Zik, Boaz

Article — Published Version

Ex-post implementation with social preferences

Social Choice and Welfare

Provided in Cooperation with:
Springer Nature

Suggested Citation: Zik, Boaz (2020) : Ex-post implementation with social preferences, Social Choice and Welfare, ISSN 1432-217X, Springer, Berlin, Heidelberg, Vol. 56, Iss. 3, pp. 467-485, https://doi.org/10.1007/s00355-020-01291-x

This Version is available at:
https://hdl.handle.net/10419/288441

Social Choice and Welfare (2021) 56:467–485
https://doi.org/10.1007/s00355-020-01291-x

ORIGINAL PAPER

Ex-post implementation with social preferences

Boaz Zik[1]

Received: 13 December 2018 / Accepted: 21 September 2020 / Published online: 30 September 2020

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, I address the question of whether deterministic mechanisms that are robust in both of these dimensions exist. I 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. I 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.

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, I study the problem of ex-post implementation of deterministic mechanisms in a simple model of

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

✉ Boaz Zik
bzik@uni-bonn.de

[1] Institute for Microeconomics, University of Bonn, Bonn, Germany

social preferences.[2] I 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, I investigate the implementation of decision rules that depend only on information about the agents' personal payoffs. I 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. I 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. I then consider ex-post implementation and show that the ex-post implementation of non-trivial decision rules is impossible in environments with social preferences.

In the second part of the paper, I 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. I 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' payoff signals.

This paper relates to 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 et al. (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 et al. (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 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. The first result on the impossibility of

[2] Similar models appear in Morgan et al. (2003) and in Brandt et al. (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.

ex-post implementation implies that it is impossible to construct mechanisms that are robust in both of these dimensions.

The construction of the social-preference-robust mechanisms in Bierbrauer and Netzer (2016), Bartling and Netzer (2016), and Bierbrauer et al. (2017) is based on two properties. The first is that under the mechanism one agent's actions cannot affect the payoff of another agent. This property is referred to in the literature as *externality-free*.[4] The second property is that the mechanism is incentive compatible in an environment where agents are selfish. Externality-free and incentive compatibility imply the social-preference-robustness of a mechanism in every model of social preferences in which agents behave selfishly whenever they cannot affect other agents' payoffs. For example, inequity aversion models, e.g., Fehr and Schmidt (1999) and models of intention-based preferences, e.g, Rabin (1993). In the second part of the proof of the impossibility theorem I show that under mild assumptions on the economic environment, that are satisfied in most of the standard settings of mechanism design, externality-freeness and ex-post incentive compatibility cannot coexist. This result is general and does not depend on the specific model of social preferences. The implication of this result is 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 and ex-post incentive compatible.

One economic environment in which the assumptions of this paper do not hold appears in Bierbrauer et al. (2017). They consider a bilateral trade environment where both the buyer and the seller have two types and present mechanisms that are social-preference-robust and payoff-information-robust by constructing mechanisms that are externality-free and ex-post incentive compatible.[5] They conduct a laboratory experiment that compares participants' behavior under a mechanism that is both social-preference-robust and payoff-information-robust and under a mechanism that is only payoff-information-robust but not social-preference-robust. They find that the first mechanism performs significantly better than the latter. The fact that the mechanisms they compare are both payoff-information-robust and differ only in their social-preference-robustness enables to account the difference in the participants' behavior under the different mechanisms to the existence of social preferences.[6] This paper relates to their work by implying that such an experiment cannot be replicated in most mechanism design environments, where externality-freeness and ex-post incentive compatibility cannot coexist, and that in order to conduct such an experiment one needs to carefully design the economic environment.

[4] There are natural economic settings, who are not subjected to design, where externality-freeness arise. For example, externality freeness may arise in environments where agents are price takers and do not internelize the effect of their actions on the market price, see Dufwenberg et al. (2011). Another example appears in Bierbrauer (2011) who analyzes a problem of income taxation and presents an optimal solution with the property that an agent's tax depends only on her income.

[5] I discuss this point further in Sect. 3.

[6] 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.

Bierbrauer and Netzer (2016) consider social-preferences-robust mechanisms that are Bayesian implementable and present an extensive possibility result that is based on the properties of externality-freeness and incentive compatibility. The reason for the difference in the possibility to achieve both externality-freeness and incentive compatibility between Bayesian and ex-post implementation is the following. Externality-freeness means that an agent's report does not affect the payoffs of other agents. This implies that the other agents' transfers should eliminate the effect of this agent report on their valuation. In addition, under the requirement of incentive compatibility, these agents' transfers must also incentivize each of them to report truthfully. Under ex-post implementation, these requirements for an agent's transfer must be satisfied for every realization of signals. I show that this cannot happen without contradictions. However, under Bayesian implementation, these requirements should only be met in expectation, and then it is possible to construct transfer schemes that satisfy these requirements.

The rest of the paper is organized as follows. In Sect. 2, I present the model. In Sect. 3, I discuss the implementation of decision rules that depend only on agents' payoff signals. I characterize the set of Bayes–Nash implementable decision rules 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. I present an impossibility result of ex-post implementation. In Sect. 4, I present an impossibility result on the ex-post implementation of decision rules that depend both on agents' payoff signals and their social preferences. I also discuss the difference between the social preferences model of this paper model and the classical interdependent values model. Section 5 concludes.

2 The model

I consider a model with two agents, $i \in I = \{1,2\}$, and two social alternatives,[7] $A = \{a,b\}$. 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 $\Pi_i = v_i(k,\theta_i) + t_i$. I 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[8] $0 \in \left[\underline{\delta_i}, \overline{\delta_i}\right]$. The signals θ_i and δ_i are the private information of agent i. I denote $\Theta := \underset{i\in I}{\times}\Theta_i$ with generic element θ, and $\mathcal{D} := \underset{i\in I}{\times}\left[\underline{\delta_i}, \overline{\delta_i}\right]$ with generic element δ. A function $q : \Theta \times \mathcal{D} \to A$ is called a *decision rule.* A *social choice function* is a

[7] 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.

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

function $s(\theta,\delta)=\big(q(\theta,\delta),t_1(\theta,\delta),t_2(\theta,\delta)\big)$, where $q(\theta,\delta)\in A$ and $t_i(\theta,\delta)\in\mathbb{R}$ for every $i\in I$. A social choice function $\big(q(\theta,\delta),t_1(\theta,\delta),t_2(\theta,\delta)\big)$ is *ex-post implementable* if for every $i\in I$, and $(\theta,\delta)\in\Theta\times D$ we have

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

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

3 Decisions that depend only on payoff signals

In this section, I discuss the implementation of decision rules that depend only on information about the personal payoffs of the agents. I 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.[9] Harsanyi, one of the greatest advocates of utilitarian theory, suggests that:

[9] 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 he is interested in maximizing her own personal payoff. Therefore, this transfer scheme incentivizes truth telling in every setup. I now show this result formally.

Proposition 1 *Consider a profile* $\Theta, \left(v_i\right)_{i\in I}$. *Let* $\left(q(\theta), t_1(\theta), t_2(\theta)\right)$ *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 social preferences and in the environment where agents are selfish.*

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

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

Consider $j, l \in \{1, 2\}$ with $j \neq l$. Now agent $j's$ expected utility as a function of her report is

$$\begin{aligned} &E_{\theta_l}\left[v_j\left(q\left(\hat{\theta}_j, \theta_l\right), \theta_j\right) + t_j'\left(\hat{\theta}_j, \theta_l\right) + \delta_j\left(v_l\left(q\left(\hat{\theta}_j, \theta_l\right), \theta_l\right) + t_l'\left(\hat{\theta}_j, \theta_l\right)\right)\right] \\ &\quad = E_{\theta_l}\left[v_j\left(q\left(\hat{\theta}_j, \theta_l\right), \theta_j\right) + t_j'\left(\hat{\theta}_j, \theta_l\right)\right] \end{aligned}$$

i.e., from agent j's perspective, the report $\hat{\theta}_j$ does not affect the expected payoff of agent l. 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 $\left(t_i(\theta)\right)_{i\in I}$ 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 action does not affect the expected payoff that he assigns to agent j. This property is referred to in the literature as *externality-free*. The second property is that the mechanism is incentive compatible. Externality-free and incentive compatibility imply the social-preference-robustness of the mechanism not only in the particular model of this paper but in every model of social preferences in which agents behave selfishly whenever they cannot affect other agents' payoffs.

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

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. In these environments, the assumption of Theorem 2 is satisfied whenever one agent is pivotal for two different types of the other agent. For example, consider a single-unit auction with two agents. For each agent the type set is $[\underline{\theta}, \overline{\theta}]$. Any deterministic decision rule $q\left(\theta_i,\theta_j\right)$ with the

property that there exist two types of agent j, $\tilde{\theta}_j$ and θ_j', for which $q(\cdot,\tilde{\theta}_j)$ and $q(\cdot,\theta_j')$ are non-trivial functions of agent i's type, is not ex-post implementable.[10]

The argument behind Theorem 2 is the following. Ex-post implementation implies that for any two signals θ_i and θ_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 θ_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 and consider some (θ_j,δ_j). For every $\theta_i,\theta_i'\in\Theta_i$; we have that $\Pi_j(\theta_j,\delta_j,\theta_i,\cdot)\overset{a.e}{=}\Pi_j(\theta_j,\delta_j,\theta_i',\cdot)$.*

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

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

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

[10] Ex-post implementability in the independent private value setting implies that both $q(\cdot,\tilde{\theta}_j)$ and $q(\cdot,\theta_j')$ have thresholds (not necessarily the same) such that agent i receives the item if and only if his reported type exceeds the threshold. Therefore, we can restrict our attention to non-trivial functions, $q(\cdot,\tilde{\theta}_j)$ and $q(\cdot,\theta_j')$, with the above threshold property. The threshold property implies that the assumption of Theorem 2 holds.

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

[12] See Krishna and Maenner (2001).

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

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

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

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

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

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

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

I 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, θ_i', θ_j and θ_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

[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 $\delta\in\left[\underline{\delta},\overline{\delta}\right]$, $\delta\in\arg\max_{\hat{\delta}\in\left[\underline{\delta},\overline{\delta}\right]}u(\delta,\hat{\delta})$ then for every $\delta\in\left[\underline{\delta},\overline{\delta}\right]$, $t(\delta)+\delta q(\delta)=t\left(\underline{\delta}\right)+\underline{\delta}\cdot q\left(\underline{\delta}\right)+\int_{\underline{\delta}}^{\delta}q(s)\,ds$.

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

a contradiction. □

Remark Theorem 2 concerns decision rules that depend only on agents' payoff signals. However, throughout the analysis, I have allowed agents' transfers to depend also on the information about the agents' social preferences. In that sense, the theorem shows that the 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 the specific model of social preferences, the second claim does not. This means that in any model of mechanism design with social preferences it is 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. Moreover, in any model of mechanism design with social preferences it suffices to show that ex-post implementability implies that externality-freeness must hold for every realization of payoff signals to prove that mechanisms that are social-preference-robust and payoff-information-robust do not exist.

Remark Bierbrauer et al. (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 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.

Remark The ex-post implementation of a decision rule $q(\theta)$ under the assumption that an agent's social preferences are privately known implies that $q(\theta)$ is implementable in a model where the profile of agents' social preferences signals δ is commonly known. Under the assumption that the profile of agents' social preferences signals δ is commonly known, the model of the paper corresponds to a model with interdependent separable valuations. Jehiel et al. (2006) show an impossibility result

of ex-post implementation in models with interdependent valuations. However, their result does not imply the impossibility of ex-post implementation in the model of this paper for the following reasons. First, Jehiel et al. (2006) result depends on the assumption that the payoff type of each agent is multi-dimensional, while I allow the agents' payoff types to be uni-dimensional. When agents' types are uni-dimensional, it is possible to implement non-trivial decision rules in models with interdependent valuations. Second, when agents' social preferences signals δ is commonly known the model of the paper corresponds to a model with interdependent separable valuations and Jehiel et al. (2006) result does not apply to models with interdependent separable valuations. Indeed, non-trivial ex-post implementation is possible in models with interdependent separable valuations.[15] I further discuss the differences between the social preferences model and the interdependent valuation model in Sect. 4.2.

4 Discussion

4.1 Decisions that depend on social preferences

In the previous section, I discussed the notion of social-preference robustness. This notion is suitable to situations where the designer does not want to condition her decision on information about the agents' social preferences. In this subsection, I 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. I 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.

I consider the 2×2 model that I presented in Sect. 2 except that now agent 2 is selfish, i.e., $u_1 = \Pi_1 + \delta_1 \cdot \Pi_2$ and $u_2 = \Pi_2$. The impossibility theorem, Proposition 6, and its proof are relegated to the Appendix. In the following, I illustrate the theorem and its proof by considering the following example of an allocation problem of a single good.

Example 4 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

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

[15] See Sect. 5.4.2 in Jehiel et al. (2006).

where $q = a$ is the allocation where agent 1 gets the item and $q = b$ is the allocation where agent 2 gets the item. The impossibility result implies that this decision rule is not ex-post implementable.

The impossibility of ex-post implementation follows from the fact that agent 2's transfer appear in the incentive compatibility conditions of agents 1 and 2 and there is no transfer function that can satisfy the IC constraints of both agents. The argument is the following. In the above example there exist $\theta_1^{'}$ and $\theta_1^{''}$ and $\theta_2^{'}$ and $\theta_2^{''}$ such that

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

and

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

for every $\delta_1 \in [0.1, 0.2]$. For example $\theta_1^{'} = 0.8$, $\theta_1^{''} = 0.4$, $\theta_2^{'} = 0.1$, and $\theta_2^{''} = 0.9$. Incentive compatibility of agent 1 implies, by a similar argument to the one that appears in the proof of Theorem 2, that

$$t_2\left(\theta_1^{'}, \delta_1, a\right) = t_2\left(\theta_1^{'}, \delta_1, \theta_2^{'}\right) \overset{a.e}{=} t_2\left(\theta_1^{''}, \delta_1, \theta_2^{'}\right) = t_2\left(\theta_1^{''}, \delta_1, a\right)$$

where $t_2\left(\theta_1^{'}, \delta_1, a\right)$ is agent 2's transfer for alternative a conditional on the report $\left(\theta_1^{'}, \delta_1\right)$ and $t_2\left(\theta_1^{''}, \delta_1, a\right)$ is agent 2's transfer for alternative a conditional on the report $\left(\theta_1^{''}, \delta_1\right)$. In an identical way we get that

$$t_2\left(\theta_1^{'}, \delta_1, b\right) = t_2\left(\theta_1^{'}, \delta_1, \theta_2^{''}\right) \overset{a.e}{=} t_2\left(\theta_1^{''}, \delta_1, \theta_2^{''}\right) = t_2\left(\theta_1^{''}, \delta_1, b\right)$$

where $t_2\left(\theta_1^{'}, \delta_1, b\right)$ is agent 2's transfer for alternative b conditional on the report $\left(\theta_1^{'}, \delta_1\right)$ and $t_2\left(\theta_1^{''}, \delta_1, b\right)$ is agent 2's transfer for alternative b conditional on the report $\left(\theta_1^{''}, \delta_1\right)$. Now there exist $\tilde{\theta}_2^{'}$ and $\tilde{\theta}_2^{''}$ such that

$$q\left(\theta_1^{'}, \delta_1, \tilde{\theta}_2^{'}\right) = a \text{ and } q\left(\theta_1^{''}, \delta_1, \tilde{\theta}_2^{'}\right) = b$$

and

$$q\left(\theta_1^{'}, \delta_1, \tilde{\theta}_2^{''}\right) = a \text{ and } q\left(\theta_1^{''}, \delta_1, \tilde{\theta}_2^{''}\right) = b$$

for every $\delta_1 \in [0.1, 0.2]$. For example $\tilde{\theta}_2^{'} = 0.6$, and $\tilde{\theta}_2^{''} = 0.5$. 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:

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

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:

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

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

$$\tilde{\theta}_2' = \beta - \alpha$$

An identical argument yields that

$$\tilde{\theta}_2'' = \beta - \alpha$$

but this contradicts the assumption that

$$\tilde{\theta}_2' \neq \tilde{\theta}_2''$$

4.2 Social preferences vs. interdependent values

In this paper, I 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 the social preferences model resembles the model in Jehiel et al. (2006), it is different from their model in the following important respect. In the social preferences 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 the social preferences model agent i's report affects his utility through the decision rule, his personal transfer, and the personal transfer of agent j.[16] That is, in the social preferences model mechanisms affect agents' incentives in a more complex way, compared to in the interdependent values model. On the one hand, since an agent's utility is affected by the other 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, consider the interdependent values model where agent i's utility function is $v_i\left(q, \theta_i\right) + \delta_i \cdot v_j\left(q, \theta_j\right) + t_i$, where $q \in A$, whereas in the social preferences model agent i's utility is $v_i\left(q, \theta_i\right) + \delta_i \cdot v_j\left(q, \theta_j\right) + z_i$, where $z_i = \delta_i t_j + t_i$. The difference between the models is that t_i depends only on agent i's reported signal and not on her actual signal, while

[16] Note that while the effect of the agent's personal transfer on his utility is independent of the realization on signals, the effect of other agents' transfers on his utility depends on the realization of signals.

z_i depends both on agent i's reported signal and on her actual signal.[17] To further illustrate the difference between the models, I 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, I present a decision rule that is not ex-post implementable in the social preferences model but is ex-post implementable in the interdependent values model. In the second example, I present decision rules that are ex-post implementable in the social preferences model but are not ex-post implementable in the interdependent values model.

Example 4 (continued) Consider the setup of Example 4 (for which it has been shown that the optimal decision rule is not ex-post implementable in the social preferences 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:

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

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

Example 5 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 $v_i\left(a,\theta_i\right)=\theta_i+c$, and his valuation if alternative b is chosen is $v_i\left(b,\theta_i\right)=\theta_i$. I analyze the possibility of implementing decision rules that depend only on information about agents' payoffs both in the social preferences model and in the interdependent values model.

I first analyze the paper's model. Consider an arbitrary decision rule $q(\theta)$. For every $i\in\{1,2\}$ I define the following transfer function:

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

Under these transfer functions any type $\left(\theta_i,\delta_i\right)$ 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.[18]

[17] Another way to try to compare the two models in to make an adaptation of the utilities in the social preferences model to the standard quasi-linear utility by separating the term that depends on agent i's private signal and the term that depends solely on her report. For this I define $V_i\left(q,\theta,t_j\right)=v_i\left(q,\theta_i\right)+\delta_i\left[v_j\left(q,\theta_j\right)+t_j\right]$ and so agent i's utility is $V_i\left(q,\theta,t_j\right)+t_i$, so the mechanism affects V_i through q and t_j, while in the interdependent values the term in agent i's utility that depends on agent i's private signal is her valuation that is affected by the mechanism only through q.

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

I 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 $(\tilde{\theta}_j, \tilde{\delta}_j)$ of agent $j, j \neq i$. Ex-post implementability implies that for every $(\theta_i, \delta_i), (\theta_j', \delta_j') \in [0,1]^2$ such that $q(\theta_i, \tilde{\theta}_j) = q(\theta_i', \tilde{\theta}_j)$ we have[19] $t_i(\theta_i, \delta_i, \tilde{\theta}_j, \tilde{\delta}_j) = t_i(\theta_i', \delta_i', \tilde{\theta}_j, \tilde{\delta}_j)$. That is, agent i's transfer function depends only on the chosen alternative; hence, we denote $t_i(\theta_i, \delta_i, \tilde{\theta}_j, \tilde{\delta}_j) := t_i(q(\theta_i, \theta_j), \tilde{\theta}_j, \tilde{\delta}_j)$. Consider a non-constant decision rule $q(\theta)$. Look at a type $(\tilde{\theta}_j, \tilde{\delta}_j)$ of agent j for which agent i is pivotal. This means that there exist two signals θ_i' and θ_i'' such that $q(\theta_i', \tilde{\theta}_j) = a$ and $q(\theta_i'', \tilde{\theta}_j) = b$. Now, ex-post implementability implies that for every $\delta_i \in [0,1]$ we have that

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

and

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

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

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

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.

5 Conclusion

I 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. I 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 social-preference-robust and payoff-information-robust. The impossibility result also shows that in any model with 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 and incentive compatible.

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

Acknowledgements I would like to thank Alex Gershkov, Ilan Kremer, Motty Perry, Phil Reny, Assaf Romm, and participants of various seminars for their valuable comments. Funding by the German Research Foundation (DFG) through CRC TR 224 (Project B01) is gratefully acknowledged.

Funding Open Access funding enabled and organized by Projekt DEAL.

Appendix

Proposition 6 *Consider a decision rule of the following form. There exist two types* θ_1', θ_1'', *and two types,* θ_2', θ_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*

$$q\left(\theta_1', \delta_1, \theta_2'\right) = q\left(\theta_1'', \delta_1, \theta_2'\right) = a$$
$$q\left(\theta_1', \delta_1, \theta_2''\right) = q\left(\theta_1'', \delta_1, \theta_2''\right) = b$$

and there exist two types, $\tilde{\theta}_2'$ *and* $\tilde{\theta}_2''$, *such that for every* $\delta_1 \in \left[\underline{\delta_1}, \overline{\delta_1}\right]$ *we have that*

$$q\left(\theta_1', \delta_1, \tilde{\theta}_2'\right) = q\left(\theta_1', \delta_1, \tilde{\theta}_2''\right) = a$$
$$q\left(\theta_1'', \delta_1, \tilde{\theta}_2'\right) = q\left(\theta_1'', \delta_1, \tilde{\theta}_2''\right) = b$$

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* $q(\theta)$ *is not ex-post implementable.*

Proof of proposition 6

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

$$\delta_1 \cdot \Pi_2\left(\theta_1, \delta_1, \theta_2\right) + t_1\left(\theta_1, \delta_1, \theta_2\right) = \delta_1 \cdot \Pi_2\left(\dot{\theta}_1, \delta_1, \theta_2\right) + t_1\left(\dot{\theta}_1, \delta_1, \theta_2\right)$$
$$:= \sigma\left(k, \delta_1, \theta_2\right)$$

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 θ_2' and that agent 1 is of type θ_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 $\delta_1, \delta_1' \in \left[\underline{\delta_1}, \overline{\delta_1}\right]$

$$\begin{aligned} &v_1\left(a, \theta_1'\right) + \delta_1 \cdot \Pi_2\left(\theta_1', \delta_1, \theta_2'\right) + t_1\left(\theta_1', \delta_1, \theta_2'\right) \\ &\quad \geq v_1\left(a, \theta_1'\right) + \delta_1 \cdot \Pi_2\left(\theta_1', \delta_1', \theta_2'\right) + t_1\left(\theta_1', \delta_1', \theta_2'\right) \end{aligned}$$

This implies that

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

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 $t_2\left(\theta_1, \delta_1, a\right)$ denote the transfer for alternative a given $\left(\theta_1, \delta_1\right)$.

Fixing $\theta_2^{''}$ 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

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

□

References

Bartling B, Netzer N (2016) An externality-robust auction: theory and experimental evidence. Games Econ Behav 97:186–204

Bergemann D, Morris S (2005) Robust mechanism design. Econometrica 73(6):1771–1813

Bierbrauer FJ (2011) On the optimality of optimal income taxation. J Econ Theory 146(5):2105–2116

Bierbrauer F, Netzer N (2016) Mechanism design and intentions. J Econ Theory 163:557–603

Bierbrauer F, Ockenfels A, Pollak A, Ruckert D (2017) Robust mechanism design and social preferences. J Public Econ 149:59–80

Blanchet D, Fleurbaey M (2006) Selfishness, altruism and normative principles in the economic analysis of social transfers. Handbook Econ Giv Altruism Reciprocity 2:1465–1503

Brandt F, Sandholm T, Shoham Y (2007) Spiteful bidding in sealed-bid auctions. IJCAI 7:1207–1214

Cooper DJ, Kagel JH (2016) Other-regarding preferences. Handbook Exp Econ 2:217

Dufwenberg M, Heidhues P, Kirchsteiger G, Riedel F, Sobel J (2011) Other-regarding preferences in general equilibrium. Rev Econ Stud 78(2):613–639
Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. Q J Econ 114(3):817–868
Goodin RE (1986) Laundering preferences. Found Soc Choice Theory 75:81–86
Harsanyi JC (1977) Morality and the theory of rational behavior. Soc Res 623–656
Jehiel P, Vehn MM, Moldovanu B, Zame WR (2006) The limits of ex post implementation. Econometrica 74(3):585–610
Krishna V, Maenner E (2001) Convex potentials with an application to mechanism design. Econometrica 69(4):1113–1119
Morgan J, Steiglitz K, Reis G (2003) The spite motive and equilibrium behavior in auctions. Contribut Econ Anal Policy 2(1)
Rabin M (1993) Incorporating fairness into game theory and economics. Am Econ Rev 1281–1302
Wilson R (1987) Game-theoretic analysis of trading processes. In: Advances in economic theory: fifth world congress. Cambridge University Press, Cambridge, pp 33–70 **(50)**