Ex-Post Implementation with Social Preferences*

Boaz Zik†

Abstract

The current literature on mechanism design in models with social preferences discusses social-preference-robust mechanisms, i.e., mechanisms that are implementable in any environment with social preferences. The literature also discusses payoff-information-robust mechanisms, i.e., mechanisms that are implementable for any belief and higher-order beliefs of the agents about the payoff types of the other agents. In the present paper we address the question of whether deterministic mechanisms that are robust in both of these dimensions exist. We consider environments where each agent holds private information about his personal payoff and about the existence and extent of his social preferences. In such environments a mechanism is robust in both dimensions only if it is ex-post implementable, i.e., only if incentive compatibility holds for every realization of payoff signals and for every realization of social preferences. We show that ex-post implementation of deterministic mechanisms is impossible in such environments; i.e., deterministic mechanisms that are both social-preference-robust and payoff-information-robust do not exist. (Keywords: Mechanism design; Social preferences; Ex-post implementation; Bayesian implementation.)

*I would like to thank Alex Gershkov, Ilan Kremer, Motty Perry, Phil Reny, Assaf Romm, and participants of various seminars for their valuable comments.

†Department of Economics and Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem.

1 Introduction

Models of mechanism design usually consider selfish agents, that is, agents whose utilities consist of their own personal payoffs. However, it is well established that in many economic environments subjects often have social preferences.[1] In these environments agents' utilities depend not only on their own personal payoff but also on the payoffs of other agents in the society. In this paper, we study the problem of ex-post implementation of deterministic mechanisms in a simple model of social preferences.[2] We consider environments where each agent holds private information about his personal payoff from allocations and about the extent of his social preferences.[3]

In the first part of the paper we investigate the implementation of decision rules that depend only on information about the agents' personal payoffs. We find that the possibility of implementing such decision rules in environments with social preferences heavily depends on the solution concept that is used for the implementation. We first consider Bayesian implementation and reestablish the result of Bierbrauer and Netzer (2016) that for each decision rule that is implementable in the environment where agents are selfish, there exists a mechanism that implements it in a Bayes–Nash equilibrium in every environment with social preferences as well as in the environment where agents are selfish. We then consider ex-post implementation and show our main result that the ex-post implementation of non-trivial decision rules is impossible in environments with social preferences.

In the second part of the paper we consider the ex-post implementation of decision rules that depend both on information about the agents' personal payoffs and on information about the agents' social preferences. We present an impossibility result on ex-post implementation in environments where there exists an agent whose utility depends on the payoff of a selfish agent. This result indicates that the difficulty of robust implementation extends beyond decision rules that depend only on the agents'

[1]There is evidence in the experimental economics literature that subjects often have such "social" preferences. See Cooper and Kagel (2016) for a survey.

[2]Similar models appears in Morgan and Steiglitz (2003) and in Brandt, Sandholm, and Shoham (2007).

[3]That is, the dependency of an agent's utility on the payoffs of other agents is a function of a signal that is privately known to the agent.

payoff signals.

One way to put this work in perspective is to consider the existing literature on implementation in models with social preferences and in particular the papers of Bierbrauer and Netzer (2016), Bartling and Netzer (2016), and Bierbrauer, Ockenfels, Pollak, and Rückert (2017). The focus of these papers is on the implementation of decision rules that depend only on agents' payoff types. They revolve around the notion of *social-preference-robust mechanisms*, i.e., mechanisms that are implementable in any setup with social preferences, including the setup where agents are selfish. Such mechanisms ensure the implementability of a decision rule even if there is no common knowledge about the existence and extent of agents' social preferences. Bartling and Netzer (2016) and Bierbrauer, Ockenfels, Pollak, and Rückert (2017) conduct experiments that show that social-preference-robust mechanisms perform significantly better than mechanisms that are suitable only to the setup where agents are selfish. These findings indicate that this notion of robustness is indeed important. Another important dimension of robustness is the robustness to the distributions of other agents' payoff signals. A mechanism is *payoff-information-robust* if it ensures the implementability of the decision rule for any belief and higher-order beliefs of the agents about the payoff types of the other agents, see Bergemann and Morris (2005). The social-preference-robust mechanisms that are considered in Bierbrauer and Netzer (2016) and Bartling and Netzer (2016) are only Bayesian implementable and are not payoff-information-robust.[4] The question arises of whether it is possible to construct a mechanism that is robust both in the dimension of the agents' payoff information and in the dimension of social preferences. Such a mechanism would require that the incentive-compatibility constraints of each agent hold for every realization of payoff signals and for every realization of social preferences. Our first result on the impossibility of ex-post implementation implies that it is impossible to construct mechanisms that are robust in both of these dimensions.[5]

[4]Bartling and Netzer (2016) investigate the trade-off between belief-robust implementation and externality-robust implementation. They examine participants' behavior both in the second-price auction, which is dominant-strategy implementable but is not robust to the existence of social preferences, and in its externality-robust counterpart, which is robust to the existence of social preferences but is only Bayesian implementable.

[5]Bierbrauer, Ockenfels, Pollak, and Rückert (2017) consider a bilateral trade problem environment

A second way to put our work in perspective is to recall the literature on the impossibility of robust implementation. Our impossibility results on ex-post implementation join several other impossibility results on implementation by robust solution concepts in the literature. In environments of private values and unrestricted preferences, Gibbard (1973) and Satterthwaite (1975) show that if the cardinality of the set of social alternatives is greater than or equal to three, then only dictatorial decision rules are implementable in dominant strategies. Most of the literature on implementation, however, focuses on environments with quasilinear preferences. In such environments an agent's utility is affected by his personal transfer in an additive manner independently of the realization of signals. The designer can use these personal transfers to assist him in aligning agents' preferences with her preferences. The analysis of robust implementation in these environments provides positive results both in the case of private values and in the case of interdependent values and single-dimensional signals. In the case of interdependent values and multidimensional signals, however, Jehiel, Meyer-ter Vehn, Moldovanu, and Zame (2006) show that for generic valuation functions only constant decision rules are ex-post implementable. In this paper, we consider environments in which an agent's utility depends on another agent's payoff. This implies that an agent's personal transfer affects not only the preferences of the agent who receives the transfer but also the preferences of other agents whose utilities depend on this agent's payoff. On the one hand, this provides the designer with more ways to align agents' preferences with her preferences with respect to the standard quasilinear environment. On the other hand, since an agent's transfer affects the incentives of other agents, this property is also confining. Our results show that ultimately environments with social preferences do not allow robust implementation.

The rest of the paper is organized as follows. In Section 2 we present the model. In Section 3 we discuss the implementation of decision rules that depend only on agents' payoff signals. We characterize the set of Bayes–Nash implementable decision rules

where both the buyer and the seller have two types and present mechanisms that are both social-preference-robust and robust to the information of others. The existence of such mechanisms is possible because the decision rules they consider do not satisfy the assumption of our impossibility theorem. We discuss this point further in Section 3.

and construct a transfer scheme that implements a decision rule that belongs to this set in every setup with social preferences as well as in the independent private values setup. We present an impossibility result of ex-post implementation. In Section 4 we present an impossibility result on the ex-post implementation of decision rules that depend both on agents' payoff signals and on their social preferences. We also discuss the difference between our model and the classical interdependent values model. Section 5 concludes. The proof of Proposition 4 is relegated to the Appendix.

2 The Model

We consider a model with two agents, $i \in I = \{1, 2\}$, and two social alternatives,[6] $A = \{a, b\}$. Each agent $i \in I$ receives a signal $\theta_i \in \Theta_i$, where Θ_i is a convex subset of a finite dimensional Euclidean space. If alternative $k \in A$ is chosen, if the signal realization is θ_i, and if agent i obtains a transfer t_i, then agent i's payoff is given by $\Pi_i = v_i(k, \theta_i) + t_i$. We assume that $v_i(k, \theta_i)$ is a convex function of θ_i for every $i \in I$. The utility of agent i depends in a linear manner on her personal payoff and on the payoff of agent j, i.e., $u_i = \Pi_i + \delta_i \cdot \Pi_j$, where $\delta_i \in \left[\underline{\delta_i}, \overline{\delta_i}\right] \subset \mathbb{R}$ with $\underline{\delta_i} < \overline{\delta_i}$ and[7] $0 \in \left[\underline{\delta_i}, \overline{\delta_i}\right]$. The signals θ_i and δ_i are the private information of agent i. We denote $\Theta \coloneqq \bigtimes_{i\in I} \Theta_i$ with generic element θ, and $\mathcal{D} \coloneqq \bigtimes_{i\in I} \left[\underline{\delta_i}, \overline{\delta_i}\right]$ with generic element δ. A function $q : \Theta \times \mathcal{D} \to A$ is called a *decision rule*. A *social choice function* is a function $s(\theta, \delta) = (q(\theta, \delta), t_1(\theta, \delta), t_2(\theta, \delta))$, where $q(\theta, \delta) \in A$ and $t_i(\theta, \delta) \in \mathbb{R}$ for every $i \in I$. We say that a social choice function $(q(\theta, \delta), t_1(\theta, \delta), t_2(\theta, \delta))$ is *ex-post implementable* if for every $i \in I$, and $(\theta, \delta) \in \Theta \times D$ we have

$$(\theta_i, \delta_i) \in \underset{(\hat{\theta}_i, \hat{\delta}_i) \in \Theta_i \times \left[\underline{\delta_i}, \overline{\delta_i}\right]}{\arg\max} v_i\left(q\left(\hat{\theta}_i, \hat{\delta}_i, \theta_j, \delta_j\right), \theta_i\right) + t_i\left(\hat{\theta}_i, \hat{\delta}_i, \theta_j, \delta_j\right) +$$

$$+\delta_i\left(v_j\left(q\left(\theta_j, \delta_j, \hat{\theta}_i, \hat{\delta}_i\right), \theta_j\right) + t_j\left(\theta_j, \delta_j, \hat{\theta}_i, \hat{\delta}_i\right)\right)$$

[6] This 2×2 model is embedded in every model with more agents and alternatives, and the impossibility result for this model therefore extends to the general model of N agents and K alternatives.

[7] We assume that 0 is in the support because we want to consider environments where both the existence and the extent of social preferences are not commonly known.

A decision rule $q(\theta, \delta)$ is *ex-post implementable* if there exists a profile of real-valued functions $(t_1(\theta, \delta), t_2(\theta, \delta))$ such that $(q(\theta, \delta), t_1(\theta, \delta), t_2(\theta, \delta))$ is ex-post implementable.

3 Decisions that Depend only on Payoff Signals

In this section we discuss the implementation of decision rules that depend only on information about the personal payoffs of the agents. We consider situations where the designer wants to implement such a decision rule irrespective of whether agents are selfish or have social preferences. A first line of such situations is agency problems in institutions with a hierarchical organizational structure. For example, consider a conglomerate's central administration that needs to choose an alternative from a set of possible alternatives. The central administration wants to choose the alternative that maximizes the conglomerate's profit, i.e., that maximizes the sum of the profits of the conglomerate's corporations. The effect of each alternative on a corporation's profit is the private information of the corporation's manager. Now in many environments managers' utilities may depend not only on the profits of their corporations but also on the profits of other corporations in the conglomerate. Such dependency may occur, for example, when a manager is a shareholder in the conglomerate and, therefore, profits from its success; when a manager is rewarded according to the relative success of her corporation with respect to the other corporations in the conglomerate; when a manager is connected in some way (say, through family, friendship, or business ties) to other managers in the conglomerate; or when a manager is invested in some other corporation of the conglomerate.

A second line of situations is that of utilitarian designers who are called upon to choose a social alternative. Consider a society some of whose members may have antisocial preferences, such as envy, spite, and so on. In such a society utilitarian theory suggests that the agents' preferences will be "laundered," i.e., that the antisocial aspects in these preferences will be removed before the preferences are incorporated into the social utility.[8] Harsanyi, one of the greatest advocates of utilitarian theory, suggests that:

[8] See, for example, Harsanyi (1977), Goodin (1986), and Blanchet and Fleurbaey (2006).

> Some preferences ... must be altogether excluded from our social-utility function. In particular we must exclude all clearly antisocial preferences such as sadism, envy, resentment and malice. ... Utilitarian ethics makes all of us members of the same moral community. A person displaying ill will toward others does remain a member of this community, but not with his whole personality. That part of his personality that harbors these hostile antisocial feelings must be excluded from membership, and has no claim to a hearing when it comes to defining our concept of social utility (Harsanyi 1977, p. 647)

Laundering preferences means that when the designer is called to choose the social alternative, she should consider only information about agents' personal payoffs and disregard information about agents' social preferences. That is, her optimal decision rule depends only on agents' payoff signals.

The question of whether it is possible to Bayesian implement decision rules that depend only on agents' payoff signals in the presence of social preferences is analyzed in Bierbrauer and Netzer (2016) and Bartling and Netzer (2016). They show that any decision rule that is Bayesian implementable in the environment where agents are selfish is also Bayesian implementable in any environment with social preferences. Moreover, there exists a mechanism that implements the decision rule in a Bayes–Nash equilibrium in every environment with social preferences as well as in the environment where agents are selfish. Such a mechanism is called a *social-preference-robust mechanism.* The construction of this mechanism is achieved by constructing a transfer scheme that eliminates the effect of agent i's report on the expected payoff of agent j. At the same time, this transfer scheme incentivizes agent i to report truthfully when she is interested in maximizing her own personal payoff. Therefore, this transfer scheme incentivizes truth telling in every setup. We now show this result formally.

Proposition 1. *Consider a profile* Θ, $(v_i)_{i\in I}$. *Let* $(q(\theta), t_1(\theta), t_2(\theta))$ *be Bayesian implementable in the environment where agents are selfish; then there exists a social choice function* $\left(q(\theta), t_1'(\theta), t_2'(\theta)\right)$ *that is Bayesian implementable in any environment with with social preferences and in the environment where agents are selfish.*

Proof. Given a transfer scheme $(t_i(\theta))_{i\in I}$ that implements $q(\theta)$ in the environment where agents are selfish, we define $\left(t_i'(\theta)\right)_{i\in I}$ to be

$$t_i'(\theta) = t_i(\theta) - E_{\tilde{\theta}_i}\left[v_i\left(q\left(\theta_j, \tilde{\theta}_i\right), \tilde{\theta}_i\right) + t_i\left(\theta_j, \tilde{\theta}_i\right)\right]$$

Let $j \in I$. The second additive term in the transfer function $t_i'(\theta)$ eliminates the effect of the report of agent j on the expected payoff of agent i from agent j's perspective. In other words, from agent j's perspective, the report $\hat{\theta}_j$ does not affect the expected payoff of agent i. It is therefore sufficient to show that $\left(t_i'(\theta)\right)_{i\in I}$ Bayesian implements $q(\theta)$ in the environment where agents are selfish. This follows from the fact that $t_i'(\theta)$ equals $t_i(\theta)$ plus additive terms that do not depend on $\hat{\theta}_i$ and that $(t_i(\theta))_{i\in I}$ Bayesian implements $q(\theta)$ in the environment where agents are selfish. □

Remark. The construction of the social-preference-robust mechanism is based on two properties. The first is that under this mechanism agent i's actions cannot affect the payoff of agent j. This property is referred to in the literature as *externality-free.* The second property is that the mechanism is incentive compatible in the environment where agents are selfish. Externality-free and incentive compatibility imply the social-preference-robustness of the mechanism not only in our particular model but in every model of social preferences in which agents behave selfishly whenever they cannot affect other agents' payoffs.[9]

3.1 The Impossibility of Ex-post Implementation

Another renowned and important dimension of robustness is robustness to the payoff information of others. A mechanism is *payoff-information-robust* if it ensures the implementability of the decision rule for any belief and higher-order beliefs of the agents about the payoff types of the other agents, see Bergemann and Morris (2005). Wilson (1987) suggests that mechanisms should be free from assumptions of common knowledge. The question then arises whether it is possible to implement decision rules in environments where there is no common knowledge of the distribution of other agents'

[9] For example, inequity aversion models (Fehr and Schmidt (,1999) and models of intention-based preferences (Rabin (,1993).

payoff signals nor of the presence and the extent of social preferences. Robustness in both of these dimensions is captured by the notion of ex-post implementation, which requires that the strategy of each agent i be optimal with respect to the strategies of the other agents for every possible realization of payoff signals and social preferences. In the following theorem we show that it is impossible to ex-post implement a decision rule that depends only on agents' payoff signals. This result implies that it is impossible to construct a mechanism that is robust in both dimensions.

Theorem 2. *Consider a profile Θ, $(v_h)_{h\in I}$ and a decision rule $q(\theta)$. If there exist two signals θ_i and $\theta_i^{'}$ and two signals θ_j and $\theta_j^{'}$, such that $q(\theta_i,\theta_j) = q\left(\theta_i,\theta_j^{'}\right) = a$ and $q\left(\theta_i^{'},\theta_j\right) = q\left(\theta_i^{'},\theta_j^{'}\right) = b$ and $v_j(a,\theta_j) - v_j(b,\theta_j) \neq v_j\left(a,\theta_j^{'}\right) - v_j\left(b,\theta_j^{'}\right)$, then $q(\theta)$ is not ex-post implementable.*

Note that Theorem 2 implies the impossibility of ex-post implementation of non-trivial deterministic decision rules in the standard settings of the mechanism design literature such as auctions and public goods environments.[10] The argument behind this result is the following. Ex-post implementation implies that for any two signals θ_i and $\theta_i^{'}$ the payoff of agent j must remain equal on a subset of measure one of the interval $\left[\underline{\delta_i},\overline{\delta_i}\right]$ for any fixed (θ_j,δ_j). Therefore, if the decision rule assigns different alternatives for θ_i and $\theta_i^{'}$, and if agent j's valuation is different for each alternative, it is left for agent j's transfer function t_j to eliminate this gap in agent j's payoff. However, t_j also plays a role in incentivizing agent j to report truthfully. These two roles of t_j lead to a contradiction and hence make ex-post implementation impossible.

Lemma 3. *Let $q(\theta)$ be ex-post implementable. Consider some (θ_j,δ_j) for every $\theta_i,\theta_i^{'}\in\Theta_i$; we have that $\Pi_j(\theta_j,\delta_j,\theta_i,\cdot) \overset{a.e}{=} \Pi_j\left(\theta_j,\delta_j,\theta_i^{'},\cdot\right)$.*

Proof of Lemma 3. Consider some (θ_j,δ_j). The payoff of agent j given (θ_j,δ_j) as a function of agent i's report, $\left(\hat{\theta}_i,\hat{\delta}_i\right)$, is $\Pi_j\left(\theta_j,\delta_j,\hat{\theta}_i,\hat{\delta}_i\right) = v_j\left(q\left(\theta_j,\hat{\theta}_i\right),\theta_2\right) + t_j\left(\theta_j,\delta_j,\hat{\theta}_i,\hat{\delta}_i\right)$. The transfer of agent i given (θ_j,δ_j) as a function of agent i's report is $t_i\left(\hat{\theta}_i,\hat{\delta}_i,\theta_j,\delta_j\right)$. Agent i's utility function given (θ_j,δ_j) is $v_i\left(q\left(\hat{\theta}_i,\theta_j\right),\theta_i\right) + \delta_i\Pi_j\left(\theta_j,\delta_j,\hat{\theta}_i,\hat{\delta}_i\right) + t_i\left(\hat{\theta}_i,\hat{\delta}_i,\theta_j,\delta_j\right)$. Now assume that agent i reports δ_i truthfully.

[10] It is enough that the decision rule is not a constant function of agent i's signals for two given types of agent j.

Ex-post implementability implies that he must report θ_i truthfully. The problem is therefore to incentivize agent i to report θ_i truthfully when his utility function is $v_i\left(q\left(\hat{\theta}_i,\theta_j\right),\theta_i\right)+\delta_i\Pi_j\left(\theta_j,\delta_j,\hat{\theta}_i,\delta_i\right)+t_i\left(\hat{\theta}_i,\delta_i,\theta_j,\delta_j\right)$. This problem is equivalent to the problem of incentivizing him to report truthfully in the environment where agents are selfish.[11] Since Θ_i is a convex subset of a finite dimensional Euclidean space and since $v_i\left(k,\theta_i\right)$ is a convex function of θ_i, revenue equivalence holds; i.e., the transfer to agent i given θ_j in any transfer scheme that implements $q\left(\theta\right)$ is unique up to a constant.[12] Hence a truthful report of θ_i implies that for every $\delta_i\in\left[\underline{\delta_i},\overline{\delta_i}\right]$ and $\theta_i\in\Theta_i$ we have

$$\delta_i\Pi_j\left(\theta_j,\delta_j,\hat{\theta}_i,\delta_i\right)+t_i\left(\hat{\theta}_i,\delta_i,\theta_j,\delta_j\right)=\varphi_i\left(\theta_i,\theta_j\right)+\sigma_i\left(\delta,\theta_j\right)\tag{1}$$

where $\varphi_i:\Theta_i\times\Theta_j\to\mathbb{R}$ and[13] $\sigma_i:\mathcal{D}\times\Theta_2\to\mathbb{R}$. On the other hand, assume that agent i reports θ_i truthfully. Ex-post implementability implies that he must report δ_i truthfully; i.e, for every $\theta_i\in\Theta_i$ and $\delta_i\in\left[\underline{\delta_i},\overline{\delta_i}\right]$ we have

$$v_i\left(q\left(\theta_i,\theta_j\right),\theta_i\right)+\delta_i\Pi_j\left(\theta_j,\delta_j,\theta_i,\delta_i\right)+t_i\left(\theta_i,\delta_i,\theta_j,\delta_j\right)\geq$$

$$v_i\left(q\left(\theta_i,\theta_j\right),\theta_i\right)+\delta_i\Pi_j\left(\theta_j,\delta_j,\theta_i,\hat{\delta}_i\right)+t_i\left(\theta_i,\hat{\delta}_i,\theta_j,\delta_j\right)$$

for every $\hat{\delta}_i\in\left[\underline{\delta_i},\overline{\delta_i}\right]$. Subtracting $v_i\left(q\left(\theta_i,\theta_j\right),\theta_i\right)$ from both sides of the inequality we have

$$\delta_i\Pi_j\left(\theta_j,\delta_j,\theta_i,\delta_i\right)+t_i\left(\theta_i,\delta_i,\theta_j,\delta_j\right)\geq\delta_i\Pi_j\left(\theta_j,\delta_j,\theta_i,\hat{\delta}_i\right)+t_i\left(\theta_i,\hat{\delta}_i,\theta_j,\delta_j\right)$$

---

[11] Define $\tilde{t}_i^{\delta}\left(\hat{\theta}_i,\theta_j\right)=\delta_i\Pi_j\left(\theta_j,\delta_j,\hat{\theta}_i,\delta_i\right)+t_i\left(\hat{\theta}_i,\delta_i,\theta_j,\delta_j\right)$ and the problem is to incentivize agent i to report θ_1 truthfully given that his utility is $v_i\left(q\left(\hat{\theta}_i,\theta_j\right),\theta_i\right)+\tilde{t}_i^{\delta}\left(\hat{\theta}_i,\theta_j\right)$.

[12] See Krishna and Maenner (2001).

[13] Revenue equivalence means that $\tilde{t}_i^{\delta}\left(\hat{\theta}_i,\theta_j\right)$ equals the sum of a function that depends on θ_i, which we denote by $\varphi_i\left(\theta_i,\theta_j\right)$, and a constant, which we denote by $\sigma_i\left(\delta,\theta_j\right)$.

for every $\hat{\delta}_i \in \left[\underline{\delta_i}, \overline{\delta_i}\right]$. This implies that[14]

$$(2)\ \delta_i \Pi_j\left(\theta_j, \delta_j, \theta_i, \delta_i\right)+t_i\left(\theta_i, \delta_i, \theta_j, \delta_j\right)=\underline{\delta_i} \Pi_j\left(\theta_j, \delta_j, \theta_i, \underline{\delta_i}\right)+t_i\left(\theta_i, \underline{\delta_i}, \theta_j, \delta_j\right)+\int_{\underline{\delta_i}}^{\delta_i} \Pi_j\left(\theta_j, \delta_j, \theta_i, s\right) ds$$

Combining equations (1) and (2) yields that for every $\delta_i \in \left[\underline{\delta_i}, \overline{\delta_i}\right]$ and every $\theta_i \in \Theta_i$, $\int_{\underline{\delta_i}}^{\delta_i} \Pi_j\left(\theta_j, \delta_j, \theta_i, s\right)\ ds = \sigma\left(\delta_i, \delta_j, \theta_j\right) - \sigma\left(\underline{\delta_i}, \delta_j, \theta_j\right)$. This implies that that for every $\theta_i, \theta_i^{'} \in \Theta_i$ we have that $\Pi_j\left(\theta_j, \delta_j, \theta_i, \cdot\right) \overset{a.e}{=} \Pi_j\left(\theta_j, \delta_j, \theta_i^{'}, \cdot\right)$ $\square$

We now complete the proof by showing that the requirements that Lemma 3 imposes on agent *j's* transfer function contradict the requirements that incentive compatibility imposes on agent *j's* transfer function.

Proof of Theorem 2. Assume that $\delta_j = 0$. According to the assumption of the theorem there exist signals θ_i, $\theta_i^{'}$, θ_j and $\theta_j^{'}$ such that $q\left(\theta_i, \theta_j\right) = q\left(\theta_i, \theta_j^{'}\right) = a$, $q\left(\theta_i^{'}, \theta_j\right) = q\left(\theta_i^{'}, \theta_j^{'}\right) = b$, and $v_j\left(a, \theta_j\right) - v_j\left(b, \theta_j\right) \neq v_j\left(a, \theta_j^{'}\right) - v_j\left(b, \theta_j^{'}\right)$. In addition, Lemma 3 implies that we can find a signal δ_i such that $\Pi_j\left(\theta_j, \delta_j, \theta_i, \delta_i\right) = \Pi_j\left(\theta_j, \delta_j, \theta_i^{'}, \delta_i\right)$ and $\Pi_j\left(\theta_j^{'}, \delta_j, \theta_i, \delta_i\right) = \Pi_j\left(\theta_j^{'}, \delta_j, \theta_i^{'}, \delta_i\right)$. This yields that

$$t_j\left(\theta_j, \delta_j, \theta_i, \delta_i\right) - t_j\left(\theta_j, \delta_j, \theta_i^{'}, \delta_i\right) \neq t_j\left(\theta_j^{'}, \delta_j, \theta_i, \delta_i\right) - t_j\left(\theta_j^{'}, \delta_j, \theta_i^{'}, \delta_i\right)$$

However, since $\delta_j = 0$ we get that for agent j to report truthfully, function t_j must assign the same transfer to signals that map the same alternative for a given report of agent i. This implies that

$$t_j\left(\theta_j, \delta_j, \theta_i, \delta_i\right) - t_j\left(\theta_j, \delta_j, \theta_i^{'}, \delta_i\right) = t_j\left(\theta_j^{'}, \delta_j, \theta_i, \delta_i\right) - t_j\left(\theta_j^{'}, \delta_j, \theta_i^{'}, \delta_i\right)$$

a contradiction. $\square$

Remark. Our result concerns decision rules that depend only on agents' payoff signals. However, throughout the analysis we have allowed agents' transfers to depend also on the information about the agents' social preferences. In that sense we showed that the

[14] This stems from the following result. Let $u(\delta, \hat{\delta}) = \delta \cdot q\left(\hat{\delta}\right) + t\left(\hat{\delta}\right)$. If for every $\delta \in [\underline{\delta}, \overline{\delta}]$, $\delta \in \underset{\hat{\delta} \in [\underline{\delta}, \overline{\delta}]}{\arg\max}\, u(\delta, \hat{\delta})$ then for every $\delta \in \left[\underline{\delta}, \overline{\delta}\right]$, $t\left(\delta\right) + \delta q\left(\delta\right) = t\left(\underline{\delta}\right) + \underline{\delta} \cdot q\left(\underline{\delta}\right) + \int_{\underline{\delta}}^{\delta} q(s)\, ds$.

implementation of non-constant **decision rules** is not robust to social preferences. The literature on mechanism design with social preferences speaks of **mechanisms** that are robust to social preferences. In such mechanisms not only the decision rule but also the agents' transfers need not depend on information about social preferences. Therefore, Theorem 2 shows a stronger result that implies the nonexistence of social-preference-robust mechanisms.

Remark. The proof of Theorem 2 is based on two claims. The first claim, which appears in Lemma 3, suggests that ex-post implementation implies that the property of externality-freeness, i.e, the property that agent i cannot affect the payoff of agent j, must hold for every realization of agent j's payoff signals. The second claim suggests that externality-freeness and ex-post incentive compatibility in the case where agent j is selfish cannot coexist. While the first claim depends on our specific model of social preferences, the second claim does not. This means that in any model of social preferences it would be impossible to construct a mechanism that is social-preference-robust and payoff-information-robust by constructing a mechanism that is both externality-free for every realization of signals and ex-post incentive compatible in the environment where agents are selfish. Therefore, in any model of social preferences, any possibility result on the existence of a mechanism that is social-preference-robust and payoff-information-robust would need to be based on a different argument. Moreover, in any model of social preferences it suffices to show that ex-post implementability implies that externality-freeness must hold for every realization of payoff signals in order to prove that mechanisms that are social-preference-robust and payoff-information-robust do not exist.

Remark. Bierbrauer, Ockenfels, Pollak, and Rückert (2017) consider a bilateral trade problem in an environment where both the buyer and the seller have two types. They present non-trivial mechanisms that are social-preference-robust and payoff-information-robust by constructing a mechanism that is both externality-free for every signals realization and ex-post incentive compatible in the environment where agents are selfish. The construction of such a mechanism is possible because the decision rules they consider do not satisfy the assumption of Theorem 2. That is, there is no agent i that is pivotal between two alternatives a and b for two different types of agent j.

4 Discussion

4.1 Decisions that Depend on Social Preferences

In the previous section we discussed the notion of social-preference robustness. This notion is suitable to situations where the designer does not want to condition her decision on the information about the agents' social preferences. In this subsection we consider the possibility of ex-post implementation of decision rules that depend both on information about the agents' payoffs and on information about the extent of the agents' social preferences. We present an impossibility result on ex-post implementation in environments where there is at least one agent whose utility relies on the payoff of a selfish agent. This result shows that at least in this important environment the possibility of conditioning decision rules on information about social preferences does not create enough freedom to enable ex-post implementation.

We consider the 2×2 model that we presented in Section 2 except that now agent 2 is selfish, i.e., $u_1 = \Pi_1 + \delta_1 \cdot \Pi_2$ and $u_2 = \Pi_2$. We now present the impossibility theorem.

Proposition 4. *Consider a decision rule of the following form. There exist two types $\theta_1^{'}$, $\theta_1^{''}$, and two types, $\theta_2^{'}$, $\theta_2^{''}$, and some positive interval $\left[\underline{\delta_1}, \overline{\delta_1}\right]$ such that for every $\delta_1 \in \left[\underline{\delta_1}, \overline{\delta_1}\right]$ we have that*

$$q\left(\theta_1^{'}, \delta_1, \theta_2^{'}\right) = a \text{ and } q\left(\theta_1^{''}, \delta_1, \theta_2^{'}\right) = a$$

$$q\left(\theta_1^{'}, \delta_1, \theta_2^{''}\right) = b \text{ and } q\left(\theta_1^{''}, \delta_1, \theta_2^{''}\right) = b$$

and there exist two signals $\tilde{\theta}_2^{'}$ and $\tilde{\theta}_2^{''}$, such that $q\left(\theta_1^{'}, \delta_1, \tilde{\theta}_2^{'}\right) = q\left(\theta_1^{'}, \delta_1, \tilde{\theta}_2^{''}\right) = a$ and $q\left(\theta_1^{''}, \delta_1, \tilde{\theta}_2^{'}\right) = q\left(\theta_1^{''}, \delta_1, \tilde{\theta}_2^{''}\right) = b$ and $v_2\left(a, \tilde{\theta}_2^{'}\right) - v_2\left(b, \tilde{\theta}_2^{'}\right) \neq v_2\left(a, \tilde{\theta}_2^{''}\right) - v_2\left(b, \tilde{\theta}_2^{''}\right)$, then $q\left(\theta\right)$ is not ex-post implementable.

The proof of Proposition 4 uses a similar argument as the proof of Theorem 2 and therefore is relegated to the Appendix. To illustrate the applicability of Proposition 4 we present the following example of an allocation problem of a single good.

Example 5. Consider a principal who is looking to allocate a single indivisible good between two agents. Each of the agents has a value for the good in $[0, 1]$ and $\delta_1 \in [0.1, 0.2]$. The principal wants to choose the allocation that provides the highest social utility. That is, the optimal decision rule is

$$q(\theta_1, \delta_1, \theta_2) = \begin{cases} 1 & \text{if } \theta_1 > (1 + \delta_1) \cdot \theta_2 \\ 2 & \text{otherwise} \end{cases}$$

where $q = 1$ is the allocation where agent 1 gets the item and $q = 2$ is the allocation where agent 2 gets the item. This decision rule satisfies the condition of Proposition 4 and therefore it is not ex-post implementable.[15]

4.2 Our Model vs. the Interdependent Values Model

In this paper we presented impossibility theorems regarding ex-post implementation in a model with social preferences. Jehiel et al. (2006) present an impossibility result on ex-post implementation in a model with interdependent values. Although our model resembles the model in Jehiel et al. (2006), it is different from their model in the following important respect. In our model an agent's utility depends on the other agent's signals and transfers, while in the interdependent values model an agent's utility depends only on the other agent's signals. In the interdependent values model agent i's report affects his utility through the decision rule and his personal transfer, while in our model agent i's report affects his utility through the decision rule, his personal transfer, and the personal transfer of agent[16] j. That is, mechanisms affect agents' incentives in a more complex way in our model than in the interdependent values model. On the one hand, since an agent's utility is affected by the other

[15] To see this set $\theta_1^{'} = 0.8$, $\theta_1^{''} = 0.4$, $\theta_2^{'} = 0.1$, $\theta_2^{''} = 0.9$, $\tilde{\theta}_2^{'} = 0.6$, and $\tilde{\theta}_2^{''} = 0.5$.

[16] While the effect of the agent's personal transfer on his utility is independent of the realization of signals, the effect of the other agent's transfer on his utility depends on the realization of his signals.

agent's transfer, mechanisms provide more tools to achieve implementation. On the other hand, since each agent's transfer also affects the incentives of the other agent, mechanisms also impose further restrictions on achieving implementation.

To illustrate the difference between the models, we analyze two examples that show that the impossibility of ex-post implementation in one model does not imply the impossibility of ex-post implementation in the other model. In the first example we present a decision rule that is not ex-post implementable in our model but is ex-post implementable in the interdependent values model. In the second example we present decision rules that are ex-post implementable in our model but are not ex-post implementable in the interdependent values model.

When we talk about interdependent values model agent i's utility function is $v_i(q,\theta_i)+\delta_i\cdot v_j(q,\theta_j)+t_i$, where $q\in A$, whereas in our model agent i's utility is $v_i(q,\theta_i)+\delta_i\cdot(v_j(q,\theta_j)+t_j)+t_i$, where $q\in A$.

Example 5 (continued). Consider the setup of Example 5 (for which we have shown that the optimal decision rule is not ex-post implementable in our model) in the interdependent values model. The optimal decision rule is ex-post implementable in the interdependent values model by applying the following transfer scheme:

$$t_1(\theta_1,\delta_1,\theta_2)=\begin{cases}-\theta_2 & \text{if } \theta_1>(1+\delta_1)\cdot\theta_2\\ 0 & \text{otherwise}\end{cases}$$

$$t_2(\theta_1,\delta_1,\theta_2)=\begin{cases}0 & \text{if } \theta_1>(1+\delta_1)\cdot\theta_2\\ -\left(\frac{\theta_1}{1+\delta_1}\right) & \text{otherwise}\end{cases}$$

Example 6. Consider the following setup where for each agent $i\in I$, $\theta_i\in[0,1]$ and $\delta_i\in[0,1]$. Agent i's valuation if alternative a is chosen is $v_i(a,\theta_i)=\theta_i+c$, and his valuation if alternative b is chosen is $v_i(b,\theta_i)=\theta_i$. We analyze the possibility of implementing decision rules that depend only on information about agents' payoffs both in our model and in the interdependent values model.

We first analyze our model. Consider an arbitrary decision rule $q(\theta)$. For every

$i \in \{1, 2\}$ we define the following transfer function:

$$t_i(\theta_i, \delta_i, \theta_j, \delta_j) = \begin{cases} -c & \text{if } q(\theta_i, \theta_j) = a \\ 0 & \text{if } q(\theta_i, \theta_j) = b \end{cases}$$

Under these transfer functions any type (θ_i, δ_i) of agent i receives the same utility, $\theta_i + \delta_i \cdot \theta_j$, irrespective of his report. Therefore, the decision rule is ex-post implementable.[17] We now analyze the interdependent values model and show that it is impossible to ex-post implement non-constant decision rules in this model. Consider an arbitrary type $\left(\tilde{\theta}_j, \tilde{\delta}_j\right)$ of agent j, $j \neq i$. Ex-post implementability implies that for every $(\theta_i, \delta_i), \left(\theta_i^{'}, \delta_i^{'}\right) \in [0,1]^2$ such that $q\left(\theta_i, \tilde{\theta}_j\right) = q\left(\theta_i^{'}, \tilde{\theta}_j\right)$ we have[18] $t_i\left(\theta_i, \delta_i, \tilde{\theta}_j, \tilde{\delta}_j\right) = t_i\left(\theta_i^{'}, \delta_i^{'}, \tilde{\theta}_j, \tilde{\delta}_j\right)$. That is, agent i's transfer function depends only on the chosen alternative; hence, we denote $t_i\left(\theta_i, \delta_i, \tilde{\theta}_j, \tilde{\delta}_j\right) := t_i\left(q(\theta_i, \theta_j), \tilde{\theta}_j, \tilde{\delta}_j\right)$. Consider a non-constant decision rule $q(\theta)$. Look at a type $\left(\tilde{\theta}_j, \tilde{\delta}_j\right)$ of agent j for which agent i is pivotal. This means that there exist two signals $\theta_i^{'}$ and $\theta_i^{''}$ such that $q\left(\theta_i^{'}, \tilde{\theta}_j\right) = a$ and $q\left(\theta_i^{''}, \tilde{\theta}_j\right) = b$. Now, ex-post implementability implies that for every $\delta_i \in [0,1]$ we have that

$$\theta_i^{'} + c + \delta_i \cdot \left(\tilde{\theta}_j + c\right) + t_i\left(a, \tilde{\theta}_j, \tilde{\delta}_j\right) \geq \theta_i^{'} + \delta_i \cdot \tilde{\theta}_j + t_i\left(b, \tilde{\theta}_j, \tilde{\delta}_j\right)$$

and

$$\theta_i^{''} + c + \delta_i \cdot \left(\tilde{\theta}_j + c\right) + t_i\left(a, \tilde{\theta}_j, \tilde{\delta}_j\right) \leq \theta_i^{''} + \delta_i \cdot \tilde{\theta}_j + t_i\left(b, \tilde{\theta}_j, \tilde{\delta}_j\right)$$

hence we get that for every $\delta_i \in [0,1]$

$$c \cdot (1 + \delta_i) = t_i\left(b, \tilde{\theta}_j, \tilde{\delta}_j\right) - t_i\left(a, \tilde{\theta}_j, \tilde{\delta}_j\right)$$

Since the left-hand side of the equation varies with δ_i and the right-hand side of the equation is constant we reach a contradiction.

[17] Ex-post implementation is possible because the assumption of Theorem 2 does not hold.

[18] Assume that $t_i\left(\theta_i, \delta_i, \tilde{\theta}_j, \tilde{\delta}_j\right) > t_i\left(\theta_i^{'}, \delta_i^{'}, \tilde{\theta}_j, \tilde{\delta}_j\right)$; then agent i of type $\left(\theta_i^{'}, \delta_i^{'}\right)$ will have a profitable deviation to (θ_i, δ_i)

5 Conclusion

We have considered the possibility of ex-post implementation in a model with social preferences where each agent holds private information about his personal payoff from allocations and about the extent of his social preferences. We presented an impossibility result on the ex-post implementation of decision rules that depend only on information about agents' payoffs. This result implies that it is impossible to construct mechanisms that are robust to the existence of social preferences and to the distribution of agents' payoff signals. Our impossibility results highlight the question of whether ex-post implementation is possible in other environments with social preferences. The environment we have considered joins several other environments in which it has been shown that implementation in robust solution concepts is impossible. This presents yet another example of the difficulty of carrying out robust implementation.

Appendix

Proof of Proposition 4

Lemma. *Let δ_1, θ_2, θ_1 and $\dot{\theta}_1$ be such that $q(\theta_1, \delta_1, \theta_2) = q\left(\dot{\theta}_1, \delta_1, \theta_2\right) = k$, $k \in \{a, b\}$, then*

$$\begin{aligned}\delta_1 \cdot \Pi_2(\theta_1, \delta_1, \theta_2) + t_1(\theta_1, \delta_1, \theta_2) &= \delta_1 \cdot \Pi_2\left(\dot{\theta}_1, \delta_1, \theta_2\right) + t_1\left(\dot{\theta}_1, \delta_1, \theta_2\right) \\ &\coloneqq \sigma(k, \delta_1, \theta_2)\end{aligned}$$

Proof. Holding δ_1 constant, the problem is equivalent to a standard ex-post implementation problem in an independent private values setting. This implies that given a fixed θ_2 the transfers to agent 1 must be equal for any θ_1 and $\dot{\theta}_1$ that result in the same alternative. □

Assume that agent 2 is of type $\theta_2^{'}$ and that agent 1 is of type $\theta_1^{'}$. Ex-post implementability implies that he must report δ_1 truthfully for every $\delta_1 \in \left[\underline{\delta_1}, \overline{\delta_1}\right]$, i.e., for every $\delta_1, \delta_1' \in \left[\underline{\delta_1}, \overline{\delta_1}\right]$

$$v_1\left(a, \theta_1'\right) + \delta_1 \cdot \Pi_2\left(\theta_1', \delta_1, \theta_2'\right) + t_1\left(\theta_1', \delta_1, \theta_2'\right) \geq$$

$$v_1\left(a, \theta_1'\right) + \delta_1 \cdot \Pi_2\left(\theta_1', \delta_1', \theta_2'\right) + t_1\left(\theta_1', \delta_1', \theta_2'\right)$$

This implies that

$$\delta_1\Pi_2\left(\theta_1', \delta_1, \theta_2'\right) + t_1\left(\theta_1', \delta_1, \theta_2'\right) = \underline{\delta_1}\Pi_2\left(\theta_1', \underline{\delta_1}, \theta_2'\right) + t_1\left(\theta_1', \underline{\delta_1}, \theta_2'\right) + \int_{\underline{\delta_1}}^{\delta_1} \Pi_2\left(\theta_1', s, \theta_2'\right) ds$$

i.e.,

$$\int_{\underline{\delta_1}}^{\delta_1} \Pi_2\left(\theta_1', s, \theta_2'\right) ds = \sigma\left(a, \delta_1, \theta_2'\right) - \sigma\left(a, \underline{\delta_1}, \theta_2'\right)$$

Fixing θ_2' and θ_1'' we get by an identical argument that

$$\int_{\underline{\delta_1}}^{\delta_1} \Pi_2\left(\theta_1'', s, \theta_2'\right) ds = \sigma\left(a, \delta_1, \theta_2'\right) - \sigma\left(a, \underline{\delta_1}, \theta_2'\right)$$

This implies that

$$\Pi_2\left(\theta_1', s, \theta_2'\right) \overset{a.e}{=} \Pi_2\left(\theta_1'', s, \theta_2'\right)$$

i.e.,

$$v_2\left(a, \theta_2'\right) + t_2\left(\theta_1', s, \theta_2'\right) \overset{a.e}{=} v_2\left(a, \theta_2'\right) + t_2\left(\theta_1'', s, \theta_2'\right)$$

which implies that

$$t_2\left(\theta_1', s, \theta_2'\right) \overset{a.e}{=} t_2\left(\theta_1'', s, \theta_2'\right)$$

and since the transfer of agent 2 for a given signal of agent 1 depends only on the chosen alternative we get that

$$t_2\left(\theta_1', s, a\right) \overset{a.e}{=} t_2\left(\theta_1'', s, a\right)$$

where $t_2(\theta_1, \delta_1, a)$ denote the transfer for alternative a given (θ_1, δ_1).
Fixing θ_2'' and applying the same analysis we get that

$$t_2\left(\theta_1', s, b\right) \overset{a.e}{=} t_2\left(\theta_1'', s, b\right)$$

Now, assume that agent 2's type is $\tilde{\theta}_2^{'}$ and that agent 1's type is $\theta_1^{'}$. Incentive compatibility implies that agent 2 does not want to report $\theta_2^{''}$; i.e., for every $\delta_1 \in \left[\underline{\delta_1}, \overline{\delta_1}\right]$ we have:

$$v_2\left(a, \tilde{\theta}_2^{'}\right) + t_2\left(\theta_1^{'}, \delta_1, a\right) \geq v_2\left(b, \tilde{\theta}_2^{'}\right) + t_2\left(\theta_1^{'}, \delta_1, b\right)$$

Assume that agent 1's type is $\theta_1^{''}$. Incentive compatibility implies that agent 2 does not want to report $\theta_2^{'}$; i.e., for every $\delta_1 \in \left[\underline{\delta_1}, \overline{\delta_1}\right]$ we have:

$$v_2\left(a, \tilde{\theta}_2^{'}\right) + t_2\left(\theta_1^{''}, \delta_1, a\right) \leq v_2\left(b, \tilde{\theta}_2^{'}\right) + t_2\left(\theta_1^{''}, \delta_1, b\right)$$

In addition we can find $\delta_1 \in \left[\underline{\delta_1}, \overline{\delta_1}\right]$ for which

$$t_2\left(\theta_1^{'}, \delta_1, b\right) - t_2\left(\theta_1^{'}, \delta_1, a\right) = t_2\left(\theta_1^{''}, \delta_1, b\right) - t_2\left(\theta_1^{''}, \delta_1, a\right) := \beta - \alpha$$

and so we get that

$$v_2\left(a, \tilde{\theta}_2^{'}\right) - v_2\left(b, \tilde{\theta}_2^{'}\right) = \beta - \alpha$$

An identical argument yields that

$$v_2\left(a, \tilde{\theta}_2^{''}\right) - v_2\left(b, \tilde{\theta}_2^{''}\right) = \beta - \alpha$$

but this contradicts the assumption that

$$v_2\left(a, \tilde{\theta}_2^{'}\right) - v_2\left(b, \tilde{\theta}_2^{'}\right) \neq v_2\left(a, \tilde{\theta}_2^{''}\right) - v_2\left(b, \tilde{\theta}_2^{''}\right) \quad \square$$

References

[1] Bartling, B., and Netzer, N. An externality-robust auction: Theory and experimental evidence. *Games and Economic Behavior 97* (2016), 186–204.

[2] Bergemann, D., and Morris, S. Robust mechanism design. *Econometrica 73*, 6 (2005), 1771–1813.

[3] Bierbrauer, F., and Netzer, N. Mechanism design and intentions. *Journal of Economic Theory 163* (2016), 557–603.

[4] Bierbrauer, F., Ockenfels, A., Pollak, A., and Rückert, D. Robust mechanism design and social preferences. *Journal of Public Economics 149* (2017), 59–80.

[5] Blanchet, D., and Fleurbaey, M. Selfishness, altruism and normative principles in the economic analysis of social transfers. *Handbook of the economics of giving, altruism and reciprocity 2* (2006), 1465–1503.

[6] Brandt, F., Sandholm, T., and Shoham, Y. Spiteful bidding in sealed-bid auctions. In *IJCAI* (2007), vol. 7, pp. 1207–1214.

[7] Cooper, D. J., and Kagel, J. H. Other-regarding preferences. *The handbook of experimental economics 2* (2016), 217.

[8] Fehr, E., and Schmidt, K. M. A theory of fairness, competition, and cooperation. *The quarterly journal of economics 114*, 3 (1999), 817–868.

[9] Gibbard, A. Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society* (1973), 587–601.

[10] Goodin, R. E. Laundering preferences. *Foundations of social choice theory 75* (1986), 81–86.

[11] Harsanyi, J. C. Morality and the theory of rational behavior. *Social Research* (1977), 623–656.

[12] Jehiel, P., Meyer-ter Vehn, M., Moldovanu, B., and Zame, W. R. The limits of ex post implementation. *Econometrica 74*, 3 (2006), 585–610.

[13] Krishna, V., and Maenner, E. Convex potentials with an application to mechanism design. *Econometrica 69*, 4 (2001), 1113–1119.

[14] Morgan, J., Steiglitz, K., and Reis, G. The spite motive and equilibrium behavior in auctions. *Contributions in Economic Analysis & Policy 2*, 1 (2003).

[15] Rabin, M. Incorporating fairness into game theory and economics. *The American economic review* (1993), 1281–1302.

[16] Satterthwaite, M. A. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory 10*, 2 (1975), 187–217.

[17] Wilson, B. Game-theoretic analysis of trading processes. *In: Advances in Economic Theory: Fifth World Congress. Cambridge University Press, Cambridge* (1987), 33 70.