, U_2)	ω_l	ω_h
ω_l	$(3.70, 3.26)$	$(3.70, 1.99)$
ω_h	$(3.72, 3.26)$	$(3.72, 1.99)$

The game induced by the Mirrleesian mechanism for $(\omega_1, \omega_2) = (\omega_h, \omega_h)$.

(U_1, U_2)	ω_l	ω_h
ω_l	$(3.70, 3.70)$	$(3.70, 3.72)$
ω_h	$(3.72, 3.70)$	$(3.72, 3.72)$

E.4 Normal form games which are induced by the Piketty mechanism

The game induced by the Piketty mechanism for $(\omega_1, \omega_2) = (\omega_l, \omega_l)$.

(U_1, U_2)	ω_l	ω_h
ω_l	$(2.32, 2.32)$	$(3.98, 1.04)$
ω_h	$(1.04, 3.98)$	$(2.94, 2.94)$

The game induced by the Piketty mechanism for $(\omega_1, \omega_2) = (\omega_l, \omega_h)$.

(U_1, U_2)	ω_l	ω_h
ω_l	$(2.32, 3.06)$	$(3.98, 3.08)$
ω_h	$(1.04, 4.38)$	$(2.94, 4.40)$

The game induced by the Piketty mechanism for $(\omega_1, \omega_2) = (\omega_h, \omega_l)$.

(U_1, U_2)	ω_l	ω_h
ω_l	$(3.06, 2.32)$	$(4.38, 1.04)$
ω_h	$(3.08, 3.98)$	$(4.40, 2.94)$

The game induced by the Piketty mechanism for $(\omega_1, \omega_2) = (\omega_h, \omega_h)$.

(U_1, U_2)	ω_l	ω_h
ω_l	$(3.06, 3.06)$	$(4.38, 3.08)$
ω_h	$(3.08, 4.38)$	$(4.40, 4.40)$

E.5 Choice data T4 and T5

		Individual 1		*Individual* 2	
	Game induced by	ω_l^1	ω_h^1	ω_l^2	ω_h^2
T4 *Mirrleesian approach*	(ω_l, ω_l)	62	1	62	1
	(ω_l, ω_h)	62	1	2	61
	(ω_h, ω_l)	2	61	62	1
	(ω_h, ω_h)	2	61	2	61
T5 *Piketty's approach*	(ω_l, ω_l)	57	6	55	8
	(ω_l, ω_h)	60	3	14	49
	(ω_h, ω_l)	7	56	61	2
	(ω_h, ω_h)	2	61	9	54