Bierbrauer, Felix; Netzer, Nick

Working Paper

Mechanism design and intentions

Preprints of the Max Planck Institute for Research on Collective Goods, No. 2016/4

Provided in Cooperation with:
Max Planck Institute for Research on Collective Goods

**Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2016/4**

Mechanism Design and Intentions

Felix Bierbrauer
Nick Netzer

MAX PLANCK SOCIETY

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2016/4

Mechanism Design and Intentions

Felix Bierbrauer / Nick Netzer

February 2016

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

Mechanism Design and Intentions*

Felix Bierbrauer
Max Planck Institute, Bonn
University of Cologne

Nick Netzer
University of Zurich

This version: February 2016
First version: July 2011

Abstract

We introduce intention-based social preferences into mechanism design. We explore information structures that differ with respect to what is commonly known about the weight that agents attach to reciprocal kindness. When the designer has no information on reciprocity types, implementability of an incentive-compatible social choice function is guaranteed if it satisfies an additional insurance property. By contrast, precise information on reciprocity types may imply that all efficient social choice functions are implementable. We show how these results extend to a two-dimensional mechanism design setting where the agents have private information about their material payoff types and their reciprocity types. We also provide a systematic account of the welfare implications of intentionality.

Keywords: Mechanism Design, Psychological Games, Social Preferences, Reciprocity.
JEL Classification: C70, C72, D02, D03, D82, D86.

*Email: bierbrauer@wiso.uni-koeln.de and nick.netzer@econ.uzh.ch. We thank Tomer Blumkin, Stefan Buehler, Antonio Cabrales, Juan Carlos Carbajal, Martin Dufwenberg, Kfir Eliaz, Florian Englmaier, Ernst Fehr, Alexander Frankel, Silvia Grätz, Hans Peter Grüner, Paul Heidhues, Martin Hellwig, Holger Herz, Benny Moldovanu, Johannes Münster, Roger Myerson, Zvika Neeman, Axel Ockenfels, Marco Ottaviani, Ariel Rubinstein, Désirée Rückert, Larry Samuelson, Klaus Schmidt, Armin Schmutzler, Alexander Sebald, Joel Sobel, Ran Spiegler, Balázs Szentes, André Volk, Roberto Weber, Philipp Weinschenk, David Wettstein, Philipp Wichardt, and seminar participants at the CESifo Area Conference on Behavioural Economics 2011, the MPI Conference on Private Information, Interdependent Preferences and Robustness 2013, ESSET 2013, Ben-Gurion University of the Negev, CERGE-EI Prague, HU and FU Berlin, ULB Brussels, MPI Bonn, LMU Munich, Tel Aviv University and the Universities of Basel, Bern, Chicago, Cologne, Heidelberg, Mannheim, St. Gallen and Zurich. Financial support by the Swiss National Science Foundation (Grant No. 100018_126603 "Reciprocity and the Design of Institutions") is gratefully acknowledged. All errors are our own.

1 Introduction

Agents with intention-based social preferences 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 (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004). The behavioral relevance of such preferences is well established (e.g. Andreoni et al., 2002; Falk et al., 2003, 2008). In this paper, we explore their implications for the theory of mechanism design. Specifically, we provide answers to the following questions:

> There is a rich literature on mechanism design that has proceeded under the assumption that agents are selfish. To what extent are these mechanisms robust to the possibility that the participants may be motivated by intention-based social preferences?
>
> How do intention-based social preferences affect the set of implementable social choice functions relative to a benchmark with selfish agents? In particular, does intentionality make it easier or more difficult to implement good outcomes?
>
> Suppose that the designer seeks not only good material outcomes but also good attitudes among the participants of the mechanism. Is there a trade-off between these objectives? Do we have to sacrifice efficiency if we want kindness among the agents, or are sensations of kindness helpful for the implementation of efficient outcomes?

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.

We approach the questions above in three different ways. We first characterize social choice functions that are *strongly implementable*. Our notion of strong implementability is attractive from the perspective of a mechanism designer who acknowledges the possibility that the agents may be motivated by intention-based social preferences, but who wishes to remain agnostic about the intensity of these preferences. A strongly implementable social choice function is implementable irrespective of the mixture between selfish and reciprocal individuals among the participants of a mechanism. We then consider social choice functions that are *weakly implementable*, i.e., which are implementable if the mechanism designer has precise information on the strength of intention-based social preferences. This concept is of interest for two different reasons. First, it allows for a clear exposition of the conceptual issues that arise due to the procedural nature of intention-based preferences. For instance, we show that the revelation principle does not hold. We also discuss alternative notions of welfare and the modelling of participation constraints in a model with intentions. Second, looking first at weakly implementable social choice functions sets the stage for our analysis of the *two-dimensional mechanism design* problem that emerges if the agents have private information both about their material payoffs and about the weight that kindness sensations have in their utility function.

Strongly Implementable Social Choice Functions. Our first main result (Theorem 1) states that a social choice function is strongly implementable if it is implementable in a model

with selfish agents and, moreover, is such that the agents cannot affect each other's payoff by unilateral deviations from truth-telling. We refer to the latter property as the *insurance property*, since it implies that the expected payoff of agent i does not depend on the type of agent j, i.e., each agent is insured against the randomness of the other agent's type. The insurance property shuts down the transmission channel for reciprocal behavior. If agent i cannot influence the payoff of agent j, then j has no reason to interpret i's behavior as kind or as unkind. Agent j thus neither has a reason nor an opportunity to reward or punish agent i, so she focusses on her own expected payoff and acts as if she was selfish. Incentive-compatibility implies that this selfish behavior implements the given social choice function.

We then turn to a characterization of social choice functions that are incentive-compatible and have the insurance property. Proposition 1 establishes that the set of efficient social choice functions contains functions which have these two properties. Proposition 2 provides a tool for the construction of social choice functions with the insurance property. As an input it requires a social choice function that is implementable if everybody is selfish. It then modifies the monetary transfers in such a way that the insurance property holds and key properties of the initial social choice function (expected payments, expected payoffs, incentive-compatibility) remain unchanged.

Theorem 1 and Propositions 1 and 2 are reassuring from the perspective of the established theory in mechanism design, which is based on the assumption that individuals are selfish. 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 robust mechanisms in which 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 implementability.

Weakly Implementable Social Choice Functions. Our analysis of weakly implementable social choice functions proceeds under the assumption that, while agents have private information about their material payoffs, the weight of kindness in their overall utility function is commonly known. This information structure makes it possible to highlight the issues that arise if one seeks to exploit intention-based social preferences for mechanism design.

With intention-based preferences, whether agent i interprets agent j as kind or as unkind depends not only on what j does, but also on what j could have done instead. Hence our analysis begins with the observation that the specification of message sets affects the set of achievable outcomes. For instance, the designer can augment a direct mechanism with additional actions that give each agent a possibility to enrich himself at the expense of the other agents. If all agents refrain from using these actions, they will interpret each other as kind. These kindness sensations then make it possible to implement social choice functions that are out of reach if everybody is selfish.[1]

[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 percent of the cake, for instance, is

Theorem 2 uses this insight to provide conditions under which indeed any efficient social choice function can be implemented. The mechanism that we construct in order to prove Theorem 2 also satisfies participation constraints and hence eliminates any tension between efficiency, incentive-compatibility and voluntary participation. This implies that famous impossibility results such as the one by Myerson and Satterthwaite (1983) are turned into possibility results.

Proposition 3 addresses the question whether Pareto-efficiency and kindness are competing objectives. It shows that, under the conditions of either Theorem 1 or 2, materially surplus-maximizing outcomes can be implemented with maximal kindness levels. Thus, one can have maximal material payoffs and at the same time reach a maximal level of kindness.

Two-Dimensional Mechanism Design. We then turn to an information structure where the agents have private information both on their material payoffs and on the weight of kindness sensations in their utility function. Both dimensions may be elicited by a mechanism. Clearly, our results on strong implementability provide a lower bound on what can be achieved in this setting, while our results on weak implementability provide an upper bound.

We start by considering the class of social choice functions that map material payoff types into economic outcomes. This class includes materially Pareto-efficient social choice functions, because reciprocity types have no direct impact on material payoffs. It does not include social choice functions under which the final allocation varies with the weight that the agents attach to sensations of kindness. Proposition 4 shows that incentive-compatibility remains a necessary condition for implementability of such social choice functions if there is a positive probability that the agents are selfish. Hence, a mere possibility of reciprocal behavior does not enlarge the set of implementable social choice functions relative to a benchmark model with selfish agents. Incentive-compatibility is not sufficient, though. Agents with strong reciprocal inclinations may be ready to deviate from truthful behavior to affect other agents. We already know from Theorem 1 that the insurance property is a sufficient condition to rule out this possibility. Proposition 5 clarifies the conditions under which the insurance property is also necessary for implementability. In particular, there must be a positive probability that agents attach a sufficiently large weight to sensations of kindness. Under these assumptions, Theorem 1 and Propositions 4 and 5 together imply that a social choice function is implementable if and only if it is incentive-compatible and has the insurance property. We also show that the resulting equilibrium kindness of zero is the highest level of kindness one can hope for when selfish types are around.

How do these findings change if we know for sure that the agents are not selfish? Under the assumption that reciprocity weights are bounded away from zero, Proposition 6 provides an extension of Theorem 2 and Proposition 3. Thus, our results reveal that the issue is not really whether reciprocity types are taken to be observable. A more important distinction is whether selfish types are possible. If they are not, our results on weakly implementable social choice functions extend to a model with private information on reciprocity types. If selfish types are possible, one has to live with incentive-compatibility and, under plausible conditions, also the insurance property. It is therefore appropriate to focus on strongly implementable social choice functions in this case.

considered very unfair if better offers such as 50 percent or 80 percent 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.

Finally, we study social choice functions under which the allocation depends not only on the agents' material payoffs but also on their reciprocity types. The dependence of outcomes on reciprocity types typically implies a loss of material efficiency, but it may generate additional degrees of freedom in incentive provision. Specifically, we consider a bilateral trade example where both agents are either selfish or attach a positive weight to sensations of kindness. We study a social choice function under which selfish types choose to enrich themselves at the expense of the other agent, while reciprocal types refrain from doing so. Equilibrium kindness becomes positive if the selfish types are sufficiently rare, and we can reach an efficient outcome with a probability close to one. Importantly, as the probability of selfish types goes to zero, we approximate a mechanism that relies on unused actions and induces efficient outcomes under the assumption of known reciprocity types, i.e., the type of mechanism that enabled us to prove Theorem 2. Hence we can interpret unused actions as the limit of actions which are rarely used, namely in the small probability event that an agent is entirely selfish.

The remainder is organized as follows. The next section contains 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 strongly implementable social choice functions, and Section 5 covers weakly implementable social choice functions. Our analysis of the two-dimensional mechanism design problem can be found in Section 6. Throughout, we illustrate our results using a simplified version of the bilateral trade problem due to Myerson and Satterthwaite (1983). As an extension, Section 7 discusses the possibility that the agents do not perceive the mechanism as exogenous but have intention-based preferences also towards the mechanism designer. Concluding remarks which discuss the applicability of our results can be found in Section 8. All proofs and some supplementary materials are relegated to the appendix.

2 Literature

Models of social preferences are usually distinguished according to whether they are outcome-based or intention-based. Prominent examples for outcome-based models are Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), while Rabin (1993) and Dufwenberg and Kirchsteiger (2004) are intention-based. An extensive experimental literature (e.g. Andreoni et al., 2002; Falk et al., 2003, 2008) has concluded that behavior is most likely influenced by both types of considerations. 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. The formal framework for modelling intentions is provided by psychological game theory (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009), which allows payoffs to depend on higher-order beliefs. The literature does not yet contain a general treatment of intention-based social preferences for

games of incomplete information.[2] Our mechanism design approach requires a general theory of intentions for Bayesian games, and we will outline such a theory in Section 3.3.

Several authors have investigated mechanism design problems with outcome-based social preferences.[3] 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. Desiraju and Sappington (2007) study a one-principal-multiple-agents-model with inequality-averse agents. They show that, if the agents' types are independently drawn, then payment schemes can be constructed so that inequality-aversion does not interfere with incentive provision by the principal. Our results on strong implementability can be viewed as a generalization of this observation that also covers models with intention-based social preferences. By contrast, von Siemens (2011) studies a model in which inequality-aversion among the employees of a firm cannot be neutralized. As a consequence, the firm deviates from the incentive scheme that would be optimal if all agents were selfish. Tang and Sandholm (2012) characterize optimal auctions in the presence of spiteful agents, who attach an exogenous negative weight to the utility of others, and Kucuksenel (2012) investigates an optimal design problem with altruistic agents, who attach an exogenous positive weight to the utility of others.

Experimental and theoretical studies have shown that the design of incentive contracts can be facilitated in environments with reciprocal agents (e.g. Fehr et al., 1997; Fehr and Falk, 2002; Englmaier and Leider, 2012; Hoppe and Schmitz, 2013; Benjamin, 2014). However, reciprocity is not necessarily a beneficial force. In Hart and Moore (2008) and Netzer and Schmutzler (2014), for instance, negative reciprocal reactions can be inevitable and generate inefficient contract outcomes. Our mechanism design approach clarifies the conditions under which reciprocity among the agents either hampers or helps in achieving good outcomes.

Several papers are related in that they study problems of mechanism design with a focus on procedural questions.[4] 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 mechanism design problem. Gradwohl (2014) studies implementation with agents who do not only care about the outcome of a mechanism but also about the extent of information revelation in equilibrium. The possibility that institutions

[2] Rabin (1993) and Dufwenberg and Kirchsteiger (2004) assume complete information. 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. Some contributions (e.g. Sebald, 2010; Aldashev et al., 2015) introduce randomization devices into psychological games, but still under the assumption of perfect observability. von Siemens (2009, 2013) contain models of intentions for two-stage games with incomplete information about the second-mover's social type.

[3] There also exist applications of outcome-based social preferences to moral hazard problems (e.g. Englmaier and Wambach, 2010; Bartling, 2011) and to labor market sceening problems (e.g. Cabrales et al., 2007; Cabrales and Calvó-Armengol, 2008; Kosfeld and von Siemens, 2011). Reciprocity is introduced into moral hazard problems by De Marco and Immordino (2014, 2013) and into a screening problem by Bassi et al. (2014). These contributions work with adaptations of the models by Rabin (1993) and Levine (1998), respectively, which effectively transform them into outcome-based models.

[4] Frey et al. (2004) provide a general discussion of procedural preferences and their role for the design of institutions. Gaspart (2003) follows an axiomatic approach to procedural fairness in implementation problems. Aside from procedural issues, the literature has investigated a range of other behavioral phenomena in mechanism design problems, among them error-prone behavior (Eliaz, 2002), emotions like fear (Caplin and Eliaz, 2003), and myopic learning (Cabrales and Serrano, 2011).

affect preferences has received some attention in general (see e.g. Bowles and Polanía-Reyes, 2012). In Alger and Renault (2006), the mechanism and its equilibrium influence the agents' intrinsic propensity to lie. This can sometimes make non-direct mechanisms optimal. de Clippel (2014) studies the problem of full implementation under complete information with agents whose behavior is described by arbitrary choice functions instead of preferences. Indirect mechanisms play a role also in this context, due to the possibility of menu-dependence. Saran (2011), in contrast, provides conditions for the revelation principle to hold in a Bayesian framework even in such cases. Antler (2015) investigates a matching problem where the agents' preferences are directly affected by the stated preferences of their potential partners.

3 The Model

3.1 Environment and Mechanisms

We focus on the conventional textbook environment with quasi-linear payoffs and independent private values (see Mas-Colell et al., 1995, ch. 23). For simplicity we consider the case of only two agents, but we deal with extensions to an arbitrary finite number of agents in Section B of the appendix.

The environment is described by $E = [A, \Theta_1, \Theta_2, p_1, p_2, \pi_1, \pi_2]$. The set of feasible allocations is denoted by A, where a typical element is a list $a = (q_1, q_2, t_1, t_2)$. Depending on the application, q_i may stand for agent i's consumption of a public or private good, or it may denote her effort or output. We will simply refer to q_i as agent i's consumption level. The monetary transfer to agent i is denoted by t_i. Formally, the set of allocations is given by $A = Q \times \mathbb{R}^2$ for some $Q \subseteq \mathbb{R}^2$. We assume that pairs of consumption levels (q_1, q_2) do not come with an explicit resource requirement. Resource costs can be captured in the payoff functions for most applications of interest. An allocation is said to achieve budget-balance if $t_1 + t_2 = 0$.

The type of agent i is the realization θ_i of a random variable $\tilde{\theta}_i$ that takes values in a finite set Θ_i. The realized type is privately observed by the agent. Types are independently distributed and p_i denotes the probability distribution of $\tilde{\theta}_i$. We also write $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2)$ and denote realizations of $\tilde{\theta}$ by $\theta = (\theta_1, \theta_2) \in \Theta = \Theta_1 \times \Theta_2$. We write $\mathbb{E}[\cdot]$ for the expectation with respect to all random variables within the squared brackets.[5]

Finally, $\pi_i : A \times \Theta_i \to \mathbb{R}$ is the material payoff function of agent i. If allocation a is selected and type θ_i has realized, then agent i obtains the material payoff $\pi_i(a, \theta_i) = v_i(q_i, \theta_i) + t_i$.[6]

The material surplus that is generated by consumption levels (q_1, q_2) if types are given by $\theta = (\theta_1, \theta_2)$ equals $v_1(q_1, \theta_1) + v_2(q_2, \theta_2)$. An allocation $a = (q_1, q_2, t_1, t_2)$ is said to be materially surplus-maximizing for type profile θ if $v_1(q_1, \theta_1) + v_2(q_2, \theta_2) \geq v_1(q_1', \theta_1) + v_2(q_2', \theta_2)$, for all $(q_1', q_2') \in Q$. An allocation a is said to be materially Pareto-efficient for type profile θ if it is materially surplus-maximizing and achieves budget-balance. A social choice function

[5] Throughout, possible realizations of a random variable $\tilde{x}$ are denoted by x, x' or $\hat{x}$. For instance, if $\tilde{x}_1$ and $\tilde{x}_2$ are random variables and g is an arbitrary function, then $\mathbb{E}[g(\tilde{x}_1, \tilde{x}_2)]$ indicates that an expectation is computed based on the joint distribution of the random variables $\tilde{x}_1$ and $\tilde{x}_2$, and $\mathbb{E}[g(\tilde{x}_1, x_2')]$ indicates that an expectation is computed based on the distribution of $\tilde{x}_1$ conditional on the event $\tilde{x}_2 = x_2'$.

[6] The type θ_i affects only agent i's material payoff. In Section 6 we will analyse a two-dimensional mechanism design problem where an agent is characterized by a material payoff type and a reciprocity type, both of which are private information.

(SCF) $f : \Theta \to A$ specifies an allocation as a function of both agents' types. We also write $f = (q_1^f, q_2^f, t_1^f, t_2^f)$. A social choice function f is said to be materially Pareto-efficient if the allocation $f(\theta)$ is materially Pareto-efficient for every type profile $\theta \in \Theta$.

A mechanism $\Phi = [M_1, M_2, g]$ contains a finite message set M_i for each agent and an outcome function $g : M \to A$, which specifies an allocation for each profile $m = (m_1, m_2) \in M = M_1 \times M_2$. We also write $g = (q_1^g, q_2^g, t_1^g, t_2^g)$. A pure strategy for agent i in mechanism Φ is a function $s_i : \Theta_i \to M_i$. The set of all such strategies of agent i is denoted S_i, and we write $S = S_1 \times S_2$. We denote by $g(s(\theta))$ the allocation that is induced if types are given by θ and individuals follow the strategies $s = (s_1, s_2)$. For later reference, we also introduce notation for first- and second-order beliefs about strategies. Since we will focus on pure strategy equilibria in which beliefs are correct, we can without loss of generality assume that agent i's belief about j's strategy puts unit mass on one particular element of S_j, which we will denote by s_i^b (we assume $j \neq i$ here and throughout the paper). Analogously, we denote by $s_i^{bb} \in S_i$ agent i's (second-order) belief about j's belief about i's own strategy.

3.2 Bayes-Nash Equilibrium

Before turning to the model of intention-based social preferences, we remind ourselves of the solution concept of a Bayes-Nash equilibrium (BNE). Given an environment E and a mechanism Φ, agent i's ex ante expected material payoff from following strategy s_i, given her belief s_i^b about the other agent's strategy, is given by

$$\Pi_i(s_i, s_i^b) = \mathbb{E}[v_i(q_i^g(s_i(\tilde{\theta}_i), s_i^b(\tilde{\theta}_j)), \tilde{\theta}_i) + t_i^g(s_i(\tilde{\theta}_i), s_i^b(\tilde{\theta}_j))].$$

Definition 1. *A BNE is a strategy profile $s^* = (s_1^*, s_2^*)$ such that, for both $i = 1, 2$,*

(a) $s_i^* \in \arg\max_{s_i \in S_i} \Pi_i(s_i, s_i^b)$, *and*

(b) $s_i^b = s_j^*$.

We say that a social choice function f can be implemented in BNE if there exists a mechanism Φ that has a BNE s^* so that, for all $\theta \in \Theta$, $g(s^*(\theta)) = f(\theta)$. The characterization of social choice functions that are implementable in BNE is facilitated by the well-known revelation principle. To state this principle, we consider the direct mechanism for a given social choice function f, i.e., the mechanism with $M_1 = \Theta_1$, $M_2 = \Theta_2$, and $g = f$. Given such a mechanism, truth-telling for agent i is the strategy s_i^T that prescribes $s_i^T(\theta_i) = \theta_i$, for all $\theta_i \in \Theta_i$. According to the revelation principle, a social choice function f is implementable in BNE if and only if truth-telling by all agents is a BNE in the corresponding direct mechanism. Equivalently, a social choice function is implementable in BNE if and only if it satisfies the following inequalities, which are known as Bayesian incentive-compatibility (BIC) constraints:

$$\mathbb{E}[v_i(q_i^f(\theta_i, \tilde{\theta}_j), \theta_i) + t_i^f(\theta_i, \tilde{\theta}_j)] \geq \mathbb{E}[v_i(q_i^f(\hat{\theta}_i, \tilde{\theta}_j), \theta_i) + t_i^f(\hat{\theta}_i, \tilde{\theta}_j)], \tag{1}$$

for both $i = 1, 2$ and all $\theta_i, \hat{\theta}_i \in \Theta_i$. In many applications, in addition to the requirement of

BIC, participation constraints (PC) have to be respected:

$$\mathbb{E}[v_i(q_i^f(\theta_i, \tilde{\theta}_j), \theta_i) + t_i^f(\theta_i, \tilde{\theta}_j)] \geq 0, \tag{2}$$

for both $i = 1, 2$ and all $\theta_i \in \Theta_i$. The interpretation is that participation in the mechanism is voluntary and that agents take their participation decision after having learned their own type, but prior to learning the other agent's type. They will participate only if the payoff they expect from participation in the mechanism is non-negative.

3.3 Bayes-Nash Fairness Equilibrium

We now adapt the model of intention-based social preferences due to Rabin (1993) to normal form games of incomplete information. The resulting solution concept will be referred to as a Bayes-Nash fairness equilibrium (BNFE). Specifically, we follow the literature on intention-based social preferences and assume that the agents have a utility function of the form

$$U_i(s_i, s_i^b, s_i^{bb}) = \Pi_i(s_i, s_i^b) + y_i \, \kappa_i(s_i, s_i^b) \, \kappa_j(s_i^b, s_i^{bb}). \tag{3}$$

The first source of utility is the expected material payoff $\Pi_i(s_i, s_i^b)$. The second source of utility is a psychological payoff $\kappa_i(s_i, s_i^b) \, \kappa_j(s_i^b, s_i^{bb})$, which is added with an exogenous weight of $y_i \geq 0$. The term $\kappa_i(s_i, s_i^b)$ captures the kindness that agent i intends to achieve toward agent j by choosing strategy s_i, given her belief s_i^b about j's strategy. The term $\kappa_j(s_i^b, s_i^{bb})$ captures the belief of agent i about the analogously defined kindness $\kappa_j(s_j, s_j^b)$ intended by j toward i. Forming this belief requires agent i to reason about agent j's first-order belief, which explains why second-order beliefs become relevant. The sign of κ_j is important for i's attitude towards j. If i expects to be treated kindly, $\kappa_j > 0$, then her utility is increasing in her own kindness. The opposite holds if i expects to be treated unkindly, $\kappa_j < 0$, in which case she wants to be unkind in return.

Kindness is determined as follows. There is an equitable reference payoff $\Pi_j^e(s_i^b)$ for agent j, which describes what agent i considers as the payoff that j deserves. If i's strategy choice yields a payoff for j that exceeds this norm, then i is kind, otherwise she is unkind. Specifically, we postulate that

$$\kappa_i(s_i, s_i^b) = h(\Pi_j(s_i, s_i^b) - \Pi_j^e(s_i^b)),$$

where

$$h(x) = \begin{cases} \bar{\kappa} & \text{if} \quad \bar{\kappa} < x, \\ x & \text{if} \quad -\bar{\kappa} \leq x \leq \bar{\kappa}, \\ -\bar{\kappa} & \text{if} \quad x < -\bar{\kappa}. \end{cases}$$

The kindness bound $\bar{\kappa} > 0$ allows us to restrict the importance of psychological payoffs relative to material payoffs, but it can also be set to $\bar{\kappa} = \infty$. Having a bound on kindness sensations will be of particular importance for our analysis of the problem to implement an SCF with maximal kindness among the agents. This problem would not be well-defined in the absence of

a kindness bound.[7] The crucial feature of models with intention-based social preferences is that equitable payoffs are menu-dependent. Following Rabin (1993), we assume that, from agent i's perspective, the relevant menu is the set of Pareto-efficient own strategies, conditional on the other agent choosing strategy s_i^b. This set is henceforth denoted $E_i(s_i^b)$.[8] To be specific, we assume that the payoff deserved by j is the average of the payoff she would get if i was completely selfish and the payoff she would get if i cared exclusively for j:

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

The restriction of the relevant menu to efficient strategies ensures that kindness is generated only by choices that involve a non-trivial trade-off between the agents. This property is important for mechanism design, as it implies that kindness cannot be manipulated by merely adding non-tempting punishment options to a mechanism.[9] Different specifications of the reference point have been explored in the literature (e.g. Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). We do not wish to argue that our assumptions are the only reasonable ones. What is crucial for the analysis that follows is the menu-dependence of the equitable reference payoff. The menus that are made available by the mechanism designer affect the interpretation of behavior. This feature of the model makes our analysis conceptually different from one in which preferences are purely outcome-based.[10]

Definition 2. *A BNFE is a strategy profile $s^* = (s_1^*, s_2^*)$ such that, for both $i = 1, 2$,*

(a) $s_i^* \in \arg\max_{s_i \in S_i} U_i(s_i, s_i^b, s_i^{bb})$,

(b) $s_i^b = s_j^*$, *and*

(c) $s_i^{bb} = s_i^*$.

The definition of BNFE becomes equivalent to the definition of BNE whenever $y_1 = y_2 = 0$, so that concerns for reciprocity are absent. Our definitions of both BNE and BNFE are based on the ex ante perspective, that is, on the perspective of agents who have not yet discovered their types but plan to behave in a type-contingent way. As is well-known, for the case of BNE there is an equivalent definition which evaluates actions from an ex interim perspective, where

[7] Dufwenberg and Kirchsteiger (2004) do not have a bound on kindness, which corresponds to $\bar{\kappa} = \infty$. Rabin (1993) adopts a normalization that implies that kindness lies in the interval $[-1, 1/2]$. Still, kindness is strictly increasing in the opponent's payoff. In our model, it is increasing only as long as the kindness bound is not binding. Whenever our bound is not binding, we can rewrite utility as $U_i(s_i, s_i^b, s_i^{bb}) = \Pi_i(s_i, s_i^b) + y_i\kappa_j(s_i^b, s_i^{bb})\Pi_j(s_i, s_i^b) - y_i\kappa_j(s_i^b, s_i^{bb})\Pi_j^e(s_i^b)$, which shows that agent i maximizes a weighted sum of both agents' material payoffs. The weight on the other agent's payoff is endogenously determined by her kindness toward i and can be negative.

[8] Conditional on s_i^b, a strategy $s_i \in S_i$ is Pareto-dominated by a strategy $s_i' \in S_i$ if $\Pi_k(s_i', s_i^b) \geq \Pi_k(s_i, s_i^b)$ for both $k = 1, 2$, with strict inequality for at least one k. A strategy is Pareto-efficient and hence contained in $E_i(s_i^b)$ if it is not Pareto-dominated by any other strategy in S_i.

[9] For an assessment of i's kindness, however, it does not matter how costly it is to generate the best outcome for j, nor does it matter how much i would gain from generating the worst outcome for j. To avoid implausible implications of this property, we will, for most of our results, impose the additional requirement of budget-balance on and off the equilibrium path, which makes it impossible to take a lot from one agent without giving it to the other agent.

[10] In Appendix E, we go through several examples to demonstrate that the logic of our analysis does not depend upon whether we model equitable payoffs as in Rabin (1993) or as in Dufwenberg and Kirchsteiger (2004).

agents have learned their own type but lack information about the types of the other agents. In Appendix C, we develop an analogous ex interim version of BNFE and provide conditions on the relation between ex ante and ex interim kindness under which the two versions are equivalent.

The solution concept of a BNFE relies on two sources of utility, material payoffs and kindness sensations. This raises the question how to treat them from a welfare perspective. The question can be formulated using the notions of decision utility and experienced utility (Kahneman et al., 1997). Our analysis is based on the assumption that behavior is as if individuals were maximizing the decision utility function U_i, but it leaves open the question whether sensations of kindness should be counted as an own source of experienced well-being.[11] We will investigate welfare based on the entire utility function (3) in Section 5. First, however, we work with the conventional notion of material Pareto-efficiency introduced above, i.e., we investigate how the behavioral implications of reciprocity affect the possibility to achieve materially efficiency outcomes.

Our definition of a BNFE presumes common knowledge of the reciprocity weights $y = (y_1, y_2)$ among the agents. Consequently, expectations have to be taken only with respect to the types θ_i and θ_j. We relax this assumption in Section 6, where we clarify the conditions under which our results extend to a setting with private information on reciprocity types. Even with the assumption of common knowledge of reciprocity types among the agents, we can still make a distinction whether or not the information about y is available also to the mechanism designer. Our notion of strong implementability assumes that the designer neither knows y nor attempts to elicit these parameters. Instead, he attempts to come up with a mechanism that reaches a given social choice function for all $y \in Y$, where Y is some pre-specified set of possible values. With weak implementability, by contrast, the designer is assumed to know y and can thus calibrate the mechanism accordingly.

Definition 3.

(a) An SCF f is strongly implementable in BNFE on $Y \subseteq \mathbb{R}^2_+$ if there exists a mechanism Φ and a profile s^* such that s^* is a BNFE for all $y \in Y$ and $g(s^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$.

(b) An SCF f is weakly implementable in BNFE on $Y \subseteq \mathbb{R}^2_+$ if, for every $y \in Y$, there exists a mechanism Φ and a profile s^* such that s^* is a BNFE and $g(s^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$.

In the following, we will simply use the term "strong implementability" when referring to strong implementability on the complete set $Y = \mathbb{R}^2_+$.

3.4 The Bilateral Trade Problem

A simplified version of the classical bilateral trade problem due to Myerson and Satterthwaite (1983) will be used repeatedly to illustrate key concepts and our main results. There is a buyer b and a seller s. The seller produces $q \in [0, 1]$ units of a good that the buyer consumes. The buyer's material payoff is given by $v_b(q, \theta_b) = \theta_b q$, so that θ_b is her marginal valuation of the good. The seller's material payoff is given by $v_s(q, \theta_s) = -\theta_s q$, so that θ_s is her marginal cost of production. Each agent's type takes one of two values from $\Theta_i = \{\underline{\theta}_i, \bar{\theta}_i\}$ with equal probability. We assume $0 \leq \underline{\theta}_s < \underline{\theta}_b < \bar{\theta}_s < \bar{\theta}_b$, so that (maximal) production is optimal except if the valuation is low

[11] See Benjamin (2014) for a similar distinction in a model of outcome-based social preferences.

and the cost is high. An SCF f specifies 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 $t_s^f(\theta_b, \theta_s)$. It is materially Pareto-efficient if and only if

$$q^f(\theta_b, \theta_s) = \begin{cases} 0 & \text{if} \quad (\theta_b, \theta_s) = (\underline{\theta}_b, \bar{\theta}_s), \\ 1 & \text{if} \quad (\theta_b, \theta_s) \neq (\underline{\theta}_b, \bar{\theta}_s), \end{cases} \tag{4}$$

and $t_s^f(\theta_b, \theta_s) = -t_b^f(\theta_b, \theta_s)$ for all $(\theta_b, \theta_s) \in \Theta$. For particular parameter constellations, e.g.

$$\underline{\theta}_s = 0, \;\; \underline{\theta}_b = 20, \;\; \bar{\theta}_s = 80, \;\; \bar{\theta}_b = 100, \tag{5}$$

this setup gives rise to a discrete-type version of the famous impossibility result by Myerson and Satterthwaite (1983): There is no SCF which is materially Pareto-efficient and satisfies both BIC and PC.

In this case, a mechanism design problem of interest is to choose an SCF f that minimizes $\mathbb{E}[t_b^f(\tilde{\theta}) + t_s^f(\tilde{\theta})]$ subject to the constraints that f has to satisfy BIC, PC, and trade has to be surplus-maximizing, i.e., q^f has to satisfy (4), but the transfers do not have to be budget-balanced. Myerson and Satterthwaite (1983) study this problem under the assumption that types are drawn from compact intervals. The solution to the problem provides a measure of how severe the impossibility result is: It gives the minimal subsidy that is required in order to make efficient trade compatible with the BIC and PC constraints. For our parameter constellation in (5), a solution f^* is given in Table 1, which provides the triple $(q^{f^*}, t_s^{f^*}, t_b^{f^*})$ for each possible type profile. Trade takes place whenever efficient, at prices 75, 50, or 25, depending on marginal cost and marginal valuation. These prices are chosen so as to guarantee BIC. The incentive-compatibility constraint (1) is binding for type $\bar{\theta}_b$ of the buyer and for type $\underline{\theta}_s$ of the seller. Respecting PC now requires a lump sum subsidy of 5/2 to be paid to each agent. Below, we will use f^* to illustrate that an SCF may be BIC but fail to be (strongly) implementable in BNFE, i.e., to show that mechanisms which are designed for selfish agents may fail to be robust to the introduction of (arbitrarily small) intention-based concerns.

	$\underline{\theta}_s$	$\bar{\theta}_s$
$\underline{\theta}_b$	$(1, 5/2 + 25, 5/2 - 25)$	$(0, 5/2, 5/2)$
$\bar{\theta}_b$	$(1, 5/2 + 50, 5/2 - 50)$	$(1, 5/2 + 75, 5/2 - 75)$

Table 1: Minimal Subsidy SCF f^*

Another SCF of interest is the one which is materially Pareto-efficient and splits the gains from trade equally between the buyer and the seller. It is denoted f^{**} and given in Table 2 for general parameter configurations. Since the transfers of f^{**} are budget-balanced, Table 2 provides only the pair $(q^{f^{**}}, t_s^{f^{**}})$ for each type profile. The resulting payoffs

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

are always non-negative, so that PC is satisfied. It is easily verified, however, that f^{**} is not BIC: It gives a high type buyer an incentive to understate her willingness to pay, and a low type

seller an incentive to exaggerate her cost. Below, we will use f^{**} to illustrate that an SCF may fail to be BIC but still be (weakly) implementable in BNFE.

	$\underline{\theta}_s$	$\bar{\theta}_s$
$\underline{\theta}_b$	$(1, (\underline{\theta}_b + \underline{\theta}_s)/2)$	$(0, 0)$
$\bar{\theta}_b$	$(1, (\bar{\theta}_b + \underline{\theta}_s)/2)$	$(1, (\bar{\theta}_b + \bar{\theta}_s)/2)$

Table 2: Equal Split SCF f^{**}

4 Strongly Implementable Social Choice Functions

4.1 Example

To motivate our analysis of strongly implementable social choice functions, we begin with the example of an SCF that can be implemented if agents are selfish but not if there are arbitrarily small concerns for reciprocity (provided that the kindness bound $\bar{\kappa}$ is not too stringent). Consider the bilateral trade example with parameters as given in (5). We know that the SCF f^* solves the minimal subsidy problem, so truth-telling $s^T = (s_b^T, s_s^T)$ is a BNE in the direct mechanism. The following observation asserts that truth-telling is not a BNFE as soon as at least one agent puts a positive weight on kindness.

Observation 1. *Consider the direct mechanism for f^* in the bilateral trade example, assuming (5) and $\bar{\kappa} > 5/2$. For every y with $y_b > 0$ and/or $y_s > 0$, the strategy profile s^T is not a BNFE.*

The proof of this observation (and of all other observations) can be found in Appendix D. It rests on two arguments. First, the structure of binding incentive constraints in f^* implies that the buyer obtains the same material payoff from truth-telling as from always declaring a low willingness to pay. This understatement reduces the seller's material payoff, however, and thus gives the buyer a costless option to punish the seller. Second, the seller's kindness in a hypothetical truth-telling equilibrium is negative: truth-telling maximizes her own payoff, while she could make the buyer better off by always announcing a low cost. The buyer therefore benefits from reducing the seller's payoff and deviates from truth-telling whenever $y_b > 0$ (and $\bar{\kappa}$ is large enough for her to still experience this payoff reduction). The symmetric reasoning applies to the seller.

The example illustrates a more general insight. The combination of two properties, both of which are satisfied by many optimal mechanisms for selfish agents, can make a mechanism vulnerable to intention-based reciprocity. First, binding incentive constraints provide costless opportunities to manipulate the other agents' payoffs. Second, BIC implies that truthful agents act selfish and therefore unkind. As a consequence, a reciprocal agent wants to use the manipulation opportunities to retaliate the other agents' unkindness.[12] The results that follow show that these situations can be avoided if an appropriate mechanism is chosen.

[12] Bierbrauer et al. (2015) generalize this argument to an even larger class of social preference models, and they discuss bilateral trade mechanisms and optimal taxation mechanisms. Their theoretical and experimental findings confirm the conjecture by Baliga and Sjöström (2011) that mechanisms in which agents 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 or kindness."

4.2 Possibility Results

We will provide sufficient conditions for the strong implementability of social choice functions in BNFE. Our analysis makes use of a measure of payoff interdependence among the agents. Given an SCF f, we define

$$\Delta_i = \max_{\theta_j \in \Theta_j} \mathbb{E}[v_i(q_i^f(\tilde{\theta}_i, \theta_j), \tilde{\theta}_i) + t_i^f(\tilde{\theta}_i, \theta_j)] - \min_{\theta_j \in \Theta_j} \mathbb{E}[v_i(q_i^f(\tilde{\theta}_i, \theta_j), \tilde{\theta}_i) + t_i^f(\tilde{\theta}_i, \theta_j)], \tag{6}$$

so that Δ_i measures the maximal impact that varying j's type has on i's expected payoff. If $\Delta_i = 0$, then the SCF f insures agent i against the randomness in agent j's type. Accordingly, we will say that f has the insurance property in the particular case where $\Delta_1 = \Delta_2 = 0$.[13]

Theorem 1. *If f is BIC and has the insurance property, it is strongly implementable in BNFE.*

In the proof, we consider the direct mechanism and verify that truth-telling is a BNFE for all $y \in \mathbb{R}^2_+$. We first show that the insurance property is equivalent to the following property: no agent can affect the other agent's expected material payoff by a unilateral deviation from truth-telling. In the hypothetical truth-telling equilibrium, kindness is therefore equal to zero, so that the agents focus only on their own material payoffs. If the given SCF is BIC, then the own payoff is maximized if the agents behave truthfully. Hence, truth-telling is in fact a BNFE.

The theorem raises the question how restrictive the insurance property is. Proposition 1 below shows that there exist materially Pareto-efficient SCFs that are both BIC and have the insurance property. Proposition 2 provides an extension to environments in which, in addition, participation constraints have to be respected, but budget-balance can be dispensed with.

We first consider a class of direct mechanisms which are known as expected externality mechanisms or AGV mechanisms, and which have been introduced by d'Aspremont and Gerard-Varet (1979) and Arrow (1979). An AGV mechanism is an SCF f with surplus-maximizing consumption levels (q_1^f, q_2^f) and transfers that are given by

$$t_i^f(\theta_i, \theta_j) = \mathbb{E}[v_j(q_j^f(\theta_i, \tilde{\theta}_j), \tilde{\theta}_j)] - \mathbb{E}[v_i(q_i^f(\tilde{\theta}_i, \theta_j), \tilde{\theta}_i)]$$

for all (θ_i, θ_j). These transfers achieve budget-balance and hence guarantee Pareto-efficiency. They also ensure that the AGV mechanism is BIC (see e.g. Mas-Colell et al., 1995, for a proof).

Proposition 1. *The AGV mechanism has the insurance property.*

The expected externality mechanism derives its name from the fact that each agent pays for the expected impact that her strategy choice has on the other agents' payoffs, assuming that the other agents tell the truth. If there are only two agents, each of them obtains the payment made by the other, which implies that a truth-telling agent is protected against changes of the other agent's strategy.[14] With more than two agents, the AGV satisfies the insurance property only under an additional symmetry condition, which we introduce and discuss in Appendix B.

[13] The literature on mechanism design with risk-averse or ambiguity-averse agents (e.g. Maskin and Riley, 1984; Bose et al., 2006; Bodoh-Creed, 2012) has explored various different insurance properties. As the following result shows, an insurance property is also useful for a characterization of economic outcomes that can be implemented if agents care about intentions.

[14] Mathevet (2010) states that the AGV "has no interdependencies between agents" (p. 414).

It is well-known that AGV mechanisms may not be admissible if participation constraints have to be respected. More generally, in many situations there does not exist any SCF which is Pareto-efficient and satisfies both BIC and PC. This generates an interest in second-best social choice functions, which satisfy BIC and PC but give up on the goal of achieving full Pareto-efficiency. They specify consumption levels that are not surplus-maximizing and/or abandon the requirement of budget-balance (as e.g. the SCF f^* in our bilateral trade example). An implication of the following proposition is that any such SCF can be modified so as to make sure that the insurance property holds.

Proposition 2. *Let f be an SCF that is BIC. Then there exists an SCF $\bar{f}$ with the following properties:*

(a) The consumption levels are the same as under f: $q_i^{\bar{f}}(\theta) = q_i^f(\theta)$ for $i = 1, 2$ and all $\theta \in \Theta$.

(b) The expected revenue is the same as under f: $\mathbb{E}[t_1^{\bar{f}}(\tilde{\theta}) + t_2^{\bar{f}}(\tilde{\theta})] = \mathbb{E}[t_1^f(\tilde{\theta}) + t_2^f(\tilde{\theta})]$.

(c) The interim payoff of every agent $i = 1, 2$ and type $\theta_i \in \Theta_i$ is the same as under f:

$$\mathbb{E}[v_i(q_i^{\bar{f}}(\theta_i, \tilde{\theta}_j), \theta_i) + t_i^{\bar{f}}(\theta_i, \tilde{\theta}_j)] = \mathbb{E}[v_i(q_i^f(\theta_i, \tilde{\theta}_j), \theta_i) + t_i^f(\theta_i, \tilde{\theta}_j)].$$

(d) $\bar{f}$ is BIC and has the insurance property.

The proof is constructive and shows that the following new transfer scheme guarantees the properties stated in the proposition:

$$t_i^{\bar{f}}(\theta_i, \theta_j) = \mathbb{E}[v_i(q_i^f(\theta_i, \tilde{\theta}_j), \theta_i) + t_i^f(\theta_i, \tilde{\theta}_j)] - v_i(q_i^f(\theta_i, \theta_j), \theta_i), \tag{7}$$

for all $(\theta_i, \theta_j) \in \Theta$. As has been shown by Bose et al. (2006) and Bodoh-Creed (2012), using transformation (7) is also useful in a model with ambiguity-averse agents. If the agents compute expected payoffs based on the "worst" prior, insurance provision can even increase expected revenues without affecting the agents' expected payoffs.

Proposition 2 can be viewed as a tool that transforms any SCF that is implementable under the assumption that all agents are selfish, into one that is behaviorally robust. It is particularly useful for problems with participation constraints, because all interim expected payoffs remain unchanged by property (*c*). Applications include the problem of partnership dissolution (Cramton et al., 1987), public-goods provision (Güth and Hellwig, 1986; Hellwig, 2003; Norman, 2004), the control of externalities (Rob, 1989), or auctions (Myerson, 1981; Bartling and Netzer, 2015).

That said, there are certain properties of the initial SCF that will not be preserved if this construction is applied. First, if the initial SCF f satisfies ex post budget balance, in the sense that $t_1^f(\theta) + t_2^f(\theta) = 0$ for all θ, we will typically not also have $t_1^{\bar{f}}(\theta) + t_2^{\bar{f}}(\theta) = 0$ for all θ. The two SCFs f and $\bar{f}$ have the same budgetary implications only in expectation. This may be a problem if the mechanism designer is unable to cover ex post deficits. Second, the transfers of $\bar{f}$ can depend on the type distribution p even though this was not the case for f. Relatedly, $\bar{f}$ is only Bayesian incentive-compatible even though f may have been incentive-compatible in dominant strategies. Bartling and Netzer (2015) apply Proposition 2 to a second-price auction. The second-price auction has a dominant strategy equilibrium but is not strongly implementable in our sense. By

contrast, the modified version of the second-price auction is strongly implementable but does not have dominant strategies. Dominant strategies ensure that a selfish agent's incentives to tell the truth are robust with respect to the agent's probabilistic beliefs about the types of the other agents. The insurance property ensures that an agent's incentives to tell the truth are robust with respect to the intensity of the agent's social preferences. Bartling and Netzer (2015) show experimentally that both auction formats achieve the same level of efficiency. This finding indicates that the two dimensions of robustness may be of equal importance for the performance of a mechanism.

4.3 Example Continued

We have shown in Section 4.1 that the SCF f^*, which minimizes the subsidy that is needed to achieve efficient trade, cannot be implemented in BNFE of the direct mechanism. We can now use Proposition 2 to construct an SCF $\bar{f}^*$ which is similar to f^* but can be strongly implemented in BNFE. Applying formula (7) we obtain $\bar{f}^*$ as given in Table 3. Trade takes place whenever efficient, at prices 60, 40, or 20, depending on marginal cost and marginal valuation. The subsidy now depends on the types and differs between the agents. The seller obtains a subsidy of 20 if both types are high or if both types are low, and a tax of 20 is collected from the buyer if costs are low and valuation is high. The expected net subsidy amounts to 5, exactly as for f^*. Proposition 2 in fact implies that $\bar{f}^*$ is an alternative solution to the second-best problem from Section 3.4, which additionally satisfies the insurance property.

	$\underline{\theta}_s$	$\bar{\theta}_s$
$\underline{\theta}_b$	$(1, 20+20, -20)$	$(0, 0, 0)$
$\bar{\theta}_b$	$(1, +40, -20-40)$	$(1, 20+60, -60)$

Table 3: Robust Minimal Subsidy SCF $\bar{f}^*$

4.4 Discussion

The proof of Theorem 1 exploits only one feature of Rabin (1993)'s model of social preferences: the agents are selfish when they lack the ability to influence the other's payoffs. This property of "selfishness in the absence of externalities" also holds in many models with outcome-based social preferences, such as altruism, spitefulness, or inequality aversion.[15] Thus, Theorem 1 provides a robust sufficient condition: The combination of Bayesian incentive-compatibility and the insurance property ensures implementability for a wide class of social preferences models that have been explored in the literature. This robustness property is particularly attractive for applications of mechanism design. Confronted with the empirically well-documented individual

[15] See Bierbrauer et al. (2015) for a formal definition of selfishness in the absence of externalities and for an investigation of the social preference models by Fehr and Schmidt (1999) and Falk and Fischbacher (2006). Similar observations, albeit not in mechanism design frameworks, have been made by Levine (1998), Fehr and Schmidt (1999), Bolton and Ockenfels (2000) or Segal and Sobel (2007). Dufwenberg et al. (2011) demonstrate the behavioral irrelevance of interdependent preferences in general equilibrium under a separability condition that is essentially equivalent to selfishness in the absence of externalities.

heterogeneity in social preferences (Fehr and Schmidt, 1999; Engelmann and Strobel, 2004; Falk et al., 2008; Dohmen et al., 2009), a designer will typically be uncertain about the most appropriate specification of preferences. The insurance property offers a way out of this problem.

5 Weakly Implementable Social Choice Functions

5.1 Example

We begin with an example that illustrates several conceptual issues that arise when the designer has precise information on the weights that kindness has in the agents' utility functions. Similar issues will reappear in the context of two-dimensional design in Section 6.

Consider again the bilateral trade example, for general parameters, not necessarily those given in (5). We argued before that the SCF f^{**}, which stipulates efficient trade and splits the gains from trade equally, is not BIC and hence not implementable in BNE. We first show that it is also not implementable in BNFE when the designer is restricted to using a direct mechanism.

Observation 2. *Consider the direct mechanism for f^{**} in the bilateral trade example. For every y_b and y_s, the truth-telling strategy profile s^T is not a BNFE.*

The logic is as follows: One can show that in a hypothetical truth-telling equilibrium both the buyer and the seller realize their equitable payoffs. This implies that all kindness terms are zero and the agents focus solely on their material payoffs. Lack of BIC then implies that truth-telling is not a BNFE. Efficient trade with an equal sharing of the surplus is thus out of reach in the direct mechanism, with or without intention-based social preferences.

Now consider a non-direct 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 $q^{g'}(m_b, m_s)$ and budget-balanced 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 4 gives the pair $(q^{g'}, t_s^{g'})$ for every possible profile of messages.

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

Table 4: Non-Direct Mechanism Φ'

The mechanism works like a direct mechanism for f^{**} as long as the message profile is 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 she gets an additional discount of δ_b whenever there is trade. 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 the price she receives is increased by δ_s whenever there is trade. Intuitively, announcing $\bar{\bar{\theta}}_s$ amounts to the claim that the cost is even higher than $\bar{\theta}_s$.

Agent i's set of strategies in mechanism Φ' is $S_i' = M_i' \times M_i'$. A generic element s_i' of S_i' is a pair 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. For both agents, the strategy set of the direct mechanism, $S_i = \Theta_i \times \Theta_i$, is a subset of the extended strategy set S_i'. The outcome of Φ' under the truth-telling strategy profile s^T is still the outcome stipulated by the SCF f^{**}. The following observation asserts that truth-telling is a BNFE for particular parameter constellations.

Observation 3. *Consider the non-direct mechanism Φ' for f^{**} in the bilateral trade example. Suppose $\bar{\kappa}$ is large and y_b, $y_s > 0$. Then, there exist numbers $\delta_b, \delta_s > 0$ so that s^T is a BNFE.*

To understand the logic of the argument, assume that $\bar{\kappa} = \infty$, so that the kindness bound can be safely ignored (the statement that $\bar{\kappa}$ must be large is made precise in Theorem 2 below). In the hypothetical truth-telling equilibrium, the buyer then chooses s_b in order to maximize

$$\Pi_b(s_b, s_s^T) + y_b \kappa_s(s^T) \Pi_s(s_b, s_s^T).$$

Truth-telling means that the seller did not insist on the very high price, i.e., she did not use an opportunity to enrich herself at the buyer's expense. Thus, in contrast to the direct mechanism where the kindness of truth-telling is zero, we now have $\kappa_s(s^T) = \delta_s/4$. If we set $\delta_s = 4/y_b$ then we obtain $\kappa_s(s^T) = 1/y_b$, and the buyer's problem becomes to maximize the sum of material payoffs $\Pi_b(s_b, s_s^T) + \Pi_s(s_b, s_s^T)$. Strategy s_b^T is a solution to this problem, because the outcome under truth-telling is the efficient SCF f^{**} which maximizes the sum of material payoffs for every $\theta \in \Theta$. Similarly, truth-telling is also a best response for the seller when $\delta_b = 4/y_s$.

Our construction is akin to a Vickrey-Clarke-Groves mechanism (Vickrey, 1961; Clarke, 1971; Groves, 1973) in that it aligns private and social interests. The key difference is that it is not based on a suitable choice of payments that the agents have to make in equilibrium, but on a suitable choice of payments that the agents refuse to make in equilibrium.

Observations 2 and 3 raise various conceptual issues. First, they show that the revelation principle does not hold. The social choice function f^{**} is not implementable in BNFE with a direct mechanism, but it is implementable with a specific indirect mechanism. Second, since f^{**} ensures non-negative material payoffs for both agents and types, in the equilibrium of the indirect mechanism the material participation constraints are satisfied. This raises the question to what extent it is possible to overcome the impossibility results for efficient outcomes that are obtained in models with selfish agents and participation constraints. Finally, the indirect mechanism implements f^{**} with specific levels of kindness. Is it possible to generate even larger levels of kindness while still implementing f^{**}? Can we even reach the upper bound $\bar{\kappa}$, which would imply that we achieve a maximum of material surplus and at the same time a maximum of kindness? The analysis that follows addresses these questions.

5.2 An Augmented Revelation Principle

The non-direct mechanism Φ' that is used to implement f^{**} in the previous section 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. The following lemma shows that if implementation of

an SCF in BNFE is possible at all, then it is also possible truthfully in the class of augmented revelation mechanisms. A mechanism is called an augmented revelation mechanism for f whenever $\Theta_i \subseteq M_i$ for $i = 1, 2$ and $g(m) = f(m)$ for all $m \in \Theta$, i.e., whenever the 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. The difference between truthful direct and augmented revelation mechanisms is the existence of unused actions in the latter.

Lemma 1. *Suppose a mechanism Φ implements an SCF f in BNFE. Then there exists an augmented revelation mechanism Φ' that truthfully implements f in BNFE.*

Augmented revelation mechanisms have first been introduced by Mookherjee and Reichelstein (1990), albeit for a different purpose. They characterize SCFs that can be implemented as the unique equilibrium outcome of some mechanism. Our proof of Lemma 1 in the appendix builds on their analysis.

5.3 A Possibility Result for Efficient SCFs

The following theorem is a generalization of Observation 3. It provides sufficient conditions for the weak implementability of materially Pareto-efficient social choice functions in BNFE. The following notation will make it possible to state the theorem in a concise way. For a given SCF f, define

$$Y^f = \{(y_1, y_2) \in \mathbb{R}^2_+ \mid y_i > 0 \text{ and } 1/y_i \leq \bar{\kappa} - \Delta_i \text{ for both } i = 1, 2\},$$

where Δ_i is given by (6). The set Y^f of reciprocity weights is non-empty if and only if $\bar{\kappa} > \Delta_i$ for both agents, i.e., the kindness bound $\bar{\kappa}$ has to be large enough compared to the measure of payoff interdependence Δ_i. If $\bar{\kappa} = \infty$, then Y^f contains all pairs of strictly positive reciprocity weights.

Theorem 2. *If f is materially Pareto-efficient, it is weakly implementable in BNFE on Y^f.*

In the proof, we start from a direct mechanism for f and introduce additional messages. Specifically, we work with a mechanism in which agent i's message set is $M_i = \Theta_i \times \{0, 1\}$, so that a message consists of a type report and a decision whether or not to "press a button" (see also Netzer and Volk, 2014, for an application of such mechanisms). The outcome of the mechanism is the one stipulated by f. In addition, if agent i presses the button, this triggers an additional (possibly negative) payment from agent i to agent j. These payments are used to manipulate the kindness associated to truth-telling, and we calibrate them to generate a degree of kindness that effectively turns each agent's best-response problem into a problem of surplus-maximization, as already illustrated by Observation 3. This can require increasing or decreasing the kindness of truth-telling in the direct mechanism, so that the redistribution triggered by i's button might have to go in either direction. Ultimately, since the SCF to be implemented is materially Pareto-efficient, truth-telling is a solution to the surplus-maximization problem, and the buttons remain unpressed.

A difficulty in the proof of Theorem 2 arises from the kindness bound $\bar{\kappa}$. The crucial step for the alignment of individual incentives with the objective of surplus-maxmization is that we can generate kindness equal to $\kappa_j(s^T) = 1/y_i$. The requirement $1/y_i \leq \bar{\kappa}$ is a necessary condition for this to be possible. The condition $1/y_i \leq \bar{\kappa} - \Delta_i$ in the definition of Y^f is even more stringent. The larger is Δ_i, the larger need to be the kindness bound $\bar{\kappa}$ and/or the reciprocity weight y_i in order to guarantee implementability of f. Intuitively, while no deviation of agent j can increase the sum of payoffs over and above truth-telling, some strategy of j might increase j's own payoff and decrease i's payoff into the region where $\kappa_j = -\bar{\kappa}$ holds. Agent j no longer internalizes all payoff consequences of such a deviation. If Δ_i is sufficiently small relative to $\bar{\kappa}$, this possibility can be excluded. If $\bar{\kappa} = \infty$, i.e., if there is no a priori bound on the intensity of kindness sensations, then every materially Pareto-efficient SCF can be implemented as soon as y_1 and y_2 are strictly positive, i.e., as soon as both agents show some concern for reciprocity.

Theorem 2 also enables us to address the question whether participation constraints are an impediment for the implementation of materially efficient SCFs in BNFE. With intention-based social preferences, participation constraints could be formulated in two different ways. First, we may impose non-negativity constraints on ex interim material payoffs, i.e., we may require that PC holds. This approach allows for a clean comparison with the impossibility results in the existing literature. Second, we may model participation as a decision that must be optimal based on the entire utility function, including psychological payoffs. This can be captured by equipping the agents with veto rights, i.e., with the opportunity to opt out of the mechanism ex interim, and by studying the conditions under which they would make use of this option. We will show in the following that voluntary participation can be assured for either criterion.

Classical papers such as Myerson and Satterthwaite (1983) and Mailath and Postlewaite (1990) have noted that, when we consider an SCF that is materially Pareto-efficient and BIC, then PC must be violated for some types of some agents. By Theorem 2, however, BIC is no longer a constraint. For instance, with sufficiently strong concerns for reciprocity, we can implement efficient SCFs that give both agents an equal share of the material surplus. More generally, with the solution concept of weak implementability in BNFE, we are able to achieve SCFs that violate BIC but are surplus-maximizing and satisfy PC.

Now consider the possibility to capture voluntary participation by means of veto rights. Consider a direct mechanism with veto rights, where $M_i^v = \Theta_i \cup \{v\}$ is the message set. The mechanism stipulates some status quo allocation $a^v \in A$ if any one agent sends the veto v. We can now add buttons to M_i^v in exactly the same way as in the proof of Theorem 2 and align individual interests with the objective of surplus-maximization. Since the social choice function under consideration is a surplus-maximizing one, the veto rights and the buttons remain unused in equilibrium. Hence, all types of both agents voluntarily decide to participate in the mechanism. The only modification required to extend the proof of Theorem 2 is that the measure of payoff interdependence Δ_i needs to be replaced by the (weakly larger) measure Δ_i^v that also takes account of the payoff interdependence due to the veto rights:

$$\Delta_i^v = \max_{m_j \in M_j^v} \mathbb{E}[v_i(q_i^g(\tilde{\theta}_i, m_j), \tilde{\theta}_i) + t_i^g(\tilde{\theta}_i, m_j)] - \min_{m_j \in M_j^v} \mathbb{E}[v_i(q_i^g(\tilde{\theta}_i, m_j), \tilde{\theta}_i) + t_i^g(\tilde{\theta}_i, m_j)].$$

5.4 Implementation with Maximal Kindness

When Rabin (1993) introduced his model of intention-based social preferences, he argued 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). In the following, we provide a formalization of this idea. We fix an SCF f that is implementable in BNFE and look for a mechanism that implements f with maximal psychological utility. The following proposition asserts that any SCF which satisfies the prerequisites of either Theorem 1 or 2 can in fact be implemented so that both agents' kindness reaches the upper bound $\bar{\kappa}$.

Proposition 3. *Suppose $\bar{\kappa} < \infty$ and $y_i > 0$ for both $i = 1, 2$. Let f be an SCF for which one of the following two conditions holds:*

(a) f is BIC and has the insurance property, or

(b) f is materially Pareto-efficient and $y \in Y^f$.

Then, there exists a mechanism that implements f in a BNFE s with $\kappa_1(s) = \kappa_2(s) = \bar{\kappa}$.

The proof of Proposition 3 again uses mechanisms with a button for each agent. The payments that are triggered if a button is pressed now have to be calibrated such that the resulting kindness equals the upper bound $\bar{\kappa}$. Showing that no agent wants to deviate from truth-telling is then more intricate than in the proof of Theorem 2. In particular, we need to allow for the possibility that the payment made by j is larger than the payment received by i, i.e., we have to allow for free disposal off the equilibrium. To see this, consider case (a) of Proposition 3 and suppose that $y_i\bar{\kappa} < 1$. Then, even with maximal kindness, agent i still places a larger weight on the own than on j's payoff, and therefore would press a button that triggers a budget-balanced transfer from j to i. Agent i will refrain from pressing the button only if the payment he receives is sufficiently smaller than the loss inflicted on j. Similar issues arise in case (b), where the loss that i can impose on j by pressing the button may have to be very large. A double-deviation, where agent i presses the button and announces the type non-truthfully, may then reduce j's payoff into the region where the lower kindness bound binds. Again, i will refrain from doing so only if the own gain is smaller than the loss to j.

Proposition 3 implies that the objectives of generating kindness and material efficiency are not in conflict with each other. It allows us to first fix an SCF f that is materially Pareto-efficient and for which (a) or (b) in Proposition 3 holds. We can then implement f in a BNFE s^* of a mechanism Φ such that $\kappa_1(s^*) = \kappa_2(s^*) = \bar{\kappa}$ holds. Such a mechanism-equilibrium pair is in fact utility Pareto-efficient, in the sense that there cannot be any other mechanism-equilibrium pair (Φ', s') that yields a strictly larger utility for one agent without giving a strictly smaller utility to the other agent.[16]

[16] With procedural preferences as in our analysis, utility efficiency is a property of mechanism-equilibrium pairs rather than social choice functions, because the efficiency of an SCF is not independent from the mechanism that implements it. The observation that the efficiency of an outcome is not separable from the game form has also been made by Ruffle (1999) in the context of psychological gift-giving games.

5.5 Discussion

The arguments that we used in order to prove Theorem 2 and Proposition 3 exploit the menu-dependence of the agents' preferences. Whether agent i interprets the behavior of agent j as kind or as unkind depends not only on what j does in equilibrium, but also on what she could have done instead. A mechanism designer can take advantage of this by a specification of message sets that lets the desired behavior appear well-intentioned. While the details of our constructions are calibrated to the specific features of the model of intention-based social preferences by Rabin (1993), alternative formulations of menu-dependent preferences would not upset the basic logic of this argument.

Our proofs rely on mechanisms with a button, which are of course an artificial construction. A first issue with such mechanisms is that kindness is generated by actions which are used with probability zero in equilibrium. On the one hand, one may wonder whether kindness can really be generated by a mere possibility of behavior that is literally never observed. On the other hand, these mechanisms may appear vulnerable to the existence of selfish agents who would prefer to enrich themselves by pressing the buttons. We deal with this issue in the following Section 6. There we allow for the possibility that agents are selfish and, moreover, we assume that they privately observe whether they are selfish or not. We will show by means of an example that if these selfish types occur not too often – so that unused actions become rarely used actions – augmented mechanisms can still achieve approximately efficient outcomes.

A second issue is the plausibility of button mechanisms for real-world applications. First and foremost, they should be interpreted as a tool for the characterization of incentive-feasible outcomes, in the same way in which direct or augmented mechanisms are typically interpreted in the literature: They provide an upper bound for the outcomes that are achievable. That said, their basic logic has a broader appeal. These mechanisms give agents additional degrees of freedom and thereby generate additional opportunities to express kindness. This logic can be related to mechanisms which are empirically plausible. For instance, Herold (2010) considers an incomplete contracting relationship where one party refrains from including provisions against misbehavior of the other party into the contract, for fear of signalling a lack of trust. Once such an incomplete contract is given and a situation arises in which the contract makes no provisions, not exploiting the incompleteness by taking an opportunistic action is akin to not pressing the button in our augmented mechanism. As another example, consider the augmented mechanism for our bilateral trade application in Table 4. Here, the designer allows for announcements of types that could be proven to be impossible. By not using all available information to convict an agent of lying, the mechanism gives every party the chance to show that she makes an effort in order to meet the needs of the other party, and this makes it possible to reach a mutually beneficial outcome. As yet another example, consider a committee that has to decide whether or not to replace a status quo by some alternative a. Suppose that two different mechanisms may be used, *simple majority voting* or *majority voting with veto rights*. Under simple majority voting, the outcome is a if and only if a majority votes for a. By contrast, under majority voting with veto rights, every committee member has the right to insist on the status quo, so that the outcome is a if and only if a majority votes for a and if no committee member exercises his veto power. At first glance, one might think that the mechanism with veto rights makes it more

difficult to move away from the status quo, which is a problem if a is the efficient outcome.[17] However, majority voting with veto rights gives each committee member the possibility to say: "I don't like a, but I refrain from imposing my preferences on the whole group. So, if a majority is in favor of a, I am willing to accept this outcome." If every member acts in this way, then the outcome will be the same as under simple majority voting, and, in addition, the unused veto rights will generate a level of kindness that could not be reached if simple majority voting was applied.[18]

6 Two-Dimensional Private Information

6.1 The Model

We now consider an information structure where the agents do not only have private information about their material payoffs, but also about the weight y_i that kindness sensations have in their utility function. This creates a problem of multi-dimensional mechanism design, because we neither assume that preferences are partially known (as in our treatment of weakly implementable social choice functions) nor forgo the possibility to elicit the intensity of social preferences (as in our analysis of strongly implementable social choice functions). As will become clear, however, our previous treatment of weak and strong implementability proves very helpful for the more demanding problem with private information on both y_i and θ_i.

We assume that each agent i's reciprocity type y_i is the privately observed realization of a random variable $\tilde{y}_i$ that takes values in a finite set $Y_i \subseteq \mathbb{R}_+$. We also write $\tilde{y} = (\tilde{y}_1, \tilde{y}_2)$ and denote its realizations by $y = (y_1, y_2) \in Y = Y_1 \times Y_2$. We assume that each $\tilde{y}_i$ is distributed independently from the material payoff variable $\tilde{\theta}_i$ and independently across agents, following a probability distribution ρ_i. We let $\underline{y}_i = \min Y_i$ and $\bar{y}_i = \max Y_i$ denote agent i's smallest and largest possible reciprocity type, and we also write $\underline{y} = (\underline{y}_1, \underline{y}_2)$ and $\bar{y} = (\bar{y}_1, \bar{y}_2)$. If $\underline{y}_i = 0$ then agent i might be selfish.

Consider a mechanism $\Phi = [M_1, M_2, g]$. A strategy for agent i in this mechanism is a function from types $Y_i \times \Theta_i$ to messages M_i. For ease of comparison to our earlier analysis, we find it useful to represent such a strategy by a collection of functions $s_i = (s_{i,y_i})_{y_i \in Y_i}$, with $s_{i,y_i} : \Theta_i \to M_i$ for each $y_i \in Y_i$. Thus, we think of each reciprocity type as having a separate strategy that maps payoff types into messages. We continue to denote by S_i the set of these functions from Θ_i to M_i. As before, upper indices b and bb indicate first- and second-order beliefs. For instance, $s_i^b = (s_{i,y_j}^b)_{y_j \in Y_j}$ is agent i's belief about j's strategy s_j, where $s_{i,y_j}^b \in S_j$ denotes i's belief about the behavior of reciprocity type y_j. To interpret our results, we will often find it helpful to assume that agent i first observes her reciprocity type y_i and then chooses a strategy in S_i. We can then compute expected payoffs and expected utility conditional on y_i. These conditional expectations resemble the expressions in the preceding sections.

Conditional on y_i, agent i's strategy s_{i,y_i} yields expected material payoffs for agent $k = 1, 2$

[17] This is an important theme of the literature on mechanism design that employs the solution concept of BNE as opposed to BNFE. According to the view of this literature, insisting on voluntary participation is, if anything, bad, because it may render the achievement of efficient allocations impossible.

[18] A detailed formal analysis of this example can be found in an earlier version of this paper that is available upon request.

given by

$$\bar{\Pi}_k(s_{i,y_i}, s_i^b) = \mathbb{E}[\Pi_k(s_{i,y_i}, s_{i,\tilde{y}_j}^b)] = \mathbb{E}[v_k(q_k^g(s_{i,y_i}(\tilde{\theta}_i), s_{i,\tilde{y}_j}^b(\tilde{\theta}_j)), \tilde{\theta}_k) + t_k^g(s_{i,y_i}(\tilde{\theta}_i), s_{i,\tilde{y}_j}^b(\tilde{\theta}_j))],$$

where Π_k is the expected material payoff as defined before, and the expression $\bar{\Pi}_k$ reflects that agent i now also averages over the reciprocity types of j. Observe that reciprocity affects material payoffs only to the extent that different reciprocity types behave differently. This implies that the Pareto-efficient subset $E_i(s_i^b) \subseteq S_i$ and the equitable payoff $\Pi_j^e(s_i^b)$, now defined based on the payoffs $\bar{\Pi}_k$, are independent of y_i. The kindness of reciprocity type y_i of agent i is given by

$$\kappa_i(s_{i,y_i}, s_i^b) = h(\bar{\Pi}_j(s_{i,y_i}, s_i^b) - \Pi_j^e(s_i^b)).$$

When forming a belief about the kindness intended by j, agent i again averages over the different realizations of $\tilde{y}_j$. Formally, i's belief about j's kindness is

$$\bar{\kappa}_j(s_i^b, s_i^{bb}) = \mathbb{E}[\kappa_j(s_{i,\tilde{y}_j}^b, s_i^{bb})].$$

The expected utility of agent i with reciprocity type y_i is then given by

$$U_{i,y_i}(s_{i,y_i}, s_i^b, s_i^{bb}) = \bar{\Pi}_i(s_{i,y_i}, s_i^b) + y_i\, \kappa_i(s_{i,y_i}, s_i^b)\, \bar{\kappa}_j(s_i^b, s_i^{bb}).$$

Definition 4. *A BNFE is a strategy profile $s^* = (s_1^*, s_2^*)$ such that, for both $i = 1, 2$,*

(a) $s_{i,y_i}^* \in \arg\max_{s_{i,y_i} \in S_i} U_{i,y_i}(s_{i,y_i}, s_i^b, s_i^{bb})$, *for all* $y_i \in Y_i$,

(b) $s_i^b = s_j^*$, *and*

(c) $s_i^{bb} = s_i^*$.

An SCF $f : Y \times \Theta \to A$ assigns an allocation to each type profile (y, θ). We will be interested in SCFs that are implementable in BNFE, i.e., for which there exists a mechanism with a BNFE s^* such that $g(s_{1,y_1}^*(\theta_1), s_{2,y_2}^*(\theta_2)) = f(y, \theta)$ for all $(y, \theta) \in Y \times \Theta$. We distinguish SCFs according to whether or not they actually condition on the profile of reciprocity types y. We say that f is y-independent if $f(y, \theta) = f(y', \theta)$ for all $y, y' \in Y$ and $\theta \in \Theta$. Since the reciprocity types y are not materially payoff-relevant, Pareto-efficiency can be achieved within the class of y-independent SCFs. We investigate y-independent SCFs in Sections 6.2 and 6.3, and we discuss y-dependent SCFs in Section 6.4.

6.2 The Insurance Property Revisited

We first adapt the definitions of BIC and the insurance property to the setting with two-dimensional private information. Let f be a y-independent SCF. We can then define an SCF $\hat{f} : \Theta \to A$ as before, by $\hat{f}(\theta) = f(y, \theta)$ for an arbitrary $y \in Y$. In the following we will say that f is BIC if $\hat{f}$ is BIC. We also define the payoff interdependence Δ_i, the insurance property, and material Pareto-efficiency of a y-independent SCF f based on the respective properties of $\hat{f}$.

Theorem 1 has established that BIC and the insurance property are jointly sufficient for strong implementability. We have argued before that a virtue of the insurance property is its

applicability to situations with multi-dimensional private information. It is straightforward to verify this claim in the present setting, i.e., a y-independent SCF f that is BIC and has the insurance property is implementable in BNFE with private information about y.[19]

The following proposition establishes that BIC is also necessary for implementability, provided that there is a positive probability that the agents are selfish. Thus, a mere possibility of reciprocal behavior does not enlarge the set of implementable SCFs relative to a model in which all agents are selfish with probability one. In the context of the bilateral trade example, for instance, the proposition implies that the SCF f^{**} which splits the gains from trade equally cannot be implemented in the given setting.

Proposition 4. *If $\underline{y} = (0,0)$, a y-independent SCF is implementable in BNFE only if it is BIC.*

The next result reveals that the insurance property is also necessary for implementability, provided that both selfish types and types with a sufficiently strong weight on kindness sensations are possible.

Proposition 5. *Suppose $\underline{y} = (0,0)$, and let f be a y-independent SCF with $\Delta_i > 0$ for both $i = 1,2$. Then there exist numbers k, x_1, x_2 so that f is not implementable in BNFE when $\bar{\kappa} \geq k$ and $\bar{y}_i \geq x_i$ for at least one $i = 1,2$.*

The proof of the proposition involves various observations. First, if f is implementable at all, then it is also implementable as the truth-telling equilibrium of the direct mechanism, i.e., unused actions do not help in the presence of selfish types. In fact, if implementation is possible, then it is also possible in the "very direct" mechanism, where the agents communicate only their material payoff type. Second, implementation in the very direct mechanism requires that, for a given material payoff type θ_i, all reciprocity types of agent i behave in the same way. Moreover, since the selfish type is among them, all reciprocity types choose to behave in a selfish way. Equilibrium kindness is therefore negative. Third, with negative kindness, agents with sufficiently large values of y_i are willing to deviate from truth-telling to make the other agent worse off. Since the insurance property is violated, such a deviation is indeed available. As a consequence, f can only be implemented if it has the insurance property.[20]

The observation that equilibrium kindness cannot be positive if selfish types are around implies that the insurance property is also desirable from a welfare perspective. Given that zero is an upper bound on equilibrium kindness, if the insurance property holds then this upper bound is actually reached.

[19] Consider the "very direct" mechanism $\Phi = [\Theta_1, \Theta_2, \hat{f}]$ for f, where the agents are asked about their material payoff type only. It is easy to see that there is a BNFE s^* with $s^*_{i,y_i}(\theta_i) = \theta_i$ for $i = 1,2$ and all $(y_i, \theta_i) \in Y_i \times \Theta_i$. This follows because all kindness terms take a value of 0, so the agents are left with the problem to maximize their own expected payoff. BIC ensures that this problem is solved by revealing the payoff type truthfully.

[20] To be precise, Proposition 5 shows that implementation of f is impossible if $\Delta_i > 0$ holds for both agents. Hence a necessary condition for implementability is that $\Delta_i = 0$ for at least one agent, while the insurance property is slightly stronger and requires $\Delta_i = 0$ for both $i = 1,2$. This obviously makes no difference with a symmetric social choice function. Moreover, the sufficient condition in Theorem 1 could also be weakened so that it requires $\Delta_i = 0$ only for at least one agent. Since the two agents' kindness terms enter the utility functions multiplicatively, if kindness is zero for one agent, kindness disappears from both agents' utility functions and both agents' equilibrium behavior is selfish. We use the slightly stronger notion of the insurance property mainly because it implies robustness beyond the intention-based model, as discussed in Section 4.4.

6.3 Theorem 2 Revisited

After dealing with the case in which selfish types are possible ($\underline{y}_i = 0$), we now address the complementary case in which the agents are known to put strictly positive weight on reciprocal kindness ($\underline{y}_i > 0$). In this case, Theorem 2 can be extended to the environment with two-dimensional private information.

Proposition 6. *Suppose $\bar{\kappa} < \infty$, and let f be a y-independent and materially Pareto-efficient SCF. If $\underline{y} \in Y^f$, then f is implementable in BNFE.*

Theorem 2 was based on the construction of an augmented mechanism so that, in equilibrium, $y_i \, \kappa_j(s^T) = 1$ for all i, i.e., the agents assign equal weights to their own and the other agent's material payoff. This approach is no longer feasible if y_i is not a known parameter. Therefore, to prove Proposition 6, we use a similar approach as for the proof of Proposition 3. Specifically, we construct an augmented mechanism in which every reciprocity type y_i of every agent i exhibits the maximum level of kindness $\bar{\kappa}$. Then, since $\underline{y} \in Y^f$, we know that every reciprocity type of every agent assigns at least a weight of $y_i \bar{\kappa} \geq \underline{y}_i \bar{\kappa} \geq 1$ to the other agent's material payoff. As in the proof of Proposition 3, all deviations from truth-telling can then be shown to become unattractive.[21] The construction not only ensures the implementability of the given SCF, but also that it is implemented with maximal kindness. Again, this demonstrates that the objectives of material efficiency and of kindness among the participants of a mechanism are not in conflict with each.

6.4 Unused Actions Revisited

We finally investigate the case of y-dependent social choice functions. Why would a mechanism designer be interested in implementing a y-dependent SCF? In terms of material efficiency, there is nothing to gain by conditioning on y. We will show in the following, however, that a designer may favor the implementation of a, possibly inefficient, y-dependent SCF because this gives her additional degrees of freedom for incentive provision. In particular, the example below shows how an SCF that satisfies the conditions of Theorem 2 can be approximated by a y-dependent SCF in the environment with private information on reciprocity types and the possibility of selfish behavior.

Specifically, consider once more the bilateral trade problem and assume that $Y_i = \{0, \bar{y}\}$ with $\bar{y} > 0$ and $\rho_i(0) = \epsilon > 0$ for both $i = b, s$. We start from the y-independent SCF f^{**} that is materially efficient and splits the gains from trade equally. As we have observed previously, f^{**} is not implementable due to the possibility of selfish types. On the other hand, according to Theorem 2 we could implement f^{**} in an augmented mechanism with unused actions if it was common knowledge that $y_b = y_s = \bar{y}$. Now consider instead the budget-balanced y-dependent SCF f^{***} for which $(q^{f^{***}}, t_s^{f^{***}})$ is given in Table 5. Whenever both agents have the positive reciprocity type, then f^{***} coincides with f^{**}. When at least one agent is selfish,

[21] As for Proposition 3, off-equilibrium buget-balance cannot generally be guaranteed with this construction. The difficulty is due to the kindness bound: "pressing the button" of the augmented mechanism might decrease j's payoff into a region where the lower bound becomes binding, so that i no longer internalizes all consequences of this behavior. This problem can be avoided when i's own benefit from pressing the button is sufficiently small (yet positive).

then f^{***} prescribes the allocations that were associated to the unused actions in the augmented mechanism Φ' in Table 4. Hence the direct mechanism (with $M_i = Y_i \times \Theta_i$) for the y-dependent SCF f^{***} is an analog to the previous non-direct mechanism Φ', when selfish agents are expected to make use of the previously unused actions.

(y_s, θ_s) / (y_b, θ_b)	$(\bar{y}, \underline{\theta}_s)$	$(\bar{y}, \bar{\theta}_s)$	$(0, \underline{\theta}_s)$	$(0, \bar{\theta}_s)$
$(0, \underline{\theta}_b)$	$(1, (\underline{\theta}_b + \underline{\theta}_s)/2 - \delta_b)$	$(0, 0)$	$(0, 0)$	$(0, 0)$
$(0, \bar{\theta}_b)$	$(1, (\underline{\theta}_b + \underline{\theta}_s)/2 - \delta_b)$	$(0, 0)$	$(0, 0)$	$(0, 0)$
$(\bar{y}, \underline{\theta}_b)$	$(1, (\underline{\theta}_b + \underline{\theta}_s)/2)$	$(0, 0)$	$(0, 0)$	$(0, 0)$
$(\bar{y}, \bar{\theta}_b)$	$(1, (\bar{\theta}_b + \underline{\theta}_s)/2)$	$(1, (\bar{\theta}_b + \bar{\theta}_s)/2)$	$(1, (\bar{\theta}_b + \bar{\theta}_s)/2 + \delta_s)$	$(1, (\bar{\theta}_b + \bar{\theta}_s)/2 + \delta_s)$

Table 5: y-dependent SCF f^{***}

Observation 4. *Consider the direct mechanism for f^{***} in the bilateral trade example. Suppose $\bar{\kappa}$ is large. Then, for ϵ small enough, there exist numbers $\delta_b, \delta_s > 0$ so that s^T is a BNFE.*

If the probability ϵ that agents are selfish is small, then the actions which remained unused under the assumptions of Theorem 2 now become actions that are "rarely used." On the upside, kindness among reciprocal agents is generated by the fact that they refrain from claiming the allocations that the egoists obtain. The allocation of the truthful direct mechanism for f^{***} converges to the y-independent efficient allocation f^{**} as $\epsilon \to 0$ (it is also shown in the proof that the rarely triggered redistributive payments δ_b and δ_s converge to the values derived in the proof of Observation 3). We can thus approximate our weak implementability result as the limit case of an environment with privately observed reciprocity types by letting the probability that agents are selfish go to zero.

6.5 Discussion

A first main insight from the model with two-dimensional private information concerns the scope of achievable outcomes. If the designer thinks that both selfish agents and strongly reciprocal agents possible, then the combination of BIC and the insurance property is necessary and sufficient for implementability. This accentuates the importance of strongly implementable SCFs as introduced in Section 4. By contrast, if the designer can be sure that the agents are not selfish, then our analysis of weakly implementable SCFs from Section 5 has a natural extension to the case with two-dimensional private information. Hence what we have shown is that the distinction between the strong and the weak notion of implementability can be traced back to the question whether or not there is a possibility that the participants of a mechanism act selfishly.

A second main insight concerns the interpretation of unused actions. We have shown in our bilateral trade application that unused actions can be thought of as the limit of rarely used actions. As long as the "buttons" of the augmented mechanism are not pressed too often, or,

equivalently, as long as selfish behavior is observed only rarely, our possibility result for weak implementation can still be achieved approximately.

7 Extension: The Designer as a Player

So far we have assumed that the agents treat the mechanism as exogenous. However, they may think of the mechanism designer as a player, and their behavior may be affected by the intentions that they attribute to the designer's choice of the mechanism. For instance, they may have a desire to sabotage the mechanism if they believe that it was chosen with the intention to extract an excessive share of their rents. As an extension, we briefly explore this idea in a simplified framework. We show that the perception of the designer as a player may drastically reduce the set of implementable outcomes, even though the designer has a direct concern for the agents' well-being.

For any SCF f, denote by $\Pi_i(f) = \mathbb{E}[v_i(q_i^f(\tilde{\theta}), \tilde{\theta}_i) + t_i^f(\tilde{\theta})]$ the expected material payoff of agent i, and let $R(f) = \mathbb{E}[-(t_1^f(\tilde{\theta}) + t_2^f(\tilde{\theta}))]$ denote the expected budget surplus. We will assume that the mechanism designer maximizes

$$W(f) = H(\Pi_1(f), \Pi_2(f), R(f)),$$

where the welfare function H is strictly increasing in all three arguments. Hence the designer cares about the agents' material payoffs and about the revenue she can extract from the agents. For instance, we could think of two firms (the agents) that have been merged and are now governed by a common headquarter (the designer).[22] The task to be implemented could be a transfer of goods or services from one profit center of the integrated firm to another, just like in our bilateral trade example.

To keep the analysis tractable, we impose a constraint on the designer's strategy set, i.e., on the set of available mechanisms. We assume that the mechanism has to be an AGV mechanism as described in Section 4.2, with an additional (possibly negative) upfront transfer $\bar{t}_i$ from the designer to agent i. The insurance property and BIC are unaffected by $\bar{t} = (\bar{t}_1, \bar{t}_2)$, so that we can safely ignore intention-based social preferences between the two agents: By Theorem 1, any such mechanism is strongly implementable in BNFE when the agents treat it as exogenous. Hence the endogeneity of the mechanism is the only conceivable impediment for implementation. Formally, the designer's problem reduces to the choice of $\bar{t}$. We write $\Pi_i(\bar{t}) = \Pi_i^{AGV} + \bar{t}_i$, where $\Pi_i^{AGV} = \mathbb{E}[v_j(q_j^*(\tilde{\theta}), \tilde{\theta}_j)]$ is agent i's expected payoff in the AGV mechanism with surplus-maximizing consumption levels (q_1^*, q_2^*) and no upfront payment. We require $\bar{t}_i \geq -\Pi_i^{AGV}$ to guarantee that no agent's payoff becomes negative. We write $R(\bar{t}) = -(\bar{t}_1 + \bar{t}_2)$ for the expected revenue and we require $R(\bar{t}) \geq \bar{R}$, where $\bar{R}$ is exogenously given and could be positive or negative. We assume $\bar{R} \leq \Pi_1^{AGV} + \Pi_2^{AGV}$ to guarantee that there exist upfront transfers which satisfy all constraints.

We now introduce an equitable reference payoff for each agent i. If, for a proposed mechanism-equilibrium-pair, agent i's expected payoff fell short of this reference, this would indicate that the mechanism designer has treated i in an unfair way. In the spirit of our earlier assumptions,

[22] We are grateful to a referee for suggesting this application.

let agent i's equitable payoff be defined as the average between her best and her worst payoff on the material payoff frontier. The best outcome for i is achieved if the designer extracts all rents from j and pays (or obtains) the difference to the revenue requirement $\bar{R}$ to (from) i, so that $\bar{t}_j = -\Pi_j^{AGV}$ and $\bar{t}_i = \Pi_j^{AGV} - \bar{R}$. The worst outcome for i arises when the designer extracts all rents from i, so that $\bar{t}_i = -\Pi_i^{AGV}$. This yields the equitable payoff

$$\Pi_i^e = \frac{1}{2}(\Pi_i^{AGV} + \Pi_j^{AGV} - \bar{R}).$$

In words, the agents consider as equitable an equal split of the expected surplus that remains after satisfying the resource requirement $\bar{R}$. Assuming $\bar{\kappa} = \infty$ for simplicity, the kindness of a designer who proposes $\bar{t}$ to agent i then is

$$\kappa_{di}(\bar{t}) = \Pi_i(\bar{t}) - \Pi_1^e = \frac{1}{2}(\Pi_i^{AGV} - \Pi_j^{AGV} + \bar{R}) + \bar{t}_i.$$

Agent i's best-response problem, given truth-telling of agent j, becomes to maximize

$$\Pi_i(s_i, s_j^T) + y_i \, \kappa_{di}(\bar{t}) \, H(\Pi_i(s_i, s_j^T), \Pi_j(s_i, s_j^T), R(\bar{t})),$$

where we omitted some additive constants that do not affect the solution to this optimization problem.

Suppose that the offered mechanism yields less than half of the surplus for agent i, i.e., $\bar{t}_i < -(\Pi_i^{AGV} - \Pi_j^{AGV} + \bar{R})/2$. In the firm organization example, this would arise if the headquarter favors unit j and/or tries to extract more than $\bar{R}$. We obtain $\kappa_{di}(\bar{t}) < 0$, because agent i is disappointed by a designer who does not come up with a mechanism that generates an appropriate payoff for herself. Hence, she would like to sabotage the designer. Since the proposed mechanism has the insurance property, $\Pi_j(s_i, s_j^T)$ is independent of s_i and agent i can influence the designer's objective only through the own payoff $\Pi_i(s_i, s_j^T)$.[23] Since truth-telling maximizes $\Pi_i(s_i, s_j^T)$ by BIC, for a sufficiently large value of y_i agent i will benefit from a deviation that reduces $\Pi_i(s_i, s_j^T)$. In the firm organization example, this may capture a situation in which performance suffers due to aggravation in one of the merged firms. In the opposite case, when $\bar{t}_i > -(\Pi_i^{AGV} - \Pi_j^{AGV} + \bar{R})/2$, the constraints on feasible transfers immediately yield $\bar{t}_j < -(\Pi_j^{AGV} - \Pi_i^{AGV} + \bar{R})/2$, and the same logic implies that agent j will deviate from truth-telling when y_j is large enough. The only AGV mechanism that remains strongly implementable in BNFE is the one with $\bar{t}_i = -(\Pi_i^{AGV} - \Pi_j^{AGV} + \bar{R})/2$ for both $i = 1, 2$. In this case we obtain $\kappa_{di}(\bar{t}) = 0$ for both $i = 1, 2$, such that truth-telling is an equilibrium for all $y \in \mathbb{R}_+^2$.

This simple example demonstrates that reciprocity towards the designer can have a substantial impact on the set of implementable outcomes. While the AGV mechanism with any upfront transfers (that respect non-negativity constraints) is strongly implementable in BNFE if the agents treat the mechanism as exogenous, only an equal sharing of the revenue requirement $\bar{R}$ can be strongly implemented when the mechanism is treated as endogenous, and thus conveys the designer's intentions. In particular, the designer is unable to extract any surplus beyond $\bar{R}$ from the agents.

[23] We rule out the degenerate case where agent i's strategy has no impact on the payoffs at all.

8 Conclusion

Economists have become increasingly aware of the fact that preferences are often context-dependent. A mechanism designer who creates the rules of a game is thus confronted with the possibility that the game has an impact on behavior beyond the usually considered incentive effects, by influencing preferences through context. The theory of intention-based social preferences is one of the few well-established models that admit context-dependence, which makes it an ideal starting point for the investigation of the problem.

To ensure a broad applicability of our results, our analysis has employed a workhorse model in mechanism design theory, the independent private values model. This model has been used to study a wide range of problems, such as the allocation of indivisible private goods (Myerson, 1981), trade between privately informed parties (Myerson and Satterthwaite, 1983), the dissolution of partnerships (Cramton et al., 1987), the regulation of externalities (Rob, 1989), the provision of pure public goods (Mailath and Postlewaite, 1990), or the provision of excludable public goods (Hellwig, 2003). The virtue of working with a generic version of the independent private values model is that our theorems and propositions cover all these applications.

Our analysis provides a strong foundation for the consideration of social choice functions that are incentive-compatible and have the insurance property. According to Theorem 1, these conditions are sufficient for implementability in the absence of any knowledge about the intensity of intention-based social preferences. Moreover, as we argued in Section 4, this finding is not tied to the model of intention-based preferences but extends to a large class of interdependent preference models. According to Propositions 4 and 5, the requirements of incentive-compatibility and insurance are also necessary for the implementability of a social choice function, whenever there is two-dimensional private information and both selfish and strongly reciprocal agents are possible. Thus, for applications of mechanism design with a concern that not only monetary incentives but also social preferences are a driving force of behavior, our theoretical analysis gives rise to a firm recommendation: Use an incentive-compatible mechanism with the insurance property. This will make sure that the intended outcome is reached and, moreover, under the assumptions of Propositions 4 and 5 there is also no hope of increasing the set of implementable outcomes by a weakening of either requirement.

Whether interdependent preferences are relevant in a given application of mechanism design is ultimately an empirical question that we do not investigate in this paper. However, our emphasis of mechanisms that are robust with respect to interdependent preferences is backed by a recent experimental literature which documents that violations of the insurance property indeed trigger deviations from the intended behavior. Fehr et al. (2015) show that mechanisms for subgame-perfect implementation, which rely crucially on endowing the agents with mutual retaliation opportunities, do not reach the desired outcome. The violation of the insurance property puts at risk the promise of these mechanisms to solve all problems that arise from non-verifiability of information. Bartling and Netzer (2015) compare a conventional second-price auction, which does not have the insurance property, to its counterpart with the insurance property. The latter is obtained by applying our Proposition 2. The second-price auction gives rise to significant overbidding, while overbidding disappears in the strongly implementable mechanism with the insurance property. They attribute this finding to the robustness of the latter

mechanism to spiteful preferences. Bierbrauer et al. (2015) study a bilateral trade problem and a problem of redistributive income taxation. They compare ex post incentive-compatible social choice functions (Bergemann and Morris, 2005) with and without the insurance property. Again, they find significant deviations from truthful behavior only in situations where the insurance property is violated.

Our analysis also alerts the applied mechanism designer to the fact that different outcome-equivalent mechanisms can and should be compared according to the attitudes that they induce among the agents. Unused or rarely used actions are one tool for engineering good attitudes. As we argued in Section 5, these actions may take the form of loopholes in incomplete contracts or of veto rights in real-world mechanisms.

Finally, our analysis raises a couple of questions for future research. First, the focus on normal form mechanisms is typically justified by the argument that 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. It is also unclear which social choice functions can be implemented as a unique fairness equilibrium outcome of some extensive form mechanism. A major obstacle to answering these questions is the lack of a general theory of intentions for extensive form games with incomplete information.

References

Aldashev, G., Kirchsteiger, G., and Sebald, A. (2015). Assignment procedure biases in randomized policy experiments. *Economic Journal*, in press.

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.

Antler, Y. (2015). Two-sided matching with endogenous preferences. *American Economic Journal: Microeconomics*, 7:241–258.

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*.

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

Bartling, B. and Netzer, N. (2015). An externality-robust auction: Theory and experimental evidence. Mimeo.

Bassi, M., Pagnozzi, M., and Piccolo, S. (2014). Optimal contracting with altruism and reciprocity. *Research in Economics*, 68:27–38.

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

Benjamin, D. (2014). Distributional preferences, reciprocity-like behavior, and efficiency in bilateral exchange. *American Economic Journal: Microeconomics*, 7:70–98.

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

Bierbrauer, F., Ockenfels, A., Rückert, D., and Pollak, A. (2015). Robust mechanism design and social preferences. Mimeo.

Bodoh-Creed, A. (2012). Ambiguous beliefs and mechanism design. *Games and Economic Behavior*, 75:518–537.

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

Bose, S., Ozdenoren, E., and Pape, A. (2006). Optimal auctions with ambiguity. *Theoretical Economics*, 1:411–438.

Bowles, S. and Polanía-Reyes, S. (2012). Economic incentives and social preferences: Substitutes or complements? *Journal of Economic Literature*, 50:368–425.

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. 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.

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

Clarke, E. (1971). Multipart pricing of public goods. *Public Choice*, 11:17–33.

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.

de Clippel, G. (2014). Behavioral implementation. *American Economic Review*, 104:2975–3002.

De Marco, G. and Immordino, G. (2013). Partnership, reciprocity and team design. *Research in Economics*, 67:39–58.

De Marco, G. and Immordino, G. (2014). Reciprocity in the principal multiple agent model. *B.E. Journal of Theoretical Economics*, 14:1–39.

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

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

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

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

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*, 4:146–183.

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., Powell, M., and Wilkening, T. (2015). Behavioral limitations of subgame-perfect implementation. 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.

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.

Gradwohl, R. (2014). Privacy in implementation. Mimeo.

Groves, T. (1973). Incentives in teams. *Econometrica*, 41:617–663.

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

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

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

Herold, F. (2010). Contractual incompleteness as a signal of trust. *Games and Economic Behavior*, 68:180–191.

Hoppe, E. and Schmitz, P. (2013). Contracting under incomplete information and social preferences: An experimental study. *Review of Economic Studies*, 80:1516–1544.

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*.

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. (2012). Behavioral mechanism design. *Journal of Public Economic Theory*, 14:767–789.

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

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. and Riley, J. (1984). Optimal auctions with risk averse buyers. *Econometrica*, 52:1473–1518.

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

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

Myerson, R. (1981). Optimal auction design. *Mathematics of Operation Research*, 6:58–73.

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

Netzer, N. and Schmutzler, A. (2014). Explaining gift-exchange – the limits of good intentions. *Journal of the European Economic Association*, 12:1586–1616.

Netzer, N. and Volk, A. (2014). Intentions and ex-post implementation. Mimeo.

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

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

Rob, R. (1989). Pollution claim settlements under private information. *Journal of Economic Theory*, 47:307–333.

Ruffle, B. J. (1999). Gift giving with emotions. *Journal of Economic Behavior and Organization*, 39:399–420.

Saran, R. (2011). Menu-dependent preferences and the revelation principle. *Journal of Economic Theory*, 146:1712–1720.

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.

Tang, P. and Sandholm, T. (2012). Optimal auctions for spiteful bidders. *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1457–1463.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance*, 16:8–37.

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.

von Siemens, F. (2013). Intention-based reciprocity and the hidden costs of control. *Journal of Economic Behavior and Organization*, 92:55–65.

A Proofs of General Results

A.1 Proof of Theorem 1

Step 1. Consider the direct mechanism for a given SCF f. As a first step, we show that $\Delta_i = 0$ if and only if $\Pi_i(s_i^T, s'_j) = \Pi_i(s_i^T, s''_j)$ for any two strategies $s'_j, s''_j \in S_j$ of agent j.

Suppose $\Pi_i(s_i^T, s'_j) = \Pi_i(s_i^T, s''_j)$ for any $s'_j, s''_j \in S_j$. We show that this implies $\Delta_i = 0$. For arbitrary types $\theta'_j, \theta''_j \in \Theta_j$, let $\bar{s}'_j$ be the strategy to always announce θ'_j and $\bar{s}''_j$ the strategy to always announce θ''_j, whatever agent j's true type. Then $\Pi_i(s_i^T, \bar{s}'_j) = \Pi_i(s_i^T, \bar{s}''_j)$ holds. Equivalently,

$$\mathbb{E}[v_i(q_i^f(\tilde{\theta}_i, \theta'_j), \tilde{\theta}_i) + t_i^f(\tilde{\theta}_i, \theta'_j)] = \mathbb{E}[v_i(q_i^f(\tilde{\theta}_i, \theta''_j), \tilde{\theta}_i) + t_i^f(\tilde{\theta}_i, \theta''_j)] \; .$$

Since our choice of $\theta'_j, \theta''_j \in \Theta_j$ was arbitrary, this implies that $\Delta_i = 0$.

Now suppose that $\Delta_i = 0$. For all strategies $s_j \in S_j$ and all types $\theta_j \in \Theta_j$, define

$$\Lambda(\theta_j | s_j) = \{\theta'_j \in \Theta_j \mid s_j(\theta'_j) = \theta_j\}.$$

For any $s_j \in S_j$, observe that

$$\begin{aligned}
\Pi_i(s_i^T, s_j) &= \mathbb{E}[v_i(q_i^f(\tilde{\theta}_i, s_j(\tilde{\theta}_j)), \tilde{\theta}_i) + t_i^f(\tilde{\theta}_i, s_j(\tilde{\theta}_j))] \\
&= \mathbb{E}_j[\mathbb{E}_i[v_i(q_i^f(\tilde{\theta}_i, s_j(\tilde{\theta}_j)), \tilde{\theta}_i) + t_i^f(\tilde{\theta}_i, s_j(\tilde{\theta}_j))]] \\
&= \hat{\mathbb{E}}_j[\mathbb{E}_i[v_i(q_i^f(\tilde{\theta}_i, \tilde{\theta}_j), \tilde{\theta}_i) + t_i^f(\tilde{\theta}_i, \tilde{\theta}_j)]],
\end{aligned}$$

where $\mathbb{E}_i$ and $\mathbb{E}_j$ denote the expectations operator with respect to only $\tilde{\theta}_i$ and $\tilde{\theta}_j$, respectively, and $\hat{\mathbb{E}}_j$ is the expectations operator with respect to $\tilde{\theta}_j$ based on the modified probability distribution $\hat{p}_j$ given by

$$\hat{p}_j(\theta_j) = \sum_{\theta'_j \in \Lambda(\theta_j | s_j)} p_j(\theta'_j)$$

for all $\theta_j \in \Theta_j$. From $\Delta_i = 0$ it follows that there exists a number ρ so that $\mathbb{E}_i[v_i(q_i^f(\tilde{\theta}_i, \theta_j), \tilde{\theta}_i) + t_i^f(\tilde{\theta}_i, \theta_j)] = \rho$ for all $\theta_j \in \Theta_j$, and hence $\Pi_i(s_i^T, s_j) = \hat{\mathbb{E}}_j[\rho] = \rho$. Since our choice of s_j was arbitrary, this implies $\Pi_i(s_i^T, s'_j) = \rho = \Pi_i(s_i^T, s''_j)$ for any two $s'_j, s''_j \in S_j$.

Step 2. Now assume that f is BIC and satisfies $\Delta_1 = \Delta_2 = 0$. Consider the truthful strategy profile $s^T = (s_1^T, s_2^T)$ in the direct mechanism, and suppose all first- and second-order beliefs are

correct. For both $i = 1, 2$ we then obtain $\Pi_i^e(s_j^b) = \Pi_i^e(s_i^T) = \Pi_i(s^T)$ according to step 1, which implies that $\kappa_j(s_i^b, s_i^{bb}) = \kappa_j(s^T) = 0$. Hence agent i's problem $\max_{s_i \in S_i} U_i(s_i, s_i^b, s_i^{bb})$ becomes $\max_{s_i \in S_i} \Pi_i(s_i, s_j^T)$. Truth-telling s_i^T is a solution to this problem by BIC, so s^T is a BNFE.

A.2 Proof of Proposition 1

Consider any AGV f. For both $i = 1, 2$ and any type realization $\theta_j \in \Theta_j$ it holds that

$$\begin{aligned}
&\mathbb{E}[v_i(q_i^f(\tilde{\theta}_i, \theta_j), \tilde{\theta}_i) + t_i^f(\tilde{\theta}_i, \theta_j)] \\
&= \mathbb{E}[v_i(q_i^f(\tilde{\theta}_i, \theta_j), \tilde{\theta}_i)] + \mathbb{E}[\mathbb{E}_j[v_j(q_j^f(\tilde{\theta}_i, \tilde{\theta}_j), \tilde{\theta}_j)]] - \mathbb{E}[v_i(q_i^f(\tilde{\theta}_i, \theta_j), \tilde{\theta}_i)0] \\
&= \mathbb{E}[v_j(q_j^f(\tilde{\theta}_i, \tilde{\theta}_j), \tilde{\theta}_j)],
\end{aligned}$$

which is independent of θ_j. Therefore $\Delta_i = 0$.

A.3 Proof of Proposition 2

Let $f = (q_1^f, q_2^f, t_1^f, t_2^f)$ be an SCF that is BIC. We construct a new payment rule $(t_1^{\bar{f}}, t_2^{\bar{f}})$ as follows. For every $i = 1, 2$ and $(\theta_i, \theta_j) \in \Theta$, let

$$t_i^{\bar{f}}(\theta_i, \theta_j) = \mathbb{E}[v_i(q_i^f(\theta_i, \tilde{\theta}_j), \theta_i) + t_i^f(\theta_i, \tilde{\theta}_j)] - v_i(q_i^f(\theta_i, \theta_j), \theta_i). \tag{8}$$

We verify that $\bar{f} = (q_1^{\bar{f}}, q_2^{\bar{f}}, t_1^{\bar{f}}, t_2^{\bar{f}})$, with $q_i^{\bar{f}} = q_i^f$ for both $i = 1, 2$, satisfies properties (a) - (d).

Property (a). This property is satisfied by construction.

Property (b). This property follows after an application of the law of iterated expectations:

$$\begin{aligned}
\sum_{i=1,2} \mathbb{E}[t_i^{\bar{f}}(\tilde{\theta})] &= \sum_{i=1,2} \mathbb{E}[\mathbb{E}_j[v_i(q_i^f(\tilde{\theta}_i, \tilde{\theta}_j), \tilde{\theta}_i) + t_i^f(\tilde{\theta}_i, \tilde{\theta}_j)] - v_i(q_i^f(\tilde{\theta}_i, \tilde{\theta}_j), \tilde{\theta}_i)] \\
&= \sum_{i=1,2} \mathbb{E}[v_i(q_i^f(\tilde{\theta}_i, \tilde{\theta}_j), \tilde{\theta}_i) + t_i^f(\tilde{\theta}_i, \tilde{\theta}_j) - v_i(q_i^f(\tilde{\theta}_i, \tilde{\theta}_j), \tilde{\theta}_i)] \\
&= \sum_{i=1,2} \mathbb{E}[t_i^f(\tilde{\theta})].
\end{aligned}$$

Property (c). This property follows since

$$\begin{aligned}
\mathbb{E}[v_i(q_i^{\bar{f}}(\theta_i, \tilde{\theta}_j), \theta_i) + t_i^{\bar{f}}(\theta_i, \tilde{\theta}_j)] &= \mathbb{E}[v_i(q_i^f(\theta_i, \tilde{\theta}_j), \theta_i) + t_i^{\bar{f}}(\theta_i, \tilde{\theta}_j)] \\
&= \mathbb{E}[\mathbb{E}[v_i(q_i^f(\theta_i, \tilde{\theta}_j), \theta_i) + t_i^f(\theta_i, \tilde{\theta}_j)]] \\
&= \mathbb{E}[v_i(q_i^f(\theta_i, \tilde{\theta}_j), \theta_i) + t_i^f(\theta_i, \tilde{\theta}_j)].
\end{aligned}$$

Property (d). We first show that $\bar{f}$ has the insurance property. From (8) it follows that for any $(\theta_i, \theta_j) \in \Theta$ we have that

$$v_i(q_i^{\bar{f}}(\theta_i, \theta_j), \theta_i) + t_i^{\bar{f}}(\theta_i, \theta_j) = \mathbb{E}[v_i(q_i^f(\theta_i, \tilde{\theta}_j), \theta_i) + t_i^f(\theta_i, \tilde{\theta}_j)],$$

which is independent of θ_j. Hence the ex post payoff of any type θ_i of agent i does not depend on agent j's type, which implies that the insurance property holds. It remains to be shown that

$\bar{f}$ is BIC. Since f is BIC, it holds that

$$\mathbb{E}[v_i(q_i^f(\theta_i,\tilde{\theta}_j),\theta_i)] + \mathbb{E}[t_i^f(\theta_i,\tilde{\theta}_j)] \geq \mathbb{E}[v_i(q_i^f(\hat{\theta}_i,\tilde{\theta}_j),\theta_i)] + \mathbb{E}[t_i^f(\hat{\theta}_i,\tilde{\theta}_j)]$$

for $i = 1,2$ and all $\theta_i, \hat{\theta}_i \in \Theta_i$. Since $q_i^f = q_i^{\bar{f}}$ and

$$\begin{aligned}
\mathbb{E}[t_i^f(\theta_i,\tilde{\theta}_j)] &= \mathbb{E}[v_i(q_i^f(\theta_i,\tilde{\theta}_j),\theta_i) + t_i^f(\theta_i,\tilde{\theta}_j) - v_i(q_i^f(\theta_i,\tilde{\theta}_j),\theta_i)] \\
&= \mathbb{E}[\mathbb{E}[v_i(q_i^f(\theta_i,\tilde{\theta}_j),\theta_i) + t_i^f(\theta_i,\tilde{\theta}_j)] - v_i(q_i^f(\theta_i,\tilde{\theta}_j),\theta_i)] \\
&= \mathbb{E}[t_i^{\bar{f}}(\theta_i,\tilde{\theta}_j)]
\end{aligned}$$

for $i = 1,2$ and all $\theta_i \in \Theta_i$, this implies

$$\mathbb{E}[v_i(q_i^{\bar{f}}(\theta_i,\tilde{\theta}_j),\theta_i)] + \mathbb{E}[t_i^{\bar{f}}(\theta_i,\tilde{\theta}_j)] \geq \mathbb{E}[v_i(q_i^{\bar{f}}(\hat{\theta}_i,\tilde{\theta}_j),\theta_i)] + \mathbb{E}[t_i^{\bar{f}}(\hat{\theta}_i,\tilde{\theta}_j)],$$

for all $\theta_i, \hat{\theta}_i \in \Theta_i$, so that $\bar{f}$ is also BIC.

A.4 Proof of Lemma 1

We first state explicitly the property of strategic equivalence of arbitrary and augmented revelation mechanisms. We start from an arbitrary mechanism $\Phi = (M_1, M_2, g)$ and a strategy profile $\tilde{s} = (\tilde{s}_1, \tilde{s}_2)$, interpreted as an equilibrium of some type. We then construct an augmented revelation mechanism $\Phi'(\Phi, \tilde{s})$ based on Φ and $\tilde{s}$, with the property that the outcome of Φ' under truth-telling is the same as the outcome of Φ under $\tilde{s}$. 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 agent. Any action from M_i that is used by $\tilde{s}_i$ is relabelled according to the type θ_i that uses it, and any unused action from M_i is kept unchanged: $M_i' = \Theta_i \cup (M_i \backslash \tilde{s}_i(\Theta_i))$. To define the outcome function g' of Φ', we first construct for every agent a surjective function $\eta_i : M_i' \to M_i$ that maps actions from M_i' back into M_i:

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

For all message profiles $m' = (m_1', m_2')$ we then define

$$g'(m') = g(\eta_1(m_1'), \eta_2(m_2')). \tag{9}$$

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 $M_i \backslash \tilde{s}_i(\Theta_i)$ 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 $g'(s^T(\theta)) = g'(\theta) = g(\tilde{s}(\theta)) = f(\theta)$ for all $\theta \in \Theta$.

Lemma 2. *The mechanisms Φ and $\Phi'(\Phi, \tilde{s})$ are strategically equivalent, in the sense that, for*

$i = 1, 2$ and any $m_j \in M_j$ and $m'_j \in M'_j$ with $m_j = \eta_j(m'_j)$, it holds that $G_i(m_j) = G'_i(m'_j)$, where

$$G_i(m_j) = \{a \in A \mid \exists m_i \in M_i \text{ so that } g(m_i, m_j) = a\}$$

and

$$G'_i(m'_j) = \{a \in A \mid \exists m'_i \in M'_i \text{ so that } g'(m'_i, m'_j) = a\}.$$

Proof. We first show that $G'_i(m'_j) \subseteq G_i(\eta_j(m'_j))$. Let $a \in G'_i(m'_j)$, so that there exists m'_i so that $g'(m'_i, m'_j) = a$. By (9), this implies that $g(\eta_i(m'_i), \eta_j(m'_j)) = a$, and hence $a \in G_i(\eta_j(m'_j))$.

We now show that $G_i(\eta_j(m'_j)) \subseteq G'_i(m'_j)$. Let $a \in G_i(\eta_j(m'_j))$, so that there exists $m_i \in M_i$ so that $g(m_i, \eta_j(m'_j)) = a$. Since η_i is surjective, there exists m'_i with $\eta_i(m'_i) = m_i$. Then (9) implies that $g'(m'_i, m'_j) = a$. Hence, $a \in G'_i(m'_j)$. □

The sets $G_i(m_j)$ and $G'_i(m'_j)$ contain all allocations that agent i can induce by varying her message, holding fixed agent j's message. According to Lemma 2, these sets are the same in both mechanisms, for any pair of messages with $m_j = \eta_j(m'_j)$. This 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 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'. This proves Lemma 1.

A.5 Proof of Theorem 2

We prove the theorem in two steps. First, we augment the direct mechanism for any SCF f by additional actions and show that the equitable payoffs associated to truth-telling can be increased or decreased to arbitrary values. Second, we use the result of the first step to show that an SCF f can be implemented in BNFE when the conditions in the theorem are satisfied, i.e., when f is materially Pareto-efficient and $y_i > 0$ and $1/y_i \leq \bar{\kappa} - \Delta_i$ holds for both $i = 1, 2$.

Step 1. Fix any SCF f and consider a mechanism $\Phi(\delta)$ for f that is parameterized by $\delta = (\delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}) \in \mathbb{R}^4$. The message sets are $M_i = \Theta_i \times \{0, 1\}$ for both $i = 1, 2$, so that a message $m_i = (m_i^1, m_i^2) \in M_i$ of agent i consists of a type $m_i^1 \in \Theta_i$ and a number $m_i^2 \in \{0, 1\}$. The outcome function $g = (q_1^g, q_2^g, t_1^g, t_2^g)$ of $\Phi(\delta)$ is defined by

$$q_i^g(m) = q_i^f(m_1^1, m_2^1)$$

and

$$t_i^g(m) = t_i^f(m_1^1, m_2^1) + m_i^2 \delta_{ii} - m_j^2 \delta_{ji}$$

for both $i = 1, 2$ and all $m = (m_1, m_2) \in M_1 \times M_2$. Parameter δ_{ik}, which can be positive or negative, describes the effect that agent $i = 1, 2$ has on the transfer of agent $k = 1, 2$ through the second message component. We require $\delta_{ii} \leq \delta_{ij}$ to ensure that the transfers are

always admissible. Mechanism $\Phi(\delta)$ becomes equivalent to the direct mechanism for f when $\delta = (0,0,0,0)$, or $\delta = 0$ in short, because the second message components are payoff irrelevant in this case. Let s_i^T be agent i's strategy that announces $s_i^T(\theta_i) = (\theta_i, 0)$ for all types $\theta_i \in \Theta_i$. The outcome of strategy profile $s^T = (s_1^T, s_2^T)$ is the SCF f, independent of δ.

We use the expressions $\Pi_i(s_i, s_i^b|\delta)$, $E_i(s_i^b|\delta)$, and $\Pi_i^e(s_j^b|\delta)$ to denote expected payoffs, efficient strategies, and equitable payoffs in $\Phi(\delta)$. We also write $s_i = (s_i^1, s_i^2) \in S_i$ for strategies, so that $s_i^1(\theta_i) \in \Theta_i$ and $s_i^2(\theta_i) \in \{0,1\}$ are the two message components announced by type θ_i under strategy s_i. Let

$$x_i(s_i) = \mathbb{E}[s_i^2(\tilde{\theta}_i)]$$

be the probability with which a strategy s_i announces $m_i^2 = 1$, for both $i = 1, 2$. Then we obtain

$$\Pi_i(s_i, s_i^b|\delta) = \Pi_i(s_i, s_i^b|0) + x_i(s_i)\delta_{ii} - x_j(s_i^b)\delta_{ji}. \tag{10}$$

Lemma 3. *If* $\operatorname{sgn}\delta_{jj} = \operatorname{sgn}\delta_{ji}$, *then*

$$\max_{s_j \in E_j(s_i^T|\delta)} \Pi_i(s_i^T, s_j|\delta) = \max_{s_j \in E_j(s_i^T|0)} \Pi_i(s_i^T, s_j|0) - \min\{\delta_{ji}, 0\} \tag{11}$$

and

$$\min_{s_j \in E_j(s_i^T|\delta)} \Pi_i(s_i^T, s_j|\delta) = \min_{s_j \in E_j(s_i^T|0)} \Pi_i(s_i^T, s_j|0) - \max\{\delta_{ji}, 0\}. \tag{12}$$

Proof. We first claim that $E_j(s_i^T|\delta) \subseteq E_j(s_i^T|0)$ holds. If $s_j \notin E_j(s_i^T|0)$, then there exists a strategy $\hat{s}_j$ such that

$$\Pi_i(s_i^T, \hat{s}_j|0) \geq \Pi_i(s_i^T, s_j|0),$$
$$\Pi_j(s_i^T, \hat{s}_j|0) \geq \Pi_j(s_i^T, s_j|0),$$

with at least one inequality being strict. Now consider strategy $\tilde{s}_j$ constructed by

$$\tilde{s}_j^1(\theta_j) = \hat{s}_j^1(\theta_j) \text{ and } \tilde{s}_j^2(\theta_j) = s_j^2(\theta_j)$$

for all $\theta_j \in \Theta_j$. Using (10) and the above inequalities, we obtain

$$\begin{aligned}
\Pi_i(s_i^T, \tilde{s}_j|\delta) &= \Pi_i(s_i^T, \tilde{s}_j|0) - x_j(\tilde{s}_j)\delta_{ji} \\
&= \Pi_i(s_i^T, \hat{s}_j|0) - x_j(s_j)\delta_{ji} \\
&\geq \Pi_i(s_i^T, s_j|0) - x_j(s_j)\delta_{ji} \\
&= \Pi_i(s_i^T, s_j|\delta),
\end{aligned}$$

and analogously for agent j (with at least one strict inequality). Hence $s_j \notin E_j(s_i^T|\delta)$, which establishes the claim.

We now go through the three possible cases in which $\operatorname{sgn}\delta_{jj} = \operatorname{sgn}\delta_{ji}$ holds (given $\delta_{jj} \leq \delta_{ji}$).

Case (a): $\delta_{jj} = \delta_{ji} = 0$. The statement in the lemma follows immediately in this case.

Case (b): $0 < \delta_{jj} \leq \delta_{ji}$. Observe that $E_j(s_i^T|\delta)$ and $E_j(s_i^T|0)$ can be replaced by S_j in the maximization problems in (11), because at least one of j's strategies that maximize i's expected payoff on the (finite) set S_j must be Pareto-efficient. Using (10), statement (11) then follows because any strategy s_j that maximizes $\Pi_i(s_i^T, s_j|\delta)$ on the set S_j must clearly satisfy $x_j(s_j) = 0$. To establish statement (12), consider a minimizing strategy $s_j^{min} \in \arg\min_{s_j \in E_j(s_i^T|0)} \Pi_i(s_i^T, s_j|0)$ that satisfies $x_j(s_j^{min}) = 1$, which exists because m_j^2 is payoff irrelevant in $\Phi(0)$. We claim that $s_j^{min} \in E_j(s_i^T|\delta)$, which then implies, again using (10), that

$$\min_{s_j \in E_j(s_i^T|\delta)} \Pi_i(s_i^T, s_j|\delta) \leq \Pi_i(s_i^T, s_j^{min}|\delta) = \Pi_i(s_i^T, s_j^{min}|0) - \delta_{ji}, \tag{13}$$

and hence a weak inequality version of (12). To establish the claim, suppose to the contrary that there exists $s_j' \in E_j(s_i^T|\delta)$ such that

$$\begin{aligned}
&\Pi_i(s_i^T, s_j'|\delta) \geq \Pi_i(s_i^T, s_j^{min}|\delta), \\
&\Pi_j(s_i^T, s_j'|\delta) \geq \Pi_j(s_i^T, s_j^{min}|\delta),
\end{aligned}$$

with a least one inequality being strict. Assuming $s_j' \in E_j(s_i^T|\delta)$ is w.l.o.g. because S_j is finite, so that at least one strategy that Pareto-dominates s_j^{min} must itself be Pareto-efficient. Using (10), these inequalities can be rearranged to

$$\begin{aligned}
&\Pi_i(s_i^T, s_j'|0) + [1 - x_j(s_j')]\delta_{ji} \geq \Pi_i(s_i^T, s_j^{min}|0), \\
&\Pi_j(s_i^T, s_j'|0) - [1 - x_j(s_j')]\delta_{jj} \geq \Pi_j(s_i^T, s_j^{min}|0).
\end{aligned}$$

If $x_j(s_j') = 1$ this contradicts $s_j^{min} \in E_j(s_i^T|0)$. Hence $x_j(s_j') < 1$ must hold, which implies

$$\begin{aligned}
&\Pi_i(s_i^T, s_j'|0) < \Pi_i(s_i^T, s_j^{min}|0), \\
&\Pi_j(s_i^T, s_j'|0) > \Pi_j(s_i^T, s_j^{min}|0),
\end{aligned}$$

where the first inequality follows from the second one due to $s_j^{min} \in E_j(s_i^T|0)$. But now we must have $s_j' \notin E_j(s_i^T|0)$, as otherwise s_j^{min} would not minimize i's payoff on $E_j(s_i^T|0)$. This contradicts $s_j' \in E_j(s_i^T|\delta)$ because $E_j(s_i^T|\delta) \subseteq E_j(s_i^T|0)$, and hence establishes the claim. The opposite weak inequality of (13) follows from

$$\begin{aligned}
\min_{s_j \in E_j(s_i^T|\delta)} \Pi_i(s_i^T, s_j|\delta) &\geq \min_{s_j \in E_j(s_i^T|0)} \Pi_i(s_i^T, s_j|\delta) \\
&= \min_{s_j \in E_j(s_i^T|0)} \left[\Pi_i(s_i^T, s_j|0) - x_j(s_j)\delta_{ji}\right] \\
&\geq \min_{s_j \in E_j(s_i^T|0)} \left[\Pi_i(s_i^T, s_j|0)\right] - \delta_{ji} \\
&= \Pi_i(s_i^T, s_j^{min}|0) - \delta_{ji},
\end{aligned}$$

where the first inequality is again due to $E_j(s_i^T|\delta) \subseteq E_j(s_i^T|0)$.

Case (c): $\delta_{jj} \leq \delta_{ji} < 0$. Statement (11) again follows after replacing $E_j(s_i^T|\delta)$ and $E_j(s_i^T|0)$ by S_j, observing that any s_j that maximizes $\Pi_i(s_i^T, s_j|\delta)$ on S_j must satisfy $x_j(s_j) = 1$. To

establish statement (12), consider a strategy $s_j^{min} \in \arg\min_{s_j \in E_j(s_i^T|0)} \Pi_i(s_i^T, s_j|0)$ that satisfies $x_j(s_j^{min}) = 0$. We claim that $s_j^{min} \in E_j(s_i^T|\delta)$, which implies the weak inequality

$$\min_{s_j \in E_j(s_i^T|\delta)} \Pi_i(s_i^T, s_j|\delta) \leq \Pi_i(s_i^T, s_j^{min}|\delta) = \Pi_i(s_i^T, s_j^{min}|0). \tag{14}$$

Suppose to the contrary that there exists $s_j' \in E_j(s_i^T|\delta)$ such that

$$\begin{aligned}
&\Pi_i(s_i^T, s_j'|\delta) \geq \Pi_i(s_i^T, s_j^{min}|\delta),\\
&\Pi_j(s_i^T, s_j'|\delta) \geq \Pi_j(s_i^T, s_j^{min}|\delta),
\end{aligned}$$

with a least one inequality being strict, which can be rearranged to

$$\begin{aligned}
&\Pi_i(s_i^T, s_j'|0) - x_j(s_j')\delta_{ji} \geq \Pi_i(s_i^T, s_j^{min}|0),\\
&\Pi_j(s_i^T, s_j'|0) + x_j(s_j')\delta_{jj} \geq \Pi_j(s_i^T, s_j^{min}|0).
\end{aligned}$$

If $x_j(s_j') = 0$ this contradicts $s_j^{min} \in E_j(s_i^T|0)$. Hence $x_j(s_j') > 0$ must hold, which implies

$$\begin{aligned}
&\Pi_i(s_i^T, s_j'|0) < \Pi_i(s_i^T, s_j^{min}|0),\\
&\Pi_j(s_i^T, s_j'|0) > \Pi_j(s_i^T, s_j^{min}|0),
\end{aligned}$$

where the first inequality follows from the second one due to $s_j^{min} \in E_j(s_i^T|0)$. Now we obtain the same contradiction as for case (b) above. The opposite weak inequality of (14) follows from

$$\begin{aligned}
\min_{s_j \in E_j(s_i^T|\delta)} \Pi_i(s_i^T, s_j|\delta) &\geq \min_{s_j \in E_j(s_i^T|0)} \Pi_i(s_i^T, s_j|\delta)\\
&= \min_{s_j \in E_j(s_i^T|0)} \left[\Pi_i(s_i^T, s_j|0) - x_j(s_j)\delta_{ji}\right]\\
&\geq \min_{s_j \in E_j(s_i^T|0)} \left[\Pi_i(s_i^T, s_j|0)\right]\\
&= \Pi_i(s_i^T, s_j^{min}|0).
\end{aligned}$$

This completes the proof of the lemma. □

The following statement is an immediate corollary of Lemma 3.

Corollary 1. *If* $sgn\,\delta_{jj} = sgn\,\delta_{ji}$, *then* $\Pi_i^e(s_i^T|\delta) = \Pi_i^e(s_i^T|0) - \delta_{ji}/2$.

Step 2. Fix a materially Pareto-efficient SCF f and assume $y_i > 0$ and $1/y_i \leq \bar{\kappa} - \Delta_i$ for both $i = 1, 2$. Consider the BNFE candidate s^T in mechanism $\Phi(\delta^*)$, where δ^* is given by

$$\delta_{ii}^* = \delta_{ij}^* = 2\left[\frac{1}{y_j} - \Pi_j(s^T|0) + \Pi_j^e(s_j^T|0)\right] \tag{15}$$

for both $i = 1, 2$. Agent i's correct belief about j's kindness is then given by

$$\begin{aligned}
\kappa_j(s^T|\delta^*) &= h(\Pi_i(s^T|\delta^*) - \Pi_i^e(s_i^T|\delta^*))\\
&= h(\Pi_i(s^T|0) - \Pi_i^e(s_i^T|\delta^*))
\end{aligned}$$

$$
\begin{aligned}
&= h(\Pi_i(s^T|0) - \Pi_i^e(s_i^T|0) + \delta_{ji}^*/2) \\
&= h(1/y_i) \\
&= 1/y_i,
\end{aligned}
$$

where the third equality follows from Corollary 1 and the last equality holds due to $1/y_i \leq \bar{\kappa}$. In the equilibrium candidate, agent $i = 1, 2$ therefore chooses s_i so as to maximize

$$\Pi_i(s_i, s_j^T|\delta^*) + h(\Pi_j(s_i, s_j^T|\delta^*) - \Pi_j^e(s_j^T|\delta^*)).$$

For $s_i = s_i^T$, this term becomes $\Pi_i(s_i^T, s_j^T|\delta^*) + \Pi_j(s_i^T, s_j^T|\delta^*) - \Pi_j^e(s_j^T|\delta^*)$, because $\Pi_j(s_i^T, s_j^T|\delta^*) - \Pi_j^e(s_j^T|\delta^*) = 1/y_j \leq \bar{\kappa}$ by our construction. To exclude that there are any profitable deviations from s_i^T, we can restrict attention to conditionally efficient strategies $s_i' \in E_i(s_j^T|\delta^*)$. We consider three possible cases.

Case (a). A strategy $s_i' \in E_i(s_j^T|\delta^*)$ with $-\bar{\kappa} \leq \Pi_j(s_i', s_j^T|\delta^*) - \Pi_j^e(s_j^T|\delta^*) \leq \bar{\kappa}$ cannot be profitable, because in that case

$$
\begin{aligned}
\Pi_i(s_i', s_j^T|\delta^*) + h(\Pi_j(s_i', s_j^T|\delta^*) - \Pi_j^e(s_j^T|\delta^*)) &= \Pi_i(s_i', s_j^T|\delta^*) + \Pi_j(s_i', s_j^T|\delta^*) - \Pi_j^e(s_j^T|\delta^*) \\
&\leq \Pi_i(s_i^T, s_j^T|\delta^*) + \Pi_j(s_i^T, s_j^T|\delta^*) - \Pi_j^e(s_j^T|\delta^*),
\end{aligned}
$$

where the inequality follows from material Pareto-efficiency of f (and $\delta_{ii}^* = \delta_{ij}^*$).

Case (b). A strategy $s_i' \in E_i(s_j^T|\delta^*)$ with $\bar{\kappa} < \Pi_j(s_i', s_j^T|\delta^*) - \Pi_j^e(s_j^T|\delta^*)$ cannot be profitable, because in that case

$$
\begin{aligned}
\Pi_i(s_i', s_j^T|\delta^*) + h(\Pi_j(s_i', s_j^T|\delta^*) - \Pi_j^e(s_j^T|\delta^*)) &= \Pi_i(s_i', s_j^T|\delta^*) + \bar{\kappa} \\
&< \Pi_i(s_i', s_j^T|\delta^*) + \Pi_j(s_i', s_j^T|\delta^*) - \Pi_j^e(s_j^T|\delta^*) \\
&\leq \Pi_i(s_i^T, s_j^T|\delta^*) + \Pi_j(s_i^T, s_j^T|\delta^*) - \Pi_j^e(s_j^T|\delta^*).
\end{aligned}
$$

Case (c). We finally show that a strategy $s_i' \in E_i(s_j^T|\delta^*)$ with $\Pi_j(s_i', s_j^T|\delta^*) - \Pi_j^e(s_j^T|\delta^*) < -\bar{\kappa}$ does not exist. By contradiction, if such a strategy existed, then

$$\min_{s_i \in E_i(s_j^T|\delta^*)} \Pi_j(s_i, s_j^T|\delta^*) - \Pi_j^e(s_j^T|\delta^*) < -\bar{\kappa}$$

would have to hold as well. Using the definition of $\Pi_j^e(s_j^T|\delta^*)$, this can be rearranged to

$$\frac{1}{2}\left[\max_{s_i \in E_i(s_j^T|\delta^*)} \Pi_j(s_i, s_j^T|\delta^*) - \min_{s_i \in E_i(s_j^T|\delta^*)} \Pi_j(s_i, s_j^T|\delta^*)\right] > \bar{\kappa},$$

and, using Lemma 3, can be rewritten as

$$\frac{1}{2}\left[\max_{s_i \in E_i(s_j^T|0)} \Pi_j(s_i, s_j^T|0) - \min_{s_i \in E_i(s_j^T|0)} \Pi_j(s_i, s_j^T|0)\right] + \frac{1}{2}|\delta_{ij}^*| > \bar{\kappa}. \tag{16}$$

If $\delta_{ij}^* \geq 0$, using (15) and the definition of $\Pi_j^e(s_j^T|0)$, inequality (16) can be rewritten as

$$\max_{s_i \in E_i(s_j^T|0)} \Pi_j(s_i, s_j^T|0) - \Pi_j(s_i^T, s_j^T|0) + \frac{1}{y_j} > \bar{\kappa}.$$

Since $\Delta_j \geq \max_{s_i \in E_i(s_j^T|0)} \Pi_j(s_i, s_j^T|0) - \Pi_j(s_i^T, s_j^T|0)$, this further implies $1/y_j > \bar{\kappa} - \Delta_j$ and contradicts $1/y_j \leq \bar{\kappa} - \Delta_j$. If $\delta_{ij}^* < 0$, using (15) and the definition of $\Pi_j^e(s_j^T|0)$, inequality (16) can be rewritten as

$$\Pi_j(s_i^T, s_j^T|0) - \min_{s_i \in E_i(s_j^T|0)} \Pi_j(s_i, s_j^T|0) - \frac{1}{y_j} > \bar{\kappa}.$$

Since $\Delta_j \geq \Pi_j(s_i^T, s_j^T|0) - \min_{s_i \in E_i(s_j^T|0)} \Pi_j(s_i, s_j^T|0)$, this further implies $-1/y_j > \bar{\kappa} - \Delta_j$ and, by $y_j > 0$, again contradicts $1/y_j \leq \bar{\kappa} - \Delta_j$.

A.6 Proof of Proposition 3

Let $\Phi = [M_1, M_2, g]$ be an arbitrary mechanism with a BNFE s that results in an SCF f. We can then construct a mechanism $\Phi'(\delta)$ based on Φ in the same way as we did in the proof of Theorem 2 based on the direct mechanism (see Step 1 in Appendix A.5 for the details). In short, $\Phi'(\delta)$ has message sets $M_i' = M_i \times \{0,1\}$, so any $m_i = (m_i^1, m_i^2) \in M_i'$ consists of a message $m_i^1 \in M_i$ from Φ and a number $m_i^2 \in \{0,1\}$. The outcome function g' of $\Phi'(\delta)$ is

$$q_i^{g'}(m) = q_i^g(m_1^1, m_2^1)$$

and

$$t_i^{g'}(m) = t_i^g(m_1^1, m_2^1) + m_i^2\delta_{ii} - m_j^2\delta_{ji}.$$

Mechanism $\Phi'(0)$ is equivalent to Φ. Observe, however, that Φ might already be an augmented revelation mechanism, possibly constructed from a direct mechanism in the exact same manner. We denote by s_i^T agent i's strategy in $\Phi'(\delta)$ given by $s_i^T(\theta_i) = (s_i(\theta_i), 0)$ for all $\theta_i \in \Theta_i$. The truth-telling interpretation becomes apparent if Φ is a (possibly augmented) revelation mechanism and s is the truth-telling strategy profile in Φ. Profile $s^T = (s_1^T, s_2^T)$ is a BNFE of $\Phi'(0)$ because s is a BNFE of Φ. The outcome of s^T in $\Phi'(\delta)$ is SCF f. Proceeding as in the proof of Theorem 2, we obtain

$$\Pi_i(s_i, s_i^b|\delta) = \Pi_i(s_i, s_i^b|0) + x_i(s_i)\delta_{ii} - x_j(s_i^b)\delta_{ji} \tag{17}$$

and

$$\Pi_i^e(s_i^T|\delta) = \Pi_i^e(s_i^T|0) - \delta_{ji}/2 \tag{18}$$

for both $i = 1, 2$, provided that $\operatorname{sgn}\delta_{jj} = \operatorname{sgn}\delta_{ji}$.

From now on suppose, for both $i = 1, 2$, that

$$0 \leq \Pi_i(s^T|0) - \Pi_i^e(s_i^T|0) < \bar{\kappa}, \tag{19}$$

which will be verified later, and let

$$\delta_{ij}^* = 2\left(\bar{\kappa} - \Pi_j(s^T|0) + \Pi_j^e(s_j^T|0)\right), \tag{20}$$

such that $0 < \delta_{ij}^* \leq 2\bar{\kappa}$. Let δ_{ii}^* by any value that satisfies $0 < \delta_{ii}^* \leq \delta_{ij}^*$, and consider the BNFE candidate s^T in $\Phi'(\delta^*)$. Agent i's correct belief about j's kindness is then

$$\kappa_j(s^T|\delta^*) = h(\Pi_i(s^T|\delta^*) - \Pi_i^e(s_i^T|\delta^*)) = h(\Pi_i(s^T|0) - \Pi_i^e(s_i^T|0) + \delta_{ji}^*/2) = \bar{\kappa},$$

where (17), (18) and (20) have been used. Agent i therefore chooses s_i so as to maximize

$$\Pi_i(s_i, s_j^T|\delta^*) + y_i\,\bar{\kappa}\,h(\Pi_j(s_i, s_j^T|\delta^*) - \Pi_j^e(s_j^T|\delta^*)).$$

Based on (17) and (18) this can be rewritten as

$$\Pi_i(s_i, s_j^T|0) + x_i(s_i)\delta_{ii}^* + y_i\,\bar{\kappa}\,h(\Pi_j(s_i, s_j^T|0) - x_i(s_i)\delta_{ij}^* - \Pi_j^e(s_j^T|0) + \delta_{ij}^*/2). \tag{21}$$

We now show that, for the two different cases in the proposition and appropriate choices of Φ and s, strategy $s_i = s_i^T$ maximizes (21) and thus s^T is a BNFE of $\Phi'(\delta^*)$ that implements f with mutual kindness of $\bar{\kappa}$.

Case (a). Suppose f is BIC and satisfies $\Delta_1 = \Delta_2 = 0$. Let Φ from above be the direct mechanism and s the truth-telling strategy profile, which is a BNFE of Φ as shown in the proof of Theorem 1. Also, $\Pi_i(s^T|0) - \Pi_i^e(s_i^T|0) = 0$ holds, which verifies (19) and implies $\delta_{ij}^* = 2\bar{\kappa}$, for both $i = 1, 2$. Then (21) can be further simplified to

$$\Pi_i(s_i, s_j^T|0) + x_i(s_i)\delta_{ii}^* + y_i\,\bar{\kappa}\,(\bar{\kappa} - x_i(s_i)2\bar{\kappa}), \tag{22}$$

because $\Pi_j(s_i, s_j^T|0) = \Pi_j^e(s_j^T|0)$ for all $s_i \in S_i$ due to $\Delta_j = 0$ as shown in the proof of Theorem 1, and the bounding function h can be omitted because $x_i(s_i) \in [0, 1]$. The first term in (22) is maximized by $s_i = s_i^T$ since f is BIC. The remainder of (22) is non-increasing in $x_i(s_i)$ whenever

$$\delta_{ii}^* \leq 2y_i\bar{\kappa}^2. \tag{23}$$

Strategy $s_i = s_i^T$, for which $x_i(s_i^T) = 0$, therefore maximizes (22) whenever δ_{ii}^* is chosen to also satisfy (23). Off-equilibrium budget balance $\delta_{ii}^* = \delta_{ij}^* = 2\bar{\kappa}$ is possible if and only if $\bar{\kappa} \geq 1/y_i$.

Case (b). Suppose f is materially Pareto-efficient and $y \in Y^f$. Let Φ from above be the augmented revelation mechanism constructed in the proof of Theorem 2 and s the truth-telling strategy profile, which is a BNFE of Φ as shown in the proof of Theorem 2 (to avoid confusion, observe that δ now describes the additional redistribution in the twice augmented mechanism $\Phi'(\delta)$, not the redistribution already possible in the once augmented mechanism Φ). Also, $\Pi_i(s^T|0) - \Pi_i^e(s_i^T|0) = 1/y_i$ holds. From $y \in Y^f$ it follows that $1/y_i \leq \bar{\kappa}$. Assume that in fact

$1/y_i < \bar{\kappa}$ for both $i = 1, 2$, since otherwise Φ does not have to be further augmented for the respective agent to achieve the desired kindness $\bar{\kappa}$. This verifies (19) and implies $\delta^*_{ij} = 2(\bar{\kappa} - 1/y_j)$, for both $i = 1, 2$.

For strategy $s_i = s_i^T$, (21) becomes

$$\Pi_i(s^T|0) + y_i\,\bar{\kappa}\,\bar{\kappa}.$$

To exclude profitable deviations, we can restrict attention to conditionally efficient strategies $s'_i \in E_i(s_j^T|\delta^*)$. Note that $E_i(s_j^T|\delta^*) \subseteq E_i(s_j^T|0)$, as shown in the proof of Theorem 2. We will verify that there are no profitable deviations in $E_i(s_j^T|0)$. Any $s'_i \in E_i(s_j^T|0)$ satisfies

$$-\bar{\kappa} < \Pi_j(s'_i, s_j^T|0) - \Pi_j^e(s_j^T|0) + \delta^*_{ij}/2 \tag{24}$$

for the given value of $\delta^*_{ij} > 0$, because $-\bar{\kappa} \leq \Pi_j(s'_i, s_j^T|0) - \Pi_j^e(s_j^T|0)$ according to Case (c) in the proof of Theorem 2. Deviations $s'_i \in E_i(s_j^T|0)$ such that $\Pi_i(s'_i, s_j^T|0) + x_i(s'_i)\delta^*_{ii} \leq \Pi_i(s^T|0)$ can clearly never be profitable. Deviations $s'_i \in E_i(s_j^T|0)$ with

$$\begin{aligned}
&\Pi_i(s'_i, s_j^T|0) + x_i(s'_i)\delta^*_{ii} > \Pi_i(s^T|0),\\
&\Pi_j(s'_i, s_j^T|0) - x_i(s'_i)\delta^*_{ij} \geq \Pi_j(s^T|0),
\end{aligned}$$

do not exist by efficiency of f. Hence denote by $\Sigma_i(\delta^*)$ the remaining set of $s'_i \in E_i(s_j^T|0)$ with

$$\begin{aligned}
&\Pi_i(s'_i, s_j^T|0) + x_i(s'_i)\delta^*_{ii} > \Pi_i(s^T|0),\\
&\Pi_j(s'_i, s_j^T|0) - x_i(s'_i)\delta^*_{ij} < \Pi_j(s^T|0).
\end{aligned}$$

Any $s'_i \in \Sigma_i(\delta^*)$ satisfies

$$\Pi_j(s'_i, s_j^T|0) - x_i(s'_i)\delta^*_{ij} - \Pi_j^e(s_j^T|0) + \delta^*_{ij}/2 < \bar{\kappa} \tag{25}$$

for the given value of δ^*_{ij}, because $\Pi_j(s'_i, s_j^T|0) - x_i(s'_i)\delta^*_{ij} - \Pi_j^e(s_j^T|0) < \Pi_j(s^T|0) - \Pi_j^e(s_j^T|0) = 1/y_j$ by definition, so that the upper kindness bound can henceforth be ignored. We now treat the subsets $\Sigma_i^0(\delta^*) = \{s_i \in \Sigma_i(\delta^*) \mid x_i(s_i) = 0\}$ and $\Sigma_i^+(\delta^*) = \{s_i \in \Sigma_i(\delta^*) \mid x_i(s_i) > 0\}$ separately.

For any $s'_i \in \Sigma_i^0(\delta^*)$, the lower kindness bound can also be ignored by (24). We claim that a deviation to any $s'_i \in \Sigma_i^0(\delta^*)$ cannot make agent i better off. By contradiction, assume that

$$\Pi_i(s'_i, s_j^T|0) + y_i\,\bar{\kappa}\,(\Pi_j(s'_i, s_j^T|0) - \Pi_j^e(s_j^T|0) + \delta^*_{ij}/2) > \Pi_i(s^T|0) + y_i\,\bar{\kappa}\,\bar{\kappa}.$$

This can be rearranged to

$$\Pi_i(s'_i, s_j^T|0) - \Pi_i(s^T|0) + y_i\,\bar{\kappa}\,(\Pi_j(s'_i, s_j^T|0) - \Pi_j^e(s_j^T|0) - 1/y_j) > 0.$$

The last term in brackets is negative, as argued before. Hence $y_i\,\bar{\kappa} > 1$ implies

$$\Pi_i(s'_i, s_j^T|0) - \Pi_i(s^T|0) + (\Pi_j(s'_i, s_j^T|0) - \Pi_j^e(s_j^T|0) - 1/y_j) > 0.$$

Substituting $1/y_j$ by $\Pi_j(s^T|0) - \Pi_j^e(s_j^T|0)$ and rearranging yields

$$\Pi_i(s_i', s_j^T|0) + \Pi_j(s_i', s_j^T|0) > \Pi_i(s^T|0) + \Pi_j(s^T|0),$$

which is a contradiction to efficiency of f.

For any $s_i' \in \Sigma_i^+(\delta^*)$, so that $x_i(s_i') > 0$, observe that

$$h(\Pi_j(s_i', s_j^T|0) - x_i(s_i')\delta_{ij}^* - \Pi_j^e(s_j^T|0) + \delta_{ij}^*/2) < h(\Pi_j(s_i', s_j^T|0) - \Pi_j^e(s_j^T|0) + \delta_{ij}^*/2),$$

because the upper bound $\bar{\kappa}$ is not binding on the LHS by (25), and the lower bound $-\bar{\kappa}$ is not binding on the RHS by (24). Let $\bar{s}_i$ be the strategy with $\bar{s}_i^1(\theta_i) = s_i'(\theta_i)$ and $\bar{s}_i^2(\theta_i) = 0$ for all $\theta_i \in \Theta_i$. For sufficiently small but strictly positive values of δ_{ii}^* it then follows that

$$\begin{aligned}
&\Pi_i(s_i', s_j^T|0) + x_i(s_i')\delta_{ii}^* + y_i\,\bar{\kappa}\,h(\Pi_j(s_i', s_j^T|0) - x_i(s_i')\delta_{ij}^* - \Pi_j^e(s_j^T|0) + \delta_{ij}^*/2) \\
&\leq \Pi_i(s_i', s_j^T|0) + y_i\,\bar{\kappa}\,h(\Pi_j(s_i', s_j^T|0) - \Pi_j^e(s_j^T|0) + \delta_{ij}^*/2) \\
&= \Pi_i(\bar{s}_i, s_j^T|0) + y_i\,\bar{\kappa}\,h(\Pi_j(\bar{s}_i, s_j^T|0) - \Pi_j^e(s_j^T|0) + \delta_{ij}^*/2).
\end{aligned}$$

Observe that $\bar{s}_i \in E_i(s_j^T|0)$, because $\bar{s}_i$ and s_i' are payoff equivalent in $\Phi'(0)$ and $s_i' \in E_i(s_j^T|0)$. Observe also that $\bar{s}_i \notin \Sigma_i^+(\delta^*)$, because $x_i(\bar{s}_i) > 0$. Hence $\bar{s}_i$ cannot be a profitable deviation by our previous arguments, so that s_i' cannot be a profitable deviation either. Since $\Sigma_i^+(\delta^*)$ is finite and weakly shrinking (in the set inclusion sense) as δ_{ii}^* comes smaller, δ_{ii}^* can be chosen small enough to render all deviations unprofitable.

A.7 Proof of Proposition 4

Let f be a y-independent SCF and let $\Phi = [M_1, M_2, g]$ be a mechanism that implements f in a BNFE $s^* = (s_1^*, s_2^*)$. The best-response condition for agent i of reciprocity type $\underline{y}_i = 0$ can be rewritten as

$$s_{i,\underline{y}_i}^* \in \arg\max_{s_{i,y_i} \in S_i} \bar{\Pi}_i(s_{i,y_i}, s_j^*).$$

In particular, standard arguments imply that

$$\begin{aligned}
&\mathbb{E}[v_i(q_i^g(s_{i,\underline{y}_i}^*(\theta_i), s_{j,\tilde{y}_j}^*(\tilde{\theta}_j)), \theta_i) + t_i^g(s_{i,\underline{y}_i}^*(\theta_i), s_{j,\tilde{y}_j}^*(\tilde{\theta}_j))] \\
&\geq \mathbb{E}[v_i(q_i^g(s_{i,y_i'}^*(\theta_i'), s_{j,\tilde{y}_j}^*(\tilde{\theta}_j)), \theta_i) + t_i^g(s_{i,y_i'}^*(\theta_i'), s_{j,\tilde{y}_j}^*(\tilde{\theta}_j))]
\end{aligned}$$

must hold for all $\theta_i, \theta_i' \in \Theta$ and all $y_i' \in Y_i$. Since s^* implements f in mechanism Φ, we have

$$g(s_{1,y_1}^*(\theta_1), s_{2,y_2}^*(\theta_2)) = \hat{f}(\theta)$$

for all (y, θ). The above inequality can therefore be rewritten as

$$\begin{aligned}
&\mathbb{E}[v_i(q_i^{\hat{f}}(\theta_i, \tilde{\theta}_j)), \theta_i) + t_i^{\hat{f}}(\theta_i, \tilde{\theta}_j)] \\
&\geq \mathbb{E}[v_i(q_i^{\hat{f}}(\theta_i', \tilde{\theta}_j)), \theta_i) + t_i^{\hat{f}}(\theta_i', \tilde{\theta}_j)]
\end{aligned}$$

for all $\theta_i, \theta_i' \in \Theta$. Thus $\hat{f}$ is BIC, which implies that f is BIC.

A.8 Proof of Proposition 5

We first establish two lemmas, which we will use subsequently to prove the proposition. We first show that, when implementing a y-independent social choice function in a situation with $\underline{y} = (0,0)$, all reciprocity types behave like the selfish type and choose a strategy that maximizes their expected material payoff. As a consequence, equilibrium kindness cannot be positive.

Lemma 4. *Suppose $\underline{y} = (0,0)$, and let $\Phi = [M_1, M_2, g]$ be a mechanism that implements a y-independent SCF f in a BNFE $s^* = (s_1^*, s_2^*)$. Then, for both $i = 1,2$ and all $y_i \in Y_i$,*

(i) $s_{i,y_i}^* \in \arg\max_{s_{i,y_i} \in S_i} \bar{\Pi}_i(s_{i,y_i}, s_j^*)$, *and*

(ii) $\kappa_i(s_{i,y_i}^*, s_j^*) \leq 0$.

Proof. Property (i). Suppose by contradiction that there exists an agent i and a type y_i such that $s_{i,y_i}^* \notin \arg\max_{s_{i,y_i} \in S_i} \bar{\Pi}_i(s_{i,y_i}, s_j^*)$, which implies that

$$\begin{aligned}&\mathbb{E}[v_i(q_i^g(s_{i,y_i}^*(\theta_i), s_{j,\tilde{y}_j}^*(\tilde{\theta}_j)), \theta_i) + t_i^g(s_{i,y_i}^*(\theta_i), s_{j,\tilde{y}_j}^*(\tilde{\theta}_j))]\\ &< \mathbb{E}[v_i(q_i^g(m_i, s_{j,\tilde{y}_j}^*(\tilde{\theta}_j)), \theta_i) + t_i^g(m_i, s_{j,\tilde{y}_j}^*(\tilde{\theta}_j))]\end{aligned}$$

for some θ_i and m_i. Since Φ implements the y-independent SCF f we know that

$$g(s_{i,y_i}^*(\theta_i), s_{j,y_j}^*(\theta_j)) = \hat{f}(\theta) = g(s_{i,0}^*(\theta_i), s_{j,y_j}^*(\theta_j))$$

for all (y, θ). Hence the above inequality can be written as

$$\begin{aligned}&\mathbb{E}[v_i(q_i^g(s_{i,0}^*(\theta_i), s_{j,\tilde{y}_j}^*(\tilde{\theta}_j)), \theta_i) + t_i^g(s_{i,0}^*(\theta_i), s_{j,\tilde{y}_j}^*(\tilde{\theta}_j))]\\ &< \mathbb{E}[v_i(q_i^g(m_i, s_{j,\tilde{y}_j}^*(\tilde{\theta}_j)), \theta_i) + t_i^g(m_i, s_{j,\tilde{y}_j}^*(\tilde{\theta}_j))],\end{aligned}$$

which contradicts the best-response condition for reciprocity type $\underline{y}_i = 0$ of agent i, and hence contradicts the assumption that s^* is a BNFE.

Property (ii). Consider any $i = 1,2$ and any $y_i \in Y_i$. From property (i) it follows that $\bar{\Pi}_i(s_{i,y_i}^*, s_j^*) \geq \bar{\Pi}_i(s_{i,y_i}, s_j^*)$ for all $s_{i,y_i} \in E_i(s_j^*)$. Pareto-efficiency then implies that $\bar{\Pi}_j(s_{i,y_i}^*, s_j^*) \leq \bar{\Pi}_j(s_{i,y_i}, s_j^*)$ for all $s_{i,y_i} \in E_i(s_j^*)$. This implies $\bar{\Pi}_j(s_{i,y_i}^*, s_j^*) \leq \Pi_j^e(s_j^*)$, from which property (ii) follows. □

The fact that all types behave in a selfish way can be used to show that there is no longer a role for unused actions. Put differently, the revelation principle holds. In the direct mechanism $\Phi^d = [Y_1 \times \Theta_1, Y_2 \times \Theta_2, f]$, revealing one's type now involves to reveal both the reciprocity type y_i and the payoff type θ_i.

Lemma 5. *Suppose $\underline{y} = (0,0)$, and suppose that a mechanism Φ implements a y-independent SCF f in BNFE. Then f is truthfully implementable in BNFE in the direct mechanism.*

Proof. Step 1. From the arguments in Section 5 it follows that an augmented revelation principle applies, i.e., it is without loss of generality to consider a mechanism $\Phi = [M_1, M_2, g]$ where

$M_i \supseteq Y_i \times \Theta_i$ for both $i = 1, 2$ and $g(m) = f(m)$ whenever $m \in Y \times \Theta$, and in which the truthful strategy profile s^T is a BNFE. It remains to be shown that s^T is still a BNFE even when we eliminate all unused actions from Φ.

Step 2. Consider agent $i = 1$ (the argument for $i = 2$ is identical). Construct a mechanism $\Phi' = [M_1', M_2', g']$ from Φ by letting $M_1' = Y_1 \times \Theta_1$, keeping $M_2' = M_2$ unchanged, and letting g' be the restriction of g to $M_1' \times M_2'$. We have only (if at all) removed unused actions for agent 1, so that s^T is still an admissible strategy profile and the kindness of agent 2 is unchanged, i.e., $\kappa_2'(s^T_{2,y_2}, s^T_1) = \kappa_2(s^T_{2,y_2}, s^T_1)$ for all $y_2 \in Y_2$ (where the prime $'$ indicates terms in mechanism Φ'). We claim that $\kappa_1'(s^T_{1,y_1}, s^T_2) \geq \kappa_1(s^T_{1,y_1}, s^T_2)$ for all $y_1 \in Y_1$, i.e., agent 1's kindness has weakly increased. To prove the claim we show that $\Pi_2^{e'}(s_2^T) \leq \Pi_2^e(s_2^T)$. Consider first the minimization part in the definition of equitable payoffs. To obtain a contradiction, assume that

$$\min_{s_{1,y_1} \in E_1'(s_2^T)} \bar{\Pi}_2(s_{1,y_1}, s_2^T) > \min_{s_{1,y_1} \in E_1(s_2^T)} \bar{\Pi}_2(s_{1,y_1}, s_2^T).$$

Let $\hat{s}_{1,y_1}$ be a strategy in $E_1(s_2^T)$ that achieves the minimum in Φ, and analogously let $\hat{s}'_{1,y_1}$ be a strategy in $E_1'(s_2^T)$ that achieves the minimum in Φ'. From the definition of Pareto-efficiency it then follows that $\hat{s}_{1,y_1}$ maximizes $\bar{\Pi}_1(s_{1,y_1}, s_2^T)$ on S_1, and $\hat{s}'_{1,y_1}$ maximizes $\bar{\Pi}_1(s_{1,y_1}, s_2^T)$ on S_1'. Observe that $s^T_{1,y_1} \in \arg\max_{s_{1,y_1} \in S_1} \bar{\Pi}_1(s_{1,y_1}, s_2^T)$ also holds, by Lemma 4 and the fact that s^T is a BNFE that implements f in Φ. This implies

$$\bar{\Pi}_1(\hat{s}_{1,y_1}, s_2^T) = \bar{\Pi}_1(s^T_{1,y_1}, s_2^T) = \bar{\Pi}_1(\hat{s}'_{1,y_1}, s_2^T),$$

where the second equality follows from $s^T_{1,y_1} \in S_1' \subseteq S_1$. Thus $\hat{s}'_{1,y_1} \in S_1' \subseteq S_1$ Pareto-dominates $\hat{s}_{1,y_1} \in S_1$, contradicting that $\hat{s}_{1,y_1} \in E_1(s_2^T)$. Hence

$$\min_{s_{1,y_1} \in E_1'(s_2^T)} \bar{\Pi}_2(s_{1,y_1}, s_2^T) \leq \min_{s_{1,y_1} \in E_1(s_2^T)} \bar{\Pi}_2(s_{1,y_1}, s_2^T)$$

must hold. Consider now the maximization part. The fact that

$$\max_{s_{1,y_1} \in E_1'(s_2^T)} \bar{\Pi}_2(s_{1,y_1}, s_2^T) \leq \max_{s_{1,y_1} \in E_1(s_2^T)} \bar{\Pi}_2(s_{1,y_1}, s_2^T)$$

follows immediately because in these problems we can replace the Pareto-efficient sets $E_1(s_2^T)$ and $E_1'(s_2^T)$ by S_1 and S_1', respectively, and $S_1' \subseteq S_1$ holds. This establishes our claim. It also follows from these arguments that $\kappa_1'(s^T_{1,y_1}, s_2^T) \leq 0$ must still be true in Φ'.

Step 3. From the previous step (applied to both agents) we know that the direct mechanism Φ^d for f generates kindness κ_i^d that satisfies $\kappa_i(s^T_{i,y_i}, s_j^T) \leq \kappa_i^d(s^T_{i,y_i}, s_j^T) \leq 0$ for both $i = 1, 2$ and all $y_i \in Y_i$, and hence $\bar{\kappa}_j(s^T) \leq \bar{\kappa}_j^d(s^T) \leq 0$ for both $j = 1, 2$. To show that s^T is indeed a BNFE in Φ^d, as it is in Φ, we show that

$$s^T_{i,y_i} \in \arg\max_{s_{i,y_i} \in S_i} \bar{\Pi}_i(s_{i,y_i}, s_j^T) + \hat{\kappa}\, y_i\, h(\bar{\Pi}_j(s_{i,y_i}, s_j^T) - \Pi_j^e(s_j^T)) \tag{26}$$

for some $\hat{\kappa} \leq 0$ implies that

$$s^T_{i,y_i} \in \arg \max_{s_{i,y_i} \in S_i} \bar{\Pi}_i(s_{i,y_i}, s^T_j) + \tilde{\kappa}\, y_i\, h(\bar{\Pi}_j(s_{i,y_i}, s^T_j) - \Pi^e_j(s^T_j))$$

for all $\tilde{\kappa} \in [\hat{\kappa}, 0]$. Condition (26) can be equivalently stated as follows: for all s'_{i,y_i} in S_i,

$$\bar{\Pi}_i(s^T_{i,y_i}, s^T_j) - \bar{\Pi}_i(s'_{i,y_i}, s^T_j) \geq \hat{\kappa}\, y_i \left[h(\bar{\Pi}_j(s'_{i,y_i}, s^T_j) - \Pi^e_j(s^T_j)) - h(\bar{\Pi}_j(s^T_{i,y_i}, s^T_j) - \Pi^e_j(s^T_j))\right].$$

It follows from Lemma 4 that the LHS of this inequality is non-negative. If the term in squared brackets on the RHS is also non-negative, then the RHS is non-positive and we have

$$\bar{\Pi}_i(s^T_{i,y_i}, s^T_j) - \bar{\Pi}_i(s'_{i,y_i}, s^T_j) \geq \tilde{\kappa}\, y_i \left[h(\bar{\Pi}_j(s'_{i,y_i}, s^T_j) - \Pi^e_j(s^T_j)) - h(\bar{\Pi}_j(s^T_{i,y_i}, s^T_j) - \Pi^e_j(s^T_j))\right]$$

for all $\tilde{\kappa} \leq 0$. If instead the term in squared brackets on the RHS is negative, then

$$\begin{aligned} \bar{\Pi}_i(s^T_{i,y_i}, s^T_j) - \bar{\Pi}_i(s'_{i,y_i}, s^T_j) &\geq \hat{\kappa}\, y_i \left[h(\bar{\Pi}_j(s'_{i,y_i}, s^T_j) - \Pi^e_j(s^T_j)) - h(\bar{\Pi}_j(s^T_{i,y_i}, s^T_j) - \Pi^e_j(s^T_j))\right] \\ &\geq \tilde{\kappa}\, y_i \left[h(\bar{\Pi}_j(s'_{i,y_i}, s^T_j) - \Pi^e_j(s^T_j)) - h(\bar{\Pi}_j(s^T_{i,y_i}, s^T_j) - \Pi^e_j(s^T_j))\right] \end{aligned}$$

for all $\tilde{\kappa} \in [\hat{\kappa}, 0]$. □

We are now in the position to prove Proposition 5. Suppose $\underline{y} = (0,0)$, and let f be a y-independent SCF with $\Delta_i > 0$ for both $i = 1, 2$. By Lemma 5 it is without loss of generality to consider the direct mechanism Φ^d for f. Suppose that the truth-telling strategy profile s^T is indeed a BNFE. By Lemma 4 we then know that $s^T_{i,y_i} \in \arg\max_{s_{i,y_i} \in S_i} \bar{\Pi}_i(s_{i,y_i}, s^T_j)$ for both $i = 1, 2$ and all $y_i \in Y_i$. As argued in the proof of Lemma 4 this implies $\bar{\Pi}_j(s^T_{i,y_i}, s^T_j) \leq \min_{s_{i,y_i} \in E_i(s^T_j)} \bar{\Pi}_j(s_{i,y_i}, s^T_j)$. The fact that $\Delta_j > 0$ implies

$$\bar{\Pi}_j(s^T_{i,y_i}, s^T_j) < \max_{s_{i,y_i} \in S_i} \bar{\Pi}_j(s_{i,y_i}, s^T_j) = \max_{s_{i,y_i} \in E_i(s^T_j)} \bar{\Pi}_j(s_{i,y_i}, s^T_j),$$

and hence $\bar{\Pi}_j(s^T_{i,y_i}, s^T_j) < \Pi^e_j(s^T_j)$ and $\kappa_i(s^T_{i,y_i}, s^T_j) < 0$ for both $i = 1, 2$ and all $y_i \in Y_i$.

For each $i = 1, 2$, choose a number $k_i > \max_{y_i \in Y_i} |\bar{\Pi}_j(s^T_{i,y_i}, s^T_j) - \Pi^e_j(s^T_j)|$, and then choose $k \geq \max\{k_1, k_2\}$. Whenever $\bar{\kappa} \geq k$, the kindness bound is not reached in the hypothetical BNFE by any agent of any type. Consider now the best-response problem of type y_i of agent i. She chooses s_{i,y_i} so as to maximize

$$\bar{\Pi}_i(s_{i,y_i}, s^T_j) + y_i\, \bar{\kappa}_j(s^T)\, h(\bar{\Pi}_j(s_{i,y_i}, s^T_j) - \Pi^e_j(s^T_j)), \tag{27}$$

where $\bar{\kappa}_j(s^T) = \mathbb{E}[\kappa_j(s^T_{j,\tilde{y}_j}, s^T_i)] < 0$ is treated as fixed. Let

$$\theta^{min}_i = \arg \min_{\theta'_i \in \Theta_i} \mathbb{E}[v_j(q^{\hat{f}}_j(\theta'_i, \tilde{\theta}_j), \tilde{\theta}_j) + t^{\hat{f}}_j(\theta'_i, \tilde{\theta}_j)],$$

and denote by s^{min}_{i,y_i} the strategy that announces $s^{min}_{i,y_i}(\theta_i) = (y_i, \theta^{min}_i)$ for all $\theta_i \in \Theta_i$. From $\Delta_j > 0$ it follows that $\bar{\Pi}_j(s^{min}_{i,y_i}, s^T_j) < \bar{\Pi}_j(s^T_{i,y_i}, s^T_j)$. This implies that there exists a number x_i so that the value of (27) is strictly larger for $s_{i,y_i} = s^{min}_{i,y_i}$ than for $s_{i,y_i} = s^T_{i,y_i}$ whenever $y_i \geq x_i$

(given that $\bar{\kappa} \geq k$), contradicting that s^T is a BNFE. Hence f is not implementable in BNFE when $\bar{y}_i \geq x_i$.

A.9 Proof of Proposition 6

Note that the assumptions of the proposition imply $\Delta_i < \bar{\kappa}$ and $\underline{y}_i \bar{\kappa} \geq 1$. Now consider the one-dimensional SCF $\hat{f}$ that is induced by f, as described before. We first follow step 1 in the proof of Theorem 2 to construct an augmented direct mechanism $\Phi(\delta)$ for $\hat{f}$. We will engineer the parameter vector δ in a way so that truth-telling ($s^T_{i,y_i}(\theta_i) = \theta_i$) becomes a BNFE. Since this behavior is y-independent, we skip all y_i-related notation and use precisely the notation from the proof of Theorem 2. The only difference is that, here, we will have to check the best-response conditions for all types $y_i \in Y_i$.

From the arguments in the proof of Theorem 2 it follows that the (y_i-independent) kindness of agent $i = 1, 2$ in the hypothetical truth-telling equilibrium is given by

$$\kappa_i(s_i^T, s_j^T|\delta) = h(\Pi_j(s_i^T, s_j^T|0) - \Pi_j^e(s_j^T|0) + \delta_{ij}/2).$$

As in the proof of Proposition 3, let

$$\delta^*_{ij} = 2(\bar{\kappa} - \Pi_j(s^T|0) + \Pi_j^e(s_j^T|0)),$$

where $\Delta_i < \bar{\kappa}$ implies $\delta^*_{ij} > 0$. Let δ^*_{ii} be any value that satisfies $0 < \delta^*_{ii} \leq \delta^*_{ij}$. We now obtain that $\kappa_i(s_i^T, s_j^T|\delta^*) = \bar{\kappa}$ holds. In the hypothetical BNFE, type y_i of agent i therefore maximizes

$$\Pi_i(s_i, s_j^T|0) + x_i(s_i)\delta^*_{ii} + y_i\,\bar{\kappa}\,h(\Pi_j(s_i, s_j^T|0) - \Pi_j(s_i^T, s_j^T|0) + \bar{\kappa} - x_i(s_i)\delta^*_{ij}), \tag{28}$$

where the definition of δ^*_{ij} has been substituted once. For strategy $s_i = s_i^T$ this expression becomes $\Pi_i(s_i^T, s_j^T|0) + y_i\,\bar{\kappa}\,h(\bar{\kappa})$. We now show that there are no strategies $s_i' \in S_i$ which yield a strictly larger value of (28) than this.

For any $s_i' \in S_i$, define

$$C_i(s_i') = \Pi_i(s_i', s_j^T|0) - \Pi_i(s_i^T, s_j^T|0) + x_i(s_i')\delta^*_{ii},$$

which captures the change in agent i's material payoff when switching from s_i^T to s_i'. Similarly,

$$C_j(s_i') = \Pi_j(s_i', s_j^T|0) - \Pi_j(s_i^T, s_j^T|0) - x_i(s_i')\delta^*_{ij}$$

is the corresponding change of the argument of function h in (28). Material Pareto-efficiency of $\hat{f}$ and $\delta^*_{ii} \leq \delta^*_{ij}$ imply that $C_i(s_i') + C_j(s_i') \leq 0$. Observe first that a strategy s_i' with $C_i(s_i') \leq 0$ can never be a profitable deviation from s_i^T, as it decreases i's own payoff in (28) and cannot increase psychological payoffs due to the binding upper bound $\bar{\kappa}$. We therefore need to consider only strategies with $C_i(s_i') > 0$ and thus $C_j(s_i') < 0$ (so that the upper kindness bound $\bar{\kappa}$ is always slack).

Among such strategies, consider first those which additionally satisfy $x_i(s_i') = 0$. From $-\bar{\kappa} < -\Delta_i \leq \Pi_j(s_i', s_j^T|0) - \Pi_j(s_i^T, s_j^T|0)$ it follows that the argument of function h in (28) is

strictly bounded away from the lower bound $-\bar{\kappa}$ across all those (finitely many) strategies. The bounding function h can therefore be omitted, and from $C_i(s_i')+C_j(s_i') \leq 0$ and $y_i\bar{\kappa} \geq \underline{y}_i\bar{\kappa} \geq 1$ it follows that no such strategy can be a profitable deviation from s_i^T. Now consider the strategies which satisfy $x_i(s_i') > 0$. The additional gain in own material payoff, compared to the otherwise identical strategy s_i'' with $x_i(s_i'') = 0$, is of size $x_i(s_i')\delta_{ii}^*$. From the above argument about strict slackness of the lower bound, it follows that there exists a number $k > 0$ such that the additional loss in psychological payoff is of size $y_i\,\bar{\kappa}\,\min\{x_i(s_i')\delta_{ij}^*, k\}$. Setting δ_{ii}^* small enough so that it also satisfies $\delta_{ii}^* \leq k$, in addition to $\delta_{ii}^* \leq \delta_{ij}^*$, ensures that none of these strategies can be a profitable deviation from s_i^T either.

B Many Agents

Extending the basic mechanism design framework to an arbitrary number n of agents is straightforward. We can then denote by s_{ij}^b agent i's belief about j's strategy, and write $s_i^b = (s_{ij}^b)_{j\neq i}$. Analogously, s_{ijk}^{bb} is agent i's belief about j's belief about k's strategy, and we also write $s_{ij}^{bb} = (s_{ijk}^{bb})_{k\neq j}$ and $s_i^{bb} = (s_{ij}^{bb})_{j\neq i}$. The psychological externalities between n agents could potentially be multilateral, but we follow the literature (e.g. Dufwenberg and Kirchsteiger, 2004) and assume for simplicity that kindness sensations arise only bilaterally. Hence the kindness that agent i experiences in her relation with agent j does not depend on the implications of j's behavior for some third agent k. Agent i's expected utility can then be stated as

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

Here, y_{ij} are (possibly relation-specific) kindness weights, $\kappa_{ij}(s_i, s_i^b) = h(\Pi_j(s_i, s_i^b) - \Pi_j^e(s_i^b))$ measures how kind i intends to be to j, and $\kappa_{ji}(s_{ij}^b, s_{ij}^{bb}) = h(\Pi_i(s_{ij}^b, s_{ij}^{bb}) - \Pi_i^e(s_{ij}^{bb}))$ is i's belief about the kindness intended by j. Equitable payoffs are determined according to

$$\Pi_j^e(s_i^b) = \frac{1}{2}\left[\max_{s_i\in E_{ij}(s_i^b)} \Pi_j(s_i, s_i^b) + \min_{s_i\in E_{ij}(s_i^b)} \Pi_j(s_i, s_i^b)\right],$$

where $E_{ij}(s_i^b)$ is the set of bilaterally Pareto-efficient strategies of agent i. We define a BNFE as a strategy profile s^* so that, for all agents i, (a) $s_i^* \in \operatorname{argmax}_{s_i\in S_i} U(s_i, s_i^b, s_i^{bb})$, (b) $s_i^b = s_{-i}^*$, and (c) $s_i^{bb} = (s_{-j}^*)_{j\neq i}$.

We first discuss how our results on strong implementability (Section 4) extend to this setting. Given an SCF f, let

$$\Delta_{ij} = \max_{\theta_j\in\Theta_j} \mathbb{E}[v_i(q_i^f(\tilde{\theta}_{-j}, \theta_j), \tilde{\theta}_i) + t_i^f(\tilde{\theta}_{-j}, \theta_j)] - \min_{\theta_j\in\Theta_j} \mathbb{E}[v_i(q_i^f(\tilde{\theta}_{-j}, \theta_j), \tilde{\theta}_i) + t_i^f(\tilde{\theta}_{-j}, \theta_j)]$$

be a measure of the maximal impact that j's type has on i's expected payoff. If the insurance property holds, which now requires $\Delta_{ij} = 0$ for all i and j, then no agent can unilaterally affect the expected payoff of any other agent in the direct mechanism. From the arguments developed earlier, it then follows that Theorem 1 can be extended: If f is BIC and satisfies the insurance property, then f is strongly implementable in BNFE.

For the case of two agents, Proposition 1 shows that the AGV mechanism satisfies the insurance property. This result does not generally extend to the case of n agents. It extends, however, under symmetry of expected externalities, which requires that, for each i and θ_i,

$$\mathbb{E}[v_j(q_j^f(\theta_i, \tilde{\theta}_{-i}), \tilde{\theta}_j)] = \mathbb{E}[v_k(q_k^f(\theta_i, \tilde{\theta}_{-i}), \tilde{\theta}_k)]$$

holds for all $j, k \neq i$. If all agents' expected consumption utilities are affected equally by agent i's type, so that the expected externalities are evenly distributed, then the AGV transfers once more guarantee the insurance property. Symmetry arises naturally if the environment is such that all agents have identical payoff functions, their types are identically distributed, and the consumption rule $(q_1^f, ..., q_n^f)$ treats them all equally. Proposition 2, by contrast, extends to the n agent setting with no further qualification. The construction of the strongly implementable version $\bar{f}$ of f is given by

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

Some of our results on weak implementability (Section 5) carry over to the n agent case in a straightforward way, others would require a more elaborate analysis that is beyond the scope of this paper. Our proof of the augmented revelation principle did not make use of arguments that are specific to the case of two agents, and hence continues to apply. Theorem 2 provides the sufficient condition $y \in Y^f$ for implementability of a materially Pareto-efficient SCF f in BNFE, where

$$Y^f = \{(y_1, y_2) \in \mathbb{R}_+^2 \mid y_i > 0 \text{ and } 1/y_i \leq \bar{\kappa} - \Delta_i \text{ for both } i = 1, 2\}.$$

If $\bar{\kappa} = \infty$, so that there are no exogenous bounds on the intensity of kindness sensations, the sufficient condition reduces to the requirement that both y_1 and y_2 are strictly positive. This statement continues to hold in the setting with n agents. If all kindness weights y_{ij} are strictly positive, then the proof of Theorem 2 can be generalized by introducing bilateral redistribution possibilities and calibrating them to support a truth-telling equilibrium. We conjecture that this logic extends to the case in which $\bar{\kappa} < \infty$, but we have to leave this question for future research. An extension would require a general characterization of the set Y^f for an environment with n agents. For this paper, this would lead us astray.

Proposition 3 provides two sufficient conditions for the possibility to implement an SCF f so that both agents experience a maximal kindness of $\bar{\kappa}$. The first one is that f is BIC and has the insurance property. This finding extends to the n agent case without complications. If $\Delta_{ij} = 0$ for all i and j, then we can, as in case (a) of the proof of Proposition 3, engineer kindness sensations of $\bar{\kappa}$ by means of side-transfers that will not take place in equilibrium. The second sufficient condition is that f is materially Pareto-efficient and $y \in Y^f$. An extension of this condition is more involved, because it would, again, require a general characterization of the set Y^f for an environment with n agents.

C Interim Fairness Equilibrium

Consider an environment E and a mechanism Φ. In this appendix, we develop the notion of an interim fairness equilibrium (IFE) and provide conditions under which a strategy profile s^* is an IFE if and only if it is a BNFE. We assume throughout that first- and second-order beliefs about strategies are not type-dependent. Since we require that beliefs are correct in IFE, this assumption is without loss of generality.

If type θ_i of agent i has belief s_i^b and chooses message m_i, this yields an expected material payoff which we denote by

$$\Pi_i^{int}(m_i, s_i^b|\theta_i) = \mathbb{E}[v_i(q_i^g(m_i, s_i^b(\tilde{\theta}_j)), \theta_i) + t_i^g(m_i, s_i^b(\tilde{\theta}_j))].$$

We denote by $\kappa_i^{int}(m_i, s_i^b|\theta_i)$ the kindness intended by type θ_i of agent i ex interim. Also, agent i forms a belief $\kappa_j^{int}(s_i^b(\theta_j), s_i^{bb}|\theta_j)$ about the interim kindness of any one type θ_j of the other agent. However, the type θ_j is privately observed by agent j. We therefore assume that i assesses the kindness intended by j according to the expected value of $\kappa_j^{int}(s_i^b(\theta_j), s_i^{bb}|\theta_j)$,

$$\bar{\kappa}_j^{int}(s_i^b, s_i^{bb}) = \mathbb{E}[\kappa_j^{int}(s_i^b(\tilde{\theta}_j), s_i^{bb}|\tilde{\theta}_j)].$$

Interim utility of type θ_i of agent i is then given by

$$U_i^{int}(m_i, s_i^b, s_i^{bb}|\theta_i) = \Pi_i^{int}(m_i, s_i^b|\theta_i) + y_i\, \kappa_i^{int}(m_i, s_i^b|\theta_i)\, \bar{\kappa}_j^{int}(s_i^b, s_i^{bb}).$$

Definition 5. *An IFE is a strategy profile $s^* = (s_1^*, s_2^*)$ such that, for both $i = 1, 2$,*

(a) $s_i^*(\theta_i) \in \arg\max_{m_i \in M_i} U_i^{int}(m_i, s_i^b, s_i^{bb}|\theta_i)$ *for all* $\theta_i \in \Theta_i$,

(b) $s_i^b = s_j^*$, *and*

(c) $s_i^{bb} = s_i^*$.

The following proposition states that, if kindness at the ex ante stage is equal to the expected value of kindness at the ex interim stage, then the concepts of IFE and BNFE are equivalent.

Proposition 7. *Suppose that, for both $i = 1, 2$, all $s_i \in S_i$, and all $s_i^b \in S_j$,*

$$\kappa_i(s_i, s_i^b) = \mathbb{E}[\kappa_i^{int}(s_i(\tilde{\theta}_i), s_i^b|\tilde{\theta}_i)]. \tag{29}$$

Then, s^* is an IFE if and only if it is a BNFE.

Proof. (29) implies that $\bar{\kappa}_j^{int}(s_i^b, s_i^{bb}) = \mathbb{E}[\kappa_j^{int}(s_i^b(\tilde{\theta}_j), s_i^{bb}|\tilde{\theta}_j)] = \kappa_j(s_i^b, s_i^{bb})$ and hence

$$U_i^{int}(m_i, s_i^b, s_i^{bb}|\theta_i) = \Pi_i^{int}(m_i, s_i^b|\theta_i) + y_i\, \kappa_i^{int}(m_i, s_i^b|\theta_i)\, \kappa_j(s_i^b, s_i^{bb}).$$

Thus,

$$\begin{aligned}\mathbb{E}[U_i^{int}(s_i(\tilde{\theta}_i), s_i^b, s_i^{bb}|\tilde{\theta}_i)] &= \mathbb{E}[\Pi_i^{int}(s_i(\tilde{\theta}_i), s_i^b|\tilde{\theta}_i)] + y_i\, \mathbb{E}[\kappa_i^{int}(s_i(\tilde{\theta}_i), s_i^b|\tilde{\theta}_i)]\, \kappa_j(s_i^b, s_i^{bb}) \\ &= \Pi_i(s_i, s_i^b) + y_i\, \kappa_i(s_i, s_i^b)\, \kappa_j(s_i^b, s_i^{bb}),\end{aligned}$$

and hence $U_i(s_i, s_i^b, s_i^{bb}) = \mathbb{E}[U_i^{int}(s_i(\tilde{\theta}_i), s_i^b, s_i^{bb}|\tilde{\theta}_i)]$. By standard arguments, since all types of agent i occur with positive probability, it then follows that $s_i^* \in \arg\max_{s_i \in S_i} U_i(s_i, s_i^b, s_i^{bb})$ if and only if $s_i^*(\theta_i) \in \arg\max_{m_i \in M_i} U_i^{int}(m_i, s_i^b, s_i^{bb}|\theta_i)$ for all $\theta_i \in \Theta_i$. □

We have not made assumptions on how the interim kindness intentions are determined. A conceivable way of modeling them is to proceed as in the body of the text, replacing all ex ante notions by their ex interim analogues. Then, there are two potential obstacles to verifying condition (29), i.e., to expressing κ_i as an expectation over the terms κ_i^{int}. First, the ex ante equitable payoff might not correspond to an expectation over the ex interim equitable payoffs, for instance because they are defined based on different sets of Pareto-efficient strategies/messages. Second, a tight kindness bound $\bar{\kappa}$ might become binding for some ex interim but not for the ex ante kindness term. In any case, the condition in Proposition 7 allows us to verify whether or not IFE and BNFE are equivalent.

D Proofs of Observations

D.1 Proof of Observation 1

Consider the bilateral trade example with parameters (5) and $5/2 < \bar{\kappa}$. In the direct mechanism for f^*, the set of strategies for agent i is $S_i = \{s_i^T, s_i^H, s_i^L, s_i^{-T}\}$, where s_i^T is truth-telling, s_i^H prescribes to announce the high type $\bar{\theta}_i$ whatever the true type, s_i^L prescribes to always announce the low type $\underline{\theta}_i$, and s_i^{-T} is the strategy of always lying, 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, for any y with $y_b > 0$ and/or $y_s > 0$. We proceed by contradiction and suppose that (s_b^T, s_s^T) is a BNFE for some such y. Beliefs are correct in the hypothetical equilibrium, which implies that $s_b^b = s_s^{bb} = s_s^T$ and $s_s^b = s_b^{bb} = s_b^T$.

The seller's equitable payoff. Given s_s^T, varying the buyer's strategies yields payoffs

$$\begin{aligned}
&\Pi_b(s_b^T, s_s^T) = 20, \;\; \Pi_s(s_b^T, s_s^T) = 20,\\
&\Pi_b(s_b^L, s_s^T) = 20, \;\; \Pi_s(s_b^L, s_s^T) = 15,\\
&\Pi_b(s_b^H, s_s^T) = 0, \;\; \Pi_s(s_b^H, s_s^T) = 25,\\
&\Pi_b(s_b^{-T}, s_s^T) = 0, \;\; \Pi_s(s_b^{-T}, s_s^T) = 20.
\end{aligned}$$

Inspection of these expressions reveals that s_b^{-T} is not conditionally Pareto-efficient, because a switch to s_b^T makes the buyer better off and leaves the seller unaffected. Similarly, s_b^L is not efficient, because a switch to s_b^T makes the seller better off and leaves the buyer unaffected. The remaining two strategies are efficient, so that the equitable payoff for the seller from the buyer's perspective is $\Pi_s^e(s_s^T) = 45/2$.

The buyer's equitable payoff. Given s_b^T, varying the seller's strategies yields

$$\begin{aligned}
&\Pi_b(s_b^T, s_s^T) = 20, \;\; \Pi_s(s_b^T, s_s^T) = 20,\\
&\Pi_b(s_b^T, s_s^L) = 25, \;\; \Pi_s(s_b^T, s_s^L) = 0,\\
&\Pi_b(s_b^T, s_s^H) = 15, \;\; \Pi_s(s_b^T, s_s^H) = 20,\\
&\Pi_b(s_b^T, s_s^{-T}) = 20, \;\; \Pi_s(s_b^T, s_s^{-T}) = 0.
\end{aligned}$$

Both s_s^{-T} and s_s^H are Pareto-dominated by s_s^T, while the other strategies are efficient. The equitable payoff for the buyer is therefore also $\Pi_b^e(s_b^T) = 45/2$.

Truth-telling is not a BNFE. In the hypothetical BNFE (s_b^T, s_s^T), we have $\kappa_b(s_s^b, s_s^{bb}) = \kappa_s(s_b^b, s_b^{bb}) = h(-5/2) = -5/2$. The buyer then prefers a deviation from s_b^T to s_b^L if and only if

$$\Pi_b(s_b^L, s_s^T) - \left(\frac{5y_b}{2}\right) h\left(\Pi_s(s_b^L, s_s^T) - \frac{45}{2}\right) > \Pi_b(s_b^T, s_s^T) - \left(\frac{5y_b}{2}\right) h\left(\Pi_s(s_b^T, s_s^T) - \frac{45}{2}\right).$$

If $y_b > 0$, this can be simplified to $h(-15/2) < h(-5/2)$, which is satisfied because $5/2 < \bar{\kappa}$. Hence (s_b^T, s_s^T) is not a BNFE. The analogous argument applies to the seller if $y_s > 0$.

D.2 Proof of Observation 2

We seek to show that (s_b^T, s_s^T) is not a BNFE in the direct mechanism for f^{**}. We again proceed by contradiction. Fix $(y_b, y_s) \in [0, \infty[^2$ and suppose that (s_b^T, s_s^T) is a BNFE. Beliefs are correct in the hypothetical equilibrium, which implies that $s_b^b = s_s^{bb} = s_s^T$ and $s_s^b = s_b^{bb} = s_b^T$.

The seller's equitable payoff. Given s_s^T, varying the buyer's strategies yields

$$\begin{aligned}
\Pi_b(s_b^T, s_s^T) &= \frac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) + \frac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) + \frac{1}{8}(\bar{\theta}_b - \bar{\theta}_s),\\
\Pi_s(s_b^T, s_s^T) &= \frac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) + \frac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) + \frac{1}{8}(\bar{\theta}_b - \bar{\theta}_s),\\
\Pi_b(s_b^L, s_s^T) &= \frac{1}{4}(\bar{\theta}_b - \underline{\theta}_s),\\
\Pi_s(s_b^L, s_s^T) &= \frac{1}{4}(\underline{\theta}_b - \underline{\theta}_s),\\
\Pi_b(s_b^H, s_s^T) &= \frac{1}{4}(\underline{\theta}_b - \underline{\theta}_s) + \frac{1}{4}(\underline{\theta}_b - \bar{\theta}_s),\\
\Pi_s(s_b^H, s_s^T) &= \frac{1}{4}(\bar{\theta}_b - \underline{\theta}_s) + \frac{1}{4}(\bar{\theta}_b - \bar{\theta}_s),\\
\Pi_b(s_b^{-T}, s_s^T) &= \frac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) + \frac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) + \frac{1}{8}(\bar{\theta}_b - \bar{\theta}_s) - \frac{1}{4}(\bar{\theta}_b - \underline{\theta}_b),\\
\Pi_s(s_b^{-T}, s_s^T) &= \frac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) + \frac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) + \frac{1}{8}(\bar{\theta}_b - \bar{\theta}_s).
\end{aligned}$$

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

$$\begin{aligned}
&\Pi_b(s_b^L, s_s^T) > \Pi_b(s_b^T, s_s^T) > \Pi_b(s_b^H, s_s^T),\\
&\Pi_s(s_b^L, s_s^T) < \Pi_s(s_b^T, s_s^T) < \Pi_s(s_b^H, s_s^T).
\end{aligned}$$

Now we can easily compute that, from the buyer's perspective, the equitable payoff for the seller is her payoff under truth-telling: $\Pi_s^e(s_s^T) = \Pi_s(s_b^T, s_s^T)$.

The buyer's equitable payoff. Given s_b^T, varying the seller's strategies yields

$$\begin{aligned}
\Pi_b(s_b^T, s_s^T) &= \frac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) + \frac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) + \frac{1}{8}(\bar{\theta}_b - \bar{\theta}_s),\\
\Pi_s(s_b^T, s_s^T) &= \frac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) + \frac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) + \frac{1}{8}(\bar{\theta}_b - \bar{\theta}_s),
\end{aligned}$$

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

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 equitable payoff for the buyer is then also $\Pi_b^e(s_b^T) = \Pi_b(s_b^T, s_s^T)$.

Truth-telling is not a BNFE. In the hypothetical BNFE (s_b^T, s_s^T) we have $\kappa_b(s_s^b, s_s^{bb}) = 0$. 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, since s_s^H yields a strictly larger payoff as shown above. Hence (s_b^T, s_s^T) is not a BNFE.

D.3 Proof of Observation 3

Consider the hypothetical truth-telling BNFE $s^T = (s_b^T, s_s^T)$ of Φ', in which beliefs are correct.

Equitable payoffs. Given s_s^T, any strategy s_b that announces $\underline{\underline{\theta}}_b$ yields the same payoff pairs as the strategy that announces $\underline{\theta}_b$ instead, except for the additional redistribution from the seller to the buyer. Since s_b^L maximizes $\Pi_b(s_b, s_s^T)$ and minimizes $\Pi_s(s_b, s_s^T)$ in the direct mechanism (see Appendix D.2), strategy $\underline{\underline{s}}_b$ with $\underline{\underline{s}}_b(\theta_b) = \underline{\underline{\theta}}_b$ 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 yields the payoffs

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

The efficient strategy which yields the highest payoff for the seller remains s_b^H. We can now immediately compute the equitable payoff $\Pi_s^e(s_s^T) = \Pi_s(s_b^T, s_s^T) - \delta_b/4$. A symmetric argument implies $\Pi_b^e(s_b^T) = \Pi_b(s_b^T, s_s^T) - \delta_s/4$.

Truth-telling becomes a BNFE. We now have $\kappa_b(s_s^b, s_s^{bb}) = h(\delta_b/4)$ and $\kappa_s(s_b^b, s_b^{bb}) = h(\delta_s/4)$ in the hypothetical truth-telling equilibrium. Suppose $\bar{\kappa} \geq \max\{1/y_b, 1/y_s\}$, and note that $y_b, y_s > 0$. Setting $\delta_b = 4/y_s$ and $\delta_s = 4/y_b$ then yields $\kappa_b(s_s^b, s_s^{bb}) = 1/y_s$ and $\kappa_s(s_b^b, s_b^{bb}) = 1/y_b$, so that the buyer maximizes

$$\Pi_b(s_b, s_s^T) + h(\Pi_s(s_b, s_s^T) - \Pi_s^e(s_s^T))$$

and the seller maximizes

$$\Pi_s(s_b^T, s_s) + h(\Pi_b(s_b^T, s_s) - \Pi_b^e(s_b^T)).$$

Suppose furthermore that

$$\bar{\kappa} \geq \max\left\{\max_{s_b \in S_b'} |\Pi_s(s_b, s_s^T) - \Pi_s^e(s_s^T)|, \max_{s_s \in S_s'} |\Pi_b(s_b^T, s_s) - \Pi_b^e(s_b^T)|\right\}.$$

Then the bound $\bar{\kappa}$ can be ignored in these problems, and both agents are maximizing the sum of expected material payoffs (given truth-telling of the other agent). Own truth-telling is a solution to these problems, because the SCF f^{**} that is realized in this case is efficient, i.e., it maximizes the sum of material payoffs for any (θ_b, θ_s). Hence s^T is a BNFE.

D.4 Proof of Observation 4

We assume $\bar{\kappa} = \infty$ throughout the proof, but it is straightforward to verify that the arguments continue to hold when $\bar{\kappa} < \infty$ is large enough. For agent i, denote by $\sigma_i^T \in S_i$ (and analogously σ_i^H, σ_i^L, and σ_i^{-T}) the y_i-type's strategy s_{i,y_i} that announces the material payoff type θ_i truthfully (and analogously always high, low, and falsely), but always announces the high reciprocity type $\bar{y}$. Consider the hypothetical truth-telling BNFE s^T of the direct mechanism for f^{***}, in which beliefs are correct.

The seller's equitable payoff. Given s_s^T, varying among the above strategies of the buyer yields the payoffs

$$\begin{aligned}
\bar{\Pi}_b(\sigma_b^T, s_s^T) &= (1-\epsilon)\Pi_b(\sigma_b^T, s_{s,\bar{y}}^T) + \epsilon\left[\frac{1}{4}(\bar{\theta}_b - \bar{\theta}_s) - \frac{1}{2}\delta_s\right],\\
\bar{\Pi}_s(\sigma_b^T, s_s^T) &= (1-\epsilon)\Pi_s(\sigma_b^T, s_{s,\bar{y}}^T) + \epsilon\left[\frac{1}{4}(\bar{\theta}_b - \underline{\theta}_s) + \frac{1}{2}\delta_s\right],\\
\bar{\Pi}_b(\sigma_b^L, s_s^T) &= (1-\epsilon)\Pi_b(\sigma_b^L, s_{s,\bar{y}}^T),\\
\bar{\Pi}_s(\sigma_b^L, s_s^T) &= (1-\epsilon)\Pi_s(\sigma_b^L, s_{s,\bar{y}}^T),\\
\bar{\Pi}_b(\sigma_b^H, s_s^T) &= (1-\epsilon)\Pi_b(\sigma_b^H, s_{s,\bar{y}}^T) + \epsilon\left[\frac{1}{2}(\underline{\theta}_b - \bar{\theta}_s) - \delta_s\right],\\
\bar{\Pi}_s(\sigma_b^H, s_s^T) &= (1-\epsilon)\Pi_s(\sigma_b^H, s_{s,\bar{y}}^T) + \epsilon\left[\frac{1}{2}(\bar{\theta}_b - \underline{\theta}_s) + \delta_s\right],\\
\bar{\Pi}_b(\sigma_b^{-T}, s_s^T) &= (1-\epsilon)\Pi_b(\sigma_b^{-T}, s_{s,\bar{y}}^T) + \epsilon\left[\frac{1}{4}(\underline{\theta}_b - \bar{\theta}_b) + \frac{1}{4}(\underline{\theta}_b - \bar{\theta}_s) - \frac{1}{2}\delta_s\right],\\
\bar{\Pi}_s(\sigma_b^{-T}, s_s^T) &= (1-\epsilon)\Pi_s(\sigma_b^{-T}, s_{s,\bar{y}}^T) + \epsilon\left[\frac{1}{4}(\bar{\theta}_b - \underline{\theta}_s) + \frac{1}{2}\delta_s\right].
\end{aligned}$$

The expressions on the RHS that are weighted by $(1-\epsilon)$ correspond exactly to the payoffs derived in the proof of Observation 2. The additional terms that are weighted by ϵ arise because the seller sometimes announces (truthfully) to be selfish, i.e., they correspond to the payoffs from playing against $s_{s,0}^T$. Inspection of these payoffs, using the values derived in the proof of Observation 2, reveals that σ_b^{-T} is not conditionally Pareto-efficient, because a switch to σ_b^T

makes the buyer better off and leaves the seller unaffected, irrespective of the values of $\epsilon, \delta_s > 0$. For the remaining three strategies, there exists a critical value $\bar{\epsilon}_b > 0$ such that

$$\bar{\Pi}_b(\sigma_b^L, s_s^T) > \bar{\Pi}_b(\sigma_b^T, s_s^T) > \bar{\Pi}_b(\sigma_b^H, s_s^T),$$
$$\bar{\Pi}_s(\sigma_b^L, s_s^T) < \bar{\Pi}_s(\sigma_b^T, s_s^T) < \bar{\Pi}_s(\sigma_b^H, s_s^T),$$

whenever $\epsilon < \bar{\epsilon}_b$, irrespective of the value of $\delta_s > 0$ (this critical value arises for the comparison of $\bar{\Pi}_b(\sigma_b^L, s_s^T)$ and $\bar{\Pi}_b(\sigma_b^T, s_s^T)$; all other comparisons are unambiguous). The strategies not yet considered sometimes announce the low reciprocity type $y_b = 0$. They yield the same payoff pairs as the (above considered) strategy that announces $(\bar{y}, \underline{\theta}_b)$ instead, except for the additional redistribution from the seller to the buyer. Since σ_b^L maximizes $\bar{\Pi}_b(s_{b,y_b}, s_s^T)$ and minimizes $\bar{\Pi}_s(s_{b,y_b}, s_s^T)$ on $\{\sigma_b^T, \sigma_b^L, \sigma_b^H\}$, provided $\epsilon < \bar{\epsilon}_b$, the strategy σ_b^{LL} with $\sigma_b^{LL}(\theta_b) = (0, \underline{\theta}_b)$ maximizes $\bar{\Pi}_b(s_{b,y_b}, s_s^T)$ and minimizes $\bar{\Pi}_s(s_{b,y_b}, s_s^T)$ on S_b and hence is efficient. It yields

$$\bar{\Pi}_s(\sigma_b^{LL}, s_s^T) = (1-\epsilon)\left[\frac{1}{4}(\underline{\theta}_b - \underline{\theta}_s) - \frac{1}{2}\delta_b\right].$$

The efficient strategy which yields the highest payoff for the seller remains σ_b^H. We can now compute the equitable payoff for the case when $\epsilon < \bar{\epsilon}_b$:

$$\Pi_s^e(s_s^T) = \bar{\Pi}_s(\sigma_b^T, s_s^T) - (1-\epsilon)\frac{1}{4}\delta_b.$$

The buyer's equitable payoff. Given s_b^T, varying the seller's strategies yields

$$\bar{\Pi}_b(s_b^T, \sigma_s^T) = (1-\epsilon)\Pi_b(s_{b,\bar{y}}^T, \sigma_s^T) + \epsilon\left[\frac{1}{4}(\bar{\theta}_b - \underline{\theta}_s) + \frac{1}{2}\delta_b\right],$$
$$\bar{\Pi}_s(s_b^T, \sigma_s^T) = (1-\epsilon)\Pi_s(s_{b,\bar{y}}^T, \sigma_s^T) + \epsilon\left[\frac{1}{4}(\underline{\theta}_b - \underline{\theta}_s) - \frac{1}{2}\delta_b\right],$$
$$\bar{\Pi}_b(s_b^T, \sigma_s^L) = (1-\epsilon)\Pi_b(s_{b,\bar{y}}^T, \sigma_s^L) + \epsilon\left[\frac{1}{2}(\bar{\theta}_b - \underline{\theta}_s) + \delta_b\right],$$
$$\bar{\Pi}_s(s_b^T, \sigma_s^L) = (1-\epsilon)\Pi_s(s_{b,\bar{y}}^T, \sigma_s^L) + \epsilon\left[\frac{1}{2}(\underline{\theta}_b - \bar{\theta}_s) - \delta_b\right],$$
$$\bar{\Pi}_b(s_b^T, \sigma_s^H) = (1-\epsilon)\Pi_b(s_{b,\bar{y}}^T, \sigma_s^H),$$
$$\bar{\Pi}_s(s_b^T, \sigma_s^H) = (1-\epsilon)\Pi_s(s_{b,\bar{y}}^T, \sigma_s^H),$$
$$\bar{\Pi}_b(s_b^T, \sigma_s^{-T}) = (1-\epsilon)\Pi_b(s_{b,\bar{y}}^T, \sigma_s^{-T}) + \epsilon\left[\frac{1}{4}(\bar{\theta}_b - \underline{\theta}_s) + \frac{1}{2}\delta_b\right],$$
$$\bar{\Pi}_s(s_b^T, \sigma_s^{-T}) = (1-\epsilon)\Pi_s(s_{b,\bar{y}}^T, \sigma_s^{-T}) + \epsilon\left[\frac{1}{4}(\underline{\theta}_s - \bar{\theta}_s) + \frac{1}{4}(\underline{\theta}_b - \bar{\theta}_s) - \frac{1}{2}\delta_b\right],$$

where the expressions weighted by $(1-\epsilon)$ are again those from the proof of Observation 2. Proceeding as before we can show that σ_s^{-T} is not conditionally Pareto-efficient, while there exists a critical value $\bar{\epsilon}_s > 0$ such that

$$\bar{\Pi}_b(s_b^T, \sigma_s^L) > \bar{\Pi}_b(s_b^T, \sigma_s^T) > \bar{\Pi}_b(s_b^T, \sigma_s^H),$$

$$\bar{\Pi}_s(s_b^T, \sigma_s^L) < \bar{\Pi}_s(s_b^T, \sigma_s^T) < \bar{\Pi}_s(s_b^T, \sigma_s^H),$$

whenever $\epsilon < \bar{\epsilon}_s$, irrespective of the value of $\delta_b > 0$. The payoffs from the not yet considered strategies can again be derived from these expressions with an additional redistribution from the buyer to the seller. It follows that σ_s^{LH} with $\sigma_s^{LH}(\theta_s) = (0, \bar{\theta}_s)$ maximizes $\bar{\Pi}_s(s_b^T, s_{s,y_s})$ and minimizes $\bar{\Pi}_b(s_b^T, s_{s,y_s})$ on S_s and hence is efficient, provided $\epsilon < \bar{\epsilon}_s$. It yields

$$\bar{\Pi}_b(s_b^T, \sigma_s^{LH}) = (1-\epsilon)\left[\frac{1}{4}(\bar{\theta}_b - \bar{\theta}_s) - \frac{1}{2}\delta_s\right].$$

The efficient strategy which yields the highest payoff for the buyer remains σ_b^L. We can now compute the equitable payoff for the case when $\epsilon < \bar{\epsilon}_s$:

$$\Pi_b^e(s_b^T) = \bar{\Pi}_b(s_b^T, \sigma_s^T) - (1-\epsilon)\frac{1}{4}\delta_s.$$

The buyer's equilibrium kindness. In the hypothetical BNFE s^T when $\epsilon < \bar{\epsilon}_b$, we obtain for the buyer with reciprocity type $y_b = \bar{y}$ a kindness of

$$\kappa_b(s_{b,\bar{y}}^T, s_s^T) = \bar{\Pi}_s(s_{b,\bar{y}}^T, s_s^T) - \Pi_s^e(s_s^T) = \bar{\Pi}_s(\sigma_b^T, s_s^T) - \Pi_s^e(s_s^T) = (1-\epsilon)\frac{1}{4}\delta_b.$$

For the buyer with reciprocity type $y_b = 0$ we obtain

$$\begin{aligned}\kappa_b(s_{b,0}^T, s_s^T) &= \bar{\Pi}_s(s_{b,0}^T, s_s^T) - \Pi_s^e(s_s^T)\\ &= \bar{\Pi}_s(\sigma_b^L, s_s^T) - (1-\epsilon)\frac{1}{2}\delta_b - \Pi_s^e(s_s^T)\\ &= (1-\epsilon)\left[\frac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) - \frac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) - \frac{1}{8}(\bar{\theta}_b - \bar{\theta}_s) - \frac{1}{4}\delta_b\right] - \epsilon\left[\frac{1}{4}(\bar{\theta}_b - \underline{\theta}_s) + \frac{1}{2}\delta_s\right].\end{aligned}$$

The seller's equilibrium expectation about these terms then becomes

$$\begin{aligned}\kappa_b(s_s^b, s_s^{bb}) &= (1-\epsilon)\kappa_b(s_{b,\bar{y}}^T, s_s^T) + \epsilon\,\kappa_b(s_{b,0}^T, s_s^T)\\ &= (1-\epsilon)(1-2\epsilon)\frac{1}{4}\delta_b - \epsilon^2\frac{1}{2}\delta_s + \Lambda_b(\epsilon),\end{aligned}$$

where

$$\Lambda_b(\epsilon) = (1-\epsilon)\epsilon\left[\frac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) - \frac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) - \frac{1}{8}(\bar{\theta}_b - \bar{\theta}_s)\right] - \epsilon^2\left[\frac{1}{4}(\bar{\theta}_b - \underline{\theta}_s)\right]$$

collects terms that do not depend on δ_b or δ_s. $\Lambda_b(\epsilon)$ is continuous in ϵ and $\lim_{\epsilon\to 0}\Lambda_b(\epsilon) = 0$.

The seller's equilibrium kindness. Proceeding analogously for the seller, for $\epsilon < \bar{\epsilon}_s$ we obtain

$$\kappa_s(s_{s,\bar{y}}^T, s_b^T) = (1-\epsilon)\frac{1}{4}\delta_s$$

and

$$\kappa_s(s_{s,0}^T, s_b^T) = (1-\epsilon)\left[\frac{1}{8}(\bar{\theta}_b - \bar{\theta}_s) - \frac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) - \frac{1}{8}(\bar{\theta}_b - \underline{\theta}_s) - \frac{1}{4}\delta_s\right] - \epsilon\left[\frac{1}{4}(\bar{\theta}_b - \underline{\theta}_s) + \frac{1}{2}\delta_b\right],$$

about which the buyer forms the expectation

$$\begin{aligned}\kappa_s(s_b^b, s_b^{bb}) &= (1-\epsilon)\kappa_s(s_{s,\bar{y}}^T, s_b^T) + \epsilon\,\kappa_s(s_{s,0}^T, s_b^T)\\ &= (1-\epsilon)(1-2\epsilon)\frac{1}{4}\delta_s - \epsilon^2\frac{1}{2}\delta_b + \Lambda_s(\epsilon),\end{aligned}$$

where

$$\Lambda_s(\epsilon) = (1-\epsilon)\epsilon\left[\frac{1}{8}(\bar{\theta}_b - \bar{\theta}_s) - \frac{1}{8}(\underline{\theta}_b - \underline{\theta}_s) - \frac{1}{8}(\bar{\theta}_b - \underline{\theta}_s)\right] - \epsilon^2\left[\frac{1}{4}(\bar{\theta}_b - \underline{\theta}_s)\right].$$

Note again that $\Lambda_s(\epsilon)$ is continuous and that $\lim_{\epsilon\to 0}\Lambda_s(\epsilon) = 0$.

Truth-telling becomes a BNFE. We now assume that $\epsilon < \min\{\bar{\epsilon}_b, \bar{\epsilon}_s\}$. We then want to choose δ_b and δ_s such that $\kappa_b(s_s^b, s_s^{bb}) = 1/\bar{y}$ and $\kappa_s(s_b^b, s_b^{bb}) = 1/\bar{y}$ in the hypothetical truth-telling equilibrium. This system of equations can be written in matrix form as $A\delta = z$, where

$$A = \begin{pmatrix} (1-\epsilon)(1-2\epsilon)/4 & -\epsilon^2/2 \\ -\epsilon^2/2 & (1-\epsilon)(1-2\epsilon)/4 \end{pmatrix} \quad \delta = \begin{pmatrix} \delta_b \\ \delta_s \end{pmatrix} \quad z = \begin{pmatrix} 1/\bar{y} - \Lambda_b(\epsilon) \\ 1/\bar{y} - \Lambda_s(\epsilon) \end{pmatrix}.$$

We have $\det A = (1-\epsilon)^2(1-2\epsilon)^2/16 - \epsilon^4/4$, which is continuous and strictly decreasing in ϵ, takes the value zero for $\epsilon = 1/3$, and satisfies $\lim_{\epsilon\to 0}\det A = 1/16$. In particular, the system has a unique solution whenever $\epsilon < 1/3$, which we also assume from now on. To apply Cramer's rule, define

$$A_b = \begin{pmatrix} 1/\bar{y} - \Lambda_b(\epsilon) & -\epsilon^2/2 \\ 1/\bar{y} - \Lambda_s(\epsilon) & (1-\epsilon)(1-2\epsilon)/4 \end{pmatrix} \quad A_s = \begin{pmatrix} (1-\epsilon)(1-2\epsilon)/4 & 1/\bar{y} - \Lambda_b(\epsilon) \\ -\epsilon^2/2 & 1/\bar{y} - \Lambda_s(\epsilon) \end{pmatrix},$$

from which we can obtain $\det A_b = (1-\epsilon)(1-2\epsilon)\left[1/\bar{y} - \Lambda_b(\epsilon)\right]/4 + \epsilon^2\left[1/\bar{y} - \Lambda_s(\epsilon)\right]/2$ and $\det A_s = (1-\epsilon)(1-2\epsilon)\left[1/\bar{y} - \Lambda_s(\epsilon)\right]/4 + \epsilon^2\left[1/\bar{y} - \Lambda_b(\epsilon)\right]/2$. These terms are continuous in ϵ and satisfy $\lim_{\epsilon\to 0}\det A_b = \lim_{\epsilon\to 0}\det A_s = 1/4\bar{y}$. Hence, for ϵ small, enough we obtain well-defined solutions $\delta_b = \det A_b/\det A > 0$ and $\delta_s = \det A_s/\det A > 0$ (which satisfy $\lim_{\epsilon\to 0}\delta_b = \lim_{\epsilon\to 0}\delta_s = 4/\bar{y}$). Given these transfers, the buyer with reciprocity type $y_b = 0$ maximizes $\bar{\Pi}_b(s_{b,0}, s_s^T)$, for which $s_{b,0} = s_{b,0}^T$ is indeed a solution, because it yields the same payoffs as σ_b^{LL} discussed earlier. The buyer with reciprocity type $y_b = \bar{y}$ now maximizes $\bar{\Pi}_b(s_{b,\bar{y}}, s_s^T) + \bar{\Pi}_s(s_{b,\bar{y}}, s_s^T)$. A solution must be contained in the subset $\{\sigma_b^T, \sigma_b^L, \sigma_b^H\} \subset S_b$, as σ_b^{-T} is Pareto-dominated and the remaining strategies only induce additional sum-neutral redistribution. Using the payoffs derived at the beginning of the proof, it follows that $s_{b,\bar{y}} = s_{b,\bar{y}}^T = \sigma_b^T$ is indeed a solution whenever ϵ is small enough. Analogous arguments show that truth-telling is also a best response for the seller when ϵ is small enough, which completes the proof.

E Unconditional Efficiency

E.1 The Unconditional Efficiency Concept

In the body of the text we define equitable payoffs as in Rabin (1993). Dufwenberg and Kirchsteiger (2004) have proposed an alternative definition. For the Dufwenberg-Kirchsteiger equi-

table payoff, we replace the set of conditionally Pareto-efficient strategies $E_i(s_i^b) \subseteq S_i$ by a set of unconditionally Pareto-efficient strategies $E_i \subseteq S_i$. Strategy s_i belongs to E_i unless there exists $s_i' \in S_i$ such that $\Pi_i(s_i', s_i^b) \geq \Pi_i(s_i, s_i^b)$ and $\Pi_j(s_i', s_i^b) \geq \Pi_j(s_i, s_i^b)$ for all $s_i^b \in S_j$, with strict inequality for at least one agent and belief 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, as the maximum of $\Pi_j(s_i, s_i^b)$ on both $E_i(s_i^b)$ and E_i always coincides with its maximum on the whole strategy set S_i.

E.2 Observation 1

We first show that $E_b = \{s_b^T, s_b^H, s_b^L\}$ and $E_s = \{s_s^T, s_s^H, s_s^L\}$. Consider the buyer (the case for the seller is analogous). The fact that s_b^T and s_b^H belong to E_b follows because both strategies are efficient conditional on $s_b^b = s_s^T$, as shown in Appendix D.1. Clearly, strategy s_b^L uniquely maximizes the buyer's payoff conditional on $s_b^b = s_s^L$, hence s_b^L belongs to E_b as well. Finally, one can easily verify that strategy s_b^{-T} does not belong to E_b: For any belief s_b^b of the buyer, strategy s_b^{-T} yields the same payoff as s_b^T for the seller, while it always yields a weakly lower payoff than s_b^T for the buyer, and a strictly lower payoff if $s_b^b = s_s^T$, as shown in Appendix D.1. The equitable payoff for the seller from the buyer's perspective is therefore $\Pi_s^e(s_s^T) = 20$. By an analogous argument we also obtain $\Pi_b^e(s_b^T) = 20$. We therefore have $\kappa_b(s_s^b, s_s^{bb}) = \kappa_s(s_b^b, s_b^{bb}) = 0$ in the hypothetical BNFE (s_b^T, s_s^T). Hence both agents focus on their own material payoffs, and truth-telling is indeed a BNFE because f^* is BIC. Observation 1 thus does not hold with Dufwenberg-Kirchsteiger equitable payoffs.

However, this is in some sense a knife-edge case. If we choose parameters differently, then we can again show that the minimal subsidy SCF f^* is not strongly implementable in BNFE. For ease of exposition, we assume again that $\bar{\kappa}$ is sufficiently large, so that it can be ignored. We also retain all other assumptions, except that now the buyer has a low valuation with probability 0.6 and a high valuation with probability 0.4. In this case, one can compute that the minimal subsidy takes a value of 1 and that trade takes place at prices of 22, 44.5, or 77.5, depending on marginal cost and marginal valuation, as illustrated in Table 6. After computing $\Pi_b(s_b, s_s)$ and $\Pi_s(s_b, s_s)$ for all strategy profiles of the direct mechanism, we find that $E_b = \{s_b^T, s_b^H, s_b^L, s_b^{-T}\}$ and $E_s = \{s_s^T, s_s^H, s_s^L\}$. Moreover, we find that both agents' kindness would be negative in a hypothetical truth-telling equilibrium. Specifically, the buyer's kindness would be equal to -1 and the seller's kindness would be equal to -0.3. Now, as soon as the weights y_b and/or y_s are positive, the agents want to deviate from truth-telling because of the desire to generate a lower payoff for the other agent. Specifically, the buyer would prefer to understate her valuation and to choose $s_b = s_b^L$, whereas the seller would prefer to exaggerate her costs and to choose $s_s = s_s^H$.

	$\underline{\theta}_s$	$\bar{\theta}_s$
$\underline{\theta}_b$	$(1, 1+22, 1-22)$	$(0, 1, 1)$
$\bar{\theta}_b$	$(1, 1+44.5, 1-44,5)$	$(1, 1+77.5, 1-77.5)$

Table 6: Minimal Subsidy SCF f^* under Asymmetry

E.3 Observation 2

One can easily verify that for both $i = b, s$ the strategy s_i^{-T} does not belong to E_i. For any strategy s_j of agent j, strategy s_i^{-T} yields the same payoff as s_i^T for j. It always yields a weakly lower payoff than s_i^T for agent i, and a strictly lower payoff if agent j chooses s_j^T (see the payoffs derived in Appendix D.2). It is also shown in Appendix D.2 that all other strategies from S_i are efficient conditional on s_j^T. Consequently, $E_b = E_b(s_s^T)$ and $E_s = E_s(s_b^T)$, so that the remaining analysis is exactly as in the proof of Observation 2 in Appendix D.2.

E.4 Observation 3

As argued in the proof of Observation 3 in Appendix D.3, 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 3 in Appendix D.3.