No. 53

U N I V E R S I T Y O F C O L O G N E

W O R K I N G P A P E R S E R I E S I N E C O N O M I C S

MECHANISM DESIGN AND INTENTIONS

FELIX BIERBRAUER
NICK NETZER

Department of Economics
University of Cologne
Albertus-Magnus-Platz
D-50923 Köln
Germany

http://www.wiso.uni-koeln.de

Mechanism Design and Intentions*

Felix Bierbrauer
University of Cologne

Nick Netzer
University of Zurich

This version: August 2012
First version: July 2011

Abstract

We introduce intention-based social preferences into a Bayesian mechanism design framework. We first show that, under common knowledge of social preferences, any tension between material efficiency, incentive compatibility, and voluntary participation can be resolved. Hence, famous impossibility results such as the one by Myerson and Satterthwaite (1983) are turned into possibility results. Second, we provide a systematic account of the welfare implications of kindness sensations. Finally, we turn to an environment without common knowledge of social preferences and introduce the notion of a psychologically robust mechanism. Such a mechanism can be implemented without information about the type or the intensity of social preferences. We show that the mechanisms which have been the focus of the conventional mechanism design literature can be modified to achieve psychological robustness.

Keywords: Mechanism Design, Psychological Games, Social Preferences, Intentions, Reciprocity, Revelation Principle.

JEL Classification: C70, C72, D02, D03, D82, D86.

*Email: bierbrauer@wiso.uni-koeln.de and nick.netzer@econ.uzh.ch. We gratefully acknowledge very helpful comments by Tomer Blumkin, Stefan Buehler, Antonio Cabrales, Juan Carlos Carbajal, Martin Dufwenberg, Kfir Eliaz, Florian Englmaier, Ernst Fehr, Silvia Grätz, Hans Peter Grüner, Paul Heidhues, Martin Hellwig, Holger Herz, Benny Moldovanu, Johannes Münster, Zvika Neeman, Axel Ockenfels, Ariel Rubinstein, Larry Samuelson, Klaus Schmidt, Armin Schmutzler, Alexander Sebald, Joel Sobel, Ran Spiegler, André Volk, Roberto Weber, Philipp Weinschenk, David Wettstein, Philipp Wichardt, and seminar participants at the CESifo Area Conference on Behavioural Economics 2011, Ben-Gurion University of the Negev, HU and FU Berlin, ULB Brussels, MPI Bonn, LMU Munich, Tel Aviv University and the Universities of Cologne, Heidelberg, Mannheim, St. Gallen and Zurich. All errors are our own.

1 Introduction

We study the problem of mechanism design under the assumption that players have intention-based social preferences. They are willing to give up own material payoffs in order to either reward behavior by others that they attribute to good intentions, or to punish behavior that they attribute to bad intentions. The behavioral relevance of such preferences is by now well established. In this paper we explore their implications for the theory of mechanism design.

The procedural nature of intention-based social preferences has a profound impact on the analysis. For an assessment of intentions, it does not only matter what players do, but also what they could have done instead. Hence, a first contribution of the paper is to develop a theory of mechanism design in which the interpretation of behavior is crucial. This implies, in particular, that the revelation principle does not hold. In our model, a truth-telling strategy may appear selfish in a direct mechanism but it may appear kind in the context of a mechanism in which the set of actions is larger than the set of conceivable payoff functions. We show that one can enlarge the set of implementable social choice functions by means of actions that ultimately remain unchosen. A second contribution of the paper is to allow for a discussion of procedural questions. We show that two mechanisms which induce the same economic outcome can be compared according to the attitudes that they induce among the players. Specifically, we formalize the problem to implement a given outcome with a maximal degree of kindness and clarify the conditions under which such an ideal mechanism exists. The third contribution of the paper is to introduce the idea of psychologically robust mechanisms, which implement economic outcomes when interdependent preferences might exist but are not common knowledge. We provide robust possibility results for many of the outcomes that the conventional mechanism design literature has focussed on.

For clarity of exposition, our analysis is based on one particular model of intention-based social preferences. Specifically, we adapt the model by Rabin (1993) to games of incomplete information and work with the solution concept of a *Bayes-Nash fairness equilibrium* in the context of an otherwise conventional *independent private values model* of mechanism design. This approach makes it possible to get very clear-cut theoretical results. We discuss the sensitivity of these results to modelling details throughout the paper.

Our formal analysis has two main parts, which differ with respect to the information available about social preferences. In the first part, we assume that individuals have private information only about their material payoffs. The process by which kindness sensations are generated and traded off against material payoffs is assumed to be common knowledge. We refer to this case as *mechanism design with known kindness generating process.* In the second part, we relax this assumption. This part of the analysis comes under the heading of *psychologically robust mechanism design.*

Part I: Mechanism Design with Known Kindness Generating Process. We first show that the revelation principle is not available in our framework. There exist social choice functions that cannot be implemented by direct mechanisms with a truth-telling Bayes-Nash fairness equilibrium, but that can be implemented by means of a non-direct mechanism. With a direct mechanism, every available message is used in a truth-telling equilibrium. Put differently, this

class of mechanism-equilibrium-pairs excludes unused actions from the analysis, which restricts the set of implementable social choice functions. We can show, by contrast, that an *augmented revelation principle* (Mookherjee and Reichelstein 1990) holds. Accordingly, it is without loss of generality to focus on mechanisms where each player's action set *includes* the set of possible types, and which possess truth-telling equilibria. Hence, while the restriction that every action must be used in equilibrium would involve a loss of generality, the restriction that every used action is a truthfully communicated type is without loss of generality.[1]

We then turn to the welfare implications of intention-based social preferences. In particular, we explore various approaches to the treatment of kindness sensations for assessments of economic welfare. First, kindness sensations and psychological payoffs might be considered as relevant from a behavioral but not from a welfare perspective. We are then left with the question what one can say about the implementability of social choice functions that are in a conventional sense efficient. We show in *Theorem 1* that *every* efficient social choice function can be implemented by an appropriately chosen augmented mechanism. Intention-based social preferences render conventional incentive-compatibility constraints innocuous and enlarge the set of implementable social choice functions in a drastic way. When intentions matter, the interpretation of equilibrium play can be influenced by adding actions to the mechanism that would trigger redistribution among players. The challenge in the design of such actions is that they must be tempting to the players but nevertheless remain unused. Our proof of Theorem 1 makes use of the possibility to engineer kindness sensations in such a way that every individual's utility function is turned into a utilitarian welfare function. The construction is akin to a Groves mechanism, in that it aligns private and social interests. The key difference is that it is not based on a suitable choice of payments that individuals have to make in equilibrium, but on a suitable choice of payments that individuals refuse to make in equilibrium. The mechanism that we construct in order to prove Theorem 1 also satisfies voluntary participation constraints, for any status quo that is not efficient. Hence it not only eliminates any tension between incentive compatibility and efficiency, but also any tension between efficiency, incentive compatibility and voluntary interim participation.

We then turn to the possibility of treating the individuals' overall utility, which aggregates material payoffs and kindness sensations, as the relevant criterion of welfare. We show that there are mechanisms that implement a trivial economic outcome – i.e. an outcome that is unresponsive to the individuals' private information, so that there is no real reason to let individuals communicate in the mechanism – but which generate positive kindness and hence also positive utility, possibly even at unbounded levels. The observation shows that, if this welfare criterion is accepted, there is scope for institutions which serve no allocative purpose but which make people feel good. From the perspective of mechanism design theory, however, such institutions are probably not the most interesting ones. As an alternative to welfare based on overall utilities, we therefore consider a lexicographic approach. We first fix a social choice function, i.e. the

[1]The empirical relevance of unchosen actions for kindness judgements has been illustrated by Andreoni et al. (2002) and Falk and Fischbacher (2006), among others. For instance, Falk and Fischbacher (2006) report on how individuals assess the kindness of proposals for the division of a cake of fixed size. They show that this assessment depends on the choice set that is available to the proposer. An offer of 20 per cent of the cake, for instance, is considered very unfair if better offers such as 50 per cent or 80 per cent were also possible. It is considered less unfair if it was the only admissible offer, and even less unfair if only worse offers were possible otherwise.

material outcome that we want to achieve. We then compare all the different mechanisms that may be used to implement this social choice function, and ask whether they can be ranked in the kindness dimension. In particular, we ask whether there exists a mechanism that implements the given social choice function with maximal kindness. For environments with two players, *Theorem 2* clarifies the conditions under which such a best mechanism exists. A necessary and sufficient condition is that the social choice function gives rise to *bilateral externalities*, implying that the expected payoff of player 1 varies with the type of player 2 and vice versa.

Part II: Psychologically Robust Mechanism Design. The previous results relied on the possibility to fine-tune out-of-equilibrium transfers so that players are either turned into maximizers of the sum of material payoffs, or, for a given material outcome, into maximizers of kindness. In the second part of the paper, we investigate the design problem under the assumption that such fine-tuning is impossible, due to lack of common knowledge about the kindness generating process.

Common knowledge can be violated in several ways. The players could have private information about their psychological payoffs, in addition to their private information about material payoffs. Even if the kindness generating process remains common knowledge among players, the designer might be uninformed about the extent of the individuals' concern for reciprocity, or the exact determination of equitable reference payoffs. Instead of trying to solve complicated multi-dimensional design problems for each of these cases, we propose a notion of *psychological robustness* that speaks to all of them simultaneously. Our formal analysis is based on a simple robustness property. Suppose that, for some social choice function, the expected payoff of any one player i does not unilaterally depend on the type of any other player j, so that each player is insured against the randomness of each other player's type. If this *insurance property* holds, then players cannot affect each other's payoff by unilateral deviations from truth-telling in the direct mechanism. If truth-telling is an equilibrium with selfish preferences, then it continues to be an equilibrium for a large class of interdependent preference models, because the insurance property renders those preferences behaviorally irrelevant. The social choice function can hence be implemented without any need to worry about the details of multidimensional design. Importantly, the insurance property implies robustness of the direct mechanism even with respect to interdependent preferences that are not modelled in this paper, such as outcome-based inequality aversion or spitefulness.

Theorem 3 states that to any social choice function that is incentive compatible in the conventional sense, there exists an "equivalent" one that also has the insurance property. Equivalence holds with respect to the decision rule, the interim expected payoffs of players, and the expected deficit or surplus. The theorem covers essentially any application of the independent private values model that has been studied in the literature. In particular, it also covers the study of optimal mechanisms with participation constraints, because interim payoffs are preserved by our construction of insurance. We show how these mechanisms can be modified to make them psychologically robust.

A limitation of Theorem 3 is that ex post budget deficits and surpluses may become unavoidable. One cannot insure individuals against the risk in the other individuals' types and

simultaneously balance the budget in each and every circumstance. *Theorem 4* therefore states sufficient conditions under which psychological robustness is compatible with ex post budget balance. At the core of this result lies the observation that the expected externality mechanism due to d'Aspremont and Gerard-Varet (1979) and Arrow (1979), which respects ex post budget balance, satisfies the insurance property under an assumption of *symmetry*. This follows by construction of the mechanism, which requires each agent to compensate all others for the expected implications of a change in his type. We conclude part II of the paper by discussing the robustness of various social choice functions that are of interest in general equilibrium theory, public finance and contract theory.

A comprehensive view on Parts I and II. Our results on psychological robustness are reassuring from the perspective of conventional mechanism design theory. Even if individuals are inclined to respond to the behavior of others in a reciprocal way, this will in many cases not upset implementability of the outcomes that have been the focus of this literature. For many applications of interest, there is a way to design mechanisms so that the transmission channel for reciprocal behavior is simply shut down. If it is shut down, then individuals are, by design, acting as selfish payoff maximizers, and incentive compatibility in the traditional sense is all that is necessary to ensure the implementability of a social choice function. By contrast, our analysis under the assumption of a known kindness generating process shows the potential of exploiting the reciprocity channel, rather than shutting it down. Every efficient social choice function becomes implementable, and there is no longer a tension between efficiency and voluntary participation. Moreover, the question whether there exists a best mechanism for a given social choice function becomes meaningful. With an analysis that is based exclusively on consequentialist preferences, it would be impossible to even ask this question.

The paper is organized as follows. The next section gives a more detailed discussion of the related literature. Section 3 states the mechanism design problem and introduces the solution concept of a Bayes-Nash fairness equilibrium. Section 4 deals with the analysis of mechanism design when the kindness generating process is known. It also contains several examples that illustrate the main results. Section 5 contains our analysis of psychologically robust mechanism design. The last section contains concluding remarks, in particular on avenues for future research. Several proofs and extensions are relegated to the Appendix.

2 Related Literature

Our work is related to several strands in the literature, (i) a literature which tries to model and to empirically identify *interdependent preferences*, (ii) a literature that studies the implications of interdependent preferences for various *applications*, and (iii) the theory of *mechanism design*.

Interdependent Preferences. Models of interdependent or social preferences are usually distinguished according to whether they are outcome-based or intention-based.[2] Prominent

[2]See Sobel (2005) for an excellent survey, with a focus on reciprocity.

examples for the first class are Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), while Rabin (1993) and Dufwenberg and Kirchsteiger (2004) belong to the second class of models. An extensive experimental literature – examples include Andreoni et al. (2002), Falk et al. (2003) and Falk et al. (2008) – has concluded that behavior is most likely influenced by both types of considerations.[3] The theoretical models proposed by Levine (1998), Charness and Rabin (2002), Falk and Fischbacher (2006) and Cox et al. (2007) combine outcomes and intentions as joint motivations for social behavior. In this paper, we consider intention-based social preferences only. We do this for a methodological reason. The distinguishing feature of intention-based preferences is their procedural nature, i.e., sensations of kindness are endogenous to the game form. This is a challenge for mechanism design theory, which is concerned with finding optimal game forms. With outcome-based social preferences, this methodological issue would not arise.[4] To keep the exposition straight, we refrain from also modelling outcome-based social preferences. That said, enriching our framework so that also outcome-based social preferences come into play would in principle be possible.

The formal framework for modelling intentions is provided by psychological game theory (Geanakoplos et al. 1989) which allows payoffs to depend on higher-order beliefs.[5] To the best of our knowledge, the literature does not yet contain a general treatment of intention-based social preferences for games of incomplete information. Rabin (1993) focusses on normal form games and Dufwenberg and Kirchsteiger (2004) consider extensive form games, but both contributions assume complete information.[6] Our mechanism design approach requires a general theory of intentions for Bayesian games, and we will outline such a theory in Section 3.2.

Applications. Interdependent preferences have been modeled theoretically and investigated empirically within a wide range of applications to contracts, firms, and other relevant fields.

There are numerous applications of outcome-based social preferences. Some focus on unobserved behavior or moral hazard. Examples include Englmaier and Wambach (2010) and Bartling (2011). Others focus on unobserved characteristics or screening problems, such as Cabrales et al. (2007), Cabrales and Calvó-Armengol (2008) or Kosfeld and von Siemens (2011). The literature on auctions has investigated, among other behavioral phenomena, spiteful preferences (Morgan et al. 2003) and their role for the empirical phenomenon of overbidding.

Intention-based social preferences have been applied in the context of moral hazard problems (Englmaier and Leider 2012, Netzer and Schmutzler 2010), as an explanation of wage rigidity (Dufwenberg and Kirchsteiger 2000), in models of political competition (Hahn 2009), and the

[3] See Stanca (2010) for an instance where only outcomes matter, and McCabe et al. (2003) for evidence in favor of purely intention-based models. Offerman (2002) distinguishes between positive and negative intentions and finds stronger evidence for the latter.

[4] Frey et al. (2004) provide a general discussion of procedural preferences and their potentially important role for the design of institutions. Gaspart (2003) provides a formalization of fair game forms.

[5] The framework of Geanakoplos et al. (1989) has been further developed by Battigalli and Dufwenberg (2009). Psychological game theory has been used to study AIDS policy in the presence of fear (Caplin and Eliaz 2003), preferences over the timing of information (Caplin and Leahy 2004), feelings of guilt (Battigalli and Dufwenberg 2007), adherence to social norms (Li 2008), or framing effects (Dufwenberg et al. 2011a).

[6] Segal and Sobel (2007) generalize the model of Rabin (1993) and provide an axiomatic foundation. They also illustrate that deleting unused actions can affect the equilibrium structure. A series of papers, including Sebald (2010) and Aldashev et al. (2010), introduce randomization devices into psychological games, but still under the assumption of perfect observability. Von Siemens (2009, online appendix) contains a model of intentions for a two-stage bargaining game with incomplete information about the second-mover's social type.

study of the hold-up problem (von Siemens 2009, Dufwenberg et al. 2011c). Nishimura et al. (2011) investigate the impact of intention-driven reciprocal responses to spiteful behavior in different auction formats under complete information.

Experimental studies have emphasized that the design of incentive contracts is often facilitated in an environment with reciprocal individuals (e.g. Fehr et al. 1997, Fehr and Falk 2002). Recent studies have revealed that, in laboratory studies of the principal-agent-problem, principals do in fact adapt their contract offers to the agents' degree of reciprocity (Fehr et al. 2007, Cabrales et al. 2010). However, reciprocity is not necessarily beneficial from the perspective of contract design. In Hart and Moore (2008), for instance, negative reciprocal reactions by at least one party can be inevitable and generate outcomes that are ex post inefficient. Experimental evidence by Fehr et al. (2011a) supports this prediction. These mixed findings motivate our theoretical treatment of the problem of mechanism design for reciprocal agents.

Mechanism Design Theory. Our underlying mechanism design framework, the independent private values model with quasilinear utilities, corresponds to standard textbook treatments (e.g. Mas-Colell et al. 1995). We limit ourselves to the basic question of mechanism design: For which social choice functions can we find a mechanism with an equilibrium that generates the social choice function?[7]

Several authors have investigated mechanism design problems under behaviorally motivated assumptions. One of the first contributions is Glazer and Rubinstein (1998), who study the problem of aggregating information across experts. Experts may not only care about consequences, but might want their own recommendation to be accepted. As in our model, this introduces procedural aspects into the design problem. Glazer and Rubinstein (1998) show that the optimal decision can be implemented in unique equilibrium if and only if such egoistic motives exist. In Alger and Renault (2006), procedural issues arise because the mechanism and its equilibrium influence the agents' propensity to lie. Intrinsically honest agents may become willing to misrepresent their private information when other agents also benefit from lying. In some situations this makes non-direct mechanisms optimal, while conventional mechanisms remain optimal in other cases.[8] In general, the possibility that institutions affect preferences has received some attention (Bowles 2008).[9] We add to this literature because in our model kindness sensations and hence the willingness to trade-off own and others' payoffs depend on the mechanism.

In an environment with symmetric information, Eliaz (2002) studies implementation under the assumption that some players behave in an unpredictable "faulty" way, and Renou and Schlag (2011) apply minimax regret equilibrium as solution concept. Yet another strand of

[7] The literature on full or unique implementation asks for which social choice functions there exists a mechanism in which *every* equilibrium generates the social choice function (see e.g. Mookherjee and Reichelstein 1990, Jackson 2001). This question is beyond the scope of this paper.

[8] Matsushima (2008a,b) also assumes that agents dislike lying, in frameworks with symmetric and private information, respectively. These papers are also related to our robustness approach in Section 5, because they share the goal of designing mechanisms which do not rely on details of the environment (Matsushima 2005).

[9] Bowles and Hwang (2008) investigate a public goods game where attitudes towards voluntary contribution may interact with the level of a subsidy. A common pool game where the shares of different social types are affected by extrinsic incentives is analyzed theoretically and experimentally by Rodriguez-Sickert et al. (2008). The contributions by Bar-Gill and Fershtman (2004, 2005) and Heifetz et al. (2007) are based on the assumption that preferences – including outcome-based social preferences in the first two papers – are affected by the institutional framework due to an underlying evolutionary adaptation process.

literature is concerned with the design of mechanisms that exhibit good learning and stability properties. Recent examples include Mathevet (2010) and Cabrales and Serrano (2011). Robustness requirements in mechanism design are of course not limited to behavioral assumptions. Bergemann and Morris (2005), for instance, require robustness with respect to the individuals' probabilistic beliefs about the types and the beliefs of other players.

A few papers have looked at mechanism design problems with outcomes-based social preferences. Desiraju and Sappington (2007) allow for inequality aversion in a model where a profit-maximizing principal faces two agents with private information about their production costs. Inequality aversion has no impact on optimal employment contracts when the agents are symmetric, but modifies their structure otherwise. A related model, with additional private information about social preferences, is presented by von Siemens (2011), who studies the optimality of excluding agents from the firm and hence the social reference group. Jehiel and Moldovanu (2006) provide a survey of papers that deal with a general structure of externalities, some of which might be viewed as resulting from interdependent or social preferences. Finally, Kucuksenel (2011) introduces altruism in an environment similar to ours. Standard tools such as the revelation principle remain available in this context. Increasing degrees of altruism help to achieve efficiency, because individual utilities already internalize social goals. With intention-based social preferences, internalization through preferences is not given by implication. Instead, it is the task of the mechanism to endogenously generate the right levels of kindness for the purpose of internalization.

3 Fairness Equilibria and Mechanism Design

3.1 Environment, Social Choice Functions, Mechanisms

An environment $E = [I, A, (\Theta_i, \pi_i)_{i \in I}, p]$ is defined by the following characteristics. There is a finite set of agents denoted by $I = \{1, \ldots, n\}$ and a set of feasible allocations A. Player i has a type θ_i which belongs to a finite set Θ_i. If an allocation $a \in A$ is chosen then player i realizes a material payoff given by $\pi_i(a, \theta)$, where $\theta = (\theta_1, \ldots, \theta_n)$ is a vector that lists the individuals' types. Types are random, described by a probability distribution p with support $\Theta = \prod_{i=1}^{n} \Theta_i$, so that $p(\theta)$ is the probability of type vector θ. We use expressions like $p(\theta_i)$ or $p(\theta_{-i})$ for the marginal distributions and $p(\theta_{-i}|\theta_i)$ for conditional distributions. Under independence we have $p(\theta_{-i}|\theta_i) = p(\theta_{-i})$ for all players and types.

We focus on the conventional textbook environment with quasilinear preferences and independent private values, by making the following assumptions.[10] First, types are independently distributed and privately observed by the agents. Second, an allocation $a = (q_1, \ldots, q_n, t_1, \ldots, t_n)$ specifies for each individual a consumption level q_i of a private or public good and a monetary transfer t_i. Finally, material payoffs take the form $\pi_i(a, \theta) = v_i(q_i, \theta_i) + t_i$. Possible production costs associated with a profile $(q_1, ..., q_n) \in Q$ are, without loss of generality, assumed to be shared equally among individuals, so that the payoff functions $v_i(q_i, \theta_i)$ are interpreted as net valuations. The set of feasible allocations is $A = Q \times T$. Here, $Q \subseteq \mathbb{R}^n$ specifies all possible

[10] That said, some of our results, for instance those on solution concepts in the next subsection, hold more generally and require neither independence nor quasilinearity.

consumption profiles. The set $T \subseteq \mathbb{R}^n$ then describes the possible transfers to the agents. If not mentioned otherwise, we rule out subsidies from outside, that is, we assume $T = \bar{T}$ where $\bar{T} = \{(t_1, ..., t_n) \in \mathbb{R}^n | \sum_{i=1}^n t_i \leq 0\}$.

A social choice function $f : \Theta \to A$ specifies an allocation as a function of the individuals' types. We also write $f = (q_1^f, \ldots, q_n^f, t_1^f, \ldots, t_n^f)$. Material efficiency of an SCF f requires the allocation $f(\theta)$ to be Pareto efficient in A, for every type profile $\theta \in \Theta$. In our quasilinear framework this is equivalent to saying that $\sum_{i=1}^n v_i(q_i, \theta_i)$ must be maximized by $q_1^f(\theta), ..., q_n^f(\theta)$ and that budget balance, $\sum_{i=1}^n t_i^f(\theta) = 0$, must hold, for every profile $\theta \in \Theta$.

A mechanism $\Phi = [M_1, \ldots, M_n, g]$ contains a message set M_i for each player i and an outcome function $g : M \to A$ which specifies an allocation for each message profile $m = (m_1, \ldots, m_n) \in M = \prod_{i=1}^n M_i$. We also write $g = (q_1^g, \ldots, q_n^g, t_1^g, \ldots, t_n^g)$. Players privately observe their types and hence can condition their message on their type. Therefore, a pure strategy for player i in a mechanism Φ is a function $s_i : \Theta_i \to M_i$. Throughout, we focus on pure strategies. The set of all pure strategies of player i is denoted by S_i. We also use the conventional notation $S = \prod_{i=1}^n S_i$ and $S_{-i} = \prod_{j \neq i} S_j$, and we write $s = (s_i, s_{-i}) \in S$ for strategy profiles.

We will shortly define what it means that a strategy profile s is a Bayes-Nash equilibrium (BNE) or a Bayes-Nash fairness equilibrium (BNFE). We then obtain the following definition of an implementable social choice function: A social choice function f is implementable in BNE/ BNFE if there exists a mechanism with a BNE/ BNFE strategy profile s^* so that, for all $\theta \in \Theta$, $g(s^*(\theta)) = f(\theta)$.

3.2 Solution Concepts

Bayes-Nash Equilibrium. As a benchmark, we first introduce the familiar solution concept of a BNE. To facilitate the comparison to BNFE, we state the definition in a way that emphasizes the role of the players' beliefs about their opponents' strategies. Since we focus on pure strategy equilibria, we can, without loss of generality, assume that player i's beliefs about player j's strategy put unit mass on a particular strategy in S_j. In the following, we denote this strategy by $s_{ij}^b \in S_j$. We denote by $s_i^b = (s_{ij}^b)_{j \neq i}$ the collection of player i's first order beliefs. Analogously, we denote by $s_{ijk}^{bb} \in S_k$ player i's belief about j's belief about k's strategy. We use $s_{ij}^{bb} = (s_{ijk}^{bb})_{k \neq j}$ as a shorthand notation for the collection of i's belief about j's first-order beliefs. Likewise, $s_i^{bb} = (s_{ij}^{bb})_{j \neq i}$ is the collection of all of player i's second order beliefs.

Given an environment E and a mechanism Φ, player i's expected material payoff from following strategy s_i, given his beliefs, then equals

$$\Pi_i(s_i, s_i^b) = \sum_{\theta \in \Theta} p(\theta) \pi_i(g(s_i(\theta_i), s_i^b(\theta_{-i})), \theta) \ .$$

Definition 1. *A BNE is a strategy profile s^* such that, for all players $i \in I$,*
(1) $s_i^* \in \arg\max_{s_i \in S_i} \Pi_i(s_i, s_i^b)$, and
(2) $s_i^b = s_{-i}^*$.

Bayes-Nash Fairness Equilibrium. Rabin (1993) has introduced the notion of a fairness equilibrium for normal form games of complete information. In the following we adapt this

solution concept to normal form games of incomplete information. In particular, we follow Rabin in that we enrich the individuals' payoff functions by psychological components which capture the desire to reward kind and punish unkind behavior.

Given an environment E and a mechanism Φ, player i's expected utility is given by

$$U_i(s_i, s_i^b, s_i^{bb}) = \Pi_i(s_i, s_i^b) + \sum_{j \neq i} y_{ij}\, \kappa_{ij}(s_i, s_i^b)\, \kappa_{ji}(s_{ij}^b, s_{ij}^{bb}) \ .$$

The first source of utility is the expected material payoff $\Pi_i(s_i, s_i^b)$. In addition, player i's interaction with any other player j gives rise to sensations of kindness (or unkindness). This is captured by $y_{ij}\, \kappa_{ij}(s_i, s_i^b)\, \kappa_{ji}(s_{ij}^b, s_{ij}^{bb})$. In this expression, y_{ij} is an exogenous parameter, interpreted as the weight that kindness in the relationship with player j has in player i's utility function. The term $\kappa_{ij}(s_i, s_i^b)$ provides a measure of how kindly player i wants to treat player j. Whether i's intended kindness enters the utility function with a positive or a negative sign depends on i's beliefs about the kindness intended by j, which is given by $\kappa_{ji}(s_{ij}^b, s_{ij}^{bb})$.

Following the literature, we treat the kindness that i intends towards j, given that he has beliefs s_i^b and behaves according to s_i, as the difference between j's actual expected material payoff and an equitable reference payoff,

$$\kappa_{ij}(s_i, s_i^b) = \Pi_j(s_i, s_i^b) - \pi_j^{e_i}(s_i^b) \ .$$

The equitable payoff $\pi_j^{e_i}(s_i^b)$ is to be interpreted as a norm, or the payoff that j deserves from i's perspective, given that the behavior of all players different from i is given by s_i^b. Following Rabin (1993), we model this reference payoff as the average of the best and the worst that player i could do to player j, i.e.

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

where $E_{ij}(s_i^b)$ is the set of bilaterally Pareto efficient strategies.[11] The restriction to efficient strategies ensures that kindness is generated only by choices that involve a non-trivial trade-off between players i and j. This is important for mechanism design, as it implies that kindness cannot be generated by merely adding non-tempting punishment options to a mechanism.

Consequently, $\kappa_{ji}(s_j, s_j^b)$ is the kindness intended by j towards i, if j behaves according to s_j and believes everybody else to behave according to s_j^b. Replacing all arguments by beliefs of higher order, we obtain $\kappa_{ji}(s_{ij}^b, s_{ij}^{bb})$ for player i's belief about the kindness of player j.

Definition 2. *A BNFE is a strategy profile s^* such that, for all $i \in I$,*
(1) $s_i^* \in \arg\max_{s_i \in S_i} U_i(s_i, s_i^b, s_i^{bb})$,
(2) $s_i^b = s_{-i}^*$, and
(3) $s_i^{bb} = (s_{-j}^*)_{j \neq i}$.

Observe that the definition of a BNFE becomes equivalent to the definition of BNE whenever $y_{ij} = 0$ for all $i, j \in I$, $j \neq i$, i.e. if concerns for reciprocity are absent.

[11] A strategy s_i belongs to $E_{ij}(s_i^b)$ if and only if there is no alternative strategy $s_i' \in S_i$ so that $\Pi_i(s_i', s_i^b) \geq \Pi_i(s_i, s_i^b)$ and $\Pi_j(s_i', s_i^b) \geq \Pi_j(s_i, s_i^b)$, with at least one of those inequalities being strict.

Discussion of Alternative Modelling Choices. In Rabin's approach the fairness norm is modeled in a particular way. Given her beliefs about the behavior of others, player i considers the Pareto-frontier of possible material payoff combinations of herself and player j. The equitable payoff for player j is then simply the average of the best and the worst outcome for j along this frontier. We will be able to derive precise and transparent results based on this approach. However, we also want to discuss some of the conceivable alternative modelling choices and their potential impact on our analysis.

For an assessment of i's kindness towards j it does not matter how painful it is for player i to generate the best outcome for j, nor does it matter how much i would gain from generating the worst outcome for j. This has some implications that may appear implausible. From a design perspective, it may appear too easy to generate kindness. A designer could modify the Pareto-frontier so that the worst outcome for j is made much worse, while player i is just given a little extra payoff. As a consequence, player j's equitable payoff decreases by a lot, and i's kindness goes up by a lot. Our results are not overly sensitive to that particular feature of the model. For instance, in our analysis of a simple version of the bilateral trade problem and in the proof of Theorem 1, we work with the additional condition of budget balance, which makes it impossible to take a lot from one player without giving it to the other player.

The literature agrees to define the equitable payoff as a value between the largest and the smallest material payoff that one player can give to another when varying the own strategy, where attention is restricted to a set of strategies that are efficient. Different authors invoke different efficiency concepts, however. Rabin (1993) defines efficiency conditional on the opponents' strategies: A strategy is efficient if it induces a Pareto efficient material payoff profile, *given* the fixed strategies of the opponents. Dufwenberg and Kirchsteiger (2004) define efficiency unconditionally: A strategy is inefficient only if it is Pareto dominated by some other strategy for *all* opponent strategies. Our efficiency notion follows Rabin (1993). However, this choice seems to be inconsequential for our results. In Appendix D we demonstrate that, for all examples in our formal analysis, we get to the same results when we apply the notion of Dufwenberg and Kirchsteiger (2004).

Rabin (1993) considers two player games only. When moving to more than two players, the efficiency concept can be applied *bilaterally* or *population-wide*. As much of the earlier literature, we model social preferences as purely bilateral. In the assessment of the kindness in his relation to player j, player i does not care about how j treats a third player k. In our formalism, this also shows up in the definition of the bilateral efficiency set $E_{ij}(s_i^b)$. Our reason for using the bilateral efficiency concept is twofold. First, this enables us to generalize some of Rabin's (1993) results in a straightforward way. Second, as argued above, it fits well with the bilateral notion of reciprocity, and we avoid the reintroduction of more complicated interpersonal effects.

In our approach, like in Dufwenberg and Kirchsteiger (2004) but in contrast Rabin (1993), player i's psychological payoff $\sum_{j\neq i} y_{ij}\,\kappa_{ij}(s_i, s_i^b)\,\kappa_{ji}(s_{ij}^b, s_{ij}^{bb})$ is not bounded relative to the material payoff. This choice is made for ease of exposition. We comment on how it affects our main results as we go along.

Finally, the definitions of a BNE and a BNFE are based on an ex ante perspective, that is, on the perspective of players who have not yet discovered their types but plan to behave in a type-

contingent way. As is well known, for the case of a BNE there is an equivalent definition which evaluates actions (as opposed to strategies) from an interim perspective, where each player has learned his own type, but lacks information about the types of the other players. In Appendix B.1 we develop an analogous ex interim version of a BNFE and provide conditions on the fairness norm under which the ex interim and the ex ante version are equivalent.

4 Mechanism Design with Known Kindness Generating Process

In the following we will study the problem of mechanism design when then kindness generating process is common knowledge among players and the mechanism designer. In particular, we consider a designer who knows how individuals compute equitable payoffs, $\pi_j^{e_i}$, to what extent they are willing to trade-off material payoffs and mutual kindness, y_{ij}, etc. We are interested in two main questions. First, are standard tools of mechanism design theory still available in a model with intention-based social preferences? Second, to what extent can kindness sensations enter efficiency considerations as an own source of individual well-being?

In models in which individuals care only about their material payoffs, the revelation principle is the main tool for the analysis of mechanism design problems. Accordingly, it is without loss of generality to focus on outcomes that can be obtained as the truth-telling equilibrium of a direct mechanism. This approach makes it possible to separate outcomes from the procedures according to which they are obtained: incentive compatibility becomes a property of a social choice function that can be investigated without recourse to specific institutions. We show that this is no longer true if psychological considerations are introduced into the model. The reason is that individuals do no longer care exclusively about the outcomes that are obtained in equilibrium, but also about alternatives that might have been chosen instead. Consequently, with the solution concept of a BNFE, procedures matter in the following two senses: First, the revelation principle does not hold. Second, as will be studied in more detail in Section 4.3 below, different mechanisms which implement the same social choice function might differ in the kindness sensations that they generate. We can therefore ask which mechanism performs best in the kindness dimension.

4.1 Failure of the Revelation Principle

We start the discussion of why the revelation principle fails for the solution concept of a BNFE by recalling why it holds for the solution concept of BNE.

Proposition 1 (Myerson 1979)**.** *Suppose a mechanism $\Phi = [M_1, \ldots, M_n, g]$ implements the social choice function f in some BNE s^*. Then the direct mechanism $\Phi^d = [\Theta_1, \ldots, \Theta_n, f]$ also implements f, in the BNE s^T where $s_i^T(\theta_i) = \theta_i$ for all $i \in I$ and $\theta_i \in \Theta_i$.*

The logic is as follows: Suppose we start with some abstract mechanism Φ that has an outcome function g and which gives rise to some Bayes-Nash equilibrium s^*. We can, as a first step, delete, for every player i, all actions in M_i which are never used, i.e. which do not belong to $s_i^*(\Theta_i)$. This yields a new mechanism Φ' and an outcome function g' which is the restriction of g to the domain $\prod_{i=1}^n s_i^*(\Theta_i)$. Now, s^* is still an equilibrium of the game induced by Φ':

If s_i^* is a best response to s_{-i}^* when player i can choose from a large set of strategies, then it must be the case that s_i^* is also a best response to s_{-i}^* when i can choose from the smaller set of strategies that includes s_i^*. As a second step, we can now construct a direct mechanism Φ^d with an outcome function g^d. If individuals communicate a type profile $\theta = (\theta_1, \ldots, \theta_n)$ under the direct mechanism, then the outcome is the one that results under Φ' if individuals send the message profile $s^*(\theta)$. Consequently, for any player i, deviating from s_i^* under Φ' has the same consequences as deviating from truthtelling under Φ^d. If there is no reason to deviate from s_i^* in Φ', then there is also no reason to deviate from truthtelling in Φ^d. The revelation principle now follows from the observation that the equilibrium allocation of the direct mechanism is the same as the equilibrium allocation of the mechanism Φ from which we started. We use the following bilateral trade example to illustrate why this logic fails with the solution concept of a BNFE.

Example 1. *There are two agents, referred to as the buyer and the seller, $I = \{b, s\}$. The buyer's material payoff is $v_b(q, \theta_b) = \theta_b q + t_b$, where $q \in [0, 1]$ is the consumption of a good that is produced by the seller. We refer to θ_b as the buyer's valuation of the good, and assume that it can take two values, $\Theta_b = \{\underline{\theta}_b, \bar{\theta}_b\}$, with equal probability. The seller's material payoff is $v_s(q, \theta_s) = -\theta_s q + t_s$. We refer to θ_s as the seller's marginal cost, and assume that it can also take two values, $\Theta_s = \{\underline{\theta}_s, \bar{\theta}_s\}$, again with equal probability. We assume that $0 \leq \underline{\theta}_s < \underline{\theta}_b < \bar{\theta}_s < \bar{\theta}_b$.*

In this setting, a social choice function f specifies, for all $(\theta_b, \theta_s) \in \Theta_b \times \Theta_s$, the amount of the good to be traded, $q^f(\theta_b, \theta_s)$, and the accompanying payments $t_b^f(\theta_b, \theta_s)$ and $t_s^f(\theta_b, \theta_s)$. Given the assumption on parameters, f is efficient if and only if

$$q^f(\theta_b, \theta_s) = \begin{cases} 0 & \text{if} \quad (\theta_b, \theta_s) = (\underline{\theta}_b, \bar{\theta}_s) \ , \\ 1 & \text{if} \quad (\theta_b, \theta_s) \neq (\underline{\theta}_b, \bar{\theta}_s) \ , \end{cases}$$

and $t_s^f(\theta_b, \theta_s) = -t_b^f(\theta_b, \theta_s)$ for all (θ_b, θ_s).

For particular parameter constellations – for instance $\underline{\theta}_s = 0$, $\underline{\theta}_b = 20$, $\bar{\theta}_s = 80$ and $\bar{\theta}_b = 100$ – this setup gives rise to a discrete-type version of the impossibility result by Myerson and Satterthwaite (1983): There is no social choice function which is efficient, implementable in BNE, and yields a non-negative expected material payoff for every type of every player.

We will be interested in a particular social choice function $f^* = (q^{f^*}, t_b^{f^*}, t_s^{f^*})$, which is efficient and splits the gains from trade equally between the buyer and the seller,

$$t_s^{f^*}(\theta_b, \theta_s) = -t_b^{f^*}(\theta_b, \theta_s) = \left(\frac{\theta_b + \theta_s}{2} \right) q^{f^*}(\theta_b, \theta_s) \ .$$

Under this social choice function we have that, for every $\theta = (\theta_b, \theta_s)$,

$$\pi_s(f^*(\theta), \theta) = \pi_b(f^*(\theta), \theta) = \left(\frac{\theta_b - \theta_s}{2} \right) q^{f^*}(\theta_b, \theta_s) \ ,$$

which implies that every agent's material payoff is always non-negative. As we will see below, this social choice function is not implementable in BNE.

Our first observation shows that, whatever the individuals' reciprocity concerns, social choice function f^* in Example 1 cannot be implemented in a truthful BNFE of a direct mechanism.

Since we have only two players, we can simplify notation and write y_i for the reciprocity weight of player i toward player $j \neq i$. We also write $y = (y_b, y_s)$ for the collection of reciprocity weights.

Observation 1. *Consider Example 1 and the direct mechanism for f^*. For every $y \in [0, \infty[^2$, the strategy profile (s_b^T, s_s^T) is not a BNFE.*

A formal of proof of this observation (and of all other observations) can be found in part C of the appendix. The logic is as follows: One can show that in a hypothetical truth-telling equilibrium both the buyer and the seller realize their equitable payoff. This implies that all kindness terms are zero, so that all players focus solely on their material payoffs. But if they focus on their material payoffs, truth-telling is not an equilibrium because the buyer would benefit from understating his valuation and the seller would benefit from exaggerating his marginal cost.

Observation 1 shows that efficient trade with an equal sharing of the trade surplus is out of reach if only direct mechanisms are considered. Since it covers the case where $y_b = y_s = 0$, the revelation principle for BNE implies that f^* cannot be implemented in BNE. We will now show that, whenever there are mutual reciprocity concerns, there exists a non-direct mechanism Φ' which implements the social choice function f^* in BNFE. More specifically, consider the mechanism $\Phi' = [M_b', M_s', g']$ in which the buyer has the extended message set $M_b' = \{\underline{\underline{\theta}}_b, \underline{\theta}_b, \bar{\theta}_b\}$ and the seller has the extended message set $M_s' = \{\underline{\theta}_s, \bar{\theta}_s, \bar{\bar{\theta}}_s\}$. The outcome of the mechanism is, for every pair of messages $(m_b, m_s) \in M_b' \times M_s'$, a decision on trade $q^{g'}(m_b, m_s) \in [0,1]$ and transfers $t_s^{g'}(m_b, m_s) = -t_b^{g'}(m_b, m_s)$, i.e. the price to be paid by the buyer. Table 1 gives the pair $(q^{g'}, t_s^{g'})$ for every possible pair of messages. The mechanism Φ' works like a direct mechanism with outcome function f^* as long as messages are in $\{\underline{\theta}_b, \bar{\theta}_b\} \times \{\underline{\theta}_s, \bar{\theta}_s\}$. If the buyer chooses the message $\underline{\underline{\theta}}_b$, the consequence is the same as when announcing a low valuation $\underline{\theta}_b$, except that, when there is trade, he gets an additional discount of δ_b. Intuitively, announcing $\underline{\underline{\theta}}_b$ amounts to the claim that the valuation is even lower than $\underline{\theta}_b$. If the seller chooses the message $\bar{\bar{\theta}}_s$, the consequence is the same as when announcing a high cost $\bar{\theta}_s$, except that, when there is trade, the price he receives is increased by δ_s. Announcing $\bar{\bar{\theta}}_s$ can, analogously, be interpreted as claiming a marginal cost somewhat above $\bar{\theta}_s$.

m_b \ m_s	$\underline{\theta}_s$	$\bar{\theta}_s$	$\bar{\bar{\theta}}_s$
$\underline{\underline{\theta}}_b$	$(1, \frac{\underline{\theta}_b + \underline{\theta}_s}{2} - \delta_b)$	$(0,0)$	$(0,0)$
$\underline{\theta}_b$	$(1, \frac{\underline{\theta}_b + \underline{\theta}_s}{2})$	$(0,0)$	$(0,0)$
$\bar{\theta}_b$	$(1, \frac{\bar{\theta}_b + \underline{\theta}_s}{2})$	$(1, \frac{\bar{\theta}_b + \bar{\theta}_s}{2})$	$(1, \frac{\bar{\theta}_b + \bar{\theta}_s}{2} + \delta_s)$

Table 1: The non-direct mechanism Φ'.

In the non-direct mechanism Φ', player i's set of pure strategies equals $S_i' = M_i' \times M_i'$. A generic element s_i' of S_i' is a tuple in which the first entry is the message chosen in case of having a low type, and the second entry is the message chosen in case of having a high type. Note that, for both players, the strategy set of the direct mechanism, $S_i = \Theta_i \times \Theta_i$, is a subset of the extended strategy set S_i', and that the outcome of Φ' under the strategy pair (s_b^T, s_s^T) is still the outcome stipulated by the SCF f^*.

Observation 2. *Consider Example 1 and the non-direct mechanism Φ' in Table 1. For every $y \in]0,\infty[^2$, there exist numbers $\delta_i > 0$, $i \in \{b,s\}$, so that (s_b^T, s_s^T) is a BNFE.*

Starting from a hypothetical equilibrium (s_b^T, s_s^T), we can, in order to derive the seller's equitable payoff, solve for the set of efficient strategies of the buyer if he believes the seller to behave according to s_s^T. The best that the buyer can do for the seller is to exaggerate his willingness to pay, which leads both to more trade and to trade at a higher price. The worst outcome for the seller is obtained if the buyer behaves according to $(\underline{\underline{\theta}}_b, \underline{\underline{\theta}}_b)$, i.e. if the buyer insists on the discount of δ_b. Straightforward computations then yield the equitable payoff

$$\pi_s^{e_b}(s_s^T) = \Pi_s(s_b^T, s_s^T) - \frac{1}{4}\delta_b \ .$$

Therefore, the buyer's kindness in the hypothetical truth-telling equilibrium, where he does not insist on the discount, is strictly positive:

$$\kappa_{bs}(s_b^T, s_s^T) = \Pi_s(s_b^T, s_s^T) - \pi_s^{e_b}(s_s^T) = \frac{1}{4}\delta_b \ .$$

A symmetric argument implies that the seller is kind when he does not use the action $\bar{\bar{\theta}}_s$ and therefore does not ask for the very high price:

$$\kappa_{sb}(s_b^T, s_s^T) = \frac{1}{4}\delta_s \ .$$

Given these observations, we can now calibrate the numbers δ_b and δ_s so as to turn every player's utility-maximization problem into a problem of welfare-maximization. To see how this works, consider the seller's problem. If the buyer chooses s_b^T, the seller chooses s_s in order to maximize

$$\Pi_s(s_b^T, s_s) + y_s \kappa_{bs}(s_b^T, s_s^T)\Pi_b(s_b^T, s_s).$$

Now let $\kappa_{bs}(s_b^T, s_s^T) = 1/y_s$, or, equivalently, $\delta_b = 4/y_s$. Then the problem becomes: choose s_s in order to maximize expected utilitarian welfare $\Pi_s(s_b^T, s_s) + \Pi_b(s_b^T, s_s)$. Since the social choice function f^* is efficient, s_s^T is a solution to this problem. Truthtelling ensures that, for every θ, the outcome is the one which maximizes the sum of material payoffs. Similarly, player b chooses s_b in order to maximize

$$\Pi_b(s_b, s_s^T) + y_b \kappa_{sb}(s_b^T, s_s^T)\Pi_s(s_b, s_s^T).$$

Now let $\kappa_{sb}(s_b^T, s_s^T) = 1/y_b$, or, equivalently, $\delta_s = 4/y_b$. Then the problem becomes: choose s_b in order to maximize expected utilitarian welfare $\Pi_b(s_b, s_s^T) + \Pi_s(s_b, s_s^T)$. Again, s_b^T solves this problem.

Observations 1 and 2 together imply the failure of the revelation principle for the solution concept of a BNFE. Efficient trade with an equal sharing of the trade surplus is not possible with a direct mechanism, but is possible with a non-direct mechanism. The actions which remain unused in the non-direct mechanism affect the interpretation of equilibrium behavior. In the truth-telling equilibrium of the non-direct mechanism, both players refrain from exploiting the possibility to enrich themselves at the expense of the other player. Consequently, truth-telling is now interpreted as kind, and this makes it possible to implement the desired outcome.

Since f^* ensures non-negative material payoffs for all players and types, Observation 2 also illustrates that the problem of voluntary participation can be overcome, provided that interim material payoffs are considered as relevant for participation decisions. We will discuss this matter in greater detail after Theorem 1 below.

4.2 The Augmented Revelation Principle

In the example above, the non-direct mechanism Φ' that can be used to implement f^* in BNFE still resembles a truthful direct mechanism. The set of messages includes the set of types and truth-telling is an equilibrium. This is not a coincidence. In the following, we will show that if implementation of a social choice function is possible at all, then it is also possible in the class of *augmented revelation mechanisms* (Mookherjee and Reichelstein 1990). A mechanism Φ is called an augmented revelation mechanism for a social choice function f whenever $\Theta_i \subseteq M_i$ for all $i \in I$, and $g(m) = f(m)$ for all $m \in \Theta$, i.e., whenever the individual message sets include the type sets and the SCF f is realized in the event that all messages are possible types. An augmented revelation mechanism Φ *truthfully implements* f in BNFE if the truth-telling profile s^T is a BNFE of Φ. Hence, the difference between truthful direct and augmented revelation mechanisms is the existence of unused actions in the latter.

Mookherjee and Reichelstein (1990) have shown that an *augmented revelation principle* holds when uniqueness of the BNE or of the equilibrium outcomes is required. In their construction, additional messages are used to destroy unwanted equilibria in the direct mechanism. In contrast, we seek to prove an augmented revelation principle for BNFE, according to which a social choice function f is implementable in BNFE if and only if there exists an augmented revelation mechanism that truthfully implements it in BNFE. It is more convenient, however, to first state explicitly the property of strategic equivalence of arbitrary mechanisms and augmented revelation mechanisms. The augmented revelation principle for BNFE – as well as for the equilibrium concepts to be considered later – will follow immediately from this property.

We start from an arbitrary mechanism $\Phi = (M_1, ..., M_n, g)$ and some strategy profile $\tilde{s}$, interpreted as an equilibrium of some type. We then construct an augmented revelation mechanism $\Phi' = \Phi'(\Phi, \tilde{s})$ based on Φ and $\tilde{s}$, with the property that the outcome of Φ' under truthtelling is the same as the outcome of Φ under $\tilde{s}$.[12] We then establish that Φ and Φ' are strategically equivalent, in the sense that any outcome that can be induced by some action under Φ can be induced by some action under Φ' and vice versa.

Formally, consider an arbitrary pair $(\Phi, \tilde{s})$ and let f be the social choice function induced by $\tilde{s}$ in Φ, i.e. $f(\theta) = g(\tilde{s}(\theta))$ for all $\theta \in \Theta$. We now construct new message sets M_i' for every player. Any action from M_i that is used by $\tilde{s}_i$ is relabelled according to the type $\theta_i \in \Theta_i$ that uses it. Any unused action from M_i is kept unchanged. Formally, let $M_i' = \Theta_i \cup (M_i \backslash \tilde{s}_i(\Theta_i))$ be the message sets of the new mechanism Φ'. To define the outcome function g' of Φ', we first construct for every player a surjective function $h_i : M_i' \to M_i$ that maps actions from M_i' back

[12] Mookherjee and Reichelstein (1990) use the same construction, starting from an arbitrary mechanism with a unique BNE or a unique BNE outcome.

into actions from M_i. Formally,

$$h_i(m_i') = \begin{cases} \tilde{s}_i(m_i') & \text{if} \quad m_i' \in \Theta_i \ , \\ m_i' & \text{if} \quad m_i' \in M_i \setminus \tilde{s}_i(\Theta_i) \ . \end{cases}$$

Now we let

$$g'(m') = g(h(m')) \tag{2}$$

for all message profiles m', which completes the definition of the new mechanism Φ'. In words, announcing a type $\theta_i \in \Theta_i$ in Φ' has the same consequences as choosing the action $\tilde{s}_i(\theta_i)$ in Φ, and choosing an action from $M_i \setminus \tilde{s}_i(\Theta_i)$ in Φ' has the same consequences as choosing that same action in Φ. Observe that Φ' is in fact an augmented revelation mechanism for f, because by construction $g'(s^T(\theta)) = g'(\theta) = g(\tilde{s}(\theta)) = f(\theta)$ for all $\theta \in \Theta$.

Proposition 2. *The mechanisms Φ and $\Phi'(\Phi, \tilde{s})$ are strategically equivalent, in the sense that, for every $i \in I$ and any two profiles $m_{-i} \in M_{-i}$ and $m'_{-i} \in M'_{-i}$ with $m_{-i} = h_{-i}(m'_{-i})$, it holds that $G_i(m_{-i}) = G'_i(m'_{-i})$, where*

$$G_i(m_{-i}) = \{a \in A \mid \exists m_i \in M_i \ \textit{ so that } \ g(m_i, m_{-i}) = a\}$$

and

$$G'_i(m'_{-i}) = \{a \in A \mid \exists m'_i \in M'_i \ \textit{ so that } \ g'(m'_i, m'_{-i}) = a\}.$$

The sets $G_i(m_{-i})$ and $G'_i(m'_{-i})$ contain all allocations that player i can induce by varying his message, holding fixed the other players' messages. According to the proposition, these sets are the same in both mechanisms, for message profiles that are associated by the function h. Proposition 2 has the following implication. If we start from an arbitrary mechanism Φ with BNFE s^* that implements an SCF f, the above construction yields an augmented revelation mechanism Φ' in which truth-telling induces f and is a BNFE as well. This conclusion follows from the observation that unilateral deviations from s^T in Φ' can achieve exactly the same outcomes as unilateral deviations from s^* in Φ. The equivalence of achievable deviation outcomes implies in particular that the kindness terms associated to s^* and all unilateral deviations in Φ are identical to those of s^T and all corresponding deviations in Φ'.

Corollary 1. *Suppose a mechanism Φ implements the social choice function f in BNFE. Then there exists an augmented revelation mechanism Φ' that truthfully implements f in BNFE.*

4.3 Efficiency and Kindness

The solution concept of a BNFE relies on two sources of utility, material payoffs and kindness sensations. This raises the question how to treat these from a welfare perspective. This question can be formulated using the notions of *decision utility* and *experienced utility* (Kahneman et al. 1997). Our whole analysis is based on the assumption that the desire to reward kind and to punish unkind behavior matters for the individuals' behavior. Hence, behavior is as if individuals were maximizing U_i. This leaves open the question whether sensations of kindness should be counted as a source of well-being in addition to the individuals' material payoffs.

If we disregard kindness sensations, we are left with a conventional notion of efficiency based on material payoffs only. The question is then how the behavioral implications of reciprocity affect the possibility to implement materially efficient outcomes. We study this question in subsection 4.3.1, and we return to it in the context of psychological robustness in Section 5.

As an alternative, we also entertain the possibility that kindness sensations are an own source of well-being, or experienced utility. We thus follow the path outlined by Rabin (1993) who argues that "welfare economics should be concerned not only with the efficient allocation of material goods, but also with designing institutions such that people are happy about the way they interact with others" (p. 1283). We explore two routes for how this can be done. We will first show that a welfare objective which puts material payoffs and kindness sensations on an equal footing may lead to paradoxical results. There exist mechanisms which implement trivial and materially inefficient social choice functions, which are unresponsive to the individuals' private information, but still achieve large (and possibly unbounded) utility levels. As an alternative, we therefore consider the *utility-efficient implementation* of a given SCF. That is, we consider the class of mechanisms which all implement the same social choice function, and we ask which of those mechanisms performs best in the kindness dimension.

4.3.1 Material Efficiency

It is a well known result that, with the solution concept of a BNE, *some* materially efficient social choice functions can be implemented (d'Aspremont and Gerard-Varet 1979, Arrow 1979). The ones that are implementable are those which are incentive compatible. Formally, for all $i \in I$ and $\theta_i, \theta_i' \in \Theta_i$ it must be true that

$$\mathbb{E}_{\theta_{-i}}[v_i(q_i^f(\theta_i, \theta_{-i}), \theta_i) + t_i^f(\theta_i, \theta_{-i})] \geq \mathbb{E}_{\theta_{-i}}[v_i(q_i^f(\theta_i', \theta_{-i}), \theta_i) + t_i^f(\theta_i', \theta_{-i})] \,, \tag{3}$$

where $\mathbb{E}_{\theta_{-i}}[.]$ is an expectation over θ_{-i} that treats θ_i as fixed. The following theorem shows that, with the solution concept of a BNFE, incentive compatibility is no longer a constraint.

Theorem 1. *Suppose that $y \in]0, \infty[^{n(n-1)}$. Then, if a social choice function f is materially efficient, it is implementable in BNFE.*

The proof of Theorem 1 is based on a generalization of Observation 2. Consider an efficient SCF and an augmented revelation mechanism with truth-telling s^T as a hypothetical equilibrium. In this equilibrium candidate, player i maximizes

$$\Pi_i(s_i, s_{-i}^T) + \sum_{j \neq i} y_{ij} \kappa_{ji}(s^T) \Pi_j(s_i, s_{-i}^T)$$

by choice of $s_i \in S_i$, where $\kappa_{ji}(s^T)$ is treated as fixed. Now, if we had $\kappa_{ji}(s^T) = 1/y_{ij}$ for all $j \neq i$, then the individual's problem would become the problem of maximizing welfare $\sum_{k=1}^n \Pi_k(s_i, s_{-i}^T)$. Since the social choice function to be implemented is materially efficient, i.e. it maximizes the sum of material payoffs for every profile $\theta \in \Theta$, truth-telling is clearly a solution to this problem. Hence, we need to manipulate equitable payoffs and equilibrium kindness to obtain $\kappa_{ji}(s^T) = 1/y_{ij}$ for all i and j. To this end, we start from a direct mechanism and introduce additional actions that would trigger redistributive payments among the individuals. We

calibrate these payments in such a way that they generate the desired degrees of kindness but ultimately remain unused. This resembles a Groves mechanism, where transfers between individuals are designed so as to align individual interests with the objective of surplus-maximization. Here, however, out-of-equilibrium payments are used for that purpose.

The construction faces the difficulty that additional messages can have a non-trivial impact on the set of bilaterally Pareto efficient strategies, and they must yield bilaterally Pareto efficient outcomes themselves, to have an effect on the kindness norm. The messages that we construct in the proof of Theorem 1 trigger real redistribution between players, as already illustrated in Example 1. The resulting augmented mechanism therefore satisfies budget balance also off the equilibrium and does not rely on implausible threats of punishment.[13]

It is crucial for our construction that the kindness terms κ_{ij} are not bounded a priori. If y_{ij} is very small, then $\kappa_{ji}(s^T)$ has to become very large. In Rabin's model, kindness sensations are bounded. If we introduced an upper bound on kindness, say $\bar{\kappa}$, we would have to add the qualification in Theorem 1 that every efficient social choice function can be implemented provided that all y_{ij} exceed the lower bound $1/\bar{\kappa}$. An analogous argument applies when out-of-equilibrium payments are bounded due to limited liability of the agents.

Theorem 1 also speaks to the issue of voluntary interim participation. Classical papers on the conflict between voluntary participation and efficiency, such as Myerson and Satterthwaite (1983) and Mailath and Postlewaite (1990), have noted that when we consider a Bayesian incentive compatible and efficient social choice function, then for some types of some players the expected material payoff might be lower than under a given status quo outcome. Theorem 1 tells us that incentive compatibility constraints are no longer relevant since every efficient social choice function is implementable. This implies, in particular, that we can also implement an efficient social choice function with the property that all players get an equal share of the material surplus, for every θ. Under such a mechanism all types of all individuals realize a higher material payoff than under the status quo, as already illustrated in Example 1. Hence, the problem of voluntary participation is solved, at least if we take material payoffs as the criterion. In our framework, however, it seems more plausible that individuals agree to play a mechanism if their overall utility, including kindness sensations, under the mechanism is larger than their overall utility under the status quo. Theorem 1 can be adapted so that voluntary participation is also guaranteed with this criterion. The proof of Theorem 1 starts from a direct mechanism – where $M_i = \Theta_i$ for all $i \in I$ – and then adds unused actions to turn every agent's objective into a utilitarian welfare function. To respect the requirement that interim participation in the mechanism is voluntary, we can as well start out from a direct mechanism with veto rights – where $M_i^v = \Theta_i \cup \{\text{veto}\}$ for all $i \in I$ – with the understanding that, if any one player exercises the veto, then the status quo outcome prevails. We can then add actions to the message sets M_i^v in exactly the same way as in the proof of Theorem 1 and align individual preferences with the objective of welfare maximization. Consequently, if the outcome of the mechanism is materially efficient and the status quo is not, then all players will refrain from exercising their veto rights and communicate

[13] Mookherjee and Reichelstein (1990) also maintain budget balance off the equilibrium, but their construction of "flags" and "counterflags" is otherwise very different from our approach. In our construction, each unused action has the same consequence as the announcement of a certain type, but with additional pairwise redistribution. We obtain $n|\Theta_i|$ as an upper bound on the cardinality of player i's message set.

their types truthfully to the mechanism.

Theorem 1 can hence be interpreted as a universal possibility result for materially efficient social choice functions. The theorem does, however, not extend to *any* social choice function. The requirement of material efficiency is important. To clarify this point, Appendix A contains an example of an inefficient social choice function that cannot be implemented in BNFE.

4.3.2 Utility-Efficiency and Implementation with Maximal Kindness

We now seek to define Pareto efficiency based on the entire utility function U_i. Developing this notion faces a difficulty. In the conventional approach one defines efficient social choice functions taking only the players' material payoffs and the economy's resource constraint into account. A separate question then is whether there exists a mechanism that implements this social choice function as the equilibrium outcome. This route is not available if we seek to define a notion of efficiency based on utilities, because the latter endogenously depend on the mechanism that is used and the equilibrium that is played. Therefore, we cannot simply define a utility-efficient social choice function. Instead, we define efficiency for mechanism-equilibrium pairs.

Definition 3. *A mechanism-equilibrium-pair (Φ, s^*) is utility-efficient if*
(1) s^* is a BNFE of Φ, and
(2) there is no pair (Φ', s') of a mechanism Φ' with BNFE s' whose equilibrium utilities Pareto dominate those of (Φ, s^*).

The following observation demonstrates that it is possible to generate unbounded utility with a mechanism that implements a degenerate social choice function, i.e. one that stipulates the same outcome for all θ.

Observation 3. *Consider any environment E, any trivial SCF f where $f(\theta) = \bar{a}$ for all $\theta \in \Theta$, and assume $y \in]0, \infty[^{n(n-1)}$. Then, it is possible to implement f in BNFE with arbitrarily large equilibrium utility levels.*

A first implication of this observation is that utility-efficient mechanism-equilibrium-pairs do not exist. To any given pair we can find another one that gives more utility to all players. This problem could possibly be overcome by imposing bounds on the kindness terms as in Rabin (1993), or bounds directly on the set of admissible out-of-equilibrium payments. More importantly, however, the possibility to generate kindness sensations does not depend on a meaningful economic allocation problem being solved in the background.[14] An efficiency notion based on the individuals' utility functions therefore faces the problem that economic allocation problems may appear insignificant.

We will therefore explore a different route for introducing kindness considerations into welfare assessments. We fix some SCF and then ask whether or not we can implement it in BNFE. If we can, we look for the mechanism and the BNFE which implements it with maximal utility levels. That is, kindness sensations play a subordinate role in our welfare analysis. They are a good thing, but only to the extent that they do not upset the material outcomes of a given implementation exercise.

[14] Rabin (1993) has already shown that there exist games in which psychological payoffs remain bounded away from zero as material payoffs converge to zero.

Definition 4. *A mechanism-equilibrium-pair (Φ, s^*) implements an SCF f utility-efficiently if*
(1) s^* is a BNFE of Φ which implements f, and
(2) there is no pair (Φ', s') of a mechanism Φ' with BNFE s' which also implements f, but with equilibrium utilities that Pareto dominate those of (Φ, s^*).

As Observation 3 has shown, the problem to implement an SCF with a maximal degree of kindness might not be well-defined in some cases, e.g. for trivial social choice functions. In general, however, fixing the SCF can yield endogenous bounds on utility levels. To illustrate this, we first characterize a mechanism which, for Example 1, implements the social choice function f^* (efficient trade and equal sharing of the surplus) utility-efficiently. For ease of exposition, we will make the more specific parameter assumptions mentioned earlier.

Observation 4. *Consider Example 1 with $\underline{\theta}_s = 0$, $\underline{\theta}_b = 20$, $\bar{\theta}_s = 80$ and $\bar{\theta}_b = 100$, and assume $y \in]0, \infty[^2$. Let Φ' be the extended mechanism in Table 1 with $\delta_b = 44/5y_s$ and $\delta_s = 44/5y_b$. Then (Φ', s^T) utility-efficiently implements f^*.*

The observation shows, in the context of Example 1, that it is possible to implement efficient trade with equal sharing of the gains in a utility-efficient way. The proof proceeds as follows. We first note that an augmented revelation principle applies: Whenever some mechanism implements f^* in BNFE, then there exists an augmented revelation mechanism that truthfully implements f^* with identical utility levels. We can therefore again restrict attention to augmented revelation mechanisms and truth-telling. Secondly, we observe that whenever we seek to implement f^*, kindness has to be bounded in equilibrium. To see why, note that in *any* augmented revelation mechanism for f^*, the strategies of always reporting the low parameter (denoted s_i^L) and of always reporting the high parameter (denoted s_i^H) are available for both players. The conditions that no player wants to deviate to any of these strategies yields upper and lower bounds on kindness. For instance, if the seller became too kind, the buyer would deviate to s_b^H, i.e. he would consent to trading at a high price, which would make the seller better off. By contrast, if the seller became too unkind, the buyer would deviate to strategy s_b^L, i.e. he would understate his willingness to pay. This would maximize his own material payoff. As a last step, we calibrate the off-equilibrium-transfers δ_i of mechanism Φ' from Table 1 so that the upper bounds on kindness are in fact reached in equilibrium.

In the following, we provide a generalization of this observation for two players, using the notion of bilateral externalities.

Definition 5. *Let $n = 2$. Given an environment E, a social choice function f exhibits bilateral externalities if, for each $i = 1, 2$ and $j \neq i$, there exist types $\theta_j', \theta_j'' \in \Theta_j$ such that*

$$\mathbb{E}_{\theta_i}\left[v_i(q_i^f(\theta_i, \theta_j'), \theta_i) + t_i^f(\theta_i, \theta_j')\right] \neq \mathbb{E}_{\theta_i}\left[v_i(q_i^f(\theta_i, \theta_j''), \theta_i) + t_i^f(\theta_i, \theta_j'')\right] . \tag{4}$$

With bilateral externalities, an agent is not indifferent with respect to the other agent's type. The social choice function f^* in Example 1 satisfies this property: The buyer's expected payoff conditional on $\theta_s = \underline{\theta}_s$ is larger than his expected payoff conditional on $\theta_s = \bar{\theta}_s$, and the seller's payoff conditional on $\theta_b = \bar{\theta}_b$ is larger than his expected payoff conditional on $\theta_b = \underline{\theta}_b$. Trivially, degenerate social choice functions as in Observation 3 do not give rise to bilateral externalities,

since the outcome does not depend on the type of any player. As the following theorem shows, this is the reason why the SCF f^* from Example 1 can be implemented utility-efficiently, while utility-efficient mechanisms do not exist for trivial SCFs.[15]

Theorem 2. *Let $n = 2$, $y \in]0, \infty[^2$, and let f be a materially efficient SCF. Then, there exists a mechanism that utility-efficiently implements f if and only if f exhibits bilateral externalities.*

4.4 Coercion and Kindness

To further illustrate the importance of unused actions in real-world mechanisms, in this section we will study a public goods example and show that the introduction of veto-rights – so that, by exercising his veto right, each individual can force a status quo outcome with no public goods provision – may increase all individuals' kindness sensations without impeding efficiency. Hence, in our framework, the introduction of veto rights may be a good thing. People enjoy kindness sensations if they are not forced to contribute to public goods that they might not like, but nevertheless voluntarily choose to contribute because they want to reciprocate the kindness of others. This stands in stark contrast to the view of participation constraints which emerges with the conventional solution concept of a Bayes-Nash equilibrium. With BNE as opposed to BNFE, participation constraints are, if anything, bad, because they may render efficient public goods provision impossible.[16]

Example 2. *There are three individuals, $I = \{1, 2, 3\}$, and a decision about the provision of an indivisible public good has to be taken. Hence $q_1 = q_2 = q_3 = q$ and either $q = 0$ or $q = 1$. Each individual values the public good either at 1 or at 3, $\Theta_i = \{1, 3\}$, with equal probabilities. The per capita cost of public goods provision equals $c = 2$. Hence $v_i(1, 3) = 1$ and $v_i(1, 1) = -1$, whereas $v_i(0, \theta_i) = 0$ for all $\theta_i \in \Theta_i$. Material payoff efficiency requires that the public good is provided if at least two individuals value it highly, and not to provide it otherwise. We seek to implement this efficient rule without additional transfers, so that costs are shared equally.*[17]

We will compare the performance of two mechanisms which we refer to as "simple majority voting" and "majority voting with veto rights". Under simple majority voting, each individual's message set is given by $M_i = \{no, yes\}$ and the public good is provided if and only if at least two individuals vote "yes". Since both the set of types and the set of messages is binary, and the voting mechanism entails no transfers, simple majority voting is equivalent to a direct mechanism for the SCF we seek to implement. Under majority voting with veto rights, each

[15] An extension of the theorem to more than two players would have to address the following complication. Any misreport in an augmented revelation mechanism might affect the opponents in opposite directions, i.e., any deviation from truth-telling might make one opponent better off and one opponent worse off. Without additional assumptions, kindness is then not necessarily bounded, because the desire to reward an increasingly kind opponent could always be outweighed by the unfortunate necessity to punish another increasingly kind opponent at the same time.

[16] In an independent private values model, efficient public goods provision is possible if there are no participation constraints, see d'Aspremont and Gerard-Varet (1979), but generally impossible if voluntary participation is required, see Güth and Hellwig (1986) or Mailath and Postlewaite (1990). Our view of participation decisions is more in line with the findings of Frey and Stutzer (2004), according to which greater political participation rights lead to increased life-satisfaction.

[17] The parameters of the example are deliberately chosen in a non-generic way, to simplify calculations. Qualitatively identical results can be obtained in more general and generic models.

individual's message set is given by $M_i' = \{no, yes, veto\}$. The public good is provided if and only if no individual chooses the action "veto" and at least two individuals choose the action "yes". Majority voting with veto rights is equivalent to an augmented revelation mechanism for our SCF.

A strategy for player i in the game induced by simple majority voting is a tuple $(v_i^1, v_i^3) \in \{no, yes\}^2$, where v_i^1 is the vote that is cast by type $\theta_i = 1$ of individual i, and v_i^3 is the vote that is cast by type $\theta_i = 3$. Likewise, a strategy in the game induced by majority voting with veto rights is an element (v_i^1, v_i^3) of $\{veto, no, yes\}^2$. A strategy of particular interest is "sincere voting" $(v_i^1, v_i^3) = (no, yes)$, which can readily be interpreted as truth-telling. For both voting mechanisms, material payoff efficiency is reached if and only if all players vote sincerely.

The following observation establishes that, whatever the intensity of the players' kindness sensations, sincere voting by all players is an equilibrium under simple majority voting.

Observation 5. *In Example 2, sincere voting is always a BNFE under simple majority voting.*

Simple majority voting makes it possible to reach material efficiency. Moreover, this comes without (positive or negative) sensations of kindness. In the given example, if one player unilaterally changes his strategy, this does not affect the other players' expected payoffs. Consequently, no player has the possibility to be kind or unkind to the other players. Therefore, the only remaining concern is the own payoff, which is maximized by sincere voting.

The next observation will show that majority voting with veto rights may yield the same outcome as simple majority voting and, moreover, generate positive kindness. These sensations are induced because players refrain from exercising their veto power. If a player has a low valuation of the public good and chooses the action "no" instead of the action "veto", he takes the risk of ending up with a payoff of -1, in case the other two players both vote "yes", instead of making sure that he gets a payoff of 0. The other players will interpret this behavior as kind, because the player in question harms himself so as to make it possible for them to benefit from public goods provision. The willingness to sacrifice own payoff is motivated by the desire to reciprocate that the other players also refrain from exercising their veto rights if they are in the same situation.

Observation 6. *In Example 2, sincere voting is a BNFE under majority voting with veto rights if and only if concerns for reciprocity are sufficiently large. Whenever sincere voting is a BNFE, then equilibrium utilities Pareto dominate those of sincere voting under simple majority voting.*

Consequently, in circumstances where both majority voting with veto rights and simple majority voting work – in the sense of generating the materially efficient outcome – majority voting with veto rights is preferable because it comes with the extra benefit of positive kindness sensations.

The explicit introduction of veto rights into the game helps to solve, paradoxically, an apparent participation problem. Sincere voting in the simple majority voting mechanism induces zero kindness, so that equilibrium utilities coincide with equilibrium material payoffs. Treating interim participation constraints in the conventional way, one would have to conclude that voluntary participation fails, because individuals with type $\theta_i = 1$ prefer to veto the mechanism. Once veto rights are included as part of the game, however, their existence enables the

appearance of an equilibrium in which they remain unused, in which utilities are increased, and in which voluntary participation is assured. This can be recast as an application of the Lucas Critique to the case of endogenous preferences (Bowles and Reyes, 2009). In our example, the introduction of veto rights changes the players' preferences towards more social behavior, and fundamentally alters their participation decisions.

However, majority voting with veto rights yields efficient outcomes only if kindness sensations carry enough weight in the players' utility functions. Simple majority voting, by contrast, works whatever those weights are. It is therefore the more robust procedure. In the following section, we will make the idea of psychological robustness more precise.

5 Psychologically Robust Mechanism Design

In the preceding section we studied mechanism design problems under the assumption that the designer and all the players know the kindness generating process and its behavioral implications. More specifically, common knowledge was assumed about how individuals compute equitable payoffs, how deviations from those equitable payoffs translate into sensations of kindness or unkindness, and finally how much own material payoff individuals are willing to sacrifice in order to reciprocate the kindness of other players. In the following, we seek to complement this analysis by asking what a designer can accomplish in case such common knowledge assumptions are not satisfied. This question is motivated by the empirically well-documented individual heterogeneity in social preferences (Fehr and Schmidt 1999, Engelmann and Strobel 2004, Falk et al. 2008, Dohmen et al. 2009). In many cases, direct observation of these preferences will be hard, if not impossible.

There are several directions in which our model could be extended to allow for less comprehensive knowledge. For instance, we could assume that the importance $y = (y_{ij})_{i,j \in I, i \neq j}$ which the players attribute to their psychological payoffs is not observable to the mechanism designer. However, this is just one of many dimensions in which robustness may be desirable. The designer could also be uninformed about other aspects of the kindness generating process, such as the exact definition of the equitable payoffs. The assumption of common knowledge among players about these aspects might also be too strong. Finally, interdependent preferences beyond those introduced in Section 3, such as outcome-based inequality aversion or spitefulness, could play a role in practice.

A rigorous treatment of any such consideration leads to a multi-dimensional design problem. Such problems are notoriously difficult to solve, and the optimal mechanisms are often complicated. We will proceed differently and investigate *psychologically robust* social choice functions. Consider a social choice function with the property that no player can affect the payoff of any other player by a unilateral deviation from the truth-telling BNE in the direct mechanism. Such a social choice function will remain implementable with interdependent preferences from a large class of models, because these preferences become behaviorally irrelevant in equilibrium. The class includes alternative assumptions on the information available on the weights y or different definitions of equitable payoffs. It also includes models with outcome-based social preferences. Finally, it includes models where players have additional private information about their social

preferences, because psychological considerations will disappear from their optimization problems in all these cases.

As our main result, we will show that most social choice functions of interest either are, or can be made, psychologically robust. Consequently, for these social choice functions there is no need to worry about the details of multidimensional design. Instead, there will be an easy solution which makes it possible to reach the outcome without knowledge of the correct behavioral model.

5.1 The Insurance Property

From a player i's perspective, the types of all other players are random quantities. The social choice function might, however, provide players with payoff insurance against that randomness. Formally, insurance can be defined in terms of expected payoffs as follows.

Definition 6. *Given an environment E, a social choice function f has the insurance property if, for all i and $j \neq i$,*

$$\mathbb{E}_{\theta_{-j}}[v_i(q_i^f(\theta'_j, \theta_{-j}), \theta_i) + t_i^f(\theta'_j, \theta_{-j})] = \mathbb{E}_{\theta_{-j}}[v_i(q_i^f(\theta''_j, \theta_{-j}), \theta_i) + t_i^f(\theta''_j, \theta_{-j})] \tag{5}$$

for any pair θ'_j, θ''_j of possible realizations of j's type.

Note that insurance is required only bilaterally, taking an expectation over the types of all remaining players. Changing the types of two or more players simultaneously can still have an impact on player i's expected payoff.[18] The following proposition provides an alternative characterization of the insurance property in game-theoretic terms.

Proposition 3. *A social choice function f has the insurance property if and only if in the corresponding direct mechanism we have that, for all i and $j \neq i$,*

$$\Pi_i(s'_j, s_{-j}^T) = \Pi_i(s''_j, s_{-j}^T) \tag{6}$$

for any pair s'_j, s''_j of possible strategies of player j.

When the insurance property is satisfied, players are unable to affect each others' payoffs by unilaterally misreporting the own type in the direct mechanism. The social choice function then remains implementable with interdependent preferences provided that it is implementable in BNE. This follows from the more general insight that in situations where players do not have the possibility to affect the payoffs of others, they will behave as if they were egoistic maximizers.[19] In the following, we will say that a social choice function is *psychologically robust* whenever it is implementable in BNE and exhibits the insurance property.

[18] With only two individuals, the insurance property is virtually the opposite of the bilateral externalities property from Definition 5. It is possible, however, that an SCF neither exhibits bilateral externalities nor satisfies the insurance property, whenever only one player is indifferent with respect to the other's type.

[19] Similar observations, albeit not in a mechanism design framework, have already been made by Levine (1998), Fehr and Schmidt (1999), Bolton and Ockenfels (2000) or Segal and Sobel (2007). In mechanism design theory, the property of "non-bossiness" (Satterthwaite and Sonnenschein, 1981) requires that no agent can ever change the outcome for another agent without also changing the own outcome. Similarly, Baliga and Sjöström (2011) conjecture that mechanisms in which players can influence their opponents' payoffs without own sacrifice "...may have little hope of practical success if agents are inclined to manipulate each others' payoffs due to feelings of spite

We want to illustrate the idea of psychological robustness by introducing a specific and simple equilibrium concept. It is a robust version of a BNFE which relaxes the assumption that the parameters y are observable to the mechanism designer.[20]

Definition 7. *A strategy profile s^* is a psychologically robust equilibrium (PRE) if it is a BNFE for all $y \in [0, \infty[^{n(n-1)}$.*

Proposition 4. *If a social choice function is implementable in BNE and has the insurance property, then it is implementable in PRE.*

In the following sections, we will make heavy use of the insurance property. It will enable us to show that, in our independent private values environment, we can make essentially any social choice function psychologically robust, provided that it is implementable in BNE and provided that we require budget balance only in expected terms. Hence, in these cases, the requirement of psychological robustness is not more demanding than the conventional notion of implementability in BNE. We will then show that, in symmetric environments, materially efficient and symmetric social choice functions can be made psychologically robust even if we insist on ex post budget balance. Finally, we will use the insurance property to show that social choice functions that can be made the outcome of a screening procedure or those that admit a decentralization via a price system are also psychologically robust.

5.2 A Possibility Result Based on Expected Budget Balance

Consider first an environment where both surplus and deficit of a mechanism are possible, that is, where $T = \mathbb{R}^n$ is the unrestricted set of transfers. Budget balance in expectation, $\mathbb{E}_\theta \left[\sum_{i=1}^n t_i^f(\theta)\right] \leq 0$, is often a reasonable constraint in such environments. The following theorem shows that, to *any* social choice function that is implementable in BNE, there exists another one that has the insurance property and is essentially equivalent otherwise. In particular, it entails the same expected transfers and utilities.

Theorem 3. *Let f be an SCF that is implementable in BNE. Then there exists an SCF $\bar{f}$ that has the following properties:*
(a) The decision rule is the same as under f: $q_i^{\bar{f}}(\theta) = q_i^f(\theta)$ for all $i \in I$ and $\theta \in \Theta$.
(b) Expected transfers are the same as under f: $\mathbb{E}_\theta \left[\sum_{i=1}^n t_i^{\bar{f}}(\theta)\right] = \mathbb{E}_\theta \left[\sum_{i=1}^n t_i^f(\theta)\right]$.
(c) Interim payoffs of every individual $i \in I$ and type $\theta_i \in \Theta_i$ are the same as under f:

$$\mathbb{E}_{\theta_{-i}}[v_i(q_i^{\bar{f}}(\theta_i, \theta_{-i}), \theta_i) + t_i^{\bar{f}}(\theta_i, \theta_{-i})] = \mathbb{E}_{\theta_{-i}}[v_i(q_i^f(\theta_i, \theta_{-i}), \theta_i) + t_i^f(\theta_i, \theta_{-i})].$$

(d) $\bar{f}$ is implementable in BNE.
(e) $\bar{f}$ has the insurance property.

or kindness." From our perspective, the following qualifications are appropriate. First, with common knowledge of the kindness generating process, giving players the opportunity to affect each others' payoffs becomes, quite to the contrary, an important design instrument. Without common knowledge, the insurance property implies robustness but it does not contain any requirement about the impact of deviations on own outcomes or payoffs. Fehr et al. (2011b) report on the behavioral non-robustness of the Moore-Repullo mechanism for subgame-perfect implementation under symmetric information. This mechanism does not satisfy what would be an appropriate modification of our insurance property for the Moore-Repullo framework.

[20] Appendix B.2 contains a detailed discussion of this solution concept. In particular, it is shown that the revelation principle holds for implementation in PRE.

The proof is constructive. We start from f and construct a new social choice function $\bar{f}$ which has the same decision rule as f but a modified transfer scheme $(t_1^{\bar{f}}, \ldots, t_n^{\bar{f}})$. This modified transfer scheme ensures that the payoff of type θ_i of individual i depends only on θ_i but not on θ_{-i}, which implies that the insurance property is satisfied.[21] Formally, denote by

$$\rho(\theta_i) = \mathbb{E}_{\theta_{-i}}[v_i(q_i^f(\theta_i, \theta_{-i}), \theta_i) + t_i^f(\theta_i, \theta_{-i})] \tag{7}$$

the ex interim expected payoff of type θ_i of player i under the initial social choice function f, and construct the new transfer scheme so that

$$v_i(q_i^f(\theta_i, \theta_{-i}), \theta_i) + t_i^{\bar{f}}(\theta_i, \theta_{-i}) = \rho(\theta_i) \ , \tag{8}$$

for all θ_{-i}. Note that equations (7) and (8) pin down $t_i^{\bar{f}}(\theta)$, for every $i \in I$ and $\theta \in \Theta$. By construction, expected payments are the same under both social choice functions, since

$$\mathbb{E}_{\theta_{-i}}[t_i^{\bar{f}}(\theta_i, \theta_{-i})] = \rho(\theta_i) - \mathbb{E}_{\theta_{-i}}[v_i(q_i^f(\theta_i, \theta_{-i}), \theta_i)] = \mathbb{E}_{\theta_{-i}}[t_i^f(\theta_i, \theta_{-i})] \ .$$

Now, since consumption levels and expected payments are, for every type of every individual, identical under the new and the initial social choice function, it follows that the new social choice function is also implementable in BNE.

Theorem 3 implies that, when budgetary implications matter only in expected terms, we can psychologically robustly implement any decision rule that would be implementable in BNE. Implementability in a world inhabited by selfish homines oeconomicii is then not only necessary but also sufficient for psychological robustness. For instance, d'Aspremont and Gerard-Varet (1979) establish the possibility to implement *any* surplus-maximizing decision rule in BNE, together with the stronger requirement of ex post budget balance. Theorem 3 then implies that any such rule can be made psychologically robust if we are allowed to replace ex post budget balance by budget balance in expectation. Theorem 3 also extends to models that include interim participation constraints, i.e. minimal values of $\mathbb{E}_{\theta_{-i}}[v_i(q_i^f(\theta_i, \theta_{-i}), \theta_i) + t_i^f(\theta_i, \theta_{-i})]$ for different players i and types θ_i. The reason is that the transfer schemes $(t_1^{\bar{f}}, ..., t_n^{\bar{f}})$ and $(t_1^f, ..., t_n^f)$ are such that the interim payoffs of all types and all players are the same. For some prominent applications this means the following:

- *Second-price auction:* A second-price auction is a special case of the above environment, where f is such that a private good is assigned to an individual with maximal valuation, who then has to pay the second highest valuation. As is well known, truth-telling is a BNE of the corresponding direct mechanism (in fact, even an equilibrium in dominant strategies), and all types of all individuals are willing to participate. For this setting, Theorem 3 implies that there exists a modified version of the second-price auction that shares all these properties and is psychologically robust.

- *Partnership dissolution:* The problem to dissolve a partnership efficiently, which has been

[21] Börgers and Norman (2009) investigate a related question, asking under which conditions an otherwise equivalent but ex post budget balanced SCF exists for a given SCF. In contrast, we are interested in insurance against other players' type realizations, not insurance against mechanism deficits.

studied by Cramton et al. (1987), is also a special case of our setup. In this application, shares of an object have to be assigned to a number of agents who have private information about their valuation of those shares, and who have pre-specified property rights. Cramton et al. (1987) require that participation constraints are satisfied, or equivalently, that an agreement is reached with unanimity, and obtain possibility results for an efficient allocation of shares. Theorem 3 implies that these results can be generalized to a setting where (anti-)social preferences matter among partners.

- *Public goods provision:* Various authors have studied the provision of excludable and non-excludable public goods in the presence of participation constraints (Güth and Hellwig 1986, Hellwig 2003, Norman 2004). Again, Theorem 3 implies that these results generalize to a setting where individuals are willing to react in a reciprocal way to the other agents' contributions to a public good.

Whether or not expected (as opposed to ex post) budget balance is a reasonable requirement will depend on the application. When it comes to public goods provision, insisting on ex post budget balance is appropriate if there is no external source of funds that may help to cover the provision costs. By contrast, an auctioneer who runs several independent auctions may be willing to accept losses on some, provided that overall there a is positive expected surplus. A focus on expected budget balance can also be justified if the number of individuals is large. With many individuals, the probability that the mechanism runs a surplus or a deficit larger than ε, for an arbitrary $\varepsilon > 0$, converges to zero if and only if expected budget balance holds, due to the law of large numbers.

5.3 A Possibility Result Based on Ex Post Budget Balance

Let us return to an environment where $T = \bar{T}$, so that mechanisms can never run a deficit. Full efficiency of an SCF then requires, in addition to the decision rule being surplus-maximizing, the transfers to satisfy ex post budget balance: $\sum_{i=1}^{n} t_i^f(\theta) = 0$, for all $\theta \in \Theta$.

To provide sufficient conditions for psychological robustness under ex post budget balance, we study the expected externality mechanism of d'Aspremont and Gerard-Varet (1979) or Arrow (1979), following the exposition in Mas-Colell et al. (1995, chapter 23). The expected externality mechanism is a direct mechanism $\Phi = [\Theta_1, ..., \Theta_n, f]$ where the SCF f is constructed as follows. First, the decision rule $(q_1^f, ..., q_n^f)$ is surplus-maximizing, so that for all $\theta \in \Theta$,

$$(q_1^f(\theta), \ldots, q_n^f(\theta)) \in \arg \max_{(q_1,\ldots,q_n)\in Q} \sum_{i=1}^{n} v_i(q_i, \theta_i).$$

The transfers $(t_1^f, ..., t_n^f)$ are such that

$$t_i^f(\theta_i, \theta_{-i}) = \mathbb{E}_{\theta_{-i}} \left[\sum_{j \neq i} v_j(q_j^f(\theta_i, \theta_{-i}), \theta_j) \right] + h_i(\theta_{-i}), \tag{9}$$

where $h_i(\theta_{-i})$ is defined as as

$$h_i(\theta_{-i}) = -\left(\frac{1}{n-1}\right)\sum_{j\neq i}\mathbb{E}_{\theta_{-j}}\left[\sum_{h\neq j} v_h(q_h^f(\theta_j,\theta_{-j}),\theta_h)\right]. \tag{10}$$

This social choice function satisfies ex post budget balance. Moreover, it satisfies the incentive compatibility constraints in (3), implying that truth-telling is a BNE in the corresponding direct mechanism. In the following we show that it also has the insurance property if $n = 2$ or if environment and decision rule jointly satisfy a property that we call symmetry.

Theorem 4. *Consider an expected externality mechanism $\Phi = [\Theta_1, ..., \Theta_n, f]$. The SCF f has the insurance property if $n = 2$ or if symmetry holds, so that for all i and $j \neq i$, $k \neq i$,*

$$\mathbb{E}_{\theta_{-i}}\left[v_j(q_j^f(\theta_i,\theta_{-i}),\theta_j)\right] = \mathbb{E}_{\theta_{-i}}\left[v_k(q_k^f(\theta_i,\theta_{-i}),\theta_k)\right] \tag{11}$$

for all types $\theta_i \in \Theta_i$.

Symmetry requires that all opponents of i obtain an identical expected payoff net of transfers, no matter which type $\theta_i \in \Theta_i$ is realized.[22] Symmetry arises naturally if the environment is such that all players have identical payoff functions, their types are independently and identically distributed, and the decision rule $(q_1^f, ..., q_n^f)$ treats all individuals identically.

The intuition for Theorem 4 is as follows: The expected externality mechanism derives its name from the fact that each player pays for the expected impact that his strategy choice has on the other players' payoffs. If there are just two players, this implies that player 1 is perfectly insured against the randomness in player 2's type, or equivalently, against changes of player 2's strategy. Under symmetry, the argument generalizes to more than two players: If each player's externalities are evenly distributed among all other players, then, once more, the insurance property is satisfied.[23] To illustrate that the insurance property will generally not be satisfied without symmetry, in section A.9 of the appendix we give a simple three player example of an efficient but non-symmetric decision rule for which the the expected externality mechanism does not satisfy the insurance property.

5.4 Extension: Robustness of Screening Mechanisms and Price Systems

As an extension, we use our formalism to assess the psychological robustness of social choice functions that are of particular interest from the perspective of general equilibrium theory, public finance and contract theory. In general equilibrium theory and public finance, one typically looks at social choice functions that admit a decentralization via a (possibly non-linear) price system, that may be shaped by the government's tax policy. A prominent topic in contract theory is the design of optimal screening or incentive schemes.

[22] In contrast to the insurance property, symmetry does *not* require the opponents' payoffs to be independent of θ_i, but rather that the opponents' payoffs are always identical and are thus affected *equally* by player i's type.

[23] For an example with two players, the property that each player's payoff is independent of the other's announcement has also been observed by Mathevet (2010, p. 414). Desiraju and Sappington (2007) show that, in their model with two agents and privately observed cost parameters, transfers can be structured so as to avoid ex post *inequality* and hence render inequality aversion irrelevant, whenever the agents are ex ante identical.

We consider an environment E as described in Section 3.1, but we now allow for allocations $a = (a_1, \ldots, a_n) \in A = A_1 \times \ldots \times A_n$ that specify for each individual i a consumption bundle $a_i \in A_i$, which can be multidimensional. The set of possible types of individual i is given by $\Theta_i = \{\theta_i^1, \ldots, \theta_i^{m_i}\}$. Payoff functions are given by $\pi_i(a_i, \theta_i)$, so that we remain within a private values framework but dispense with the assumption of quasi-linearity. A social choice function $f = (a_1^f, \ldots, a_n^f)$ specifies for each individual a bundle $a_i^f(\theta) \in A_i$ for each vector of types. We focus on social choice functions that are *simple* in the following sense: For each individual i there exist bundles $\bar{a}_i^1, \ldots, \bar{a}_i^{m_i}$ so that

$$a_i^f(\theta_i^k, \theta_{-i}) = \bar{a}_i^k$$

for any $k = 1, ..., m_i$, independently of θ_{-i}. The term "simple" is borrowed from Dierker and Haller (1990). A simple social choice function does not make use of the possibility to make the outcome of individual i dependent on the types of the other individuals. Rather, the outcome for i is a function of i's type only.[24]

There are many applications where one is interested in simple social choice functions. Examples include the the study of insurance markets with adverse selection á la Rothschild and Stiglitz (1976), the study of optimal monopoly regulation in the tradition of Baron and Myerson (1982), mechanism design approaches to the Mirrlees (1971)-problem of optimal income taxation, such as Stiglitz (1982), or mechanism design approaches to problems of non-linear pricing such as Mussa and Rosen (1978).

Simple social choice functions obviously have the insurance property. The property of simplicity is in fact much stronger than the insurance property. Hence a simple SCF is psychologically robust if and only if it is implementable in BNE. Implementability in BNE in turn holds if and only if the following incentive compatibility constraints are satisfied: For each individual i and all $k, l \in \{1, ..., m_i\}$,

$$\pi_i(\bar{a}_i^k, \theta_i^k) \geq \pi_i(\bar{a}_i^l, \theta_i^k). \tag{12}$$

The following proposition summarizes these observations.

Proposition 5. *A simple SCF is psychologically robust if and only if it satisfies the incentive compatibility constraints in (12).*

A related question is whether outcomes that can be decentralized by means of a price system are psychologically robust. In many applications one is actually interested in such social choice functions. This is true for any application of general equilibrium theory. It is also true for any model of public finance which rests on the assumption that the tax system shapes an individual's budget set, and that, given those budget sets, individuals solve consumer choice problems. For

[24] Simple social choice functions raise two questions. First, there is the question of what the appropriate resource constraint looks like. Models that work with simple SCFs typically require feasibility in expectation. With many individuals, this may be justified with an appeal to the law of large numbers. The second question is whether simple social choice functions are optimal from a normative perspective. Bierbrauer (2011) provides a discussion of this question in the context of a Mirrleesian model of optimal redistributive income taxation, and provides conditions which imply an optimality of simple social choice functions.

instance, this holds for any model of taxation using the framework of Ramsey (1927). It is also true for the original formulation of the optimal income tax problem by Mirrlees (1971).

We will use an argument by Hammond (1979) to demonstrate the psychological robustness of all these social choice functions. For simplicity, consider a symmetric version of the environment above, so that all individuals have the same consumption set $A_i = \bar{A}$, the same set of possible types $\Theta_i = \bar{\Theta}$, and the same payoff function $\pi_i = \pi$. A simple SCF f then specifies bundles $\bar{a}^1, \ldots, \bar{a}^m \in \bar{A}$ so that

$$a_i^f(\theta^k, \theta_{-i}) = \bar{a}^k$$

for all i and $k = 1, ..., m$. Hammond (1979) shows that such a social choice function satisfies the incentive compatibility constraints in (12) if and only if it can be decentralized by means of a (possibly non-linear) budget set, i.e., if and only if there exists some set $B \subseteq \bar{A}$ such that

$$\bar{a}^k \in \arg\max_{b \in B} \pi(b, \theta^k) \tag{13}$$

for all $k = 1, ..., m$. Proposition 5 therefore implies, in particular, that any social choice function that can be decentralized by some budget set is psychologically robust. For instance, whatever the inclination of individuals to reward kind and to punish unkind behavior of others, a competitive equilibrium allocation or an allocation that is induced by some tax system is implementable provided that it is implementable in a model with selfish individuals. Dufwenberg et al. (2011b) arrive at a similar conclusion of behavioral irrelevance of other-regarding preferences in competitive equilibrium under complete information, for a comprehensive model of social preferences that may depend both on outcomes and consumption opportunities.

6 Conclusions

This paper went through the workhorse model of mechanism design theory, known as the independent private values model, under the assumption that individuals have intention-based social preferences. The analysis had two main parts, which differ with respect to the degree of psychological sophistication. In the first part, we assumed common knowledge about all the details of how kindness sensations emerge, so that the designer can exploit this knowledge when designing an allocation mechanism. This informational requirement is demanding, and it has to be seen in applied and experimental work to what extent our insights can be used in practice. In any case, the implications are drastic: The revelation principle fails and the design of choice sets becomes a non-trivial part of mechanism design. In addition, all outcomes that are, in a conventional sense, efficient, become implementable, so that several of the classical impossibility results for which mechanism design theory is famous are turned into possibility results. Finally, it becomes possible to compare different mechanisms which all achieve the same economic outcome according to the kindness sensations that they induce. There may even be a best mechanism for a given economic outcome. The second part of the paper dispensed with the common knowledge assumption. We asked whether a mechanism designer can still resort to the mechanisms and outcomes which have attracted the attention of conventional mechanism

design theory. Put differently, we asked whether these mechanisms are psychologically robust, in the sense that they yield the desired outcome even though they were derived from a (possibly) misspecified model of behavior, namely selfish payoff maximization. Our answer to this question is basically "yes". We show that conventional mechanisms – such as a second-price auction, an expected externality mechanism, a screening mechanism, or a price system – need to be adjusted only slightly, if at all, in order to achieve psychological robustness.

There are many interesting and important questions that could be addressed by future research. First, one might want to go beyond the independent private values case, or examine a framework with complete, but non-verifiable information (Maskin 1999). Second, modelling the mechanism designer as a player, to whose behavior the agents attribute good or bad intentions, can be a relevant exercise for applications such as auction design or the design of optimal public policy. Finally, working out the details of psychologically robust mechanisms for many of the relevant applications strikes us as important and promising.

In addition, the following issues deserve closer scrutiny. First, with the conventional solution concept of a Bayes-Nash equilibrium, the focus on normal form mechanisms is typically justified as follows: Any equilibrium in an extensive form mechanism remains an equilibrium in the corresponding normal form, so that moving from normal to extensive form mechanisms can only reduce the set of implementable social choice functions. It is unclear whether this is also true with intention-based social preferences. A major obstacle to answering this question is the lack of a general theory of intentions in extensive form games with incomplete information. Second, there is the question which social choice functions can be implemented as the unique equilibrium outcome of some mechanism. Again, it is an open question whether the answers which have been given in the classical approach extend to models with interdependent preferences.

Finally, several of our results lend themselves to experimental testing. First and foremost, this concerns the role of unused actions as a design instrument. Second, our analysis has explored, theoretically, the possibility to rank different mechanisms that implement the same material outcome in the kindness dimension. This raises the question whether differences in kindness perceptions across outcome-equivalent mechanisms can also be identified empirically.

References

Aldashev, G., Kirchsteiger, G., and Sebald, A. (2010). How (not) to decide: Procedural games. Mimeo, University of Copenhagen.

Alger, I. and Renault, R. (2006). Screening ethics when honest agents care about fairness. *International Economic Review*, 47:59–85.

Andreoni, J., Brown, P., and Vesterlund, L. (2002). What makes an allocation fair? some experimental evidence. *Games and Economic Behavior*, 40:1–24.

Arrow, K. (1979). The property rights doctrine and demand revelation under incomplete information. In Boskin, M. J., editor, *Economics and Human Welfare*. Academic Press, New York.

Baliga, S. and Sjöström, T. (2011). Mechanism design: Recent developments. In Blume, L. and Durlauf, S., editors, *The New Palgrave Dictionary of Economics*.

Bar-Gill, O. and Fershtman, C. (2004). Law and preferences. *Journal of Law Economics and Organization*, 20:331–352.

Bar-Gill, O. and Fershtman, C. (2005). Public policy with endogenous preferences. *Journal of Public Economic Theory*, 7:841–857.

Baron, D. and Myerson, R. (1982). Regulating a monopolist with unknown costs. *Econometrica*, 50:911–930.

Bartling, B. (2011). Relative performance or team evaluation? Optimal contracts for other-regarding agents. *Journal of Economic Behavior and Organization*, 79:183–193.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97:170–176.

Battigalli, P. and Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, 144:1–35.

Benjamin, D. (2010). Social preferences and the efficiency of bilateral exchange. mimeo.

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.

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

Bowles, S. (2008). Policies designed for self-interested citizens may undermine "the moral sentiments": Evidence from economic experiments. *Science*, 320:1605–1609.

Bowles, S. and Hwang, S.-H. (2008). Social preferences and public economics: Mechanism design when social preferences depend on incentives. *Journal of Public Economics*, 92:1811–1820.

Bowles, S. and Reyes, S. (2009). Economic incentives and social preferences: A preference-based lucas critique of public policy. CESifo Working Paper No. 2734.

Börgers, T. and Norman, P. (2009). A note on budget balance under interim participation constraints: The case of independent types. *Economic Theory*, 39:477–489.

Cabrales, A. and Calvó-Armengol, A. (2008). Interdependent preferences and segregating equilibria. *Journal of Economic Theory*, 139:99–113.

Cabrales, A., Calvó-Armengol, A., and Pavoni, N. (2007). Social preferences, skill segregation, and wage dynamics. *Review of Economic Studies*, 74:1–33.

Cabrales, A., Miniaci, R., Piovesan, M., and Ponti, G. (2010). Social preferences and strategic uncertainty: An experiment on markets and contracts. *American Economic Review*, 100:2261–2278.

Cabrales, A. and Serrano, R. (2011). Implementation in adaptive better-response dynamics: Towards a general theory of bounded rationality in mechanisms. *Games and Economic Behavior*, 73:360–374.

Caplin, A. and Eliaz, K. (2003). Aids policy and psychology: A mechanism-design approach. *RAND Journal of Economics*, 34:631–646.

Caplin, A. and Leahy, J. (2004). The supply of information by a concerned expert. *Economic Journal*, 114:487–505.

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

Cox, J., Friedman, D., and Gjerstad, S. (2007). A tractable model of reciprocity and fairness. *Games and Economic Behavior*, 59:17–45.

Cramton, P., Gibbons, R., and Klemperer, P. (1987). Dissolving a partnership efficiently. *Econometrica*, 55:615–632.

d'Aspremont, C. and Gerard-Varet, L.-A. (1979). Incentives and incomplete information. *Journal of Public Economics*, 11:25–45.

Desiraju, R. and Sappington, D. (2007). Equity and adverse selection. *Journal of Economics and Mangement Strategy*, 16:285–318.

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

Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2009). Homo reciprocans: Survey evidence on behavioral outcomes. *Economic Journal*, 119:592–612.

Dufwenberg, M., Gächter, S., and Hennig-Schmidt, H. (2011a). The framing of games and the psychology of play. *Games and Economic Behavior*, 73:459–478.

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

Dufwenberg, M. and Kirchsteiger, G. (2000). Reciprocity and wage undercutting. *European Economic Review*, 44:1069–1078.

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

Dufwenberg, M., Smith, A., and Van Essen, M. (2011c). Hold-up: With a vengeance. *Economic Inquiry*, pages 1–13.

Eliaz, K. (2002). Fault tolerant implementation. *Review of Economic Studies*, 69:589–610.

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

Englmaier, F. and Leider, S. (2012). Contractual and organizational structure with reciprocal agents. *American Economic Journal: Microeconomics*, forthcoming.

Englmaier, F. and Wambach, A. (2010). Optimal incentive contracts under inequity aversion. *Games and Economic Behavior*, 69:312–328.

Falk, A., Fehr, E., and Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry*, 41:20–26.

Falk, A., Fehr, E., and Fischbacher, U. (2008). Testing theories of fairness - intentions matter. *Games and Economic Behavior*, 62:287–303.

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

Fehr, E. and Falk, A. (2002). Psychological foundations of incentives. *European Economic Review*, 46:687–724.

Fehr, E., Gächter, S., and Kirchsteiger, G. (1997). Reciprocity as a contract enforcement device: Experimental evidence. *Econometrica*, 65:833–860.

Fehr, E., Hart, O., and Zehnder, C. (2011a). Contracts as reference points - experimental evidence. *American Economic Review*, 101:493–525.

Fehr, E., Klein, A., and Schmidt, K. (2007). Fairness and contract design. *Econometrica*, 75:121–154.

Fehr, E., Powell, M., and Wilkening, T. (2011b). Handing out guns at a knife fight: Behavioral limitations to the moore-repullo mechanism. mimeo.

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

Frey, B., Benz, M., and Stutzer, A. (2004). Introducing procedural utility: Not only what, but also how matters. *Journal of Institutional and Theoretical Economics*, 160:377–401.

Frey, B. and Stutzer, A. (2004). Beyond outcomes: Measuring procedural utility. *Oxford Economic Papers*, 57:90–111.

Gaspart, F. (2003). A general concept of procedural fairness for one-stage implementation. *Social Choice and Welfare*, 21:311–322.

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

Glazer, A. and Rubinstein, A. (1998). Motives and implementation: On the design of mechanisms to elicit opinions. *Journal of Economic Theory*, 79:157–173.

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

Hahn, V. (2009). Reciprocity and voting. *Games and Economic Behavior*, 67:467–480.

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

Hart, O. and Moore, J. (2008). Contracts as reference points. *Quarterly Journal of Economics*, 123:1–48.

Heifetz, A., Segev, E., and Talley, E. (2007). Market design with endogenous preferences. *Games and Economic Behavior*, 58:121–153.

Hellwig, M. (2003). Public-good provision with many participants. *Review of Economic Studies*, 70:589–614.

Jackson, M. (2001). A crash course in implementation theory. *Social Choice and Welfare*, 18:655–708.

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*.

Jehiel, P., Moldovanu, B., and Stachetti, E. (1996). How (not) to sell nuclear weapons. *American Economic Review*, 86:814–829.

Kahneman, D., Wakker, P., and Sarin, R. (1997). Back to Bentham? explorations of experienced utility. *Quarterly Journal of Economics*, 112:375–405.

Kosfeld, M. and von Siemens, F. (2011). Competition, cooperation, and corporate culture. *RAND Journal of Economics*, 42:23–43.

Kucuksenel, S. (2011). Behavioral mechanism design. *Journal of Public Economic Theory*, forthcoming.

Levine, D. (1998). Modelling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1:593–622.

Li, J. (2008). The power of conventions: A theory of social preferences. *Journal of Economic Behavior and Organization*, 65:489–505.

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

Mas-Colell, A., Whinston, M., and Greene, J. (1995). *Microeconomic Theory*. Oxford University Press, USA.

Maskin, E. (1999). Nash equilibrium and welfare optimality. *Review of Economic Studies*, 66:23–38.

Mathevet, L. (2010). Supermodular mechanism design. *Theoretical Economics*, 5:403–443.

Matsushima, H. (2005). On detail-free mechanism design and rationality. *Japanese Economic Review*, 56:41–54.

Matsushima, H. (2008a). Behavioral aspects of implementation theory. *Economics Letters*, 100:161–164.

Matsushima, H. (2008b). Role of honesty in full implementation. *Journal of Economic Theory*, 139:353–359.

McCabe, K., Rigdon, M., and Smith, V. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization*, 52:267–275.

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

Mookherjee, D. and Reichelstein, S. (1990). Implemention via augmented revelation mechanisms. *Review of Economic Studies*, 57:453–475.

Morgan, J., Steiglitz, K., and Reis, G. (2003). The spite motive and equilibrium behavior in auctions. *Contributions to Economic Analysis & Policy*, 2:1–25.

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

Myerson, R. (1979). Incentive compatibility and the bargaining problem. *Econometrica*, 47:61–73.

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

Netzer, N. and Schmutzler, A. (2010). Rotten kids with bad intentions. SOI Discussion Paper No. 0919.

Nishimura, N., Cason, T., Saijo, T., and Ikeda, Y. (2011). Spite and reciprocity in auctions. *Games*, 2:365–411.

Norman, P. (2004). Efficient mechanisms for public goods with use exclusion. *Review of Economic Studies*, 71:1163–1188.

Offerman, T. (2002). Hurting hurts more than helping helps. *European Economic Review*, 46:1423–1437.

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

Ramsey, F. (1927). A contribution to the theory of taxation. *Economic Journal*, 37:47–61.

Renou, L. and Schlag, K. (2011). Implementation in minimax regret equilibrium. *Games and Economic Behavior*, 71:527–533.

Rodriguez-Sickert, C., Guzman, R., and Cardenas, J. (2008). Institutions influence preferences: Evidence from a common pool resource experiment. *Journal of Economic Behavior and Organization*, 67:215–227.

Rothschild, M. and Stiglitz, J. (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics*, 90:629–650.

Satterthwaite, M. and Sonnenschein, H. (1981). Strategy-proof allocation mechanisms at differentiable points. *Review of Economic Studies*, 48:587–597.

Sebald, A. (2010). Attribution and reciprocity. *Games and Economic Behavior*, 68:339–352.

Segal, U. and Sobel, J. (2007). Tit for tat: Foundations of preferences for reciprocity in strategic settings. *Journal of Economic Theory*, 136:197–216.

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

Stanca, L. (2010). How to be kind? Outcomes versus intentions as determinants of fairness. *Economics Letters*, 106:19–21.

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

von Siemens, F. (2009). Bargaining under incomplete information, fairness, and the hold-up problem. *Journal of Economic Behavior and Organization*, 71:486–494.

von Siemens, F. (2011). Heterogeneous social preferences, screening, and employment contracts. *Oxford Economic Papers*, 63:499–522.

A Appendix

A.1 Proof of Proposition 1

We first show that $G_i'(m_{-i}') \subseteq G_i(h_{-i}(m_{-i}'))$. Let $a \in G_i'(m_{-i}')$. Hence, there exists m_i' so that $g'(m_i', m_{-i}') = a$. By (2), this implies that $g(h_i(m_i'), h_{-i}(m_{-i}')) = a$, and hence $a \in G_i(h_{-i}(m_{-i}'))$. We now show that $G_i(h_{-i}(m_{-i}')) \subseteq G_i'(m_{-i}')$. Let $a \in G_i(h_{-i}(m_{-i}'))$. Hence, there exists $m_i \in M_i$ so that $g(m_i, h_{-i}(m_{-i}')) = a$. Since the function h is surjective there exists m_i' with $h_i(m_i') = m_i$. Then (2) implies that $g'(m_i', m_{-i}') = a$. Hence, $a \in G_i'(m_{-i}')$.

A.2 Proof of Theorem 1

Idea and structure of proof. We first deal with the case of 2 agents, $n = 2$. Below, we explain how to generalize the argument for an arbitrary number of players.

Fix an efficient SCF f and consider the direct mechanism $\Phi = [M_1, M_2, g]$ for f, that is, $M_1 = \Theta_1$, $M_2 = \Theta_2$ and $g = f$. We denote by S_i the set of player i's pure strategies, and by $\Pi_i : S_1 \times S_2 \to \mathbb{R}$ player i's ex ante expected material payoff. Let $s^T = (s_1^T, s_2^T)$ be the truth-telling profile, with its associated kindness values $\kappa_{21}(s^T)$ and $\kappa_{12}(s^T)$. As argued in the body of the text, efficiency of f implies that s^T is a BNFE in Φ when $\kappa_{21}(s^T) = 1/y_1$ and $\kappa_{12}(s^T) = 1/y_2$. Now, in general this condition on kindness values will not be satisfied by s^T in the direct mechanism. We will therefore show how to extend Φ to an augmented revelation mechanism $\tilde{\Phi}$, by introducing additional messages for each player j to adjust $\kappa_{ji}(s^T)$ to the desired value $1/y_i$. Once $\kappa_{ji}(s^T) = 1/y_i$ holds in $\tilde{\Phi}$ for both players, efficiency of f again implies that s^T is a BNFE in $\tilde{\Phi}$, and the additional messages will actually remain unused.

We first study the problem of adding messages for player 1 to manipulate

$$\kappa_{12}(s^T) = \Pi_2(s^T) - \pi_2^{e_1}(s^T).$$

This kindness value can either be increased by making the equitable payoff $\pi_2^{e_1}(s^T)$ smaller, or be decreased by making $\pi_2^{e_1}(s^T)$ larger. We will show that, starting from the direct mechanism, $\pi_2^{e_1}(s^T)$ can in fact be adjusted to any arbitrary value by an appropriate choice of $\tilde{\Phi}$.

In the following we denote by $E_1(s_2^T)$ the set of conditionally efficient strategies of player 1 in the direct mechanism, and by

$$\pi_2^{e_1}(s_2^T) = \frac{1}{2}\left[\min_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T) + \max_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T)\right]$$

the equitable payoff in the direct mechanism.

Decreasing $\pi_2^{e_1}(s_2^T)$. Let $s_1^{min} \in \arg\min_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T)$ be a strategy that minimizes player 2's payoff among efficient strategies in the direct mechanism Φ. Then $s_1^{min}(\Theta_1) \subseteq \Theta_1$ is the range of s_1^{min}, and $r = |s_1^{min}(\Theta_1)|$ is its cardinality. Let $\sigma : \{d_1, d_2, ..., d_r\} \to s_1^{min}(\Theta_i)$ be an arbitrary bijective function, which assigns to every element of the set $\{d_1, d_2, ..., d_r\}$ a distinct type from the range $s_1^{min}(\Theta_i)$. The inverse of σ is denoted σ^{-1}. We now construct the extended mechanism $\tilde{\Phi} = [\tilde{M}_1, \tilde{M}_2, \tilde{g}]$ from Φ as follows. Let $\tilde{M}_1 = \Theta_1 \cup \{d_1, d_2, ..., d_r\}$ and $\tilde{M}_2 = \Theta_2$, i.e. we add r new messages for player 1 and keep player 2's message set unchanged. Hence we have new strategy sets $\tilde{S}_1 \supset S_1$ and $\tilde{S}_2 = S_2$. We construct $\tilde{g}$ from g by keeping $\tilde{g}(m) = g(m)$ whenever $m \in \Theta_1 \times \Theta_2$. When $m = (d_j, \theta_2)$ for some $j \in \{1, ..., r\}$ and $\theta_2 \in \Theta_2$, we define $\tilde{g}(m) = (q_1^{\tilde{g}}(m), q_2^{\tilde{g}}(m), t_1^{\tilde{g}}(m), t_2^{\tilde{g}}(m))$ by

$$q_1^{\tilde{g}}(d_j, \theta_2) = q_1^g(\sigma(d_j), \theta_2), \quad q_2^{\tilde{g}}(d_j, \theta_2) = q_2^g(\sigma(d_j), \theta_2)$$

and

$$t_1^{\tilde{g}}(d_j, \theta_2) = t_1^g(\sigma(d_j), \theta_2) + \epsilon, \quad t_2^{\tilde{g}}(d_j, \theta_2) = t_2^g(\sigma(d_j), \theta_2) - \delta$$

for some pre-specified ϵ and δ with $0 < \epsilon \leq \delta$. Hence announcing d_j has the same consequences as announcing type $\sigma(d_j) \in s_1^{min}(\Theta_1) \subseteq \Theta_1$, except for additional transfers from player 2 to player 1. With slight abuse of notation, we denote player i's ex ante expected payoffs in $\tilde{\Phi}$ by

$\Pi_i : \tilde{S}_1 \times \tilde{S}_2 \to \mathbb{R}$ as well, because payoffs in Φ and $\tilde{\Phi}$ coincide on $S_1 \times S_2$, so that we can simply extend Π_i to the new strategy sets. Let $\tilde{E}_1(s_2^T)$ be the set of efficient strategies for player 1 in $\tilde{\Phi}$, conditional on player 2 still telling the truth.

We first prove that the maximization part in the definition of $\pi_2^{e_1}(s_2^T)$ remains unaffected by the mechanism extension.

Lemma 1. *For any arbitrary ϵ, δ with $0 < \epsilon \leq \delta$, it holds that*

$$\max_{s_1 \in \tilde{E}_1(s_2^T)} \Pi_2(s_1, s_2^T) = \max_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T).$$

Proof. Step 1. We first claim that, for the maximization in Φ, we can replace $E_1(s_2^T)$ by S_1, i.e.,

$$\max_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T) = \max_{s_1 \in S_1} \Pi_2(s_1, s_2^T).$$

The inequality $\max_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T) \leq \max_{s_1 \in S_1} \Pi_2(s_1, s_2^T)$ immediately follows from $E_1(s_2^T) \subseteq S_1$. To obtain a contradiction, assume $\max_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T) < \max_{s_1 \in S_1} \Pi_2(s_1, s_2^T)$, and let $s_1^{max} \in \arg\max_{s_1 \in S_1} \Pi_2(s_1, s_2^T)$, which implies $s_1^{max} \notin E_1(s_2^T)$. Then there exists, by finiteness of S_1, a strategy $s_1' \in E_1(s_2^T)$ that Pareto dominates s_1^{max} and satisfies $\Pi_2(s_1', s_2^T) \geq \Pi_2(s_1^{max}, s_2^T) > \max_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T)$, a contradiction.

The same argument holds for the extended mechanism $\tilde{\Phi}$, i.e. we can replace $\tilde{E}_1(s_2^T)$ by $\tilde{S}_1$ when maximizing Π_2. Thus the lemma follows when we have established

$$\max_{s_1 \in \tilde{S}_1} \Pi_2(s_1, s_2^T) = \max_{s_1 \in S_1} \Pi_2(s_1, s_2^T).$$

Step 2. The inequality $\max_{s_1 \in \tilde{S}_1} \Pi_2(s_1, s_2^T) \geq \max_{s_1 \in S_1} \Pi_2(s_1, s_2^T)$ follows from $S_1 \subset \tilde{S}_1$. For the other inequality, observe that for every $\tilde{s}_1 \in \tilde{S}_1 \backslash S_1$ we can construct an associated strategy $s_1 \in S_1$ by replacing the announcement of any $d_j \in \{d_1, d_2, ..., d_r\}$ by $\sigma(d_j)$. Formally, $s_1(\theta_1) = \tilde{s}_1(\theta_1)$ whenever $\tilde{s}_1(\theta_1) \in \Theta_1$, and $s_1(\theta_1) = \sigma(\tilde{s}_1(\theta_1))$ whenever $\tilde{s}_1(\theta_1) \in \{d_1, d_2, ..., d_r\}$. It follows that $\Pi_2(\tilde{s}_1, s_2^T) < \Pi_2(s_1, s_2^T)$, so that $\max_{s_1 \in \tilde{S}_1} \Pi_2(s_1, s_2^T) \leq \max_{s_1 \in S_1} \Pi_2(s_1, s_2^T)$. □

We now examine the minimization part in the definition of $\pi_2^{e_1}(s_2^T)$. We show that the minimum can be decreased to any arbitrary value, by an appropriate choice of δ.

Lemma 2. *For any arbitrary ϵ, δ with $0 < \epsilon \leq \delta$, it holds that*

$$\min_{s_1 \in \tilde{E}_1(s_2^T)} \Pi_2(s_1, s_2^T) = \min_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T) - \delta.$$

Proof. Step 1. Consider strategy $s_1^{min} \in E_1(s_2^T) \subseteq S_1$ from above, and construct the associated strategy $\tilde{s}_1 \in \tilde{S}_1$ by replacing every announcement of a type $\bar{\theta}_1 \in s_1^{min}(\Theta_1)$ by the associated new message $\sigma^{-1}(\bar{\theta}_1)$. Formally, $\tilde{s}_1(\theta_1) = \sigma^{-1}(s_1^{min}(\theta_1))$ for all $\theta_1 \in \Theta_1$. Then it follows that

$$\Pi_1(\tilde{s}_1, s_2^T) = \Pi_1(s_1^{min}, s_2^T) + \epsilon, \quad \Pi_2(\tilde{s}_1, s_2^T) = \Pi_2(s_1^{min}, s_2^T) - \delta. \tag{14}$$

Step 2. We claim that $\tilde{s}_1 \in \tilde{E}_1(s_2^T)$, which then implies

$$\min_{s_1 \in \tilde{E}_1(s_2^T)} \Pi_2(s_1, s_2^T) \leq \min_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T) - \delta.$$

To obtain a contradiction, suppose $\tilde{s}_1 \notin \tilde{E}_1(s_2^T)$, so that there exists $\tilde{s}_1' \in \tilde{S}_1$ such that

$$\Pi_1(\tilde{s}_1', s_2^T) \geq \Pi_1(\tilde{s}_1, s_2^T), \quad \Pi_2(\tilde{s}_1', s_2^T) \geq \Pi_2(\tilde{s}_1, s_2^T), \tag{15}$$

with at least one of the inequalities being strict. Starting from $\tilde{s}_1'$, construct $s_1' \in S_1$ by replacing the announcement of any $d_j \in \{d_1, d_2, ..., d_r\}$ by $\sigma(d_j)$. Formally, $s_1'(\theta_1) = \tilde{s}_1'(\theta_1)$ whenever $\tilde{s}_1'(\theta_1) \in \Theta_1$, and $s_1'(\theta_1) = \sigma(\tilde{s}_1'(\theta_1))$ whenever $\tilde{s}_1'(\theta_1) \in \{d_1, d_2, ..., d_r\}$. We obtain

$$\Pi_1(s_1', s_2^T) = \Pi_1(\tilde{s}_1', s_2^T) - x\epsilon, \quad \Pi_2(s_1', s_2^T) = \Pi_2(\tilde{s}_1', s_2^T) + x\delta, \tag{16}$$

where $x \in [0, 1]$ is the probability of announcements from $\{d_1, d_2, ..., d_r\}$ under $\tilde{s}_1'$. Conditions (14), (15) and (16) together imply

$$\Pi_1(s_1', s_2^T) \geq \Pi_1(s_1^{min}, s_2^T) + (1 - x)\epsilon, \quad \Pi_2(s_1', s_2^T) \geq \Pi_2(s_1^{min}, s_2^T) - (1 - x)\delta, \tag{17}$$

where at least one inequality is strict. If $x = 1$, this contradicts $s_1^{min} \in E_1(s_2^T)$. Hence assume $x < 1$, so that $\Pi_1(s_1', s_2^T) > \Pi_1(s_1^{min}, s_2^T)$. From $s_1^{min} \in E_1(s_2^T)$ it then follows that $\Pi_2(s_1', s_2^T) < \Pi_2(s_1^{min}, s_2^T)$ must hold. Now, if $s_1' \in E_1(s_2^T)$, this contradicts that s_1^{min} minimizes $\Pi_2(s_1, s_2^T)$ on $E_1(s_2^T)$. Otherwise, if $s_1' \notin E_1(s_2^T)$, there exist strategies in S_1 that Pareto dominate s_1'. Any such strategy $s_1'' \in S_1$ must, however, still satisfy $\Pi_2(s_1'', s_2^T) < \Pi_2(s_1^{min}, s_2^T)$, since otherwise it would also Pareto dominate s_1^{min}. Finiteness of S_1 then implies that there exists $s_1'' \in E_1(s_2^T)$ with $\Pi_2(s_1'', s_2^T) < \Pi_2(s_1^{min}, s_2^T)$, which is the final contradiction.

Step 3. The lemma follows when we can also establish the opposite inequality

$$\min_{s_1 \in \tilde{E}_1(s_2^T)} \Pi_2(s_1, s_2^T) \geq \min_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T) - \delta.$$

To obtain a contradiction, assume

$$\min_{s_1 \in \tilde{E}_1(s_2^T)} \Pi_2(s_1, s_2^T) < \min_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T) - \delta, \tag{18}$$

where the right hand side equals $\Pi_2(\tilde{s}_1, s_2^T)$ according to (14). Let $\tilde{s}_1^{min} \in \arg\min_{s_1 \in \tilde{E}_1(s_2^T)} \Pi_2(s_1, s_2^T)$. Efficiency of $\tilde{s}_1^{min}$ then requires

$$\Pi_1(\tilde{s}_1^{min}, s_2^T) > \Pi_1(\tilde{s}_1, s_2^T). \tag{19}$$

Let $s_1^{min'} \in S_1$ be obtained from $\tilde{s}_1^{min}$ by, again, replacing the announcement of any $d_j \in \{d_1, d_2, ..., d_r\}$ by $\sigma(d_j)$. This implies

$$\Pi_1(s_1^{min'}, s_2^T) = \Pi_1(\tilde{s}_1^{min}, s_2^T) - y\epsilon, \quad \Pi_2(s_1^{min'}, s_2^T) = \Pi_2(\tilde{s}_1^{min}, s_2^T) + y\delta, \tag{20}$$

where $y \in [0,1]$ is the probability of announcements from $\{d_1, d_2, ..., d_r\}$ under $\tilde{s}_1^{min}$. Conditions (14), (18), (19) and (20) together then imply

$$\Pi_1(s_1^{min'}, s_2^T) > \Pi_1(s_1^{min}, s_2^T), \quad \Pi_2(s_1^{min'}, s_2^T) < \Pi_2(s_1^{min}, s_2^T).$$

Now, if $s_1^{min'} \in E_1(s_2^T)$ this contradicts that s_1^{min} minimizes $\Pi_2(s_1, s_2^T)$ on $E_1(s_2^T)$. Otherwise, we obtain a contradiction exactly as in step 2. □

Lemmas 1 and 2 together imply that, starting from the direct mechanism, we can decrease the equitable payoff arbitrarily, with the help of the examined mechanism extension.

Increasing $\pi_2^{e_1}(s_2^T)$. The mechanism extension used to increase the equitable payoff works analogously. Let $s_1^{max} \in \arg\max_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T)$ and let $r = |s_1^{max}(\Theta_1)|$ be the cardinality of the range of s_1^{max}. Fix any bijection $\sigma : \{d_1, d_2, ..., d_r\} \to s_1^{max}(\Theta_1)$ and extend Φ to $\tilde{\Phi}$ by letting $\tilde{M}_1 = \Theta_1 \cup \{d_1, d_2, ..., d_r\}$ and $\tilde{M}_2 = \Theta_2$. The function $\tilde{g}$ again coincides with g on $\Theta_1 \times \Theta_2$. For any $m = (d_j, \theta_2)$, let $\tilde{g}(m)$ be given by

$$q_1^{\tilde{g}}(d_j, \theta_2) = q_1^g(\sigma(d_j), \theta_2), \quad q_2^{\tilde{g}}(d_j, \theta_2) = q_2^g(\sigma(d_j), \theta_2)$$

and

$$t_1^{\tilde{g}}(d_j, \theta_2) = t_1^g(\sigma(d_j), \theta_2) - \delta, \quad t_2^{\tilde{g}}(d_j, \theta_2) = t_2^g(\sigma(d_j), \theta_2) + \epsilon$$

with $0 < \epsilon \leq \delta$. Here, using messages from $\{d_1, d_2, ..., d_r\}$ redistributes from player 1 to player 2. We have $\tilde{S}_1 \supset S_1$ and $\tilde{S}_2 = S_2$, we let Π_i again denote player i's ex ante expected payoffs defined on the extension $\tilde{S}_1 \times \tilde{S}_2$, and we write $\tilde{E}_1(s_2^T)$ for the efficient strategies in $\tilde{\Phi}$, conditional on s_2^T. Observe that, if $s_1 \notin E_1(s_2^T)$ for some $s_1 \in S_1$, then $s_1 \notin \tilde{E}_1(s_2^T)$ holds as well, because enlarging the strategy set cannot make a previously inefficient strategy efficient.

We first prove that the minimization part in the definition of $\pi_2^{e_1}(s_2^T)$ remains unaffected by the mechanism extension.

Lemma 3. *For any arbitrary ϵ, δ with $0 < \epsilon \leq \delta$, it holds that*

$$\min_{s_1 \in \tilde{E}_1(s_2^T)} \Pi_2(s_1, s_2^T) = \min_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T).$$

Proof. Step 1. Let $s_1^{min} \in \arg\min_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T)$. We claim that $s_1^{min} \in \tilde{E}_1(s_2^T)$, which then implies

$$\min_{s_1 \in \tilde{E}_1(s_2^T)} \Pi_2(s_1, s_2^T) \leq \min_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T).$$

To obtain a contradiction, assume $s_1^{min} \notin \tilde{E}_1(s_2^T)$. This implies that there exists $\tilde{s}_1 \in \tilde{S}_1 \backslash S_1$ with

$$\Pi_1(\tilde{s}_1, s_2^T) \geq \Pi_1(s_1^{min}, s_2^T), \quad \Pi_2(\tilde{s}_1, s_2^T) \geq \Pi_2(s_1^{min}, s_2^T), \tag{21}$$

with at least one of the inequalities being strict. Let $s_1' \in S_1$ be the strategy obtained from $\tilde{s}_1$

by replacing the announcement of any $d_j \in \{d_1, d_2, ..., d_r\}$ by $\sigma(d_j)$. It follows that

$$\Pi_1(s_1', s_2^T) > \Pi_1(\tilde{s}_1, s_2^T), \quad \Pi_2(s_1', s_2^T) < \Pi_2(\tilde{s}_1, s_2^T). \tag{22}$$

From (21) and (22), together with $s_1^{min} \in E_1(s_2^T)$, it then follows that

$$\Pi_1(s_1', s_2^T) > \Pi_1(s_1^{min}, s_2^T), \quad \Pi_2(s_1', s_2^T) < \Pi_2(s_1^{min}, s_2^T).$$

This is a contradiction to s_1^{min} minimizing $\Pi_2(s_1, s_2^T)$ on $E_1(s_2^T)$, with the same argument as in the proof of Lemma 2.

Step 2. To establish the other inequality

$$\min_{s_1 \in \tilde{E}_1(s_2^T)} \Pi_2(s_1, s_2^T) \geq \min_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T),$$

assume to the contrary that

$$\min_{s_1 \in \tilde{E}_1(s_2^T)} \Pi_2(s_1, s_2^T) < \min_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T).$$

Let $\tilde{s}_1^{min} \in \arg\min_{s_1 \in \tilde{E}_1(s_2^T)} \Pi_2(s_1, s_2^T)$. Since $s_1 \notin E_1(s_2^T)$ implies $s_1 \notin \tilde{E}_1(s_2^T)$, we must have $\tilde{s}_1^{min} \in \tilde{S}_1 \backslash S_1$. Similarly to above, let $s_1' \in S_1$ be the strategy obtained from $\tilde{s}_1^{min}$ by replacing the announcement of any $d_j \in \{d_1, d_2, ..., d_r\}$ by $\sigma(d_j)$. It follows that

$$\Pi_1(s_1', s_2^T) > \Pi_1(\tilde{s}_1^{min}, s_2^T), \quad \Pi_2(s_1', s_2^T) < \Pi_2(\tilde{s}_1^{min}, s_2^T).$$

If $s_1' \in \tilde{E}_1(s_2^T)$, we have obtained a contradiction against $\tilde{s}_1^{min} \in \arg\min_{s_1 \in \tilde{E}_1(s_2^T)} \Pi_2(s_1, s_2^T)$. Otherwise, there exists $s_1'' \in \tilde{E}_1(s_2^T)$ that Pareto dominates s_1' but, due to $\tilde{s}_1^{min} \in \tilde{E}_1(s_2^T)$, still satisfies $\Pi_2(s_1'', s_2^T) < \Pi_2(\tilde{s}_1^{min}, s_2^T)$, again a contradiction. □

We now examine the maximization part in the definition of $\pi_2^{e_1}(s_2^T)$. We show that the maximum can be increased to any arbitrary value, by an appropriate choice of ϵ.

Lemma 4. *For any arbitrary ϵ, δ with $0 < \epsilon \leq \delta$, it holds that*

$$\max_{s_1 \in \tilde{E}_1(s_2^T)} \Pi_2(s_1, s_2^T) = \max_{s_1 \in E_1(s_2^T)} \Pi_2(s_1, s_2^T) + \epsilon.$$

Proof. As shown in the proof of Lemma 1, step 1, the statement follows when we have established

$$\max_{s_1 \in \tilde{S}_1} \Pi_2(s_1, s_2^T) = \max_{s_1 \in S_1} \Pi_2(s_1, s_2^T) + \epsilon,$$

where the right hand side equals $\Pi_2(s_1^{max}, s_2^T) + \epsilon$. Starting from s_1^{max}, construct the associated strategy $\tilde{s}_1 \in \tilde{S}_1$ by replacing every announcement of a type $\bar{\theta}_1 \in s_1^{max}(\Theta_1)$ by the associated new message $\sigma^{-1}(\bar{\theta}_1)$. Formally, $\tilde{s}_1(\theta_1) = \sigma^{-1}(s_1^{max}(\theta_1))$ for all $\theta_1 \in \Theta_1$. It follows that $\Pi_2(\tilde{s}_1, s_2^T) =$

$\Pi_2(s_1^{max}, s_2^T) + \epsilon$, which establishes the first inequality

$$\max_{s_1 \in \tilde{S}_1} \Pi_2(s_1, s_2^T) \geq \max_{s_1 \in S_1} \Pi_2(s_1, s_2^T) + \epsilon.$$

The opposite inequality holds as well, because for any $\tilde{s}_1 \in \tilde{S}_1 \backslash S_1$ we can construct the associated $s_1 \in S_1$ by replacing the announcement of any $d_j \in \{d_1, d_2, ..., d_r\}$ by $\sigma(d_j)$, to obtain

$$\Pi_2(\tilde{s}_1, s_2^T) - \Pi_2(s_1, s_2^T) = x\epsilon,$$

where $x \in]0, 1]$ is the probability of announcements from $\{d_1, d_2, ..., d_r\}$ under $\tilde{s}_1$. □

Lemmas 3 and 4 together imply that, starting from the direct mechanism, we can increase the equitable payoff arbitrarily, with the help of the examined mechanism extension.

Synthesis and generalization to an arbitrary number of players. The above construction can be done equivalently for player 2. For profiles m where both players use an additional message from the extended mechanism, the outcome $\tilde{g}(m)$ can be specified arbitrarily, because we only need to address unilateral deviations from s^T. Hence we can achieve $\kappa_{ji}(s^T) = 1/y_i$ for both players in $\tilde{\Phi}$, which implies that s^T is a BNFE in $\tilde{\Phi}$, which in turn implements the SCF f.

If the number of players exceeds 2, the construction above can be done for each pair of players separately. The actions $\{d_1^{ij}, \ldots, d_{r_{ij}}^{ij}\}$ of player i that are added to manipulate $\pi_j^{e_i}(s_{-i}^T)$ simply have to be chosen such that they do not affect the outcomes for all other players. Hence, for every pair i and j we can make sure that $\kappa_{ji}(s^T) = 1/y_{ij}$, so that all players become welfare maximizers and our arguments about truth-telling apply unaltered.

A.3 A Social Choice Function Which Cannot Be Implemented in BNFE

Let $I = \{1, 2\}$ and $\Theta_i = \{1, 2, 3\}$ for both $i = 1, 2$, where all types are equally likely. We consider a social choice function which only specifies transfers. The consumption levels q_i are set equal to zero to all players and all types. Specifically, consider the SCF f in the following table, which gives the transfers (t_1^f, t_2^f) for every realization of types. This SCF is not materially efficient, because it does not satisfy budget balance. It generates a strictly positive surplus whenever the players' types coincide.

θ_1 \ θ_2	1	2	3
1	$(-3, -3)$	$(0, 0)$	$(0, 0)$
2	$(0, 0)$	$(-2, -2)$	$(0, 0)$
3	$(0, 0)$	$(0, 0)$	$(-1, 0)$

According to the augmented revelation principle, if f is implementable in BNFE, then also in an augmented revelation mechanism with truth-telling being a BNFE. In any such mechanism, feasible strategies for player 1 include the strategy of always announcing type 2, denoted s_1^M, and the strategy of always announcing type 3, denoted s_1^H. Observe that a deviation from s_1^T to s_1^H makes both players better off, given s_2^T, and thus requires strict unkindness for not being

used. Strategy s_1^M, by contrast, leaves the deviator unaffected but makes the opponent worse off, and will thus be preferred to s_1^T by player 1 whenever there is strict unkindness. Formally, for truth-telling s_1^T being weakly preferred to s_1^H we would need $\kappa_{21}(s^T) < 0$ in the hypothetical equilibrium s^T, while for s_1^T being weakly preferred to s_1^M we must have $\kappa_{21}(s^T) \geq 0$. Hence no matter how unused actions are introduced to manipulate $\kappa_{21}(s^T)$, these deviations cannot simultaneously be made unattractive and the SCF cannot be implemented in BNFE.

A.4 Proof of Theorem 2

We first prove the "if" part: a utility-efficient mechanism exists when f exhibits bilateral externalities. To do so, we first establish that kindness values must be bounded in that case. We then show how these bounds can be reached. In the second step, we prove the "only if" part: without bilateral externalities, kindness terms can be increased arbitrarily in equilibrium. Throughout, we use notation from the proof of Theorem 1.

Step 1. Suppose f is materially efficient and exhibits bilateral externalities, and consider the direct mechanism Φ for f, with strategy sets S_i, for $i = 1, 2$. We partition S_i into three subsets $\overline{S}_i$, $\hat{S}_i$ and $\underline{S}_i$ as follows. Let $\overline{S}_i$ be the set of strategies s_i that satisfy $\Pi_j(s_i, s_j^T) > \Pi_j(s_i^T, s_j^T)$. Bilateral externalities imply that $\overline{S}_i$ is nonempty. It contains, for instance, the strategy of always announcing the type θ_i that maximizes $\mathbb{E}_{\theta_j}[v_j(q_j^f(\theta_j, \theta_i), \theta_j) + t_j^f(\theta_j, \theta_i)]$. Let $\underline{S}_i$ be the set of strategies s_i that satisfy $\Pi_j(s_i, s_j^T) < \Pi_j(s_i^T, s_j^T)$. Again, $\underline{S}_i$ is nonempty due to bilateral externalities. Finally, $\hat{S}_i$ is the set of strategies for which $\Pi_j(s_i, s_j^T) = \Pi_j(s_i^T, s_j^T)$, so that $s_i^T \in \hat{S}_i$ and $\hat{S}_i$ is also nonempty.

As argued in the proof of Theorem 1, s_i^T maximizes $\Pi_i(s_i, s_j^T) + \Pi_j(s_i, s_j^T)$ among all $s_i \in S_i$, by material efficiency of f. Hence we have $\Pi_i(s_i, s_j^T) < \Pi_i(s_i^T, s_j^T)$ for all $s_i \in \overline{S}_i$ and $\Pi_i(s_i, s_j^T) \leq \Pi_i(s_i^T, s_j^T)$ for all $s_i \in \hat{S}_i$. Furthermore, denoting by $\kappa_{ji}(s^T)$ the kindness terms associated to the truth-telling profile s^T in Φ, s_i^T is in fact a best response to s_j^T for player i when $\kappa_{ji}(s^T) = 1/y_i$.

Consider any $s_i \in \overline{S}_i$. Condition $\Pi_i(s_i^T, s_j^T) + y_i\kappa_{ji}\Pi_j(s_i^T, s_j^T) \geq \Pi_i(s_i, s_j^T) + y_i\kappa_{ji}\Pi_j(s_i, s_j^T)$, i.e., that a deviation from truth-telling to s_i is not attractive, can be rearranged to

$$\kappa_{ji} \leq \frac{1}{y_i} \frac{\Pi_i(s_i^T, s_j^T) - \Pi_i(s_i, s_j^T)}{\Pi_j(s_i, s_j^T) - \Pi_j(s_i^T, s_j^T)},$$

where the right hand side of the inequality is strictly positive by definition of $\overline{S}_i$. This defines an upper bound

$$\kappa_{ji}^{max} = \min_{s_i \in \overline{S}_i} \left(\frac{1}{y_i} \frac{\Pi_i(s_i^T, s_j^T) - \Pi_i(s_i, s_j^T)}{\Pi_j(s_i, s_j^T) - \Pi_j(s_i^T, s_j^T)} \right),$$

so that, in the direct mechanism, all deviations from truth-telling to strategies in $\overline{S}_i$ are unattractive if and only if $\kappa_{ji}(s^T) \leq \kappa_{ji}^{max}$. From the previous arguments we must have $1/y_i \leq \kappa_{ji}^{max}$. The analogous argument for $\underline{S}_i$ yields the lower bound

$$\kappa_{ji}^{min} = \max_{s_i \in \underline{S}_i} \left(\frac{1}{y_i} \frac{\Pi_i(s_i, s_j^T) - \Pi_i(s_i^T, s_j^T)}{\Pi_j(s_i^T, s_j^T) - \Pi_j(s_i, s_j^T)} \right),$$

so that all deviations from truth-telling to strategies in $\underline{S}_i$ are unattractive if and only if $\kappa_{ji}(s^T) \geq$

κ_{ji}^{min}. We must have $\kappa_{ji}^{min} \leq 1/y_i$, but κ_{ji}^{min} can be positive or negative. Deviations from truth-telling to strategies in $\hat{S}_i$ can never be attractive. Altogether, truth-telling s^T is a BNFE in the direct mechanism if and only if $\kappa_{ji}^{min} \leq \kappa_{ji}(s^T) \leq \kappa_{ji}^{max}$ for $i = 1, 2$ and $j \neq i$.

Since any augmented revelation mechanism $\tilde{\Phi}$ with strategy sets $\tilde{S}_i$, for $i = 1, 2$, satisfies $S_i \subseteq \tilde{S}_i$, the bounds κ_{ji}^{min} and κ_{ji}^{max} are also bounds on the kindness values of any truth-telling BNFE in any augmented revelation mechanism, because deviations to $\overline{S}_i$ or $\underline{S}_i$ are available in $\tilde{\Phi}$ as well. By the augmented revelation principle, this implies that κ_{ji}^{min} and κ_{ji}^{max} are bounds for every mechanism-equilibrium pair that implements f. Thus $\max\{\kappa_{21}^{min} \cdot \kappa_{12}^{min}, \kappa_{21}^{max} \cdot \kappa_{12}^{max}\}$ is an upper bound on the product of kindness terms for every mechanism-equilibrium pair that implements f.

Case 1a. Suppose $\kappa_{21}^{max} \cdot \kappa_{12}^{max} \geq \kappa_{21}^{min} \cdot \kappa_{12}^{min}$. We will construct an augmented revelation mechanism which truthfully implements f and in which the bounds κ_{ji}^{max} and κ_{ij}^{max} are reached for both players, so that the mechanism implements f utility-efficiently.

Using the construction given in the proof of Theorem 1, we can add messages to the direct mechanism Φ for each player i to adjust $\kappa_{ij}(s^T)$ to the desired level κ_{ij}^{max}. It remains to be shown that s^T is a BNFE in the resulting augmented revelation mechanism $\tilde{\Phi}$. It follows from the above derivation of the kindness bounds that deviations from truth-telling to strategies $s_i \in S_i \subseteq \tilde{S}_i$ are not attractive. Hence we only need to show that no player i wants to deviate to a strategy from $\tilde{S}_i \backslash S_i$.

If no messages have been added for player i in the construction of $\tilde{\Phi}$, that is, if $\kappa_{ij}(s^T) = \kappa_{ij}^{max}$ already in the direct mechanism, this holds trivially because $\tilde{S}_i \backslash S_i$ is empty.

If messages have been added for player i to decrease $\pi_j^{e_i}(s^T)$, that is, if $\kappa_{ij}(s^T) < \kappa_{ij}^{max}$ in the direct mechanism, then for any strategy $\tilde{s}_i \in \tilde{S}_i \backslash S_i$ there exists an associated strategy $s_i' \in S_i$ such that

$$\Pi_i(\tilde{s}_i, s_j^T) = \Pi_i(s_i', s_j^T) + x\epsilon \quad \text{and} \quad \Pi_j(\tilde{s}_i, s_j^T) = \Pi_j(s_i', s_j^T) - x\delta,$$

where $x \in]0, 1]$ is the probability of messages from $\{d_1, d_2, ..., d_r\}$ under $\tilde{s}_i$, $\delta > 0$ is determined by the requirement to achieve κ_{ij}^{max}, and ϵ is arbitrary with $0 < \epsilon \leq \delta$. Hence we have

$$\Pi_i(\tilde{s}_i, s_j^T) + y_i \kappa_{ji}^{max} \Pi_j(\tilde{s}_i, s_j^T) = \Pi_i(s_i', s_j^T) + y_i \kappa_{ji}^{max} \Pi_j(s_i', s_j^T) - x \left[y_i \kappa_{ji}^{max} \delta - \epsilon \right].$$

The last term in squared brackets is weakly positive since $y_i \kappa_{ji}^{max} \geq 1$ and $\delta \geq \epsilon$, so that all strategies from $\tilde{S}_i \backslash S_i$ are weakly less attractive than the associated strategies from S_i. Deviations from truth-telling to $\tilde{S}_i \backslash S_i$ are therefore also not attractive.

If messages have been added for player i to increase $\pi_j^{e_i}(s^T)$, that is, if $\kappa_{ij}(s^T) > \kappa_{ij}^{max}$ in the direct mechanism, then for any $\tilde{s}_i \in \tilde{S}_i \backslash S_i$ there exists an associated $s_i' \in S_i$ such that

$$\Pi_i(\tilde{s}_i, s_j^T) + y_i \kappa_{ji}^{max} \Pi_j(\tilde{s}_i, s_j^T) = \Pi_i(s_i', s_j^T) + y_i \kappa_{ji}^{max} \Pi_j(s_i', s_j^T) - x \left[\delta - y_i \kappa_{ji}^{max} \epsilon \right],$$

where $x \in]0, 1]$, $\epsilon > 0$ is determined by the requirement to achieve κ_{ij}^{max}, and δ is arbitrary as long as $\delta \geq \epsilon$. Hence for any $\tilde{s}_i \in \tilde{S}_i \backslash S_i$ we can choose δ large enough to make $\tilde{s}_i$ less attractive than the associated s_i', so that, by finiteness of $\tilde{S}_i$, for large enough values of δ no deviation from truth-telling to $\tilde{S}_i \backslash S_i$ is attractive.

Therefore, s^T is a BNFE in the augmented revelation mechanism $\tilde{\Phi}$, which reaches the upper

bound $\kappa_{21}^{max} \cdot \kappa_{12}^{max}$ for psychological payoffs and thus implements f utility-efficiently.

Case 1b. Suppose $\kappa_{21}^{max} \cdot \kappa_{12}^{max} < \kappa_{21}^{min} \cdot \kappa_{12}^{min}$, which requires $\kappa_{21}^{min} < 0$ and $\kappa_{12}^{min} < 0$. Suppose further that, in the direct mechanism, $\kappa_{ji}(s^T) \geq \kappa_{ji}^{min}$ holds for both players. Then we can use the standard construction of $\tilde{\Phi}$ to increase $\pi_j^{e_i}(s^T)$ for both players and achieve the (negative) lower bounds on kindness. Deviations to strategies $s_i \in S_i \subseteq \tilde{S}_i$ are again not attractive for any $i = 1, 2$, by definition of κ_{ji}^{min}. For any $\tilde{s}_i \in \tilde{S}_i \backslash S_i$, there exists an associated strategy $s_i' \in S_i$ such that

$$\Pi_i(\tilde{s}_i, s_j^T) + y_i \kappa_{ji}^{min} \Pi_j(\tilde{s}_i, s_j^T) = \Pi_i(s_i', s_j^T) + y_i \kappa_{ji}^{min} \Pi_j(s_i', s_j^T) - x \left[\delta - y_i \kappa_{ji}^{min} \epsilon\right],$$

where $x \in]0, 1]$ and $0 < \epsilon \leq \delta$. The last term in squared brackets is positive, because $\kappa_{ji}^{min} < 0$, so that strategies from $\tilde{S}_i \backslash S_i$ are less attractive than those from S_i, which implies that s^T is a BNFE which reaches the upper bound $\kappa_{21}^{min} \cdot \kappa_{12}^{min}$ for psychological payoffs and thus implements f utility-efficiently.

Case 1c. The remaining case is characterized by $\kappa_{21}^{max} \cdot \kappa_{12}^{max} < \kappa_{21}^{min} \cdot \kappa_{12}^{min}$ and, for at least one player i, $\kappa_{ij}(s^T) < \kappa_{ij}^{min}$ holds in the direct mechanism. We claim that, in this case, whenever s^T is a BNFE in any augmented revelation mechanism for f, we must have $\kappa_{ji}(s^T) > 0$ in this equilibrium. The claim implies that $\kappa_{21}^{max} \cdot \kappa_{12}^{max}$, and not $\kappa_{21}^{min} \cdot \kappa_{12}^{min}$, is in fact an upper bound on psychological payoffs in this case, which can then be reached as shown for case 1a above.

To establish the claim, observe again that $\kappa_{21}^{max} \cdot \kappa_{12}^{max} < \kappa_{21}^{min} \cdot \kappa_{12}^{min}$ requires $\kappa_{21}^{min} < 0$ and $\kappa_{12}^{min} < 0$. If $\kappa_{ji}^{min} < 0$, then we must have $\Pi_i(s_i, s_j^T) < \Pi_i(s_i^T, s_j^T)$ for all $s_i \in \underline{S}_i$, so that all strategies from $\underline{S}_i$ yield Pareto inefficient outcomes conditional on s_j^T. This implies that, in the direct mechanism Φ,

$$\Pi_j(s_i^T, s_j^T) = \min_{s_i \in E_i(s_j^T)} \Pi_j(s_i, s_j^T),$$

because $s_i^T \in E_i(s_j^T)$ clearly holds due to material efficiency of f. For s^T to be a BNFE in an augmented revelation mechanism $\tilde{\Phi}$, we must achieve a kindness level of at least κ_{ij}^{min}, which requires the equitable payoff $\pi_j^{e_i}(s^T)$ to be strictly smaller in $\tilde{\Phi}$ than in the direct mechanism Φ. We cannot have $\max_{s_i \in \tilde{E}_i(s_j^T)} \Pi_j(s_i, s_j^T) < \max_{s_i \in E_i(s_j^T)} \Pi_j(s_i, s_j^T)$, because adding messages cannot decrease the maximal payoff for player j (see the proof of Lemma 1). Hence we must have

$$\min_{s_i \in \tilde{E}_i(s_j^T)} \Pi_j(s_i, s_j^T) < \min_{s_i \in E_i(s_j^T)} \Pi_j(s_i, s_j^T) = \Pi_j(s_i^T, s_j^T),$$

i.e., there must exist a strategy $\tilde{s}_i \in \tilde{E}_i(s_j^T)$ such that $\Pi_j(\tilde{s}_i, s_j^T) < \Pi_j(s_i^T, s_j^T)$, and, by bilateral Pareto efficiency, $\Pi_i(\tilde{s}_i, s_j^T) > \Pi_i(s_i^T, s_j^T)$. But a deviation from s_i^T to this strategy $\tilde{s}_i$ is clearly profitable whenever $\kappa_{ji}(s^T) \leq 0$, which establishes the claim and completes the proof of the "if" statement.

Step 2. Suppose f is materially efficient and does not exhibit bilateral externalities, so there exists at least one player, say player 1, such that $\mathbb{E}_{\theta_1}[v_1(q_1^f(\theta_1, \theta_2), \theta_1) + t_1^f(\theta_1, \theta_2)]$ is independent of θ_2. It follows that $\Pi_1(s_1^T, s_2)$ is independent of $s_2 \in S_2$ in the direct mechanism Φ (see Proposition 3 for a more general statement of this fact). Material efficiency of f then implies that $\Pi_2(s_1^T, s_2) \leq \Pi_2(s_1^T, s_2^T)$ for all $s_2 \in S_2$.

We now construct an augmented revelation mechanism $\tilde{\Phi}$, again as described in the proof

of Theorem 1, in which $\kappa_{21}(s^T) = 1/y_1$ is achieved, by adding messages for player 2. Efficiency of f then implies that s_1^T is a best response to s_2^T, irrespective of how unused messages are designed for player 1. Specifically, we can now add unused messages to decrease $\pi_2^{e_1}(s^T)$ and hence increase $\kappa_{12}(s^T)$ arbitrarily, letting $\kappa_{21}(s^T) \cdot \kappa_{12}(s^T) = 1/y_1 \cdot \kappa_{12}(s^T)$ grow without bounds. We only need to show that this is possible in a way such that s_2^T remains a best response for player 2.

Since any $s_2 \in S_2$ satisfies $\Pi_1(s_1^T, s_2) = \Pi_1(s_1^T, s_2^T)$ and $\Pi_2(s_1^T, s_2) \leq \Pi_2(s_1^T, s_2^T)$, deviations from s_2^T to any $s_2 \in S_2$ are never profitable for player 2, irrespective of the size of $\kappa_{12}(s^T) > 0$. Now consider strategies $\tilde{s}_2 \in \tilde{S}_2 \backslash S_2$, i.e., strategies that use messages which have been introduced to achieve $\kappa_{21}(s^T) = 1/y_1$. The case where $\tilde{S}_2 \backslash S_2 = \varnothing$ is trivial. If the equitable payoff $\pi_1^{e_2}(s_1^T)$ has been decreased, strategies $\tilde{s}_2 \in \tilde{S}_2 \backslash S_2$ are unprofitable whenever $\kappa_{12}(s^T) \geq 1/y_2$, with the same argument as for case 1a above. When the equitable payoff $\pi_1^{e_2}(s_1^T)$ has been increased, for every value of $\kappa_{12}(s^T)$ we can choose δ for player 2 large enough to again make all deviations to $\tilde{s}_2 \in \tilde{S}_2 \backslash S_2$ unprofitable, as shown for case 1a above. Hence, letting δ grow to infinity together with $\kappa_{12}(s^T)$, we can ensure that s^T remains a BNFE.

A.5 Proof of Proposition 3

Step 1. We first show that (6) implies (5). If (6) holds then it must be true that

$$\Pi_i(s_j^{\theta_j'}, s_{-j}^T) = \Pi_i(s_j^{\theta_j''}, s_{-j}^T),$$

where θ_j' and θ_j'' are arbitrary types from Θ_j, $s_j^{\theta_j'}$ is the strategy where player j always announces θ_j', whatever his true type, and $s_j^{\theta_j''}$ is the strategy where j always announces θ_j''. This condition is equivalent to

$$\sum_{\theta_{-j} \in \Theta_{-j}} p(\theta_{-j}) \left[v_i(q_i^f(\theta_j', \theta_{-j}), \theta_i) + t_i^f(\theta_j', \theta_{-j}) \right] = \sum_{\theta_{-j} \in \Theta_{-j}} p(\theta_{-j}) \left[v_i(q_i^f(\theta_j'', \theta_{-j}), \theta_i) + t_i^f(\theta_j'', \theta_{-j}) \right],$$

so that condition (5) holds.

Step 2. We now show that (5) implies (6). For an arbitrary strategy s_j and an arbitrary type $\theta_j \in \Theta_j$, define

$$\Lambda(\theta_j | s_j) = \{\theta_j' \in \Theta_j \mid s_j(\theta_j') = \theta_j\},$$

and observe that

$$\begin{aligned} \Pi_i(s_j, s_{-j}^T) &= \textstyle\sum_{\theta_j \in \Theta_j} \left[\left(\sum_{\theta_j' \in \Lambda(\theta_j | s_j)} p(\theta_j') \right) \sum_{\theta_{-j} \in \Theta_{-j}} p(\theta_{-j}) [v_i(q_i^f(\theta_j, \theta_{-j}), \theta_i) + t_i^f(\theta_j, \theta_{-j})] \right] \\ &= \textstyle\sum_{\theta_j \in \Theta_j} \left[\left(\sum_{\theta_j' \in \Lambda(\theta_j | s_j)} p(\theta_j') \right) \mathbb{E}_{\theta_{-j}} [v_i(q_i^f(\theta_j, \theta_{-j}), \theta_i) + t_i^f(\theta_j, \theta_{-j})] \right]. \end{aligned}$$

Now, if (5) holds then there is a number ρ so that

$$\mathbb{E}_{\theta_{-j}} [v_i(q_i^f(\theta_j, \theta_{-j}), \theta_i) + t_i^f(\theta_j, \theta_{-j})] = \rho$$

for all $\theta_j \in \Theta_j$. Hence

$$\Pi_i(s_j, s_{-j}^T) = \rho \sum_{\theta_j \in \Theta_j} \left(\sum_{\theta_j' \in \Lambda(\theta_j|s_j)} p(\theta_j') \right) = \rho \, .$$

Since our choice of s_j was arbitrary, this shows that, for all s_j' and s_j'',

$$\Pi_i(s_j', s_{-j}^T) = \Pi_i(s_j'', s_{-j}^T) = \rho$$

and hence (6) holds.

A.6 Proof of Proposition 4

Let f be a social choice function that is implementable in BNE and that has the insurance property. We show that the direct mechanism truthfully implements f in PRE. Given the direct mechanism and given that f has the insurance property, Proposition 3 implies that for all i and $j \neq i$, there exists a number $\rho_i(s_{-j}^T)$ so that

$$\Pi_i(s_j, s_{-j}^T) = \rho_i(s_{-j}^T)$$

for any strategy $s_j \in S_j$. Now consider the truth-telling BNE s^T and suppose all first- and second-order beliefs to be correct. Then, trivially, $\pi_i^{e_j}(s_{-j}^T) = \rho_i(s_{-j}^T)$ and $\kappa_{ji}(s_j, s_{-j}^T) = 0$ for all s_j, and thus $\kappa_{ji}(s^T) = 0$. Consequently, truth-telling is a best response of player i, for all parameter values $(y_{ij})_{j \neq i}$, if s_i^T is a maximizer of $\Pi_i(s_i, s_{-i}^T)$. This holds because s^T is a BNE, so that s^T is also a PRE.

A.7 Proof of Theorem 3

Step 1. For any arbitrary SCF $f = (q_1^f, \ldots, q_n^f, t_1^f, \ldots, t_n^f)$, we define the following expressions:

$$T_i^f(\theta_i) = \mathbb{E}_{\theta_{-i}}[t_i^f(\theta_i, \theta_{-i})]$$

are the expected transfers to i conditional on type θ_i, and

$$V_i^f(\theta_i) = \mathbb{E}_{\theta_{-i}}[v_i(q_i^f(\theta_i, \theta_{-i}), \theta_i)]$$

are, analogously, i's conditional expected payoffs net of transfers.

Step 2. Now, starting from $f = (q_1^f, \ldots, q_n^f, t_1^f, \ldots, t_n^f)$ as given in the theorem, we construct a payment scheme $(t_1^{\bar{f}}, \ldots, t_n^{\bar{f}})$ as follows. For every $i \in I$, $\theta_i \in \Theta_i$, and $\theta_{-i} \in \Theta_{-i}$, we let

$$t_i^{\bar{f}}(\theta_i, \theta_{-i}) = V_i^f(\theta_i) + T_i^f(\theta_i) - v_i(q_i^f(\theta_i, \theta_{-i}), \theta_i).$$

Now consider $\bar{f} = (q_1^f, \ldots, q_n^f, t_1^{\bar{f}}, \ldots, t_n^{\bar{f}})$, which has the same decision rule as f but the new payment rule. We claim that $\bar{f}$ satisfies, for all $i \in I$ and $\theta_i \in \Theta_i$,

$$T_i^{\bar{f}}(\theta_i) = T_i^f(\theta_i), \tag{23}$$

that is, the expected payment to every type of every individual is the same under f and $\bar{f}$. In fact, it holds that

$$\begin{aligned} T_i^{\bar{f}}(\theta_i) &= \mathbb{E}_{\theta_{-i}}[t_i^{\bar{f}}(\theta_i, \theta_{-i})] \\ &= \mathbb{E}_{\theta_{-i}}[V_i^f(\theta_i) + T_i^f(\theta_i) - v_i(q_i^f(\theta_i, \theta_{-i}), \theta_i)] \\ &= V_i^f(\theta_i) + T_i^f(\theta_i) - \mathbb{E}_{\theta_{-i}}[v_i(q_i^f(\theta_i, \theta_{-i}), \theta_i)] \\ &= T_i^f(\theta_i). \end{aligned}$$

Step 3. We now verify that $\bar{f} = (q_1^f, \ldots, q_n^f, t_1^{\bar{f}}, \ldots, t_n^{\bar{f}})$ satisfies properties (a) - (e). Property (a) is satisfied by construction. Property (b) follows from (23) above, after an application of the law of iterated expectations:

$$\begin{aligned} \mathbb{E}_\theta\left[\textstyle\sum_{i=1}^n t_i^{\bar{f}}(\theta)\right] &= \textstyle\sum_{i=1}^n \sum_{\theta_i \in \Theta_i} p(\theta_i) T_i^{\bar{f}}(\theta_i) \\ &= \textstyle\sum_{i=1}^n \sum_{\theta_i \in \Theta_i} p(\theta_i) T_i^{f}(\theta_i) \\ &= \mathbb{E}_\theta\left[\textstyle\sum_{i=1}^n t_i^{f}(\theta)\right]. \end{aligned}$$

Properties (a) and (23) together also immediately imply property (c). We next turn to property (d). The revelation principle for BNE implies that $\bar{f}$ is implementable in BNE if and only if the following incentive compatibility constraints are satisfied:

$$\mathbb{E}_{\theta_{-i}}[v_i(q_i^{\bar{f}}(\theta_i, \theta_{-i}), \theta_i)] + T_i^{\bar{f}}(\theta_i) \geq \mathbb{E}_{\theta_{-i}}[v_i(q_i^{\bar{f}}(\theta_i', \theta_{-i}), \theta_i)] + T_i^{\bar{f}}(\theta_i')$$

for all $i \in I$ and $\theta_i, \theta_i' \in \Theta_i$. Because of property (a) and (23), this can also be written as

$$\mathbb{E}_{\theta_{-i}}[v_i(q_i^{f}(\theta_i, \theta_{-i}), \theta_i)] + T_i^{f}(\theta_i) \geq \mathbb{E}_{\theta_{-i}}[v_i(q_i^{f}(\theta_i', \theta_{-i}), \theta_i)] + T_i^{f}(\theta_i'),$$

which is satisfied because f is implementable in BNE. We complete the proof by establishing the insurance property (e). From the definition of $t_i^{\bar{f}}(\theta_i, \theta_{-i})$ and property (a) it follows that

$$v_i(q_i^{\bar{f}}(\theta_i, \theta_{-i}), \theta_i) + t_i^{\bar{f}}(\theta_i, \theta_{-i}) = V_i^f(\theta_i) + T_i^f(\theta_i)$$

for all θ_i and θ_{-i}. Hence, for any $j \neq i$,

$$\mathbb{E}_{\theta_{-j}}[v_i(q_i^{\bar{f}}(\theta_j, \theta_{-j}), \theta_i) + t_i^{\bar{f}}(\theta_j, \theta_{-j})] = \mathbb{E}_{\theta_{-j}}[V_i^f(\theta_i) + T_i^f(\theta_i)] = \mathbb{E}_{\theta_i}[V_i^f(\theta_i) + T_i^f(\theta_i)]$$

is independent of θ_j, which is the insurance property.

A.8 Proof of Theorem 4

Step 1. We first show that for all players $i, j \in I$, $j \neq i$,

$$\Pi_j(s_i, s_{-i}^T) = \mathbb{E}_\theta\left[v_j(q_j^f(s_i(\theta_i), \theta_{-i}), \theta_j) - \frac{1}{n-1}\sum_{h \neq i} v_h(q_h^f(s_i(\theta_i), \theta_{-i}), \theta_h)\right] + \xi_{ji}, \tag{24}$$

where ξ_{ji} is independent of s_i. To see that this is true, note that, given the definition of $(t_1^f, ..., t_n^f)$ in the expected externality mechanism, we have

$$\begin{aligned}\Pi_j(s_i, s_{-i}^T) &= \mathbb{E}_\theta\left[v_j(q_j^f(s_i(\theta_i), \theta_{-i}), \theta_j)\right] + \mathbb{E}_\theta\left[t_j^f(s_i(\theta_i), \theta_{-i})\right] \\ &= \mathbb{E}_\theta\left[v_j(q_j^f(s_i(\theta_i), \theta_{-i}), \theta_j)\right] \\ &\quad + \mathbb{E}_\theta\left[\mathbb{E}_{\theta_{-j}}\left[\sum_{h\neq j} v_h(q_h^f(\theta_j, \theta_{-j}), \theta_h)\right]\right] + \mathbb{E}_\theta\left[h_j(s_i(\theta_i), \theta_{-ij})\right],\end{aligned}$$

where θ_{-ij} denotes the type profile of all players except i and j. The first term in this expression corresponds to the first term in (24). The second term is independent of s_i and can thus be subsumed into ξ_{ji}. Now consider the third term. Again using the definition of the expected externality mechanism we obtain

$$\begin{aligned}&h_j(s_i(\theta_i), \theta_{-ij}) = \\ &\quad -\frac{1}{n-1}\left\{\mathbb{E}_{\theta_{-i}}\left[\sum_{h\neq i} v_h(q_h^f(s_i(\theta_i), \theta_{-i}), \theta_h)\right] + \sum_{l\neq j,i}\mathbb{E}_{\theta_{-l}}\left[\sum_{h\neq l} v_h(q_h^f(\theta_l, \theta_{-l}), \theta_h)\right]\right\}.\end{aligned}$$

The second term is again independent of s_i and can be subsumed into ξ_{ji}. The first term, in turn, becomes the second term in (24) after taking the expectation with respect to θ.

Step 2. First, assume that there are only two players ($n = 2$). The term in squared brackets in (24) then cancels out, which implies that

$$\Pi_j(s_i', s_{-i}^T) = \Pi_j(s_i'', s_{-i}^T)$$

for any pair of strategies s_i' and s_i'' of player i. Proposition 3 then implies that the insurance property is satisfied. Now, suppose that symmetry holds. Under this assumption it is also true that the term in squared brackets in (24) vanishes. Again, this implies that the insurance property holds.

A.9 An Asymmetric Expected Externality Mechanism

The following example illustrates how lack of symmetry leads to a violation of the insurance property and to non-robustness of the expected externality mechanism.

Example 3. *Consider the problem of sharing one unit of a private good among three players $I = \{1, 2, 3\}$. Each player's type is from the set $\Theta_i = \{0, 1\}$. Both types are equally likely. Let $Q = \{(q_1, q_2, q_3) \in [0, 1]^3 | q_1 + q_2 + q_3 = 1\}$, so that q_i denotes the share of the private good that is allocated to player i. Preferences are given by $v_i(q_i, \theta_i) = \theta_i q_i$. We consider the expected externality mechanism for the efficient decision rule (q_1^f, q_2^f, q_3^f) detailed in Table 2, where each row specifies the outcome for one possible type profile $\theta = (\theta_1, \theta_2, \theta_3)$. The transfers (t_1^f, t_2^f, t_3^f) in Table 2 are those of the expected externality mechanism.*

The environment of Example 2 is symmetric, but the decision rule is not: it allocates the good entirely to player 1 whenever $\theta_1 = 1$, or when $\theta_i = 0$ for all $i \in I$. Otherwise, the good is symmetrically allocated between players 2 and 3. While not being symmetric, it is still efficient,

θ_1	θ_2	θ_3	q_1^f	q_2^f	q_3^f	t_1^f	t_2^f	t_3^f
0	0	0	1	0	0	0	0	0
0	0	1	0	0	1	1/16	1/16	-1/8
0	1	0	0	1	0	1/16	-1/8	1/16
0	1	1	0	1/2	1/2	1/8	-1/16	-1/16
1	0	0	1	0	0	-3/4	3/8	3/8
1	0	1	1	0	0	-11/16	7/16	1/4
1	1	0	1	0	0	-11/16	1/4	7/16
1	1	1	1	0	0	-5/8	5/16	5/16

Table 2: An asymmetric expected externality mechanism

because it allocates a positive share of the private good only to those players with a maximal valuation.

Observation 7. *The SCF f in Table 2 violates the insurance property.*

Proof. We can derive the players' payoffs in the direct mechanism for f, both for the truth-telling profile s^T and for the unilateral deviations s_i^H, s_i^L and s_i^{-T}. The derivations are tedious but straightforward, and the results are given in Table 3. Proposition 3 now implies that f violates the insurance property, because player 2, for instance, can affect his opponents' payoffs by unilateral deviations.

s_1	s_2	s_3	Π_1	Π_2	Π_3
s_1^T	s_2^T	s_3^T	3/16	11/32	11/32
s_1^L	s_2^T	s_3^T	3/16	11/32	11/32
s_1^H	s_2^T	s_3^T	-3/16	11/32	11/32
s_1^{-T}	s_2^T	s_3^T	-3/16	11/32	11/32
s_1^T	s_2^L	s_3^T	5/32	7/32	3/8
s_1^T	s_2^H	s_3^T	7/32	9/32	5/16
s_1^T	s_2^{-T}	s_3^T	3/16	5/32	11/32
s_1^T	s_2^T	s_3^L	5/32	3/8	7/32
s_1^T	s_2^T	s_3^H	7/32	5/16	9/32
s_1^T	s_2^T	s_3^{-T}	3/16	11/32	5/32

Table 3: Expected payoffs

□

Consider a possible behavioral implication of Observation 7. Based on the payoffs in Table 3, we obtain the efficiency sets

$$E_{12}(s_2^T, s_3^T) = \{s_1^T, s_1^L\}, E_{13}(s_2^T, s_3^T) = \{s_1^T, s_1^L\}, E_{21}(s_1^T, s_3^T) = \{s_2^T, s_2^H\},$$
$$E_{23}(s_1^T, s_3^T) = \{s_2^T, s_2^L\}, E_{31}(s_1^T, s_2^T) = \{s_3^T, s_3^H\}, E_{32}(s_1^T, s_2^T) = \{s_3^T, s_3^L\} ,$$

and the equitable payoffs

$$\pi_2^{e_1}(s_2^T, s_3^T) = 11/32, \pi_3^{e_1}(s_2^T, s_3^T) = 11/32, \pi_1^{e_2}(s_1^T, s_3^T) = 13/64,$$
$$\pi_3^{e_2}(s_1^T, s_3^T) = 23/64, \pi_1^{e_3}(s_1^T, s_2^T) = 13/64, \pi_2^{e_3}(s_1^T, s_2^T) = 23/64.$$

Based on these results, we obtain the kindness values

$$\kappa_{12}(s^T) = \kappa_{13}(s^T) = 0, \;\; \kappa_{21}(s^T) = \kappa_{23}(s^T) = \kappa_{31}(s^T) = \kappa_{32}(s^T) = -1/64.$$

Player 2, for instance, feels treated neutrally by player 1 but unkindly by player 3. A unilateral deviation from s_2^T to s_2^H would reduce player 3's payoff. Hence player 2 prefers to deviate to s_2^H whenever the reciprocity weight y_{23} is sufficiently large.

B Appendix

B.1 Interim Fairness Equilibrium

In this appendix, we adopt an interim perspective on fairness equilibria in games of incomplete information. Given some mechanism Φ, we assume that type θ_i of player i chooses an action m_i after having observed the own type, while still perceiving the other individuals' types as random quantities. We define an *ex interim fairness equilibrium* (IFE) and provide a condition on ex ante and ex interim fairness norms – formalized via ex ante and ex interim notions of equitable payoffs – so that, for our independent private values model, the sets of IFE and of BNFE coincide.

As a first step we develop an ex interim notion of kindness. In particular, consider the expected material payoff of player j from the perspective of type θ_i of player i. If this type of player i has beliefs s_i^b and chooses action m_i, this yields an expected payoff for player j which we denote with some abuse of notation by $\Pi_j(m_i, s_i^b)$. The assumption of independent private values implies that the type θ_i itself is irrelevant for this expression. We now define i's kindness towards j from an ex interim perspective as

$$\kappa_{ij}^{int}(m_i, s_i^b|\theta_i) = \Pi_j(m_i, s_i^b) - \pi_j^{int_i}(s_i^b|\theta_i),$$

where $\pi_j^{int_i}(\cdot|\theta_i)$ is the equitable payoff of player j from the perspective of type θ_i of player i.

Analogously player i has beliefs about how kind type θ_j of player j wants to be to himself, $\kappa_{ji}^{int}(s_{ij}^b(\theta_j), s_{ij}^{bb}|\theta_j)$. Due to the independent values assumption, these beliefs are the same for every type of player i. The expected value of these terms, from player i's perspective, is henceforth denoted by

$$\bar{\kappa}_{ji}(s_{ij}^b, s_{ij}^{bb}) = \sum_{\theta_j \in \Theta_j} p(\theta_j) \kappa_{ji}^{int}(s_{ij}^b(\theta_j), s_{ij}^{bb}|\theta_j) \; .$$

If type θ_i of player i chooses action m_i, his ex interim expected utility equals

$$U_i^{int}(m_i, s_i^b, s_i^{bb}|\theta_i) = \Pi_i(m_i, s_i^b|\theta_i) + \sum_{j \neq i} y_{ij} \, \kappa_{ij}^{int}(m_i, s_i^b|\theta_i) \, \bar{\kappa}_{ji}(s_{ij}^b, s_{ij}^{bb}) \; ,$$

where

$$\Pi_i(m_i, s_i^b|\theta_i) = \sum_{\theta_{-i}\in\Theta_{-i}} p(\theta_{-i})\,\pi_i(g(m_i, s_i^b(\theta_{-i})), \theta_i)\ .$$

Definition 8. *An IFE is a strategy profile s^* such that, for all players $i \in I$,*
(1) $s_i^*(\theta_i) \in \arg\max_{m_i\in M_i} U_i^{int}(m_i, s_i^b, s_i^{bb}|\theta_i)$ for all $\theta_i \in \Theta_i$,
(2) $s_i^b = s_{-i}^*$, and
(3) $s_i^{bb} = (s_{-j}^*)_{j\neq i}$.

We now introduce an assumption which will imply that BNFE and IFE are identical.

Assumption 1. *Suppose that, for all $i, j \in I$, $j \neq i$, and all $s_i^b \in S_{-i}$,*

$$\sum_{\theta_i\in\Theta_i} p(\theta_i)\,\pi_j^{int_i}(s_i^b|\theta_i) = \pi_j^{e_i}(s_i^b)\ ,$$

where $\pi_j^{e_i}(s_i^b)$ is the equitable payoff defined formally in (1).

The assumption requires that the expected value of i's interim views on what j deserves is equal to player i's ex ante view on what j deserves. This holds, in particular, if there is no updating of equitable payoffs, so that j's equitable payoff does not depend on player i's type.

Proposition 6. *Suppose Assumption 1 holds. Then, s^* is an IFE if and only if it is a BNFE.*

Proof. We show that, for any $i \in I$, $s_i = s_i^*$ is a maximizer of

$$U_i(s_i, s_i^b, s_i^{bb}) = \Pi_i(s_i, s_i^b) + \sum_{j\neq i} y_{ij}\ \kappa_{ij}(s_i, s_i^b)\ \kappa_{ji}(s_{ij}^b, s_{ij}^{bb})$$

if and only if, for every $\theta_i \in \Theta_i$, $m_i = s_i^*(\theta_i)$ maximizes

$$U_i^{int}(m_i, s_i^b, s_i^{bb}|\theta_i) = \Pi_i(m_i, s_i^b|\theta_i) + \sum_{j\neq i} y_{ij}\ \kappa_{ij}^{int}(m_i, s_i^b|\theta_i)\ \bar\kappa_{ji}(s_{ij}^b, s_{ij}^{bb})\ .$$

First, one can verify that, under Assumption 1, for every i and j,

$$\kappa_{ij}(s_i, s_i^b) = \sum_{\theta_i\in\Theta_i} p(\theta_i)\,\kappa_{ij}^{int}(s_i(\theta_i), s_i^b|\theta_i)\ .$$

Next, one can verify that

$$\kappa_{ji}(s_{ij}^b, s_{ij}^{bb}) = \bar\kappa_{ji}(s_{ij}^b, s_{ij}^{bb})$$

for every i and j. Finally, one can verify that, for any strategy $s_i \in S_i$, we have that

$$U_i(s_i, s_i^b, s_i^{bb}) = \sum_{\theta_i\in\Theta_i} p(\theta_i)\,U_i^{int}(s_i(\theta_i), s_i^b, s_i^{bb}|\theta_i)\ .$$

By standard arguments, since all types of player i occur with positive probability, a strategy s_i maximizes $U_i(s_i, s_i^b, s_i^{bb})$ if and only if $s_i(\theta_i)$ maximizes $U_i^{int}(s_i(\theta_i), s_i^b, s_i^{bb}|\theta_i)$, for all θ_i. □

B.2 Psychologically Robust Equilibrium

The solution concept of PRE is interesting for several reasons. First, implementability of an SCF in PRE makes it possible for the mechanism designer to remain ignorant about the intensity of kindness sensations. Furthermore, PRE is a refinement of BNE, since we require s^* to be a BNFE even when $y_{ij} = 0$ for all $i, j \in I$, in which case BNFE coincides with BNE. Hence we can start from established results on Bayesian incentive compatibility and investigate their robustness in a standard refinement sense. Finally, we will see that the revelation principle comes back for implementation in PRE.

We first present a characterization of PRE that generalizes an earlier result from Rabin (1993). As a preliminary step, the following lemma states that kindness between any two players cannot be positive in a conventional BNE.[25]

Lemma 5. *Let s^* be a BNE. Then it holds that $\kappa_{ij}(s^*) \leq 0$ for all $i, j \in I$, $j \neq i$.*

Proof. Consider any $i \in I$ and $j \neq i$. We claim that

$$\Pi_j(s^*) \leq \min_{s_i \in E_{ij}(s^*_{-i})} \Pi_j(s_i, s^*_{-i}),$$

which implies $\Pi_j(s^*) \leq \pi_j^{e_i}(s^*_{-i})$ and thus $\kappa_{ij}(s^*) \leq 0$. By definition of BNE, s_i^* maximizes $\Pi_i(s_i, s^*_{-i})$. Specifically, $\Pi_i(\tilde{s}_i, s^*_{-i}) \leq \Pi_i(s_i^*, s^*_{-i})$ for all $\tilde{s}_i \in E_{ij}(s^*_{-i})$. Bilateral efficiency then implies $\Pi_j(s_i^*, s^*_{-i}) \leq \Pi_j(\tilde{s}_i, s^*_{-i})$ for all $\tilde{s}_i \in E_{ij}(s^*_{-i})$, which proves the claim. □

In any BNE and hence in any PRE, every player is maximizing the own material payoff. This behavior will not be considered strictly kind by any opponent, as, with a conditional and bilateral concept of efficiency, positive kindness requires giving up payoffs for someone else's benefit. This observation provides the basis for the following characterization.

Proposition 7. *A BNE s^* is a PRE if and only if, for all $i, j \in I$, $j \neq i$,*

$$s_i^* \in \arg\max_{s_i \in S_i} \ \kappa_{ji}(s^*)\Pi_j(s_i, s^*_{-i}). \tag{25}$$

Proof. Step 1. Suppose that s^* is a BNE and suppose that condition (25) holds. We seek to show that this implies that s^* is a PRE, i.e. that for every player i, the strategy s_i^* is maximizing

$$\Pi_i(s_i, s^*_{-i}) + \sum_{j \neq i} y_{ij}\kappa_{ji}(s^*)\Pi_j(s_i, s^*_{-i}), \tag{26}$$

independently of the size of the parameters $(y_{ij})_{j \neq i}$. Since s^* is a BNE, s_i^* is a maximizer of $\Pi_i(s_i, s^*_{-i})$. Condition (25) implies that s_i^* is also a maximizer of every summand in the second term of (26). Hence s^* is a PRE.

Step 2. We now show that if s^* is a PRE (and hence a BNE), then condition (25) holds. Suppose that s^* is a PRE. Lemma 5 then implies that $\kappa_{ji}(s^*) \leq 0$ holds, for all $i, j \in I$, $i \neq j$. Condition (25) trivially holds for any pair i, j such that $\kappa_{ji}(s^*) = 0$. Hence assume $\kappa_{ji}(s^*) < 0$

[25] See Netzer and Schmutzler (2010) for a similar result, in the context of dynamic games between one materialistic and one reciprocal player.

but $s_i^* \notin \arg\min_{s_i \in S_i} \Pi_j(s_i, s_{-i}^*)$. Then, there exists $\tilde{s}_i$ so that $\Pi_i(\tilde{s}_i, s_{-i}^*) \leq \Pi_i(s_i^*, s_{-i}^*)$ and $\kappa_{ij}(\tilde{s}_i, s_{-i}^*) < \kappa_{ij}(s_i^*, s_{-i}^*)$. Consider a profile y where $y_{ij} > 0$ and $y_{ik} = 0$ for all $k \neq i, j$. Player i has an incentive to deviate from s_i^* to $\tilde{s}_i$ when

$$y_{ij} > \frac{\Pi_i(s_i^*, s_{-i}^*) - \Pi_i(\tilde{s}_i, s_{-i}^*)}{\kappa_{ji}(s^*)\left[\kappa_{ij}(\tilde{s}_i, s_{-i}^*) - \kappa_{ij}(s_i^*, s_{-i}^*)\right]},$$

which contradicts the assumption that s^* is a PRE. □

Proposition 7 says that a BNE s^* is a PRE if and only if every player i's strategy s_i^* maximizes the expression $\kappa_{ji}(s^*)\Pi_j(s_i, s_{-i}^*)$ for any opponent j. Observe that the coefficient $\kappa_{ji}(s^*)$ is a constant which does not depend on i's choice. By Lemma 5, it can either take a value of zero or a negative one. In the former case, trivially, any strategy $s_i \in S_i$ is a maximizer of $\kappa_{ji}(s^*)\Pi_j(s_i, s_{-i}^*)$. In the latter case, s_i^* is a maximizer if and only it is a minimizer of $\Pi_j(s_i, s_{-i}^*)$. Hence if, in any bilateral relation, a player experiences strictly negative kindness in a BNE s^*, then robustness requires that he minimizes the other's payoff. In the context of two player normal form games, Rabin (1993) calls a strategy profile "mutual-min" (p. 1290) when this is satisfied for both players. He proves that a mutual-min Nash equilibrium is always a fairness equilibrium, and hence robust in our sense. Proposition 7 implies that the analogous result is true in our Bayesian setting. Furthermore, Proposition 7 applies to an arbitrary number of players, and it provides a condition that is necessary *and* sufficient for robustness.[26]

Armed with Lemma 5 and Proposition 7, we can now prove the revelation principle for PRE.

Proposition 8. *Suppose a mechanism Φ implements the social choice function f in PRE. Then f is truthfully implementable in PRE in the corresponding direct mechanism.*

Proof. Step 1. First, we state the augmented revelation principle for PRE. Consider a mechanism $\Phi = (M_1, ..., M_n, g)$ with a PRE s^* that implements a social choice function f. From Proposition 2 it follows that there exists a strategically equivalent augmented revelation mechanism $\Phi' = (M_1', ..., M_n', g')$ that truthfully implements f in PRE, i.e. in which s^T is a PRE. Moreover, we can write $M_i' = \Theta_i \cup M_{i-}$ where $M_{i-} = M_i \backslash s_i^*(\Theta_i)$ are the unused actions from Φ. The strategy sets in Φ' are denoted S_i'. Lemma 5 implies that $\kappa_{ij}'(s^T) \leq 0$ holds in the PRE s^T of Φ', for all $i, j \in I$, $i \neq j$.

Step 2. The proof is completed if $M_{i-} = \varnothing$ for all $i \in I$, so that Φ' is a direct mechanism. Hence assume $M_{i-} \neq \varnothing$ for some i, and let $m_i' \in M_{i-} \subset M_i'$. Construct $\Phi'' = (M_1'', ..., M_n'', g'')$ from Φ' by letting $M_i'' = M_i' \backslash \{m_i'\}$ and keeping $M_j'' = M_j'$ for all $j \neq i$. Let g'' be the restriction of g' to $M_1'' \times ... \times M_n''$. The strategy sets in Φ'' are $S_i'' \subset S_i'$ and $S_j'' = S_j'$ for all $j \neq i$. We have only removed an unused action of player i, so s^T is still a BNE of Φ''. Lemma 5 thus implies that $\kappa_{ij}''(s^T) \leq 0$ still holds for all $j \neq i$. The kindness of all other players is completely unaffected by the removal.

Step 3. To prove that s^T is still a PRE of Φ'', we need to show that condition (25) is still satisfied. For player i this is immediate, since $s_i^T \in S_i'' \subset S_i'$ and $\kappa_{ji}''(s^T) = \kappa_{ji}'(s^T)$ holds for all

[26] Rabin (1993) also proves a robustness result for two player "mutual-max" Nash equilibria, where, phrased in terms of our notation, $s_i^* \in \arg\max_{s_i \in S_i} \Pi_j(s_i, s_{-i}^*)$ for both players. This result also follows from our proposition, because the mutual-max property implies that $\kappa_{ji}(s_j^*, s_{-j}^*) = 0$ holds for both players.

opponents $j \neq i$, i.e. the removal has left condition (25) unaffected. Then consider any player $j \neq i$, for whom $S_j'' = S_j'$. The only way in which (25) could be violated in Φ'', while not being violated in Φ', is that $\kappa_{ij}'(s^T) = 0$ held in Φ' but $\kappa_{ij}''(s^T) < 0$ holds in Φ''. We will show that this is impossible. From the proof of Lemma 5 we know that $\Pi_j(s^T) \leq \min_{s_i \in E_{ij}'(s_{-i}^T)} \Pi_j(s_i, s_{-i}^T)$, where $E_{ij}'(s_{-i}^T)$ are the bilaterally efficient strategies in Φ', due to s^T being a BNE in Φ'. Thus $\kappa_{ij}'(s^T) = 0$ requires

$$\min_{s_i \in E_{ij}'(s_{-i}^T)} \Pi_j(s_i, s_{-i}^T) = \Pi_j(s^T) = \max_{s_i \in E_{ij}'(s_{-i}^T)} \Pi_j(s_i, s_{-i}^T).$$

But the same two equalities must then hold in Φ'', implying $\kappa_{ij}''(s^T) = 0$ as well. For the maximization, we can always replace $E_{ij}'(s_{-i}^T)$ by S_i' and $E_{ij}''(s_{-i}^T)$ by S_i'', without changing the result. Thus since s_i^T maximizes $\Pi_j(s_i, s_{-i}^T)$ on S_i', and $s_i^T \in S_i'' \subset S_i'$, we have established the second equality for Φ''. The first equality for Φ'' then follows again from the proof of Lemma 5, together with the fact that s^T is a BNE in Φ''. Hence s^T still satisfies condition (25) in Φ'', and thus is a PRE.

Step 4. Iterating steps 2 and 3, we can remove all unused actions until arriving at a direct mechanism in which s^T is a PRE. □

The logic of the argument is as follows. As we have shown above, in any PRE s^* the strategy s_i^* essentially minimizes j's expected payoff, for every pair of players i and j. The revelation principle then follows from the fact that an action remains unused only if it is not needed to minimize the other players' payoffs, so that a removal of unused actions does not alter the equilibrium structure.

C Appendix

This appendix contains the proofs for Observations 1, 2, 3, 4, 5 and 6.

C.1 Proof of Observation 1

In the direct mechanism, the set of pure strategies for $i \in \{b, s\}$ is $S_i = \{s_i^T, s_i^H, s_i^L, s_i^{-T}\}$, where s_i^T denotes truth-telling, s_i^H prescribes to announce the high type $\bar{\theta}_i$ whatever the true type, s_i^L requires to always announce a low type, and s_i^{-T} is a strategy which requires to lie always, i.e., $s_i^{-T}(\underline{\theta}_i) = \bar{\theta}_i$ and $s_i^{-T}(\bar{\theta}_i) = \underline{\theta}_i$. We seek to show that (s_b^T, s_s^T) is not a BNFE. We proceed by contradiction. Hence fix some $y \in [0, \infty[^2$ and suppose that (s_b^T, s_s^T) is a BNFE. In the hypothetical equilibrium, beliefs are correct, which implies that $s_b^b = s_{sb}^{bb} = s_s^T$ and $s_s^b = s_{bs}^{bb} = s_b^T$.

The buyer's efficient strategies and the seller's equitable payoff. Solving for the efficient strategies of the buyer, given that the seller plays s_s^T, requires, in a first step, to look at how the two players' payoffs are affected as the buyer varies his strategy. Straightforward computations yield

$$\Pi_b(s_b^T, s_s^T) = \tfrac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \bar{\theta}_s),$$
$$\Pi_s(s_b^T, s_s^T) = \tfrac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \bar{\theta}_s),$$

$$\Pi_b(s_b^L, s_s^T) = \tfrac{1}{4}(\bar{\theta}_b - \underline{\theta}_s),$$
$$\Pi_s(s_b^L, s_s^T) = \tfrac{1}{4}(\underline{\theta}_b - \underline{\theta}_s),$$

$$\Pi_b(s_b^H, s_s^T) = \tfrac{1}{4}(\underline{\theta}_b - \underline{\theta}_s) + \tfrac{1}{4}(\underline{\theta}_b - \bar{\theta}_s),$$
$$\Pi_s(s_b^H, s_s^T) = \tfrac{1}{4}(\bar{\theta}_b - \underline{\theta}_s) + \tfrac{1}{4}(\bar{\theta}_b - \bar{\theta}_s),$$

$$\Pi_b(s_b^{-T}, s_s^T) = \tfrac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \bar{\theta}_s) - \tfrac{1}{4}(\bar{\theta}_b - \underline{\theta}_b),$$
$$\Pi_s(s_b^{-T}, s_s^T) = \tfrac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \bar{\theta}_s).$$

Inspection of these expressions reveals that strategy s_b^{-T} is not efficient, because a switch to s_b^T makes the buyer better off and leaves the seller unaffected. All other strategies are efficient since

$$\Pi_b(s_b^L, s_s^T) > \Pi_b(s_b^T, s_s^T) > \Pi_b(s_b^H, s_s^T),$$

$$\Pi_s(s_b^L, s_s^T) < \Pi_s(s_b^T, s_s^T) < \Pi_s(s_b^H, s_s^T).$$

Now we can easily compute that, from the buyer's perspective, the seller's equitable payoff is his payoff under truth-telling, i.e. $\pi_s^{e_b}(s_s^T) = \Pi_s(s_b^T, s_s^T)$.

The seller's efficient strategies and the buyer's equitable payoff. Analogously, the seller can induce the following payoff pairs by varying s_s:

$$\Pi_b(s_b^T, s_s^T) = \tfrac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \bar{\theta}_s),$$
$$\Pi_s(s_b^T, s_s^T) = \tfrac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \bar{\theta}_s),$$

$$\Pi_b(s_b^T, s_s^L) = \tfrac{1}{4}(\underline{\theta}_b - \underline{\theta}_s) + \tfrac{1}{4}(\bar{\theta}_b - \underline{\theta}_s),$$
$$\Pi_s(s_b^T, s_s^L) = \tfrac{1}{4}(\bar{\theta}_b - \bar{\theta}_s) + \tfrac{1}{4}(\underline{\theta}_b - \bar{\theta}_s),$$

$$\Pi_b(s_b^T, s_s^H) = \tfrac{1}{4}(\bar{\theta}_b - \bar{\theta}_s),$$
$$\Pi_s(s_b^T, s_s^H) = \tfrac{1}{4}(\bar{\theta}_b - \underline{\theta}_s),$$

$$\Pi_b(s_b^T, s_s^{-T}) = \tfrac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \bar{\theta}_s),$$
$$\Pi_s(s_b^T, s_s^{-T}) = \tfrac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) + \tfrac{1}{8}(\bar{\theta}_b - \bar{\theta}_s) - \tfrac{1}{4}(\bar{\theta}_s - \underline{\theta}_s).$$

Again, s_s^{-T} is Pareto dominated by s_s^T, while all other strategies are efficient due to

$$\Pi_b(s_b^T, s_s^L) > \Pi_b(s_b^T, s_s^T) > \Pi_b(s_b^T, s_s^H),$$

$$\Pi_s(s_b^T, s_s^L) < \Pi_s(s_b^T, s_s^T) < \Pi_s(s_b^T, s_s^H).$$

The buyers's equitable payoff is therefore also $\pi_b^{e_s}(s_b^T) = \Pi_b(s_b^T, s_s^T)$.

Verify that truth-telling is not a BNFE. In the hypothetical BNFE (s_b^T, s_s^T) we would have $\kappa_{bs}(s_{sb}^b, s_{sb}^{bb}) = 0$. This implies that the seller chooses $s_s \in S_s$ in order to maximize $\Pi_s(s_b^T, s_s)$. But s_s^T is not a solution to this problem, because s_s^H yields a strictly larger expected material payoff for the seller, as shown above. Hence (s_b^T, s_s^T) is not a BNFE.

C.2 Proof of Observation 2

Fix $y \in]0, \infty[^2$. We claim that truth-telling is an equilibrium for $\delta_b = 4/y_s$ and $\delta_s = 4/y_b$.

The buyer's kindness in a hypothetical truth-telling equilibrium. If the seller plays s_s^T, then any strategy s_b that announces $\underline{\theta}_b$ yields the same payoffs as the strategy that announces $\underline{\theta}_b$ instead, except for the additional redistribution from the seller to the buyer. Since $s_b = s_b^L$ maximized $\Pi_b(s_b, s_s^T)$ and minimized $\Pi_s(s_b, s_s^T)$ in the direct mechanism, strategy $s_b = \underline{s}_b$, with $\underline{s}_b(\theta_b) = \underline{\theta}_b$ for all θ_b, now maximizes $\Pi_b(s_b, s_s^T)$ and minimizes $\Pi_s(s_b, s_s^T)$ in Φ', and hence is efficient. It induces the payoffs

$$\Pi_b(\underline{s}_b, s_s^T) = \tfrac{1}{4}(\bar{\theta}_b - \underline{\theta}_s) + \tfrac{1}{2}\delta_b,$$
$$\Pi_s(\underline{s}_b, s_s^T) = \tfrac{1}{4}(\underline{\theta}_b - \underline{\theta}_s) - \tfrac{1}{2}\delta_b.$$

The efficient strategy which yields the highest payoff for the seller remains strategy s_b^H. Using the expression derived in the proof of Observation 1, we can compute the seller's equitable payoff

$$\pi_s^{e_b}(s_s^T) = \Pi_s(s_b^T, s_s^T) - \frac{1}{4}\delta_b.$$

Consequently, in the hypothetical truth-telling equilibrium we have

$$\kappa_{bs}(s_b^T, s_s^T) = \Pi_s(s_b^T, s_s^T) - \pi_s^{e_b}(s_s^T) = \frac{1}{4}\delta_b = \frac{1}{y_s}.$$

The seller's kindness in hypothetical truth-telling equilibrium. A symmetric argument implies

$$\pi_b^{e_s}(s_b^T) = \Pi_b(s_b^T, s_s^T) - \frac{1}{4}\delta_s$$

and

$$\kappa_{sb}(s_b^T, s_s^T) = \frac{1}{4}\delta_s = \frac{1}{y_b}$$

in the hypothetical truth-telling equilibrium.

Verify that truth-telling is a BNFE. We need to verify that $s_s = s_s^T$ is a maximizer of

$$\Pi_s(s_b^T, s_s) + y_s\kappa_{bs}(s_b^T, s_s^T)\Pi_b(s_b^T, s_s) = \Pi_s(s_b^T, s_s) + \Pi_b(s_b^T, s_s)\ ,$$

and that $s_b = s_b^T$ is a maximizer of

$$\Pi_b(s_b, s_s^T) + y_b\kappa_{sb}(s_b^T, s_s^T)\Pi_s(s_b, s_s^T) = \Pi_b(s_b, s_s^T) + \Pi_s(s_b, s_s^T)\ .$$

This follows since the social choice function f^* that is implemented if both players tell the truth is efficient, i.e. it maximizes the sum of material payoffs for any (θ_b, θ_s).

C.3 Proof of Observation 3

Consider a mechanism where each player has the message set $M_i = \{n_i, d_i\}$. The outcome function g has the following properties. If all agents choose their action n_i, or if at least two agents choose their action d_i, then the outcome is allocation $\bar{a} = (q_1^{\bar{a}}, \ldots, q_n^{\bar{a}}, t_1^{\bar{a}}, \ldots, t_n^{\bar{a}})$. This implies, in particular, that the outcome stipulated by the social choice function is implemented. If one agent i chooses d_i and all other agents $j \neq i$ choose n_j, then all consumption levels are still as in $\bar{a}$, $q_k^g(d_i, n_{-i}) = q_k^{\bar{a}}$ for all $k \in I$, but agent i receives an additional transfer $\epsilon_i > 0$,

so that $t_i^g(d_i, n_{-i}) = t_i^{\bar{a}} + \epsilon_i$, while every other agent $j \neq i$ has to pay and additional amount $\delta_{ij} > 0$, so that $t_j^g(d_i, n_{-i}) = t_j^{\bar{a}} - \delta_{ij}$. We assume that $\epsilon_i \leq \sum_{j \neq i} \delta_{ij}$ holds for all $i \in I$, so that the mechanism is feasible.

Denote by s_i^n the strategy where agent i chooses the message n_i whatever his type. Likewise, we denote by s_i^d the strategy which stipulates that he always chooses the action d_i. Consider a hypothetical BNFE $s^n = (s_1^n, \ldots, s_n^n)$. Given s_{-i}^n, the best that player i can do for any player $j \neq i$ is to play s_i^n, so that j has to pay δ_{ij} with probability 0. The worst (and also efficient) strategy for player j is s_i^d, where j would have to pay δ_{ij} with probability 1. Thus, from i's perspective the equitable payoff for player j is given by $\pi_j^{e_i}(s_{-i}^n) = -\delta_{ij}/2$. Player i's kindness towards j then equals $\kappa_{ij}(s_i^n, s_{-i}^n) = \delta_{ij}/2$ if he chooses s_i^n, and $\kappa_{ij}(s_i^d, s_{-i}^n) = -\delta_{ij}/2$ if he chooses s_i^d. We only need to consider the two strategies s_i^n and s_i^d, because if s_i^n is preferred over s_i^d then s_i^n is also preferred over any strategy that yields a payoff which is a convex combination of the payoffs associated with s_i^d and s_i^n, respectively. By symmetry it then follows immediately that s^n is a BNFE if and only if, for every player i, $\epsilon_i \leq (\sum_{j \neq i} y_{ij}\delta_{ij}\delta_{ji})/2$. Since $y \in]0, \infty[^{n(n-1)}$, we can always choose all $\epsilon_i > 0$ small enough so that this inequality is satisfied. Moreover, we can achieve arbitrarily large equilibrium utility levels by choosing δ_{ij} large enough, for all i and j.

C.4 Proof of Observation 4

We prove the observation in two steps. First, we show that there exist upper bounds on utilities for any (Φ, s^*) that implements f^*. Second, we show that (Φ', s^T), with parameters as given in the observation, reaches these bounds.

Step 1. By Proposition 2, for any mechanism that implements f^* in BNFE, we can find an augmented revelation mechanism that truthfully implements f^* in BNFE with identical utilities, i.e., an augmented revelation principle applies to utility-efficient implementation of an SCF. Hence consider w.l.o.g. a pair (Φ, s^T) where Φ is an augmented revelation mechanism for f^* and s^T is the truthful BNFE. In this BNFE, we have $s_b^b = s_{sb}^{bb} = s_s^T$ and $s_s^b = s_{bs}^{bb} = s_b^T$.

Bounds for the seller's kindness. For player b, s_b^H and s_b^L are viable strategies in Φ. Using the payoffs from the proof of Observation 1, the condition for not wanting to deviate to s_b^H (holding fixed s_s^T) can be rearranged to the upper bound on kindness

$$\kappa_{sb}(s_{bs}^b, s_{bs}^{bb}) \leq \kappa_{sb}^{max} = \frac{11}{5y_b}.$$

The condition for not wanting to deviate to s_b^L yields the lower bound

$$\kappa_{sb}(s_{bs}^b, s_{bs}^{bb}) \geq \kappa_{sb}^{min} = \frac{3}{5y_b}.$$

Bounds for the buyer's kindness. For player s, the condition for not wanting to deviate to s_s^L yields the upper bound

$$\kappa_{bs}(s_{sb}^b, s_{sb}^{bb}) \leq \kappa_{bs}^{max} = \frac{11}{5y_s}.$$

The condition for not wanting to deviate to s_s^H yields the lower bound

$$\kappa_{bs}(s_{sb}^b, s_{sb}^{bb}) \geq \kappa_{bs}^{min} = \frac{3}{5y_s}.$$

Utility bounds. Given the previous results it follows that $\Pi_i(s^T) + y_i\kappa_{bs}^{max}\kappa_{sb}^{max}$ is an upper bound on player i's utility in any pair (Φ, s^T) that implements f^*.

Step 2. Consider the extended mechanism Φ'. Recall from the proof of Observation 2 that in the hypothetical BNFE s^T we have $\kappa_{sb}(s_{bs}^b, s_{bs}^{bb}) = \delta_s/4$ and $\kappa_{bs}(s_{sb}^b, s_{sb}^{bb}) = \delta_b/4$. Choosing δ_b and δ_s as given in the observation then implies that $\kappa_{sb}(s_{bs}^b, s_{bs}^{bb}) = \kappa_{sb}^{max}$ and $\kappa_{bs}(s_{sb}^b, s_{sb}^{bb}) = \kappa_{bs}^{max}$, i.e. the upper bounds are reached. It remains to be shown that s^T is a BNFE of Φ' under these parameters.

Best responses. If player b chooses s_b^T, player s chooses s_s in order to maximize

$$\Pi_s(s_b^T, s_s) + y_s\kappa_{bs}^{max}\Pi_b(s_b^T, s_s) = \Pi_s(s_b^T, s_s) + \Pi_b(s_b^T, s_s) + \frac{6}{5}\Pi_b(s_b^T, s_s).$$

By construction of κ_{bs}^{max}, s_s^T and s_s^L yield the same value of this expression. From the payoffs derived in the proofs of Observations 1 and 2 it follows that any other strategy $s_s \in S_s' \setminus \{s_s^T, s_s^L\}$ yields a weakly lower value than s_s^T for both $\Pi_s(s_b^T, s_s) + \Pi_b(s_b^T, s_s)$ and $\Pi_b(s_b^T, s_s)$. Hence s_s^T is a best response. A symmetric argument implies that s_b^T is a best response for the buyer if the seller plays s_s^T.

C.5 Proof of Observation 5

We hypothesize (i) that players 2 and 3 vote sincerely, and (ii) that all players believe all other players to vote sincerely, and (iii) that all players have correct beliefs about the beliefs of the other players. We show that this implies that it is a best response for player 1 to also vote sincerely. A symmetric reasoning applies to the other players.

We leave it to the reader to verify that, given that players 2 and 3 vote sincerely, the expected material payoff of players 2 and 3 is given by

$$\Pi_2(s_1, (no, yes), (no, yes)) = \Pi_3(s_1, (no, yes), (no, yes)) = \frac{1}{4},$$

for all $s_1 \in \{no, yes\}^2$. Since player 1 cannot affect the expected material payoff of the other players, this trivially implies that the kindness of player 1 towards players 2 and 3 must satisfy $\kappa_{12} = \kappa_{13} = 0$ in the hypothetical equilibrium. By symmetry, this also implies that $\kappa_{21} = \kappa_{31} = 0$ in the hypothetical equilibrium. Given that all terms involving kindness sensations are equal to zero, player 1 chooses s_1 in order to maximize $\Pi_1(s_1, (no, yes), (no, yes))$. We leave it to the reader to verify that $s_1 = (no, yes)$ is the unique solution to this problem.

C.6 Proof of Observation 6

We hypothesize (i) that players 2 and 3 vote sincerely, and (ii) that all players believe all other players to vote sincerely, and (iii) that all players have correct beliefs about the beliefs of the other players. We show that, under these assumptions, it is a best response for player 1 to vote

sincerely if and only if $\sum_{j\neq 1} y_{1j} \geq 16$. A symmetric reasoning applies to the other players.

We leave it to the reader to verify that, given that players 2 and 3 vote sincerely, player 1 has two efficient strategies, (no, yes) and $(veto, yes)$. If player 1 chooses (no, yes) the associated expected material payoffs are

$$\Pi_1((no, yes), (no, yes), (no, yes)) = \frac{1}{4}$$

and

$$\Pi_2((no, yes), (no, yes), (no, yes)) = \Pi_3((no, yes), (no, yes), (no, yes)) = \frac{1}{4}.$$

If he chooses $(veto, yes)$, the payoffs are

$$\Pi_1((veto, yes), (no, yes), (no, yes)) = \frac{3}{8}$$

and

$$\Pi_2((veto, yes), (no, yes), (no, yes)) = \Pi_3((veto, yes), (no, yes), (no, yes)) = \frac{1}{8}.$$

Given that players 2 and 3 vote sincerely, their equitable payoffs are therefore $\pi_2^{e_1} = \pi_3^{e_1} = (1/4 + 1/8)/2 = 3/16$. Consequently, in a hypothetical equilibrium where all players vote sincerely, the kindness of player 1 towards players 2 and 3 equals $\kappa_{12} = \kappa_{13} = 1/4 - 3/16 = 1/16$. By symmetry, this also implies that $\kappa_{21} = \kappa_{31} = 1/16$ in the hypothetical equilibrium. Given that player 1 expects the other players to be kind, he will choose from his set of efficient strategies (as he has no incentive to sacrifice own payoff in order to harm others). Hence he will either choose $s_1 = (veto, yes)$ or $s_1 = (no, yes)$, depending on which of the two yields a larger value of

$$\Pi_1(s_1, (no, yes), (no, yes)) + \frac{1}{16}\left(y_{12}\Pi_2(s_1, (no, yes), (no, yes)) + y_{13}\Pi_3(s_1, (no, yes), (no, yes))\right).$$

It is straightforward to verify that the optimal choice is $s_1 = (no, yes)$ if and only if $y_{12} + y_{13} \geq 16$.

To complete the proof, observe that, whenever sincere voting is an equilibrium, all players have strictly positive kindness sensations, since, for all i and j, $\kappa_{ij} = \kappa_{ji} = 1/16$. By contrast, in a sincere voting equilibrium under simple majority voting (recall the proof of Observation 5) we have $\kappa_{ij} = \kappa_{ji} = 0$ for all i and j.

D Appendix

In the body of the text we have defined equitable payoffs in the spirit of Rabin (1993). Dufwenberg and Kirchsteiger (2004) have proposed an alternative definition. In this appendix, we show that Observations 1 to 6 carry through if we work with this alternative definition.

Formally, when defining the Dufwenberg-Kirchsteiger equitable payoff, we replace the set of conditionally and bilaterally Pareto efficient strategies $E_{ij}(s_i^b) \subseteq S_i$ by a set of unconditionally and globally Pareto efficient strategies $E_i \subseteq S_i$, so that

$$\pi_j^{e_i}(s_i^b) = \frac{1}{2}\left(\max_{s_i \in E_i} \Pi_j(s_i, s_i^b) + \min_{s_i \in E_i} \Pi_j(s_i, s_i^b)\right).$$

The set E_i is defined as follows. Strategy s_i belongs to E_i unless there exists $s_i' \in S_i$ such that for all $j \in I$,

$$\Pi_j(s_i', s_i^b) \geq \Pi_j(s_i, s_i^b)$$

for all possible $s_i^b \in S_{-i}$, with strict inequality for at least one j and s_i^b.

Note that the maximization part in the definition of equitable payoffs does not depend on whether we use Rabin's or Dufwenberg-Kirchsteiger's definition. The maximum of $\Pi_j(s_i, s_i^b)$ on both $E_{ij}(s_i^b)$ and E_i always coincides with its maximum on the whole strategy set S_i.

D.1 Observation 1

One can easily verify that, for both $i \in \{b, s\}$, the strategy s_i^{-T} does not belong to the set E_i. For any strategy s_j of the opponent $j \neq i$, strategy s_i^{-T} yields the same payoff as s_i^T for player j. It always yields a weakly lower payoff than s_i^T for player i, and a strictly lower payoff if the other player chooses s_j^T, as shown in the proof of Observation 1 in Appendix C.1. We have also shown in proof of Observation 1 that all other strategies are efficient conditional on the opponent being truthful. Consequently, $E_b = E_{bs}(s_s^T)$ and $E_s = E_{sb}(s_b^T)$, so that the remaining analysis is exactly as in the proof of Observation 1 in Appendix C.1.

D.2 Observation 2

As argued in the proof of Observation 2 in Appendix C.2, strategy $\underline{\underline{s}}_b$ uniquely minimizes the seller's and maximizes the buyer's expected material payoff, conditional on the seller playing s_s^T. Hence $\underline{\underline{s}}_b \in E_b$. Likewise, $\bar{\bar{s}}_s$ uniquely minimizes the buyer's and maximizes the seller's expected material payoff, conditional on the buyer playing s_b^T. Hence $\bar{\bar{s}}_s \in E_s$. The remaining analysis is thus exactly as in the proof of Observation 2 in Appendix C.2.

D.3 Observation 3

The analysis in Appendix C.3 is unaffected by the different efficiency concept.

D.4 Observation 4

Given the argument in Appendix D.2 above, the analysis in Appendix C.4 is unaffected by the different efficiency concept.

D.5 Observation 5

The analysis in Appendix C.5 is unaffected by the different efficiency concept.

D.6 Observation 6

Given that players 2 and 3 vote sincerely, $s_1 = (veto, veto)$ minimizes their payoffs. In fact, $(veto, veto)$ yields payoffs of zero for all three players. We claim that $(veto, veto) \in E_1$. Straightforward calculations reveal that, conditional on sincere voting of players 2 and 3, the strategies (yes, no), $(no, veto)$ and $(yes, veto)$ all yield strictly negative payoffs to player 1 and therefore

cannot dominate $(veto, veto)$. Conditional on $s_2 = s_3 = (yes, no)$, however, player 1's remaining strategies (no, yes), (yes, yes), (no, no), $(veto, no)$ and $(veto, yes)$ all yield strictly negative payoffs to players 2 and 3, and therefore also do not dominate $(veto, veto)$. It follows that $(veto, veto) \in E_1$. Otherwise arguing as in the proof of Observation 6 in Appendix C, equitable payoffs become $\pi_2^{e1} = \pi_3^{e1} = 1/8$, and the kindness terms in the hypothetical sincere equilibrium become $\kappa_{12} = \kappa_{13} = 1/4 - 1/8 = 1/8$. Symmetry then implies $\kappa_{21} = \kappa_{31} = 1/8$ in the equilibrium candidate. The rest of the analysis is affected by the different efficiency concept only in that we obtain the modified condition $y_{12} + y_{13} \geq 8$ for sincere voting to be a BNFE.