Bierbrauer, Felix; Netzer, Nick

Working Paper

Mechanism design and intentions

Working Paper, No. 66 [rev.]

Provided in Cooperation with:
Department of Economics, University of Zurich

Suggested Citation: Bierbrauer, Felix; Netzer, Nick (2014) : Mechanism design and intentions, Working Paper, No. 66 [rev.], University of Zurich, Department of Economics, Zurich, https://doi.org/10.5167/uzh-61224

This Version is available at:
https://hdl.handle.net/10419/111186

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 66

Mechanism Design and Intentions

Felix Bierbrauer and Nick Netzer

Revised version, April 2014

Mechanism Design and Intentions*

Felix Bierbrauer
University of Cologne

Nick Netzer
University of Zurich

This version: April 2014
First version: July 2011

Abstract

We introduce intention-based social preferences into a mechanism design framework with independent private values and quasilinear payoffs. For the case where the designer has no information about the intensity of social preferences, we provide conditions under which mechanisms which have been designed under the assumption that agents are selfish can still be implemented. For the case where precise information about social preferences is available, we show that any tension between efficiency, incentive-compatibility, and voluntary participation may disappear. Impossibility results such as the one by Myerson and Satterthwaite (1983) are then turned into possibility results. We also provide a systematic account of the welfare implications of kindness sensations.

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 gratefully acknowledge very helpful comments by 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, Zvika Neeman, Axel Ockenfels, Marco Ottaviani, Ariel Rubinstein, Désirée Rückert, Larry Samuelson, Klaus Schmidt, Armin Schmutzler, Alexander Sebald, Joel Sobel, Ran Spiegler, André Volk, Roberto Weber, Philipp Weinschenk, David Wettstein, Philipp Wichardt, and seminar participants at the CESifo Area Conference on Behavioural Economics 2011, 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, 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.

The procedural nature of intention-based social preferences has a profound impact on the analysis. For an assessment of intentions, it does not only matter what agents do, but also what they could have done instead. Hence, a first contribution of the paper is to develop a theory of mechanism design in which the interpretation of behavior is crucial. In our model, a truth-telling strategy may appear selfish in a direct mechanism, but it may appear kind in the context of a mechanism in which the set of actions is larger than the set of conceivable payoff functions. This implies, in particular, that the revelation principle does not hold. A second contribution of the paper is to allow for a discussion of procedural questions. We show that two mechanisms which induce the same economic outcome can be compared according to the attitudes that they induce among the agents. Specifically, we formalize the problem to implement a given outcome with a maximal degree of kindness, and we clarify the conditions under which such an ideal mechanism exists. A third contribution of the paper is to introduce the idea of mechanisms that are robust in the sense that they implement an economic outcome irrespective of whether or not the agents are motivated by social preferences. We call such social choice functions strongly implementable, so as to distinguish them from those which can be implemented only with prior information about the weight that kindness sensations have in the agents' utility functions. The latter are termed weakly implementable social choice functions.

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. Rabin's analysis has focussed on environments with two agents. We follow the same route, but we show that many of our results hold for an arbitrary number of agents.

We begin with an investigation of strongly implementable social choice functions. Suppose that, for some social choice function, the expected payoff of agent i does not depend on the type of agent j, so that each agent is insured against the randomness of the other agent's type. If this insurance property holds, then the agents cannot affect each other's payoff by unilateral deviations from truth-telling in the direct mechanism. If truth-telling is an equilibrium with selfish preferences, then it continues to be an equilibrium for a large class of interdependent preference models, including the intention-based model among many others. The insurance property renders these preferences behaviorally irrelevant. Thus, our Theorem 1 asserts that if a social choice function has the insurance property and is incentive-compatible under the assumption of selfish preferences, then it is strongly implementable. Propositions 1 and 2 describe classes of social choice functions that are incentive-compatible and have the insurance property. Proposition 1 establishes the existence of strongly implementable social choice functions that are surplus-maximizing and ex post budget balanced. It is based on the observation that the

expected externality mechanism due to d'Aspremont and Gerard-Varet (1979) and Arrow (1979) satisfies the insurance property. This follows by construction of the mechanism, which requires each agent to compensate the other for the expected implications of a change in her type. Proposition 2 states that to any social choice function that is incentive-compatible if agents are selfish, there exists an essentially equivalent one that also has the insurance property. Equivalence holds with respect to the decision rule, the interim expected payoffs, and the expected deficit or surplus of the mechanism. The proof is constructive and shows how an incentive-compatible social choice function can be modified so as to make it strongly implementable. The proposition covers essentially any application of the independent private values model that has been studied in the literature, ranging from bilateral trade problems and auctions to the provision of public-goods. In particular, it also covers the study of optimal mechanisms with participation constraints, because interim payoffs are preserved by our construction.

We then turn to a characterization of weakly implementable social choice functions. We first show that the revelation principle does not hold in our framework. There exist social choice functions that cannot be implemented by direct mechanisms with a truth-telling Bayes-Nash fairness equilibrium, but that can be implemented by means of a non-direct mechanism. With a direct mechanism, every available message is used in a truth-telling equilibrium. Put differently, this class of mechanism-equilibrium-pairs excludes unused actions, which restricts the set of implementable social choice functions. We can show, by contrast, that an augmented revelation principle (Mookherjee and Reichelstein, 1990) holds. Accordingly, it is without loss of generality to focus on mechanisms where each agent's action set includes the set of possible types, and which possess truth-telling equilibria. Hence, while the restriction that every action must be used in equilibrium would involve a loss of generality, the restriction that every used action is a truthfully communicated type is without loss of generality.[1] Theorem 2 then provides conditions under which any efficient social choice function can be implemented by an appropriately chosen augmented mechanism. When intentions matter, the interpretation of equilibrium play can be influenced by adding actions to the mechanism that, if taken, would trigger redistribution among the agents. The challenge in the design of such actions is that they must be tempting to the agents but nevertheless remain unused. Our proof of Theorem 2 makes use of the possibility to engineer kindness sensations in such a way that every agent's utility function is turned into a utilitarian welfare function. The construction is akin to a Groves mechanism (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. The mechanism that we construct in order to prove Theorem 2 also satisfies voluntary participation constraints, and hence eliminates the tension between efficiency, incentive-compatibility and voluntary participation.

The analysis up to here focussed on social choice functions that are in a conventional sense efficient, treating kindness sensations and psychological payoffs as relevant from a behavioral

[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 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.

but not from a welfare perspective. We next turn to the possibility of defining efficiency and welfare based on the agents' overall utility, which aggregates material and pychological payoffs. Proposition 4 provides sufficient conditions under which material surplus-maximizing outcomes can be implemented with maximal kindness levels. Thus, we show that there is generally no conflict between the desire to achieve large material payoffs and the desire to generate intense kindness sensations.

Our results on strongly implementable social choice functions are reassuring from the perspective of conventional mechanism design theory. Even if individuals are inclined to respond to the behavior of others in a reciprocal way, this will in many cases not upset implementability of the outcomes that have been the focus of this literature. For many applications of interest, there is a way to design mechanisms so that the transmission channel for reciprocal behavior is simply shut down. By contrast, our analysis of weakly implementable social choice functions shows the potential of exploiting the reciprocity channel, rather than shutting it down. This enlarges the set of social choice functions that are implementable, and also alleviates the tension between efficiency and voluntary participation that is a key concern in the traditional mechanism design literature. Moreover, the question whether there exists a best mechanism to implement a given social choice function becomes meaningful. With an analysis that is based exclusively on consequentialist preferences, it would be impossible to even ask this question.

The 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. Throughout, we illustrate our results with a bilateral trade application. Section 6 then discusses the concept of utility efficiency. Section 7 contains extensions to an arbitrary number of agents and to a case where the mechanism designer is a player with its own intentions. The last section contains concluding remarks. Proofs and some additional extensions 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.

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.

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) and von Siemens (2011) study models in which agents are inequality-averse. Tang and Sandholm (2012) solve the optimal auction problem with spiteful agents. Kucuksenel (2012) investigates a mechanism design problem under the assumption that agents are altruistic, i.e., they attach a positive weight to the utility of others irrespective of their behavior.

Several papers study mechanism design with other behaviorally motivated assumptions. Here we focus only on models that exhibit a procedural component.[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. In Alger and Renault (2006), procedural issues arise because the mechanism and its equilibrium influence the agents' propensity to lie. Intrinsically honest agents may become willing to misrepresent their private information when other agents also benefit from lying. In some situations this makes non-direct mechanisms optimal. The possibility that institutions affect preferences has generally received some attention (see Bowles and Polanía-Reyes, 2012). Antler (2012) investigates a matching problem where the agents' preferences are affected by the stated preferences of their potential partners. de Clippel (2014) studies the problem of full

[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., 2010) 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 (2012, 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. Other important contributions to behavioral mechanism design theory include Eliaz (2002), Caplin and Eliaz (2003) and Cabrales and Serrano (2011).

implementation under complete information with agents whose behavior is described by arbitrary choice functions instead of preferences. Augmented revelation 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.

Three recent contributions build upon and extend the present paper. Bartling and Netzer (2013) apply our results on strongly implementable social choice functions to an auction setting and test them experimentally. Bierbrauer et al. (2014) combine the requirement of strong implementability with a robustness requirement on the agents' probabilistic beliefs (see Bergemann and Morris, 2005). Their main application is a bilateral trade problem, and they also provide an experimental test of the resulting mechanism. Netzer and Volk (2014) propose a notion of ex post implementation for the intention-based framework developed here.

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 comment on the extension to any finite number of agents in Section 7.2.

The environment is described by $E = [A, \Theta_1, \Theta_2, p_1, p_2, \pi_1, \pi_2]$. A denotes the set of feasible allocations, where an allocation 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 $\mathbb{E}_i$ for the expectation with respect to $\tilde{\theta}_i$ based on p_i. We write $\mathbb{E}$, without subscript, for the expectation with respect to $\tilde{\theta}$ based on the joint distribution $p = p_1 \times p_2$. 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 $\pi_i(a, \theta_i) = v_i(q_i, \theta_i) + t_i$.

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 (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 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

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

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

Before turning to the model of intention-based social preferences, we remind ourselves of the solution concept of a Bayes-Nash equilibrium (BNE).

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:

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

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:

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

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 individuals 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 an intended 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 $\bar{\kappa} = \infty$.[5] 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 $E_i(s_i^b)$.[6] To be specific, we assume that the payoff deserved by j is the average of the payoff she

[5] Dufwenberg and Kirchsteiger (2004) do not have a bound on kindness, which corresponds to $\bar{\kappa} = \infty$. Rabin (1993) imposes a bound, although in a somewhat different functional form. 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 (see Segal and Sobel, 2007).

[6] Conditional on s_i^b, a strategy $s_i \in S_i$ is Pareto-dominated by a strategy $s_i' \in S_i$ if $\Pi_k(s_i', s_i^b) \geq \Pi_k(s_i, s_i^b)$

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.[7] 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.[8]

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 agents have learned their own type but lack information about the types of the other agents. In Appendix B, 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. We will investigate welfare based on the entire utility function (3) in Section 6. First, however, we work with the conventional notion of material Pareto-efficiency introduced above, i.e., we investigate how the behavioral

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.

[7] 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. 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.

[8] In Appendix D, we go through all our bilateral trade examples so as 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).

implications of reciprocity affect the possibility to achieve materially efficiency outcomes. We explore different notions of implementability, which differ by how much a priori information on the weights $y = (y_1, y_2)$ in the agents' utility functions can be used for mechanism design.

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 $\theta \in \Theta$.

If f is strongly implementable on Y, then there exists a mechanism that implements f for all weights $y \in Y$. In particular, strong implementability on the complete set $Y = \mathbb{R}^2_+$, also simply referred to as strong implementability, is relevant for a mechanism designer who acknowledges the possibility that the agents' strategy choices may be influenced by intention-based social preferences but who has no information whatsoever on the strength of this influence. Strong implementability in BNFE clearly implies implementability in BNE. With a weakly implementable SCF, by contrast, the mechanism that is used for implementation can be made dependent on the weights $y \in Y$ in the agents' utility functions. Obviously, strong implementability on Y implies weak implementability on Y. Given the information requirements for weak implementability, the set of SCFs which are weakly implementable may be too large for many practical applications. However, since at least some information about the intensity of social preferences will be available in many applications, the set of SCFs which are strongly implementable may be too small. In the following, we use the notion of strong implementability to get a lower bound and the notion of weak implementability to get an an upper bound on what can be achieved in the presence of intention-based social preferences.

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 that $0 \leq \underline{\theta}_s < \underline{\theta}_b < \bar{\theta}_s < \bar{\theta}_b$, so that (maximal) production is optimal except if the valuation is low 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 $t^f_b(\theta_b, \theta_s)$ and $t^f_s(\theta_b, \theta_s)$. 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 $t^f_s(\theta_b, \theta_s) = -t^f_b(\theta_b, \theta_s)$ 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 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 C. 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. The downward lie 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 to understatement 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.[9] The results that follow show that these situations can be avoided if an appropriate mechanism is chosen.

4.2 Possibility Results

We will provide sufficient conditions for the strong implementability of social choice functions in BNFE. Specifically, we provide conditions under which a direct mechanism strongly implements f on $Y = \mathbb{R}^2_+$, i.e., for all conceivable reciprocity weights. Our analysis makes use of a measure

[9] Bierbrauer et al. (2014) generalize this argument to an even larger class of social preference models. Fehr et al. (2011) indeed report on the behavioral non-robustness of the Moore-Repullo mechanism for subgame-perfect implementation, and Bierbrauer et al. (2014) demonstrate systematic deviations from truth-telling in a mechanism that would be ex post incentive-compatible for selfish agents. These 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."

of payoff interdependence among the agents. Given an SCF f, we define

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

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 $\Delta_1 = \Delta_2 = 0$. 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.

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

Proof. See Appendix A.1. □

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

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

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.*

Proof. See Appendix A.2. □

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.[10]

[10] 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 budget 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:

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

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

Proof. See Appendix A.3. □

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

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

for all $(\theta_i, \theta_j) \in \Theta$. Note that, by this construction, SCF $\bar{f}$ may depend on the prior p even if this was not the case for f. An example is the application to the second-price auction in Bartling and Netzer (2013). Also, if the initial SCF f satisfies budget-balance (in the ex post sense), this property will not be preserved by the construction. The two SCFs have the same budgetary implications only if evaluated from an ex ante perspective. If the mechanism designer is interested in expected revenues, this is not a problem. For instance, Bose et al. (2006) and Bodoh-Creed (2012) use the same construction for models with ambiguity-averse agents, in which the agents and the designer act on the basis of different prior distributions. The construction then has the potential to increase expected revenues without hurting the agents, which can make mechanisms with insurance optimal.

Proposition 2 is particularly useful for problems with participation constraints, because all interim expected payoffs remain unchanged by property (c). Possible 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, 2013). In Section 4.3 below we apply the result in the context of the bilateral trade problem.

The insurance property implies robustness even beyond the class of intention-based social preferences. The proof of Theorem 1 exploits only one feature of these preferences: the agents are selfish when they lack the ability to influence the others' 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.[11] Within the class of these models, the insurance property in combination with BIC remains a sufficient condition for implementability of a social choice function. This robustness property is attractive in the light of the empirically well-documented individual heterogeneity in social preferences (Fehr and Schmidt, 1999; Engelmann and Strobel, 2004; Falk et al., 2008; Dohmen et al., 2009). In many cases, direct observation of these preferences will be difficult. An alternative approach in that case is to solve a multi-dimensional design problem, where the agents also have to report their private information about their social type. With the insurance property, there is no need to worry about the details of multi-dimensional design. Instead, there is an easy solution which makes it possible to achieve prespecified material outcomes without much knowledge of the correct behavioral model.

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 subject to BIC and PC, 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 SCF 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 $\bar{f}^*$

[11] See Bierbrauer et al. (2014) 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.

5 Weakly Implementable Social Choice Functions

5.1 Example

We begin with an example that illustrates several conceptual issues that arise in the context of weak implementation, i.e., when the designer has precise information on the weights that kindness has in the agents' utility functions. We will discuss to what extent standard insights from mechanism design theory have to be qualified, such as (i) the revelation principle, or (ii) the tension between material Pareto-efficiency, incentive-compatibility and voluntary participation.

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. For y_b, y_s and $\bar{\kappa}$ large enough, there exist numbers $\delta_b, \delta_s > 0$ so that s^T is a BNFE.*

If the buyer believes the seller to behave according to s_s^T, the best she can do for the seller is to exaggerate her willingness to pay, which leads to more trade and to trade at a higher price. The worst (but still Pareto-efficient) outcome for the seller is obtained if the buyer behaves according to $(\underline{\underline{\theta}}_b, \underline{\underline{\theta}}_b)$, i.e., if she insists on the discount of δ_b. Suppose for simplicity that $\bar{\kappa} = \infty$, so that the kindness bound can be safely ignored (the statement that y_b, y_s and $\bar{\kappa}$ must be "large enough" is made precise in Theorem 2 below). Straightforward computations then show that the buyer's kindness in the hypothetical truth-telling equilibrium s^T, where she does not insist on the discount, becomes positive: $\kappa_b(s^T) = \delta_b/4$. A symmetric argument implies that the seller is kind when she does not use the action $\bar{\bar{\theta}}_s$ and does not ask for the very high price: $\kappa_s(s^T) = \delta_s/4$. Whenever $y_b > 0$ and $y_s > 0$, we can now calibrate the numbers δ_b and δ_s so as to turn both agents' utility maximization problems into problems of welfare maximization. The buyer, for instance, chooses s_b in order to maximize

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

For $\delta_s = 4/y_b$ we obtain $\kappa_s(s^T) = 1/y_b$, and the problem becomes to choose s_b in order to maximize the sum of expected 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 a best response for the seller when $\delta_b = 4/y_s$.

Observations 2 and 3 together show that (i) a revelation principle is not available for the solution concept BNFE, because the actions that remain unused in the non-direct mechanism affect the interpretation of equilibrium behavior. Truth-telling becomes kind when both agents refrain from enriching themselves at the expense of the other agent. Hence outcomes can no longer be separated from the procedures according to which they are obtained. Since f^{**} ensures non-negative material payoffs for both agents and types, the analysis also shows that (ii) voluntary participation can be guaranteed, provided that material payoffs are considered as relevant for participation considerations. We will now discuss these issues more generally.

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. In the following, we show 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. Augmented revelation mechanisms have first been introduced by Mookherjee and Reichelstein (1990). They play an important role for implementation with the additional requirement that there is a unique equilibrium or a unique equilibrium outcome.

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 $\tilde{s}$.[12] We then establish that Φ and Φ' are strategically equivalent, in the sense that any outcome that can be induced by some action under Φ can be induced by some action under Φ' and vice versa. Formally, consider an arbitrary pair $(\Phi, \tilde{s})$ and let f be the social choice function induced by $\tilde{s}$ in Φ, i.e., $f(\theta) = g(\tilde{s}(\theta))$ for all $\theta \in \Theta$. We now construct new message sets M_i' for every 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

$$g'(m') = g(\eta_1(m_1'), \eta_2(m_2')). \tag{8}$$

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 $\theta \in \Theta$.

Proposition 3. *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*

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

and

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

Proof. See Appendix A.4. □

The sets $G_i(m_j)$ and $G_i'(m_j')$ contain all allocations that agent i can induce by varying

[12] Mookherjee and Reichelstein (1990) use the same construction, albeit for a different purpose. Unused actions enable them to destroy unwanted equilibria and to attain equilibrium uniqueness.

her message, holding fixed agent j's message. According to the proposition, these sets are the same in both mechanisms, for any pair of messages with $m_j = \eta_j(m'_j)$. Proposition 3 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'.

Corollary 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.*

5.3 A Possibility Result

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 interdependence measure 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.*

Proof. See Appendix A.5. □

In the proof, we start from a direct mechanism for f and introduce additional messages that would trigger budget-balanced redistribution among the agents. 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 for the given profile of reported types, plus possible redistributive payments initiated by an agent who presses her button. 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.[13]

[13] Mookherjee and Reichelstein (1990) also maintain out-of-equilibrium budget-balance, but their construction of "flags" and "counterflags" is otherwise very different from our "buttons". Our approach amounts to introducing $|\Theta_i|$ unused messages for agent i. More parsimonious constructions are possible, as the bilateral trade example illustrates, but come at the cost of more complicated notation.

Our construction resembles a Groves mechanism, where transfers between agents are designed so as to align individual interests with the social objective. Here, however, out-of-equilibrium payments are used for that purpose.

A difficulty in the proof of Theorem 2 arises from the kindness bound $\bar{\kappa}$. The crucial step for the alignment of incentives 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 speaks to the issue of voluntary participation. 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 for some types of some agents the expected material payoff will typically be lower than under a given status quo outcome. Since BIC is no longer a constraint by Theorem 2, we can, for instance, implement an efficient SCF that gives both agents an equal share of the material surplus (provided that y_1, y_2, and $\bar{\kappa}$ are sufficiently large). More generally, with the solution concept of weak implementability in BNFE, we may be able to achieve SCFs that are surplus-maximizing and satisfy PC but violate BIC. This solves the participation problem based on the criterion of material payoffs. However, the requirement of non-negative material payoffs may be questionable if agents have social preferences. It may seem more plausible that they agree to play a mechanism if their overall utility, including kindness sensations, is larger than under the status quo. Theorem 2 can be adapted to guarantee voluntary participation also with this criterion. Instead of adding unused messages to the direct mechanism, where $M_i = \Theta_i$, we can as well start out from a direct mechanism with veto rights, where $M_i^v = \Theta_i \cup \{v\}$ and which stipulates some status quo allocation $a^v \in A$ if any one agent sends the veto v. We can add messages 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. Both the veto rights and the additional messages then remain unused in equilibrium, which implies that all types of both agents participate voluntarily. The only modification required to extend the proof of Theorem 2 is to replace each value Δ_i by the (weakly larger) value Δ_i^v that takes into account agent j's impact on agent i's expected payoff by means of the veto:

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

We add a word of caution: Our proof of Theorem 2 relies on the use of a direct mechanism with a button, which is of course an artificial construction. It should be interpreted as a tool

for the characterization of feasible outcomes, in the same way augmented or direct mechanisms are typically interpreted in the literature. Still, the logic may be related to mechanisms which are empirically more 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. Not taking an opportunistic action in such an incomplete contract is akin to not pressing the button in our augmented mechanism.

6 Utility Pareto-Efficiency

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. As a first step, 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 4. *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 $\kappa_1(s) = \kappa_2(s) = \bar{\kappa}$.

Proof. See Appendix A.6. □

The proof relies on an augmentation of the mechanisms used in the proofs of Theorems 1 and 2. The forgone redistribution now has to be specified so that the resulting kindness equals the upper bound $\bar{\kappa}$. The crucial step in the proof is to show that, even in the face of this larger temptation, no agent prefers to deviate from truth-telling.[14]

Proposition 4 is of its own interest, as it provides a result on how to achieve a given material outcome with maximal kindness. In addition, it now allows us to turn to a notion of efficiency that is based on the entire utility functions U_i, as opposed to the agents' material payoffs only. The concept of utility Pareto-efficiency gives rise to a conceptual difficulty. In a consequentialistic approach, the definition of efficiency of an SCF is independent from the investigation of the mechanisms that implement it. This separation is not possible in our approach, because utilities are procedural and depend on the mechanism and its equilibrium. Hence utility Pareto-efficiency needs to be defined as a property of mechanism-equilibrium pairs rather than of social

[14] For instance, case (a) of Proposition 4 allows $y_i\bar{\kappa} < 1$. Even with maximal kindness, both agents then still place a larger weight on their own than on the other agent's payoff, and would thus prefer to press a button that triggers budget-balanced redistribution. As a consequence, off-equilibrium budget-balance can no longer be guaranteed in the proof of Proposition 4.

choice functions.[15] We can, however, apply Proposition 4 to construct utility Pareto-efficient mechanism-equilibrium pairs. The first step is to fix an SCF f that is materially Pareto-efficient and for which (a) or (b) in Proposition 4 applies. The second step is to implement f in a BNFE s^* of a mechanism Φ such that $\kappa_1(s) = \kappa_2(s) = \bar{\kappa}$ holds, which is possible according to Proposition 4. Then the mechanism-equilibrium pair (Φ, s) is utility Pareto-efficient, i.e., there cannot be any other pair (Φ', s') that yields a strictly larger utility for one agent without giving a strictly smaller utility to the other agent. This holds irrespective of the material outcome of (Φ', s'), due to material Pareto-efficiency of f and the fact that (Φ, s) achieves maximal kindness for both agents.

7 Extensions

7.1 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 an own 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 a first extension, we briefly explore this idea in a simplified model framework. We show that the perception of the designer as a player may drastically reduce the set of implementable outcomes, even if the designer does not have a genuine own interest in the allocation but attempts to maximize a weighted average of the agents' material payoffs.

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 ex ante expected material payoff of agent i. We assume that the mechanism designer cares about welfare

$$W(f) = \gamma\Pi_1(f) + (1-\gamma)\Pi_2(f),$$

where $0 < \gamma < 1$ determines the relative weights of the agents in the objective. For instance, in the bilateral trade example we think of the mechanism designer as a benevolent regulator who cares about a weighted average of consumer and producer surplus. 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 $\bar{t}$ from agent 1 to agent 2. Note that the entire ex ante material payoff frontier can be traced out this way. The insurance property and BIC are unaffected by $\bar{t}$, so that we can safely ignore intention-based social preferences between the 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 a choice of $\bar{t}$, and we write

$$\Pi_1(\bar{t}) = \Pi_1^{AGV} - \bar{t}, \quad \Pi_2(\bar{t}) = \Pi_2^{AGV} + \bar{t},$$

[15] See Ruffle (1999) for similar welfare arguments in the context of psychological gift-giving games. In a model of outcome-based social preferences, Benjamin (2014) also distinguishes between material and utility efficiency.

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 $-\Pi_2(0) \leq \bar{t} \leq \Pi_1(0)$ to guarantee that no agent's material ex ante payoff becomes negative.

We now introduce an equitable reference payoff for each agent. 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, let agent i's equitable payoff be defined as the average between her best and her worst payoff on the material payoff frontier. For both $i = 1, 2$, this yields

$$\Pi_i^e = \frac{1}{2}\left[\max_{\bar{t}} \Pi_i(\bar{t}) + \min_{\bar{t}} \Pi_i(\bar{t})\right] = \frac{1}{2}(\Pi_1^{AGV} + \Pi_2^{AGV}).$$

In words, the agents consider as equitable an equal split of the expected surplus. Consider agent 1 first, and assume $\bar{\kappa} = \infty$ for simplicity. The kindness of a designer who proposes $\bar{t}$ then is

$$\kappa_{d1}(\bar{t}) = \Pi_1(\bar{t}) - \Pi_1^e = \frac{1}{2}(\Pi_1^{AGV} - \Pi_2^{AGV}) - \bar{t},$$

and agent 1's best response problem, given truth-telling of agent 2, becomes to maximize

$$\Pi_1(s_1, s_2^T) + y_1\, \kappa_{d1}(\bar{t})\, \gamma \Pi_1(s_1, s_j^T),$$

where all terms that do not depend on s_1 have been omitted.[16] Suppose that the offered mechanism yields less than half of the surplus for agent 1, i.e., $\bar{t} > (\Pi_1^{AGV} - \Pi_2^{AGV})/2$. In the bilateral trade example, when agent 1 is the seller, this could correspond to a regulator who puts more weight on consumer surplus than on producer surplus ($\gamma < 1/2$) and hence would like to make $\bar{t}$ as large as possible. We obtain $\kappa_{d1}(\bar{t}) < 0$, because agent 1 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, she can influence the designer's objective only through her own well-being, and, for a sufficiently large value of y_1, will attempt to minimize $\Pi_1(s_1, s_j^T)$. Truth-telling maximizes $\Pi_1(s_1, s_j^T)$ by BIC and is not a solution to this problem. In the opposite case, when $\bar{t} < (\Pi_1^{AGV} - \Pi_2^{AGV})/2$, the same logic implies that agent 2 will deviate from truth-telling when y_2 is large enough. The only AGV mechanism that remains strongly implementable in BNFE is the one with $\bar{t} = (\Pi_1^{AGV} - \Pi_2^{AGV})/2$. In this case we obtain $\kappa_{di}(\bar{t}) = 0$ for both $i = 1, 2$, such that the agents care only about their own material payoffs and truth-telling is an equilibrium for all $y \in \mathbb{R}_+^2$.

This simple example demonstrates that reciprocity towards the (benevolent) designer can have a substantial impact on the set of implementable outcomes. While the AGV mechanism with any lump sum redistribution is strongly implementable in BNFE if the agents treat the mechanism as exogenous, only the equal split distribution can be strongly implemented when the mechanism is treated as endogenous, and thus conveys the designer's intentions.

[16] Both agent 2's payoff $(1-\gamma)\Pi_2(s_1, s_2^T)$ and the designer's equitable payoff can be omitted in the kindness of agent 1 toward the designer, the former due to the insurance property, the latter because it is an additive constant.

7.2 Arbitrary Number of 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) $s_i^* \in \text{argmax}_{s_i\in S_i}\, U(s_i, s_i^b, s_i^{bb})$, (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

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

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,

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

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

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

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 4 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 4, 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.

8 Conclusion

Economists have become increasingly more 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 the problem. Our results in the first part of the paper show how to eliminate a potential impact of the context on preferences. This is relevant for a designer who wishes to refrain from calibrating the mechanism to the details of a specific behavioral model. We have shown that such a designer can still rely on many results that have been provided by the rich literature on mechanism design under the (possibly misspecified) as-

sumption of selfish behavior. Our results in the second part of the paper show how to exploit a potential impact of the context on preferences. The design of choice sets then becomes a non-trivial part of mechanism design, and efficient outcomes that are out of reach with selfish agents become implementable.

There are several open questions for future research already within our specific framework of intention-based social preferences. 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. Second, several of our results lend themselves to experimental testing. This concerns, for instance, the role of unused actions as a design instrument, or the problem whether differences in kindness perceptions across outcome-equivalent mechanisms can be identified empirically. Of course, the more general field of context-dependent mechanism design offers an even wider range of important and fascinating open questions.

References

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

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. (2012). Two sided matching with intrinsic preferences over stated rankings. Mimeo.

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. (2013). 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*, forthcoming.

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

Bierbrauer, F., Ockenfels, A., Rückert, D., and Pollak, A. (2014). Robust mechanism design and social preferences. University of Cologne, Department of Economics, Working Paper.

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*, forthcoming.

De Marco, G. and Immordino, G. (2012). Reciprocity in the principal multiple agent model. CSEF Working Paper 314.

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

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. (2011). Handing out guns at a knife fight: Behavioral limitations to the moore-repullo mechanism. Mimeo.

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

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

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.

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*, forthcoming.

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.

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,

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

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 the expectations operator $\hat{\mathbb{E}}_j$ is based on the probability distribution $\hat{p}_j$ given by

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

for all $\theta_j \in \Theta_j$, instead of p_j as for $\mathbb{E}_j$. 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

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

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

$$t_i^{\bar{f}}(\theta_i, \theta_j) = \mathbb{E}_j[v_i(q_i^f(\theta_i, \tilde{\theta}_j), \theta_i) + t_i^f(\theta_i, \tilde{\theta}_j)] - v_i(q_i^f(\theta_i, \theta_j), \theta_i). \tag{9}$$

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

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

Property (d). We first show that $\bar{f}$ has the insurance property. From (9) it follows that for any $(\theta_i, \theta_j) \in \Theta$ we have that

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

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

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

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

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

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

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

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

A.4 Proof of Proposition 3

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 (8), 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 (8) implies that $g'(m_i', m_j') = a$. Hence, $a \in G_i'(m_j')$.

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

$$x_i(s_i) = \sum_{\theta_i \in \Theta_i} p(\theta_i) s_i^2(\theta_i)$$

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 1. *If* $sgn\,\delta_{jj} = sgn\,\delta_{ji}$, *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

$$\Pi_i(s_i^T, s'_j|\delta) \geq \Pi_i(s_i^T, s_j^{min}|\delta),$$
$$\Pi_j(s_i^T, s'_j|\delta) \geq \Pi_j(s_i^T, s_j^{min}|\delta),$$

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

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

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

$$\Pi_i(s_i^T, s'_j|0) < \Pi_i(s_i^T, s_j^{min}|0),$$
$$\Pi_j(s_i^T, s'_j|0) > \Pi_j(s_i^T, s_j^{min}|0),$$

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 1.

Corollary 2. *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

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

where the third equality follows from Corollary 2 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 1, 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 4

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.

B 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

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

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 $\kappa_j^{int}(s_i^b(\theta_j), s_i^{bb}|\theta_j)$,

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

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

$$\kappa_i(s_i, s_i^b) = \mathbb{E}_i[\kappa_i^{int}(s_i(\tilde{\theta}_i), s_i^b|\tilde{\theta}_i)]. \tag{26}$$

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

Proof. (26) implies that $\bar{\kappa}_j^{int}(s_i^b, s_i^{bb}) = \mathbb{E}_j[\kappa_j^{int}(s_i^b(\tilde{\theta}_j), s_i^{bb}|\tilde{\theta}_j)] = \kappa_j(s_i^b, s_i^{bb})$ 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,

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

and hence $U_i(s_i, s_i^b, s_i^{bb}) = \mathbb{E}_i[U_i^{int}(s_i(\tilde{\theta}_i), s_i^b, s_i^{bb}|\tilde{\theta}_i)]$. 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 (26), 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 5 allows us to verify whether or not IFE and BNFE are equivalent.

C Proofs of Observations

C.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

$$\Pi_b(s_b^T, s_s^T) = 20, \;\; \Pi_s(s_b^T, s_s^T) = 20,$$

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

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

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

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 $y_s > 0$.

C.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

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

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

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

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

Now we can easily compute that, from the buyer's perspective, the 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

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

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.

C.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 $\underline{\theta}_b$ 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 C.2), strategy $\underline{\underline{s}}_b$ with $\underline{\underline{s}}_b(\theta_b) = \underline{\theta}_b$ for all θ_b now maximizes $\Pi_b(s_b, s_s^T)$ and minimizes $\Pi_s(s_b, s_s^T)$ in Φ', and hence is efficient. It 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 $y_b > 0$, $y_s > 0$ and $\bar{\kappa} \geq \max\{1/y_b, 1/y_s\}$. 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 Unconditional Efficiency

D.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. In this appendix, we show how our observations are affected by this alternative definition. For the Dufwenberg-Kirchsteiger equitable 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.

D.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 C.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 C.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 5. 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 5: Minimal Subsidy SCF f^* under Asymmetry

D.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 C.2). It is also shown in Appendix C.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 C.2.

D.4 Observation 3

As argued in the proof of Observation 3 in Appendix C.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 C.3.