Intentions and Ex-Post Implementation*

Nick Netzer André Volk
University of Zurich

January 2014

Abstract

Intention-based reciprocity is an important motivation for human behavior, and it can be exploited in the design of economic allocation mechanisms. In this paper, we address questions of robustness that arise in the context of asymmetric information about intentions. We propose allocation mechanisms that eliminate uncertainty about the players' intentions, by making all types of each player equally kind, and we investigate a first notion of ex-post fairness implementation, based on the property that learning about a player's type does not change the perception of that player's intention in such mechanisms. We show that efficient social choice functions which provide payoff insurance to the agents can be implemented in this way, with or without voluntary participation constraints.

Keywords: Mechanism Design, Intention-Based Social Preferences, Ex-Post Implementation.
JEL Classification: C70, C72, D02, D03, D82, D86.

*Email: nick.netzer@econ.uzh.ch and andre.volk@econ.uzh.ch. We are grateful for very helpful comments by Armin Schmutzler, Ron Siegel and seminar participants at the Zurich Workshop in Economics 2012. We acknowledge the financial support of the Swiss National Science Foundation (Grant No. 100018_126603 on "Reciprocity and the Design of Institutions"). All errors are our own.

1 Introduction

Private information about payoffs constitutes a key ingredient to the literature on mechanism design. We consider a mechanism design model where behavior is driven by material as well as psychological motives. In particular, we follow the framework of Bierbrauer and Netzer (2012) and equip agents with intention-based preferences for reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004). Private information about material payoffs then gives rise to private information about intentions: players cannot fully determine the other players' intended kindness, as they are lacking information about these players' types.[1] We contribute to the existing literature by addressing two problems that may arise in this context.

First, our understanding of the attribution of intentions under uncertainty is still limited. We might, for instance, not be able to discard the hypothesis that opponent types with bad intentions are more salient than those with good intentions, and hence are overweighted in the process of belief formation.[2] Therefore, while we could assume that players treat intentions like material payoffs and form an expected value at the interim stage, here we take a broader approach. We propose an equilibrium notion that requires intentions to be type-invariant, and we construct mechanisms that implement certain materially efficient social choice functions in this way. An equilibrium where each type of a player displays the same kindness remains robust for a large class of assumptions about the attribution of intentions under asymmetric information.

Second, if players were informed about all private information after taking their actions, then they could infer the other players' true intentions and may regret their decisions, as they may have acted differently had they known these intentions beforehand. Such psychological regret can be a concern in the same way and in addition to material ex-post regret, linked to the uncertainty about material payoff types, as addressed in the theory of ex-post implementation.[3] We therefore propose a notion of ex-post fairness implementation which requires that all players would want to stick to their interim decisions even if they were informed about all private information ex-post. We again utilize the property of type-invariance to eliminate psychological regret, as it renders intentions fully transparent in the first place. Our notion of ex-post fairness equilibrium rules out material ex-post regret at the same time. The mechanisms that we construct align individual and social objectives both on the interim and on the ex-post stage.

In spite of the strong demands implied by our notions of robustness, we show that materially Pareto efficient social choice functions can be implemented whenever they provide insurance to

[1] Models of intention-based preferences rely on the framework of psychological game theory (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009). Most of the papers that develop or apply models where intentions matter do not explicitly consider asymmetric information (e.g. Rabin, 1993; Dufwenberg and Kirchsteiger, 2000, 2004; Charness and Rabin, 2002; Falk and Fischbacher, 2006; Cox et al., 2007; Segal and Sobel, 2007, 2008; Hahn, 2009; Sebald, 2010; Nishimura et al., 2011; Dufwenberg et al., 2011, 2013; Hoffmann and Kolmar, 2013; Netzer and Schmutzler, 2013). Bierbrauer and Netzer (2012) and von Siemens (2009, 2013) explicitly model intention-based social preferences under asymmetric information. Models within the framework of Levine (1998) rely on asymmetric information and signalling to generate reciprocity via type-dependent preferences.

[2] A large experimental literature has provided evidence for the general importance of intentions for behavior, see e.g. Blount (1995) for an early and Falk et al. (2008) for a more recent contribution. However, already Blount (1995) has pointed out that "[f]indings on attributions and the lability of preferences in social context are particularly applicable to the relationship between games of incomplete and imperfect information" (p. 142f).

[3] See e.g. Bergemann and Morris (2005) and Jehiel et al. (2006), and the discussion therein. Filiz-Ozbay and Ozbay (2007) argue that a psychological concern to avoid anticipated material ex-post regret can lead to overbidding in auctions.

the agents. This requires that the expected material payoff of any player i is invariant to the private information of any other player j, where expectation is taken over the private information of all players except j.[4] In a direct revelation mechanism, the insurance property would imply that j cannot affect the payoff of i if all players except j were to tell the truth (Bierbrauer and Netzer, 2012; Bartling and Netzer, 2013). This property provides the basis for type-invariant kindness: It implies that truth-telling in the direct mechanism is associated with zero kindness for all types of all players, as nobody has the option to make anyone else better or worse off. We then exploit the reference-dependence of intention-based social preferences and augment the direct mechanism by messages which trigger additional budget-balanced transfers, but remain unused in equilibrium. This allows us to manipulate reference points and increase kindness to levels that guarantee truth-telling of all players, by turning them into maximizers of the sum of expected material payoffs, without violating the property of type-invariance. If players became informed about their opponents' types ex-post, kindness perceptions would not change and truth-telling would still be a best response, as players remain maximizers of the sum of ex-post material payoffs.

We also address the issue of voluntary participation, by considering mechanisms that give all players the right to enforce some status-quo allocation. Such veto rights complicate the construction of type-invariant kindness, as their execution might be kindness-relevant for some but not for other player types. Veto rights can thereby compromise the property of type-invariant kindness despite the insurance property of a social choice function. However, we show that such concerns can be addressed by a more complicated construction of a mechanism that uses different out-of-equilibrium transfers depending on whether or not a player profits in expectation from the execution of the veto right.

Our results complement those by Bierbrauer and Netzer (2012), who first modelled intention-based social preferences in a mechanism design framework.[5] They show that any materially Pareto efficient social choice function can be (voluntarily) implemented in (ex-ante or ex-interim) Bayes-Nash fairness equilibrium. Their argument rests on a manipulation of equitable payoffs very similar to the one employed here, but it does not guarantee the property of type-invariant kindness, which is central to the present paper. The insurance property is used by Bierbrauer and Netzer (2012) to guarantee robustness of implementation results with respect to non-selfish motives of the agents.[6] The analysis in this paper combines these arguments about robustness and about the possibility to manipulate social preferences by choice of a mechanism.

[4] Our notion of insurance is related to similar concepts in the literature on auctions and mechanism design with risk averse bidders (Maskin and Riley, 1984) or under ambiguity (Bose et al., 2006; Bodoh-Creed, 2012). We will discuss this in Section 4 below.

[5] The growing literature on behavioral mechanism design has also investigated procedural motives (Glazer and Rubinstein, 1998), robustness to non-equilibrium behavior (Eliaz, 2002), honesty (e.g. Alger and Renault, 2006), state-dependent and endogenous preferences (e.g. Bowles and Polanía-Reyes, 2012), level-k reasoning (Crawford et al., 2009), learning (e.g. Mathevet, 2010; Cabrales and Serrano, 2011), lack of common knowlegde of rationality (Renou and Schlag, 2011), irrational choice functions (e.g. de Clippel, 2012) and loss aversion (Eisenhuth, 2012). Distributional preferences have been investigated by Desiraju and Sappington (2007), Kucuksenel (2012) and von Siemens (2011). Jehiel and Moldovanu (2006) survey the large literature on externalities in mechanism design. De Marco and Immordino (2013) examine reciprocity in a team design problem, based on a model that differs from Rabin (1993) and does not exhibit reference-dependence.

[6] Bartling and Netzer (2013) investigate the insurance property in an auction setting and provide experimental evidence in favour of the theoretical robustness prediction.

The remainder of the paper is organized as follows. The general framework is introduced in section 2. Section 3 discusses our notions of robustness. The main results are presented in section 4. Section 5 presents an application to a public goods example, and section 6 concludes.

2 General Framework

2.1 Environment

The analysis builds on the formal framework of Bierbrauer and Netzer (2012). We fix an environment $E = [I, A, (\Theta_i, \pi_i)_{i\in I}, p]$, where $I = \{1, \ldots, n\}$ is a set of agents, A is a set of feasible allocations, Θ_i is a finite set of types for agent i, π_i denotes the material payoff function for agent i, and p represents a probability distribution with support $\Theta = \times_{i\in I}\Theta_i$. We employ the notation $p(\theta_i)$ and $p(\theta_{-i})$ for marginal distributions with respect to the types of subsets of agents.

As for material payoffs, we consider a quasilinear environment with independent private values. Formally, $A = Q \times T$ where Q represents a set of possible decisions and $T = \{(t_1, \ldots, t_n) \in \mathbb{R}^n | \sum_{i\in I} t_i \leq 0\}$ the set of feasible transfers. Each player's material payoff is a function $\pi_i : A \times \Theta_i \to \mathbb{R}$, given by $\pi_i(a, \theta_i) = v_i(q, \theta_i) + t_i$. Types are independent, so $p(\theta) = \prod_{i\in I} p(\theta_i)$.

A social choice function f is a mapping $f : \Theta \to A$. When referring to its specific parts, we employ the notation $f = (q^f, t_1^f, \ldots, t_n^f)$. A social choice function is materially Pareto efficient if (i) its decision rule q^f is value maximizing, $q^f(\theta) \in \arg\max_{q\in Q} \sum_{i\in I} v_i(q, \theta_i)$ for all $\theta \in \Theta$, and (ii) the transfer scheme is ex-post budget balanced, $\sum_{i\in I} t_i^f(\theta) = 0$ for all $\theta \in \Theta$. Throughout the paper, we will restrict attention to investigating the implementability of efficient social choice functions.

2.2 Mechanism

A mechanism $\Phi = [M_1, \ldots, M_n, g]$ prescribes a finite set of messages M_i for every agent $i \in I$, and an outcome function $g : M \to A$ where $M = \times_{i\in I} M_i$. When referring to specific parts of the outcome function, we use the notation $g = (q^g, t_1^g, \ldots, t_n^g)$.

A mechanism Φ and the environment E jointly induce a Bayesian game, where player i's pure strategy is a function $s_i : \Theta_i \to M_i$. Denote by S_i the set of all pure strategies for player i. Let the first-order point belief of player i about player j's strategy be denoted by $s_{ij}^b \in S_j$. A complete first-order belief profile of player i is denoted by $s_i^b = (s_{ij}^b)_{j\neq i} \in S_i^b = \times_{j\neq i} S_j$. A second-order point belief of player i concerning j's first-order point belief about player k's strategy is denoted by $s_{ijk}^{bb} \in S_k$. Player i's second-order belief about j's complete first-order belief profile shall be denoted by $s_{ij}^{bb} = (s_{ijk}^{bb})_{k\neq j} \in S_j^b$. Finally, a complete second-order belief profile of player i is $s_i^{bb} = (s_{ij}^{bb})_{j\neq i} \in S_i^{bb} = \times_{j\neq i} S_j^b$.

2.3 Utility

We first presume every player to submit his message at the *interim* stage, where each player is informed about the own type θ_i while, at the same time, remains uninformed about the realization of the other players' types θ_{-i}. We denote the interim expected material payoff of

player i from submitting m_i, given type θ_i and belief s_i^b about the other players' strategies, as

$$\Pi_i(m_i, s_i^b|\theta_i) = \mathbb{E}_{\theta_{-i}}\left[\pi_i(g(m_i, s_i^b(\theta_{-i})), \theta_i)\right].$$

Analogously, we let

$$\Pi_j(m_i, s_i^b) = \mathbb{E}_{\theta_{-i}}\left[\pi_j(g(m_i, s_i^b(\theta_{-i})), \theta_j)\right]$$

denote the material payoff that i expects to give to j when sending message m_i, given belief s_i^b.

Next, we follow the definition of interim utility proposed by Bierbrauer and Netzer (2012, Appendix B.1), which translates the concept of Rabin (1993) to Bayesian games. Accordingly, in addition to material payoffs $\Pi_i(m_i, s_i^b|\theta_i)$, each player is motivated by psychological reciprocity payoffs. We denote player i's interim belief about his kindness towards player j by $\kappa_{ij}(m_i, s_i^b|\theta_i)$, and his interim belief about j's kindness towards himself by $\lambda_{ji}(s_{ij}^b, s_{ij}^{bb})$. Below we will define κ_{ij} and λ_{ji} formally, in a way such that these terms take on positive values if associated with kind behavior and negative values if associated with unkind behavior. Reciprocity is captured by the assumption that mutual kindness as well as mutual unkindness increase interim utility:

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

where $y_{ij} \geq 0$ indicates the degree to which other-regarding concerns matter for individual i in relation to individual j. We will also write $y_i = (y_{ij})_{j\neq i}$ and $y = (y_i)_{i\in I}$.

2.4 Kindness

We measure interim kindness of player i towards some other player j as the difference between the expected material payoff which player i believes to give to player j and the *equitable payoff*, $\Pi_j^{e_i}$, the reference point for the evaluation of kindness. In other words, we presume that, given his belief s_i^b, type θ_i of player i believes to be kind (unkind) towards j if his message m_i yields a higher (lower) material payoff for j than equitable. Formally,

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

The equitable payoff equals a value in between the largest and smallest payoff that type θ_i of player i can give to player j, conditional on belief s_i^b:

$$\Pi_j^{e_i}(s_i^b|\theta_i) = \alpha\left[\max_{m_i\in M_i}\Pi_j(m_i, s_i^b)\right] + (1-\alpha)\left[\min_{m_i\in E_{ij}(s_i^b|\theta_i)}\Pi_j(m_i, s_i^b)\right]$$

for some $\alpha \in (0, 1)$.[7] The set of messages relevant for the minimization contains only messages which induce bilaterally Pareto efficient payoff pairs: $E_{ij}(s_i^b|\theta_i) = \{m_i \in M_i | \nexists\, m_i' \in M_i$ with $\Pi_i(m_i', s_i^b|\theta_i) \geq \Pi_i(m_i, s_i^b|\theta_i)$ and $\Pi_j(m_i', s_i^b) \geq \Pi_j(m_i, s_i^b)$, with at least one strict inequality$\}$.

[7] Bierbrauer and Netzer (2012, Appendix B.1) do not explicitly specify the interim equitable payoff. Instead, they provide a condition on interim equitable payoffs such that their concept of Bayes-Nash fairness equilibrium, which is based on an ex-ante perspective, is identical to their notion of interim fairness equilibrium, which takes the interim perspective.

This assumption guarantees that messages which hurt player j without benefitting player i do not influence the equitable payoff and hence the kindness perception of i's behavior.

The message m_i submitted by type θ_i of player i determines his intended interim kindness $\kappa_{ij}(m_i, s_i^b|\theta_i)$ towards player j, given his belief s_i^b. Suppose player i knew j's type at the interim stage. Given beliefs s_{ij}^b and s_{ij}^{bb}, he could then derive a belief $\kappa_{ji}(s_{ij}^b(\theta_j), s_{ij}^{bb}|\theta_j)$ about j's intended kindness towards himself. On the interim stage, however, player i is uninformed about θ_j and therefore cannot put himself into player j's interim shoes in order to figure out the intended kindness precisely. Following Bierbrauer and Netzer (2012), we first proceed under the assumption that player i forms his belief about j's kindness by taking the expectation over θ_j,

$$\lambda_{ji}(s_{ij}^b, s_{ij}^{bb}) = \sum_{\theta_j \in \Theta_j} p(\theta_j) \kappa_{ji}(s_{ij}^b(\theta_j), s_{ij}^{bb}|\theta_j). \tag{1}$$

2.5 Equilibrium

We can now provide a definition of interim fairness equilibrium, adapted to the present notation from Bierbrauer and Netzer (2012, p. 53):

Definition 1. *An interim fairness equilibrium (IFE) is a profile s^* such that, for all $i \in I$,*
(i) $s_i^*(\theta_i) \in \arg\max_{m_i \in M_i} U_i(m_i, s_i^b, s_i^{bb}|\theta_i)$ for all $\theta_i \in \Theta_i$, and
(ii) $s_{ij}^b = s_j^*$ for all $j \neq i$, and
(iii) $s_{ijk}^{bb} = s_k^*$ for all $j \neq i, k \neq j$.

A social choice function f is implementable in IFE if there exists a mechanism Φ with an IFE s^* such that $g(s^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$.

3 Notions of Robustness

3.1 Type-Invariance of Kindness

Picking up on the discussion in the introduction, the fact that kindness $\kappa_{ij}(s_i^*(\theta_i), s_{-i}^*|\theta_i)$ in IFE will generally depend on θ_i may be troublesome. A mechanism designer might not be confident that (1) correctly reflects the way players form beliefs about the others' intentions on the interim stage. We therefore refine the concept of IFE in order to guarantee that interim equilibrium intentions are fully transparent to every player, despite the presence of private information. In particular, we require a strategy profile s^* not only to be an IFE but also to generate type-invariant equilibrium kindness values. Formally,

$$k_{ij}(s_i^*(\theta_i), s_{-i}^*|\theta_i) = k_{ij}(s_i^*(\tilde{\theta}_i), s_{-i}^*|\tilde{\theta}_i) =: k_{ij}(s^*) \quad \text{for all} \quad \theta_i, \tilde{\theta}_i \in \Theta_i \quad \text{and} \quad i, j \in I, j \neq i. \tag{2}$$

An IFE s^* that additionally satisfies condition (2) is called *interim fairness equilibrium with type-invariant kindness* (IFE-TI). Implementability in IFE-TI is defined accordingly.

For instance, suppose beliefs about intentions and kindness are formed in a way that deviates from the simple expectation formulation in (1). Let

$$\Delta_{ji}(s_{ij}^b, s_{ij}^{bb}) = \left[\min_{\theta_j \in \Theta_j} \kappa_{ji}(s_{ij}^b(\theta_j), s_{ij}^{bb}|\theta_j), \ \max_{\theta_j \in \Theta_j} \kappa_{ji}(s_{ij}^b(\theta_j), s_{ij}^{bb}|\theta_j) \right]$$

denote the interval spanned by the smallest and the largest values of j's equilibrium interim kindness towards i. An IFE-TI remains an equilibrium if we replace (1) by the assumption that $\lambda_{ji}(s_{ij}^b, s_{ij}^{bb}) \in \Delta_{ji}(s_{ij}^b, s_{ij}^{bb})$, without specifying any additional details. For instance, if players use arbitrary weights $w_{ij}(\theta_j) \geq 0$ with $\sum_{\theta_j \in \Theta_j} w_{ij}(\theta_j) = 1$ instead of $p(\theta_j)$ to calculate λ_{ji}, or even focus on one of the extremes such as the least kind type of the opponent, IFE-TI remains robust as it collapses Δ_{ji} to a single value.

3.2 Ex-Post Fairness Implementation

Our notion of ex-post fairness implementation shall provide robustness in the sense that every player should stick to his interim decision even if he were informed ex-post about the others' types. Such additional information would allow each player i to update his beliefs about each opponent j's actual interim intention, thus moving from $\lambda_{ji}(s_{ij}^b, s_{ij}^{bb})$ to $\kappa_{ji}(s_{ij}^b(\theta_j), s_{ij}^{bb}|\theta_j)$. Notice that κ_{ji} is still based on an expectation over θ_{-j}, which reflects the information under which type θ_j actually made his choice. We then define player i's ex-post utility as

$$U_i(m_i, s_i^b, s_i^{bb}|\theta) = \Pi_i(m_i, s_i^b|\theta) + \sum_{j \neq i} y_{ij}\kappa_{ij}(m_i, s_i^b|\theta)\kappa_{ji}(s_{ij}^b(\theta_j), s_{ij}^{bb}|\theta_j), \tag{3}$$

where $\Pi_i(m_i, s_i^b|\theta) = \pi_i(g(m_i, s_i^b(\theta_{-i})), \theta_i)$ are the ex-post material payoffs and $\kappa_{ij}(m_i, s_i^b|\theta)$ is i's ex-post kindness toward j, defined as the difference between $\Pi_j(m_i, s_i^b|\theta) = \pi_j(g(m_i, s_i^b(\theta_{-i})), \theta_j)$ and some equitable payoff.[8]

We say that a social choice function f is ex-post fairness implementable if it is implementable in an IFE-TI s^* and if, for each player i and every type profile $\theta \in \Theta$, $s_i^*(\theta_i)$ still constitutes a best-response in terms of ex-post utility (3), given beliefs fixed on s^*. This definition captures the above stated robustness concern, since every player would stick to his equilibrium interim decision even if he observed the others' private information on the ex-post stage. Now observe that implementation of a materially Pareto efficient social choice function in an IFE-TI s^* implies ex-post fairness implementation if s^* gives rise to the kindness values $\kappa_{ji}(s_j^*(\theta_j), s_{-j}^*|\theta_j) = 1/y_{ij}$ for all pairs of players.[9] To see the point in more detail, substitute these kindness values into (3) and note that $s_i^*(\theta_i)$ then constitutes a best response in ex-post utility terms if and only if it maximizes the sum of all players' ex-post material payoffs. By presupposition, strategy profile s^* results in an efficient, i.e., payoff-sum maximizing allocation $g(s^*(\theta)) = f(\theta)$ for any type profile $\theta \in \Theta$, so that this is indeed the case. We will address in the following section whether and how these conditions can be achieved.

The proposed notion of ex-post fairness implementation is still restrictive. In particular, the kindness that i attributes to θ_j's interim behavior in (3) corresponds to the true kindness of θ_j in a mechanism where all choices are made on the interim stage. If a mechanism systematically grants players the right to reconsider their decisions ex-post, and this is anticipated on the interim stage, then the interim kindness of a given message might be different in the first place.

[8] Since the equitable payoff does not play a role for the present purposes, we omit its exact specification.

[9] Any strategy profile that results in a materially Pareto efficient social choice function and satisfies this particular condition on type-invariant kindness values must in fact be an IFE-TI, as all players are then maximizing the sum of expected material payoffs at the interim stage. However, not every IFE-TI exhibits these particular values. See the discussion following Proposition 1 below, and the example in Section 5.

An analysis of games with multiple stages of announcements is currently impeded by the lack of a theory of intentions for general extensive-form games. Our main intuition for ex-post fairness implementation parallels the intuition for ex-post Nash equilibrium provided by Crémer and McLean (1985, p. 349): "Of course, in our model, a bidder can never observe the types of the other bidders. Thus, the concept of ex post Nash equilibrium corresponds to the following reasoning by agent i. "If I believe that the other bidders are using [their equilibrium strategies], then even if I observed their actions, I would have no incentive to change mine"".[10] Besides this general intuition, ex-post fairness implementation can also be appropriate when the anticipation of regret affects interim decisions, in the spirit of Filiz-Ozbay and Ozbay (2007). Finally, our construction also applies when agents do in fact observe the others' types and can revise their decisions ex-post, but do not anticipate this on the interim stage.

4 General Results

4.1 Insurance and Implementation

A concept which will be very important is the *insurance property*. Intuitively, a social choice function gives rise to this property if each player i is insured against the realization of the type (or the report in a direct mechanism) of any other player j, provided an expectation is taken over the types of all other players (or provided that all other players report truthfully). Hence unilateral deviations from truth-telling in the direct mechanism will not affect any other player's payoff when the insurance property is satisfied (Bierbrauer and Netzer, 2012; Bartling and Netzer, 2013).

Definition 2. *Given an environment E and social choice function f, the insurance property holds if for each $i \in I$ there exists $P_i \in \mathbb{R}$ such that*

$$\mathbb{E}_{\theta_{-j}}\left[\pi_i\big(f(\tilde{\theta}_j, \theta_{-j}), \theta_i\big)\right] = P_i$$

for all $j \neq i$ and $\tilde{\theta}_j \in \Theta_j$.

Related notions of insurance exist in the literature on optimal auctions with risk averse bidders (e.g. Maskin and Riley, 1984) or with ambiguity (e.g. Bose et al., 2006). Maskin and Riley (1984) define a *perfect insurance* auction (p. 1491) where each bidder's payment is deterministic and depends only on the own type and the event of winning or losing the auction, with marginal utilities of income being equated across these two cases. In our framework with material payoffs that are linear in transfers, this is satisfied by a large class of mechanisms, such as first-price or all-pay auctions which do not satisfy our insurance property based on overall payoffs.[11] Bose et al. (2006) define a *full insurance* mechanism (p. 416) where each bidder's ex-post payoff depends only on the own type, and they show that full insurance is optimal with ambiguity

[10] Crémer and McLean's concept corresponds to the earlier notion of uniform equilibrium proposed by d'Aspremont and Gerard-Varet (1979), which builds on the concept of "complete ignorance" (see e.g. Luce and Raiffa, 1957).

[11] See Eisenhuth (2012) and Eisenhuth and Ewers (2012) for an analysis of such mechanisms with loss averse bidders. Maskin and Riley (1984) show that a perfect insurance auction will typically not be optimal with risk averse bidders, when the auctioneer can use risk to relax incentive constraints.

averse bidders.[12] The property of full insurance is stronger than our insurance property, as we require invariance of payoffs with respect to another player's type only from an ex-ante expected perspective. We can now state our first main result:

Proposition 1. *Assume that $y \in]0, \infty[^{n(n-1)}$. If a social choice function f is materially Pareto efficient and the insurance property is satisfied, then f is implementable in an IFE-TI s^* in which $\kappa_{ij}(s^*) = 1/y_{ji}$ holds for all pairs of players.*

Proof. We first prove the result for $n = 2$. We comment on the case where $n > 2$ afterwards. Throughout, we fix a social choice function f that is efficient and we suppose that the insurance property holds. We also assume $y_{12}, y_{21} > 0$, and we proceed in two steps. First, we construct a specific mechanism Φ for f. Second, we show that Φ has an IFE-TI in which f is realized with the desired kindness levels.

Step 1. Construct mechanism $\Phi = [M_1, M_2, g]$ as follows. For both $i = 1, 2$ we let $M_i = \Theta_i \times \{0, 1\}$, so that a message $m_i = (\eta_i, \gamma_i) \in M_i$ contains an announced type $\eta_i \in \Theta_i$ and an announced number $\gamma_i \in \{0, 1\}$. Given a message profile $m = (m_1, m_2) \in M$, we also write $\eta = (\eta_1, \eta_2) \in \Theta$ for the profile of announced types, and $\gamma = (\gamma_1, \gamma_2) \in \{0, 1\}^2$ for the profile of announced numbers. The outcome function g is defined as follows. For all $m \in M$, let $q^g(m) = q^f(\eta)$, i.e. only the announced types η matter for the decision rule, which follows f. For all $m \in M$, $i = 1, 2$ and $j \neq i$, let $t_i^g(m) = t_i^f(\eta) + r_i(\gamma)$, where

$$r_i(\gamma) = \begin{cases} +e_{ij} & \text{if } \gamma_i = 1, \gamma_j = 0, \\ -e_{ji} & \text{if } \gamma_i = 0, \gamma_j = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, transfers also depend on the announced types η according to f, plus an additional term that depends on the announced numbers γ. If i announces $\gamma_i = 1$ and j announces $\gamma_j = 0$, then i takes an additional amount of e_{ij} from j, and vice versa. In the following, use of i and j always presumes $j \neq i$. Since f is efficient and the additional transfers r_i always sum to zero across players, mechanism Φ is budget balanced for all profiles $m \in M$. As long as the announcements satisfy $\gamma_i = \gamma_j = 0$, the mechanism corresponds to a direct mechanism for f.

Step 2. Consider strategy profile $s^T = (s_1^T, s_2^T)$ where $s_i^T(\theta_i) = (\theta_i, 0)$ for all $\theta_i \in \Theta_i$. The profile s^T corresponds to truth-telling in a direct mechanism. Under s^T we have $g(s^T(\theta)) = f(\theta)$ for all $\theta \in \Theta$, so that f is realized. We will show that, for appropriately chosen values of e_{12} and e_{21}, strategy profile s^T is an IFE-TI of Φ, and the desired kindness levels arise. In the hypothetical equilibrium s^T, beliefs are correct: $s_i^b = s_j^{bb} = s_j^T$ for $i = 1, 2$.

We first derive the kindness term $\kappa_{ij}((\theta_i, 0), s_j^T|\theta_i)$. Given the definition of Φ, choice of m_i by player i induces the following pair of payoffs:

$$\left[\Pi_i(m_i, s_j^T|\theta_i), \Pi_j(m_i, s_j^T)\right] = \begin{cases} \left[P_i(\eta_i, \theta_i), P_j\right] & \text{for } m_i = (\eta_i, 0) \text{ with } \eta_i \in \Theta_i, \\ \left[P_i(\eta_i, \theta_i) + e_{ij}, P_j - e_{ij}\right] & \text{for } m_i = (\eta_i, 1) \text{ with } \eta_i \in \Theta_i, \end{cases}$$

where $P_i(\eta_i, \theta_i) = \mathbb{E}_{\theta_j}[\pi_i(f(\eta_i, \theta_j), \theta_i)]$, and $P_j = \mathbb{E}_{\theta_j}[\pi_j(f(\eta_j, \theta_j), \theta_j)]$ is a constant because of

[12] Perfect and full insurance coincide for certain classes of risk preferences, see Bose et al. (2006) for a discussion.

the insurance property.

Fix $e_{ij} = 1/[(1-\alpha)y_{ji}]$, implying $\max_{m_i \in M_i} \Pi_j(m_i, s_j^T) = P_j$. Let $\eta_i^* \in \arg\max_{\eta_i \in \Theta_i} P_i(\eta_i, \theta_i)$. Since any message $(\eta_i^*, 1)$ simultaneously maximizes $\Pi_i(m_i, s_j^T|\theta_i)$ and minimizes $\Pi_j(m_i, s_j^T)$, we have $(\eta_i^*, 1) \in E_{ij}(s_j^T|\theta_i)$, and thus $\min_{m_i \in E_{ij}(s_j^T|\theta_i)} \Pi_j(m_i, s_j^T) = P_j - 1/[(1-\alpha)y_{ji}]$. It follows that $\Pi_j^{e_i}(s_j^T|\theta_i) = P_j - 1/y_{ji}$ and therefore $\kappa_{ij}((\theta_i, 0), s_j^T|\theta_i) = 1/y_{ji}$ as required. Replicating the same argument for player j and $e_{ji} = 1/[(1-\alpha)y_{ij}]$ yields the type-invariant kindness $\kappa_{ji}((\theta_j, 0), s_i^T|\theta_j) = 1/y_{ij}$. We thus have $\lambda_{ij}(s^T) = 1/y_{ji}$ and $\lambda_{ji}(s^T) = 1/y_{ij}$ in the hypothetical equilibrium.

Consider now player i's interim utility for any $\theta_i \in \Theta_i$:

$$U_i(m_i, s_j^T, s_i^T|\theta_i) = \Pi_i(m_i, s_j^T|\theta_i) + y_{ij}\kappa_{ij}(m_i, s_j^T|\theta_i) \cdot (1/y_{ij}).$$

Omitting terms that do not depend on m_i, $m_i = (\theta_i, 0)$ is a maximizer of this expression if and only if it is a maximizer of

$$\Pi_i(m_i, s_j^T|\theta_i) + \Pi_j(m_i, s_j^T) = \mathbb{E}_{\theta_j}\left[v_i(q^g(m_i, s_j^T(\theta_j)), \theta_i) + v_j(q^g(m_i, s_j^T(\theta_j)), \theta_j)\right],$$

where the equality holds due to budget balance of Φ. Since f is efficient, so that q^f is value maximizing, $m_i = (\theta_i, 0)$ is a solution to the interim utility maximization problem, for any $\theta_i \in \Theta_i$. Replicating the argument for player j, we can conclude that s^T is an IFE-TI.

The case of $n > 2$. The arguments for $n = 2$ can be generalized to the case of $n > 2$, by defining message sets $M_i = \Theta_i \times [\{0\} \cup (I \backslash \{i\})]$. The announcement of $m_i = (\eta_i, \gamma_i)$ corresponds to the announcement of type η_i in a direct mechanism, but player i obtains an additional transfer e_{ij} from player j if and only if $\gamma_i = j$ and $\gamma_k = 0$ for all $k \neq i$. The above arguments can then be applied analogously for each pair of players, and bilateral type-invariant kindness of truth-telling $s_i^T(\theta_i) = (\theta_i, 0)$ can be adjusted by choice of the additional transfers payments so that each player's goal becomes the maximization of the sum of material payoffs. □

The mechanism that we construct in the proof of Proposition 1 works like a direct mechanism, where each player announces a type, but with the new feature that each player i can claim an additional payment of e_{ij} from any opponent j.[13] Truthful revelation s^* without claiming such payments then becomes kind behavior. We show that the (budget-balanced) payments e_{ij} can be adjusted so that the type-invariant interim kindness $\kappa_{ij}(s^*) = 1/y_{ji}$ is achieved for each pair of players. The *insurance property* is crucial for this to be possible. It implies that the realized and revealed type is irrelevant for kindness; all that matters is the fact that no additional payment is claimed. Each player then becomes a maximizer of the sum of expected material payoffs on the interim stage. Since the social choice function f is *efficient*, truth-telling is then a best-response to truth-telling of the opponents, which implies that the mechanism implements f in IFE-TI. Due to the specific values of equilibrium kindness, our arguments from Section 3.2 imply that it also implements f in ex-post fairness equilibrium.

[13] This mechanism is isomorphic to an *augmented revelation mechanism* (Mookherjee and Reichelstein 1994). The reason why it is formally not an augmented revelation mechanism is that we define message sets with a product structure, $M_i = \Theta_i \times D_i$ for some set D_i, instead of defining it as a union $M_i = \Theta_i \cup D_i$ so that $\Theta_i \subseteq M_i$. See also Bierbrauer and Netzer (2012) and de Clippel (2012). Saran (2011) discusses the validity of the revelation principle for general menu-dependent preferences.

If only implementation in IFE-TI but not ex-post fairness implementation was required, a much simpler construction would suffice. In fact, the direct revelation mechanism implements any efficient social choice function with the insurance property in IFE-TI. This is true because efficiency and insurance jointly imply standard Bayesian incentive-compatibility (see Lemma 1 below) and Bayesian incentive-compatibility and the insurance property jointly imply that f is implemented by a fairness equilibrium with (type-invariant) kindness levels of zero in the direct mechanism (see Bierbrauer and Netzer, 2012).

Lemma 1. *If f is materially Pareto efficient and satisfies the insurance property, then f is Bayesian incentive-compatible.*

Proof. From material Pareto efficiency of f it follows that

$$\sum_{i=1}^{n} \pi_i(f(\theta_j, \theta_{-j}), \theta_i) \geq \sum_{i=1}^{n} \pi_i(f(\hat{\theta}_j, \theta_{-j}), \theta_i)$$

for all $j \in I$, $(\theta_j, \theta_{-j}) \in \Theta$ and $\hat{\theta}_j \in \Theta_j$. Taking expectation with respect to θ_{-j}, this becomes

$$\sum_{i=1}^{n} \mathbb{E}_{\theta_{-j}}[\pi_i(f(\theta_j, \theta_{-j}), \theta_i)] \geq \sum_{i=1}^{n} \mathbb{E}_{\theta_{-j}}[\pi_i(f(\hat{\theta}_j, \theta_{-j}), \theta_i)].$$

Due to the insurance property of f, we have that

$$\mathbb{E}_{\theta_{-j}}[\pi_i(f(\tilde{\theta}_j, \theta_{-j}), \theta_i)] = P_i$$

is independent of $\tilde{\theta}_j$ for all agents $i \neq j$, so that we can simplify the inequality to

$$\mathbb{E}_{\theta_{-j}}[\pi_j(f(\theta_j, \theta_{-j}), \theta_j)] \geq \mathbb{E}_{\theta_{-j}}[\pi_j(f(\hat{\theta}_j, \theta_{-j}), \theta_j)],$$

which is the conventional Bayesian incentive-compatibility condition. □

As the next section shows, even IFE-TI implementation (without the additional requirement of ex-post implementability) will become more difficult when voluntary participation is required.

4.2 Voluntary Participation

The analysis in the previous section ignored the question whether or not some type of some player would prefer to opt out of the mechanism (see Myerson and Satterthwaite, 1983, for the classical impossibility result). To show that *voluntary participation* can be guaranteed as well, we now require that the mechanism used to implement f admits veto rights: every player must have a message which enforces a fixed status quo allocation $\bar{a} = (\bar{q}, \bar{t}_1, \ldots, \bar{t}_n) \in A$. We assume that $\bar{a}$ is budget balanced, $\sum_{i \in I} \bar{t}_i = 0$, but allow it to be chosen arbitrarily otherwise.[14] If IFE-TI implementation of a social choice function f is possible in such a mechanism, we say that f is voluntarily implementable in IFE-TI.

[14] Our assumption implies that the mechanism remains budget balanced out-of-equilibrium, which simplifies the proof. Our result would continue to hold, however, if $\sum_{i \in I} \bar{t}_i < 0$ was true.

Voluntary implementation raises several novel issues compared to the previous section. First, the direct mechanism with veto rights no longer implements f in IFE with zero kindness, despite efficiency and the insurance property, because some types of some players might prefer to opt out of the mechanism. Second, veto rights generally complicate the problem of achieving type-invariant kindness. Execution of the veto may induce bilaterally Pareto efficient payoff pairs for some but not for other types, so that the veto is relevant for the computation of equitable payoffs in the former but not in the latter case. Equitable payoffs and kindness can therefore vary with the realized type despite the insurance property of f, because the insurance property constrains payoffs derived from type reports but not from the exercise of a veto. Finally, truth-telling in a direct mechanism with veto rights goes along with different kindness values than in a direct mechanism without veto rights. Hence our construction of off-equilibrium payments must be different, and, in particular, it can become necessary to decrease equilibrium kindness. Nevertheless, we can establish the following result:

Proposition 2. *Assume that $y \in]0, \infty[^{n(n-1)}$. If a social choice function f is materially Pareto efficient and the insurance property is satisfied, then f is voluntarily implementable in an IFE-TI s^* in which $\kappa_{ij}(s^*) = 1/y_{ji}$ holds for all pairs of players.*

Proof. As before, we first prove the result for $n = 2$ and comment on the case $n > 2$ afterwards. We fix a social choice function f that is efficient and we suppose that the insurance property holds. We also fix an arbitrary budget balanced status quo $\bar{a} = (\bar{q}, \bar{t}_1, \bar{t}_2) \in A$. We proceed in two steps. First, we construct a mechanism Φ for f which admits veto rights. Second, we show that Φ has an IFE-TI in which f is realized with the desired kindness levels.

Step 1. Construct mechanism $\Phi = [M_1, M_2, g]$ as follows. Let $M_i = (\Theta_i \cup \{\nu\}) \times \{0, 1\}$, so that a message $m_i = (\eta_i, \gamma_i) \in M_i$ again comprises two components. First, $\eta_i \in \Theta_i \cup \{\nu\}$ allows player i to report either a type from Θ_i or to exercise a veto ν. Second, player i announces a number $\gamma_i \in \{0, 1\}$. Given a profile $m = (m_1, m_2) \in M$, we again write $\eta = (\eta_1, \eta_2)$ and $\gamma = (\gamma_1, \gamma_2)$. The outcome function g is defined differently for two cases. First, if m has $\eta_i = \nu$ for at least one $i = 1, 2$, we let $q^g(m) = \bar{q}$ and $t_i^g(m) = \bar{t}_i + \bar{r}_i(\gamma)$, where

$$\bar{r}_i(\gamma) = \begin{cases} +d_{ij} & \text{if } \gamma_i = 1, \gamma_j = 0, \\ -d_{ji} & \text{if } \gamma_i = 0, \gamma_j = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Hence allocation $\bar{a}$ is chosen, with possible additional transfers depending on γ. Second, if m has $\eta \in \Theta$, we let $q^g(m) = q^f(\eta)$ and $t_i^g(m) = t_i^f(\eta) + r_i(\gamma)$, where

$$r_i(\gamma) = \begin{cases} +e_{ij} & \text{if } \gamma_i = 1, \gamma_j = 0, \\ -e_{ji} & \text{if } \gamma_i = 0, \gamma_j = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the outcome function selects allocation $f(\eta)$ where additional transfers may occur in accordance with γ. Since f is efficient, $\bar{a}$ is budget balanced, and the additional transfers $\bar{r}_i$ and r_i always sum to zero across players, Φ is budget balanced for all profiles $m \in M$. If the

announcements satisfy $\gamma_i = \gamma_j = 0$, then the allocations induced by Φ are equivalent to the allocations given by a direct mechanism with additional veto rights for every player.

Step 2. Consider $s^T = (s_1^T, s_2^T)$ where $s_i^T(\theta_i) = (\theta_i, 0)$ for all $\theta_i \in \Theta_i$, so that the veto rights remain unused and $g(s^T(\theta)) = f(\theta)$ for all $\theta \in \Theta$. We will show that s^T is an IFE-TI in which the desired kindness levels arise for appropriate values of e_{12}, e_{21}, d_{12} and d_{21}.

We first derive $\kappa_{ij}((\theta_i, 0), s_j^T|\theta_i)$. Player i with type θ_i can induce the following payoff pairs:

$$\left[\Pi_i(m_i, s_j^T|\theta_i), \Pi_j(m_i, s_j^T)\right] = \begin{cases} \left[P_i(\eta_i, \theta_i), P_j\right] & \text{for } m_i = (\eta_i, 0) \text{ with } \eta_i \in \Theta_i, \\ \left[P_i(\eta_i, \theta_i) + e_{ij}, P_j - e_{ij}\right] & \text{for } m_i = (\eta_i, 1) \text{ with } \eta_i \in \Theta_i, \\ \left[\bar{P}_i(\theta_i), \bar{P}_j\right] & \text{for } m_i = (\nu, 0), \\ \left[\bar{P}_i(\theta_i) + d_{ij}, \bar{P}_j - d_{ij}\right] & \text{for } m_i = (\nu, 1), \end{cases}$$

where $P_i(\eta_i, \theta_i)$ and P_j are defined as in the proof of Proposition 1, $\bar{P}_i(\theta_i) = \pi_i(\bar{a}, \theta_i)$ is player i's material payoff in $\bar{a}$ and $\bar{P}_j = \mathbb{E}_{\theta_j}[\pi_j(\bar{a}, \theta_j)]$ is player j's expected material payoff from $\bar{a}$. Define $\delta_j = \bar{P}_j - P_j$, which does not depend on any type and thus is known to the mechanism designer, who can distinguish between the following three cases:

(a): $\delta_j \geq 0$. Let $e_{ij} = [1+\alpha y_{ji}\delta_j]/[(1-\alpha)y_{ji}]$ and $d_{ij} = [1+y_{ji}\delta_j]/[(1-\alpha)y_{ji}]$, so $e_{ij}, d_{ij} > 0$. We obtain

$$\bar{P}_j \geq P_j > P_j - e_{ij} = \bar{P}_j - d_{ij}$$

and hence $\max_{m_i \in M_i} \Pi_j(m_i, s_j^T) = \bar{P}_j$. Player i's own payoff can be maximized only by either a message $m_i = (\eta_i^*, 1)$ for $\eta_i^* \in \arg\max_{\eta_i \in \Theta_i} P_i(\eta_i, \theta_i)$, or by message $m_i = (\nu, 1)$. Hence, one of these messages must belong to $E_{ij}(s_j^T|\theta_i)$. All of them yield the same minimal payoff for j, so $\min_{m_i \in E_{ij}(s_j^T|\theta_i)} \Pi_j(m_i, s_j^T) = \bar{P}_j - d_{ij}$. As a result, $\Pi_i^{e_i}(s_j^T|\theta_i) = P_j - 1/y_{ji}$ and $\kappa_{ij}((\theta_i, 0), s_j^T|\theta_i) = 1/y_{ji}$.

(b): $-1/[(1-\alpha)y_{ji}] < \delta_j < 0$. Let $e_{ij} = 1/[(1-\alpha)y_{ji}]$ and $d_{ij} = [1+(1-\alpha)y_{ji}\delta_j]/[(1-\alpha)y_{ji}]$, so again $e_{ij}, d_{ij} > 0$. We obtain

$$P_j > \bar{P}_j > P_j - e_{ij} = \bar{P}_j - d_{ij}.$$

Thus, $\max_{m_i \in M_i} \Pi_j(m_i, s_j^T) = P_j$ and $\min_{m_i \in E_{ij}(s_j^T|\theta_i)} \Pi_j(m_i, s_j^T) = P_j - e_{ij}$, by the same argument as in case (a). This again implies $\Pi_i^{e_i}(s_j^T|\theta_i) = P_j - 1/y_{ji}$ and $\kappa_{ij}((\theta_i, 0), s_j^T|\theta_i) = 1/y_{ji}$.

(c): $\delta_j \leq -1/[(1-\alpha)y_{ji}]$. Let $e_{ij} = -\delta_j$ and $d_{ij} = [1 + y_{ji}\delta_j]/[\alpha y_{ji}]$, so $e_{ij} > 0$ and $d_{ij} < 0$. We obtain

$$\bar{P}_j - d_{ij} \geq P_j > \bar{P}_j = P_j - e_{ij}$$

and $\max_{m_i \in M_i} \Pi_j(m_i, s_j^T) = \bar{P}_j - d_{ij}$. Player i's own payoff can be maximized only by a message $(\eta_i^*, 1)$ where $\eta_i^* \in \arg\max_{\eta_i \in \Theta_i} P_i(\eta_i, \theta_i)$, or by message $(\nu, 0)$. Hence one of these messages must be contained in $E_{ij}(s_j^T|\theta_i)$. All of them yield the same minimal payoff for j, so that $\min_{m_i \in E_{ij}(s_j^T|\theta_i)} \Pi_j(m_i, s_j^T) = \bar{P}_j$, $\Pi_i^{e_i}(s_j^T|\theta_i) = P_j - 1/y_{ji}$, and $\kappa_{ij}((\theta_i, 0), s_j^T|\theta_i) = 1/y_{ji}$.

We have shown that $\kappa_{ij}((\theta_i, 0), s_j^T|\theta_i) = 1/y_{ji}$ can be achieved by an appropriate choice of e_{ij} and d_{ij} in any case. The remainder of the proof is analogous to the proof of Proposition 1.

The case of $n > 2$. As for Proposition 1, the arguments for $n = 2$ can be generalized to the

case of $n > 2$, now using message sets $M_i = (\Theta_i \cup \{\nu\}) \times [\{0\} \cup (I \backslash \{i\})]$. □

The mechanism constructed in the proof can be interpreted as follows: We first fix the direct revelation mechanism for f and extend it with veto rights. Analogous to the construction for Proposition 1, we then allow each player to claim extra payments from any other player, over and above the transfers of f or the status quo allocation. The goal of these extra payments is again to manipulate kindness of truth-telling to values that turn players into maximizers of the sum of expected material payoffs. The amount of these payments depends on whether the claims are coupled with either a type report or the execution of the veto right. When enforcement of the status quo benefits or does not hurt player j too strongly (cases (a) and (b) in the proof), then the transfers that i can claim from j in addition to executing the veto are designed to give the same minimal payoff to player j as when i reports a type and claims the associated additional transfers. This minimum is therefore independent of whether or not the status quo is bilaterally Pareto efficient from the perspective of type θ_i of player i, which yields the desired type-invariance. If enforcement of the status quo hurts player j strongly (case (c) in the proof), then execution of the veto defines the minimal payoff that i can give to j. The transfers that i can claim in addition to reporting a type are tailored to induce that same minimum, which again yields type-invariance. The transfers that i can claim in addition to the veto become negative and are used to adjust the maximal payoff that i can give to j until the desired equitable payoff is obtained. This construction stabilizes equitable payoffs and therefore kindness such that it does not vary with the realized type in equilibrium.

5 Public Goods Example

5.1 Environment

In this section, we will illustrate the relevance of the insurance property and provide examples of the mechanisms used in our proofs. We work with a simple public goods application.[15] Consider an environment with two players, $I = \{1, 2\}$, and the problem of whether or not to provide a public good, $Q = \{0, 1\}$. Each player can be of either high or low type, $\Theta_i = \{\theta_i^L, \theta_i^H\}$, both of which are equally likely. Types capture the players' willingness to pay for the public good, so we have $v_i(1, \theta_i) = \theta_i - c$ and $v_i(0, \theta_i) = 0$, where $c > 0$ is the per capita cost of providing the public good, assumed to be shared equally by default. We assume

$$c < \theta_1^L < \theta_1^H \text{ and } \theta_2^L < c < \theta_2^H,$$

which implies that player 1 would always like to have the public good provided, but player 2 only if he has the high type. We also assume that

$$\theta_1^L + \theta_2^L < 2c < \theta_1^H + \theta_2^L,$$

which implies that Pareto efficiency requires to provide the good except if $\theta = (\theta_1^L, \theta_2^L)$.

[15] This example application was also used in Bierbrauer and Netzer (2011), an earlier version of Bierbrauer and Netzer (2012).

5.2 Expected Externality Mechanism

The expected externality mechanism (AGV) for an efficient decision rule (d'Aspremont and Gerard-Varet, 1979; Arrow, 1979) is ex-post budget balanced, hence materially Pareto efficient, and Bayesian incentive-compatible. As Bierbrauer and Netzer (2012) have shown, it always satisfies the insurance property for the case of two players. Table 1 summarizes this mechanism for our example, by stating the decision rule q^f and player 1's transfer t_1^f. The transfer for player 2 is given by $t_2^f = -t_1^f$.

	θ_2^L	θ_2^H
θ_1^L	$(0, (\theta_2^H - \theta_1^H)/2)$	$(1, (\theta_2^H - \theta_1^H - \theta_1^L + c)/2)$
θ_1^H	$(1, (\theta_2^H - \theta_1^H + \theta_2^L - c)/2)$	$(1, (\theta_2^H - \theta_1^H + \theta_2^L - \theta_1^L)/2)$

Table 1: AGV (q^f, t_1^f).

Ex-ante expected payoffs of the two players in the truth-telling Bayes-Nash equilibrium are

$$P_1 = \frac{1}{2}\theta_2^H + \frac{1}{4}\theta_2^L - \frac{3}{4}c \quad \text{and} \quad P_2 = \frac{1}{2}\theta_1^H + \frac{1}{4}\theta_1^L - \frac{3}{4}c.$$

Due to the insurance property, each player $i = 1, 2$ obtains the same expected payoff P_i even if the other player $j \neq i$ deviates from truth-telling to any of the other possible strategies. This implies that the truth-telling strategy profile is associated with type-invariant kindness levels of zero in this mechanism. Psychological concerns are therefore irrelevant to both players, and Bayesian incentive-compatibility implies that truth-telling is also an IFE-TI.

5.3 Ex-post Fairness Implementation

The previous result can still be seen as a corollary of the general robustness arguments in Bierbrauer and Netzer (2012). The AGV does, however, not guarantee ex-post fairness implementation. To see why, assume that both players have followed a truth-telling strategy ex-interim, have correct beliefs about this fact, and type profile $\theta = (\theta_1^L, \theta_2^L)$ has realized. Since updating the interim kindness values to the new information still results in mutual kindness of zero, the maximization of ex-post utility (3) boils down to a maximization of own material ex-post payoffs for both players. The condition for player 1 wanting to deviate ex-post to the non-truthful type announcement θ_1^H becomes $(\theta_1^L - c) + (\theta_2^L - c)/2 > 0$. With

$$\theta_1^L = 3/2,\ \theta_1^H = 2,\ \theta_2^L = 1/4,\ \theta_2^H = 2,\ c = 1, \tag{4}$$

for instance, we can verify that this is true, so that the expected externality mechanism is not ex-post fairness incentive-compatible.

To achieve ex-post fairness implementation, we can instead apply the construction provided in the proof of Proposition 1 and augment the AGV by giving player $i = 1, 2$ the option to take

an additional amount of $e_{ij} = 1/[(1-\alpha)y_{ji}]$ from player $j \neq i$. With

$$\alpha = 1/2,\ y_{12} = 1,\ y_{21} = 1, \tag{5}$$

for instance, we obtain $e_{12} = e_{21} = 2$. Truthful type revelation without claiming this additional payment is then still an IFE-TI, but the associated kindness values are now $\kappa_{ij}(s^*) = 1/y_{ji}$ instead of zero. Updating leaves these values unchanged, so that ex-post utility coincides with the sum of ex-post material payoffs. No player therefore regrets having helped to induce a materially Pareto efficient allocation, or not having taken more money from the other player.

5.4 Voluntary Participation

Consider finally the possibility that each player can veto the AGV mechanism on the interim stage, thereby inducing the null allocation $\bar{a} = (0, 0)$ instead of the allocations described in Table 1. Without social preferences, player 2 of type θ_2^L would strictly prefer to do so whenever $(\theta_1^H - \theta_2^H)/2 + (\theta_2^L - c)/4 < 0$, which is again the case for the parameters introduced in (4) above. The same holds with intention-based social preferences. It is easily verified that both types of player 1 suffer in material terms from using the veto, provided that player 2 always tells the truth. Since $P_2 > 0$, the veto also hurts player 2's expected material payoff, which makes it an inefficient action for both types of player 1. Inefficient actions do not influence equitable payoffs, so truth-telling of player 1 remains associated with a type-invariant kindness of zero, from our earlier arguments. Player 2 then cares for material payoffs only, and still prefers to opt out of the mechanism if type θ_2^L has realized.

To guarantee voluntary participation, we can use the construction provided in the proof of Proposition 2. With the parameters given in (4) and (5), case (b) from the proof applies. As before, we obtain the payments $e_{12} = e_{21} = 2$ that each player can claim from the other if none of them makes use of the veto. We obtain the smaller payments $d_{12} = 11/8$ and $d_{21} = 27/16$ that can be claimed when at least one player makes use of the veto. The value of d_{ij} is defined by the equality $-d_{ij} = P_j - e_{ij}$ and ensures that the maximal damage that player i can do to j by using a veto strategy is the same as by using a type-announcement strategy.

6 Conclusions

We have studied notions of robustness in implementation for a mechanism design framework where agents are characterized by intention-based social preferences. Within this model, players are uncertain about their opponents' intentions, as they are uncertain about their material payoff types. We have firstly addressed robustness with regard to assumptions about how agents accommodate the uncertainty about others' intentions. Our concept of implementation in IFE-TI provides robustness in this regard, as it renders intentions transparent despite the presence of asymmetric information. We have secondly proposed a notion of ex-post fairness implementation, which provides robustness to the extent that no player would want to change his interim decision even if he were informed about the others' private information ex-post. This concept rules out any ex-post regret.

As our main result, we have established that any materially Pareto-efficient social choice function which provides insurance can be implemented under both robustness concepts, even if participation in the mechanism is voluntary. The insurance property is essential to our construction, because it facilitates the property of type-invariant kindness. The mechanisms that we construct allow the designer to manipulate reference points for kindness perceptions in order to align individual and social motives both on the interim stage and on the ex-post stage, even if participation is voluntary.

References

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

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.

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

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

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

Bierbrauer, F. and Netzer, N. (2011). Mechanism design and intentions. Mimeo.

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

Blount, S. (1995). When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, 63:131–144.

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

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 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.

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

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

Crawford, V., Neeman, Z., Kugler, T., and Pauzner, A. (2009). Behaviorally optimal auction design: Examples and observations. *Journal of the European Economic Association*, 7:377–387.

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

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

de Clippel, G. (2012). Behavioral implementation. Mimeo.

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.

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

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

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

Dufwenberg, M., Smith, A., and Van Essen, M. (2013). Hold-up: With a vengeance. *Economic Inquiry*, 51:896–908.

Eisenhuth, R. (2012). Reference dependent mechanism design. Mimeo.

Eisenhuth, R. and Ewers, M. (2012). Auctions with loss averse bidders. Mimeo.

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

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.

Filiz-Ozbay, E. and Ozbay, E. (2007). Auctions with anticipated regret: Theory and experiment. *American Economic Review*, 97:1407–1418.

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.

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

Hoffmann, M. and Kolmar, M. (2013). Intention-based fairness preferences in two-player contests. *Economics Letters*, 120:276–279.

Jehiel, P., Meyer-Ter-Vehn, M., Moldovanu, B., and Zame, W. (2006). The limits of ex post implementation. *Econometrica*, 74:585–610.

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*.

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.

Luce, R. D. and Raiffa, H. (1957). *Games and Decisions*. John Wiley and Sons, Inc., 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. and Satterthwaite, M. (1983). Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, 28:265–281.

Netzer, N. and Schmutzler, A. (2013). Explaining gift-exchange – the limits of good intentions. *Journal of the European Economic Association*, forthcoming.

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

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

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

Saran, R. (2011). Menu-dependent preferences and 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.

Segal, U. and Sobel, J. (2008). A characterization of intrinsic reciprocity. *International Journal of Game Theory*, 36:571–585.

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.