Efficient Incentives with Social Preferences*

Thomas Daske‡ Christoph March§

December 19, 2023

We explore mechanism design with outcome-based social preferences. Agents' social preferences and private payoffs are all subject to asymmetric information. We assume quasi-linear utility and independent types. We show how the asymmetry of information about agents' social preferences can be operationalized to satisfy agents' participation constraints. Our main result is a possibility result for groups of *at least three* agents: Any such group can resolve any given allocation problem with an ex-post budget-balanced mechanism that is Bayesian incentive-compatible, interim individually rational, and ex-post Pareto-efficient.

JEL Classification: C72; C78; D62; D82

Keywords: Mechanism design; social preferences; Bayesian implementation; participation constraints; participation stimulation; money pump

*An earlier version circulated under the title "Efficient incentives in social networks: Gamification and the Coase theorem" and is available under http://hdl.handle.net/10419/222527.
For their helpful comments and critical remarks, we thank Claude d'Aspremont, Jacques Crémer, Benny Moldovanu, Marco Sahm, Klaus Schmidt, Johannes Schneider, Roland Strausz, and Robert von Weizsäcker as well as participants of the European Winter Meeting of the Econometric Society in Milan, the World Congress of the Game Theory Society in Maastricht, the Annual Meeting of the Association for Public Economic Theory in Strasbourg, the Annual Congress of the International Institute of Public Finance in Glasgow, the European Meeting of the Econometric Society in Manchester, the Annual Congress of the German Economic Association in Leipzig, the virtual Econometric Society World Congress, and the Annual Congress of the Society for the Advancement of Economic Theory in Canberra. We are particularly grateful to several anonymous referees for their patience and very helpful suggestions.

‡Technical University of Munich, TUM School of Management, Arcisstr. 21, 80333 Munich, Germany. Email: thomas.daske@tum.de

§University of Bamberg, Department of Economics, Feldkirchenstr. 21, 96047 Bamberg, Germany; and CESifo, Poschingerstr. 5, 81679 Munich, Germany. Email: christoph.march@uni-bamberg.de

1 Introduction

How can allocation problems be resolved in an efficient and mutually acceptable way? The literature on mechanism design has postulated four desirable properties of incentive mechanisms: Incentive compatibility, ex-post Pareto efficiency, ex-post budget balance, and interim individual rationality. Bayesian implementation is suitable to achieve the first three of these properties (see, e.g., Arrow, 1979; d'Aspremont and Gérard-Varet, 1979). Often, however, Bayesian mechanisms violate agents' participation constraints.[1]

Bayesian mechanisms that reconcile all four properties exist if agents' private signals (or, types) are sufficiently *correlated*: Crémer and McLean (1985, 1988) show that the designer can exploit this correlation to validate the agents' reports, extract all information rents, and ensure participation en passant.[2] Mezzetti (2004) shows that the logic of Crémer and McLean (1985, 1988) can be extended to the case of independent private signals if the designer is permitted to implement a two-stage mechanism: The allocation problem can be resolved with unanimous participation by *sequentially* administering a social alternative and transfers, with agents first reporting their preference types and then their satisfaction with the chosen alternative before finally receiving (or paying) transfers.

The present study enriches the set of possibility results. We assume that types are independent (in contrast to Crémer and McLean, 1985, 1988) and that there is only one round of reporting (in contrast to Mezzetti, 2004). Specifically, we consider agents with outcome-based social preferences that are privately known (next to privately known preferences for consumption). That is, agents care about the overall distributive effects of a mechanism, and their distributive preferences are private information. We show how this kind of information asymmetry can be operationalized to satisfy agents' participation constraints.

Our main result, Theorem 1, states that any group of *at least three* agents can resolve any given allocation problem with an ex-post budget-balanced mechanism that is Bayesian incentive-compatible, interim individually rational, and ex-post Pareto-efficient. It builds on the following insights: In quasi-linear environments, a mechanism can be designed such that the incentives to reveal payoff types and social types are separated. While the allocation problem can be resolved through payoff-type conditional budget-balanced transfers, participation can be stimulated through additional budget-balanced

[1] For settings with independent private signals see, e.g., Myerson and Satterthwaite (1983), Mailath and Postlewaite (1990), Williams (1999), and Segal and Whinston (2016).

[2] Likewise, McAfee and Reny (1992), McLean and Postlewaite (2004), Kosenok and Severinov (2008).

transfers that condition on agents' social types. The latter is possible for more than two agents when leveraging the differences in agents' other-regarding concerns. Technically, we exploit that each agent's utility is a linear combination of all agents' private payoffs, which are weighted according to that agent's other-regarding concerns. This linearity enables us to render the agents' social types strategically inoperative in the payoff-type conditional mechanism, so we can use them in a separate, social-type conditional mechanism to cross-subsidize the former. In this manner, our solution bundles two strategically independent mechanisms. (Our bundling of two mechanisms resembles Mezzetti (2004). We detail the differences between his and our study in Section 6.4.)

Until recently, the literature on efficient design has either neglected social preferences altogether or assumed them to be common knowledge.[3] An exception is Bierbrauer and Netzer (2016), who study mechanism design when agents have privately known intention-based social preferences. They show that this sort of social preferences allows for efficient, individually rational design if and only if all agents are (commonly known to be) conditionally pro-social. Our study differs from theirs in the kind of social preferences under consideration as well as in the conditions for and the driving forces behind the possibility result: First, we consider unconditional outcome-based rather than intention-based social preferences, and next to altruism and selfishness we allow for anti-social preferences such as spite.[4] Second, the revelation principle holds in our setup but not in Bierbrauer and Netzer's (2016), as their agents' preferences depend on the set of actions (i.e., messages) available in the mechanism. Indeed, the independence of agents' preferences from the mechanism distinguishes our paper from various others on mechanism design with intention-based social preferences (e.g., Antler, 2015; Kozlovskaya and Nicoló, 2019). Finally, the possibility result of Bierbrauer and Netzer (2016) exploits the mechanism-dependence of preferences by introducing additional messages that are not chosen in equilibrium but manipulate the kindness of truth-telling; this construction only works in the absence of selfish types. In contrast, our result exploits the asymmetry of information about agents' social preferences.

Notably, our study relates to the literature on *money pumps* (or, *dutch books*). This literature has a long tradition in individual-choice theory. It shows how *non-rational*

[3]See, e.g., Desiraju and Sappington (2007), Kucuksenel (2012), Tang and Sandholm (2012).

[4]The behavioral relevance of unconditional outcome-based social preferences has been well established. For evidence on altruism and selfishness, see Andreoni and Miller (2002), Charness and Rabin (2002), and Bruhin, Fehr, and Schunk (2019). For evidence on spite, see Saijo and Nakamura (1995), Fehr, Hoff, and Kshetramade (2008), and Prediger, Vollan, and Herrmann (2014).

individual-decision making can be exploited to pull agents into transactions they stand to lose from (see, e.g., Border and Segal, 1994, and Rubinstein and Spiegler, 2008; for a survey, see Yaari, 1998). In the multi-agent version, a group of agents is subject to a money pump if an outside party is "able to extract money from the agents without putting any money at risk" (Nau, 1992, p. 380). Our study shows that a group of at least three agents with privately known social preferences can be offered an ex-post budget-balanced (non-zero) transfer scheme that all of them accept ex interim. This implies that a transfer scheme can be constructed that extracts money from the group via participation fees and is still unanimously accepted – and, thus, becomes a money pump. While the literature has focused on *non-rational* expectations (see, e.g., Eliaz and Spiegler, 2007, 2009; Chen, Micali, and Pass, 2015; Werner, 2022), we show that multi-agent money pumps can be grounded in non-standard *rational* preferences. As in Antler (2023) for non-rational expectations, we require *sufficiently many* agents, at least three in our case.

The following example provides a basic intuition for how asymmetric information about agents' social preferences can be exploited to generate a money pump: Consider two agents each of whom is either *selfish* (caring only about her private payoff) or *altruistic* (weighting the other's payoff half as much as her own). Types are independent and equally likely. If both report *selfish* (*altruistic*), each is taxed (rewarded) one dollar; if they report opposite types, the *altruist* must pay the *selfish* two dollars. Clearly, reporting *selfish* always yields a higher private payoff, incentivizing truth-telling for selfish agents. Reporting *altruist* always yields a considerably larger payoff to the opponent than reporting *selfish*, incentivizing truth-telling for altruists. As unanimous participation yields each type an interim-expected utility gain (as compared to a status quo of zero-transfers), agents are willing to pay for playing this game. Thus, a third agent can offer to finance the game by balancing the budget (i.e., to tax or reward according to the rules) in return for a uniform participation fee. As transfers are zero *ex ante*, a sufficiently small fee guarantees that all three agents are wanting *ex interim* to participate in the extended game. Conforming this scenario to the quoted money-pump definition of *extracting money from the agents without putting any money at risk* resembles government selling a casino license: An outside party may enter the scene and offer our 'third agent' the platform on which she can let others play our game – in return for half of participation fees.

In this example, when looking at the actual players (selfish or altruistic), money is redistributed ex interim to those agents who *care least* about others. On the other

hand, a *pro-social* agent interim-expects to impose a positive monetary externality on her opponent, and this externality overcompensates her emotionally for interim-expected monetary losses. These distributive effects are a general feature of the various 'money pumps' we develop in this paper, although the notions of *caring least* and *pro-sociality* will bear more intricate meanings. (We present this example in more detail in Section 5.1.)

The paper proceeds as follows. Section 2 outlines the model framework. Section 3 states and interprets our main result. Section 4 details the proof. Section 5 illustrates the intuition behind our participation-stimulating transfers. Finally, Section 6 reflects upon the assumptions that are critical to our result, distinguishes our mechanism from Mezzetti's (2004), and illustrates how participation stimulation can be implemented in practice.

2 The Model

2.1 The Allocation Problem

There is a group $\mathcal{I} = \{1, \ldots, n\}$ of $n \geq 2$ agents and there is a finite set K of social alternatives. From alternative $k \in K$ and a transfer $t_i \in \mathbb{R}$, agent i gains a *private payoff* $\Pi_i(k, t_i \,|\, \theta_i) = \pi_i(k \,|\, \theta_i) + t_i$, with $\pi_i : K \times \Theta_i \to \mathbb{R}$. Agent i's *payoff type* θ_i belongs to a finite set Θ_i, with $|\Theta_i| \geq 2$. The collection of agents' payoff types is denoted by $\theta = (\theta_i, \theta_{-i}) \in \Theta = \prod_i \Theta_i$, where $\theta_{-i} = (\theta_j)_{j \neq i}$. Agents exhibit social preferences in the form of altruism or spite: From the collection of private payoffs $(\Pi_j)_{j \in \mathcal{I}}$, agent i derives ex-post *utility*

$$u_i(k, (t_j)_{j \in \mathcal{I}}, \theta_{-i} \,|\, \theta_i, \delta_i) \;\; = \;\; \sum_{j \in \mathcal{I}} \delta_{ij} \, \Pi_j(k, t_j \,|\, \theta_j),$$

where the value δ_{ij} that i assigns to j's payoff, $j \neq i$, belongs to a closed (proper) interval $\Delta_{ij} = [\delta_{ij}^{\min}, \delta_{ij}^{\max}] \subset (-1/(n-1); 1)$, while $\delta_{ii} = 1$ for all i. We refer to δ_{ij} as i's *degree of altruism towards* j, to the collection $\delta_i = (\delta_{ij})_{j \neq i} \in \Delta_i = \prod_{j \neq i} \Delta_{ij}$ as i's *social type*, and to the pair (θ_i, δ_i) as i's *type*.

The information structure is as follows: Each agent is privately informed about her payoff type and social type. Hence, there is a type distribution on $\Theta \times \Delta$ (where $\Delta = \prod_i \Delta_i$) with strictly positive variance of payoff types and social types. Type realizations are independent across agents. An agent's payoff type and social type realize

independently according to strictly positive densities, but the various degrees of altruism determining this agent's social type may correlate. We assume that agents will observe each other's payoffs *ex post*. (We make the implicit assumption of continuous social-type distributions to keep the exposition simple, but all results are equally valid if a social-type set contains mass points; see Section 5.1.)

A few remarks are appropriate: First, the interval $(-1/(n-1);1)$ is the maximum range of altruism, or spite, for which agents care about overall material efficiency while still being selfish to the extent that every one of them prefers a dollar to be her own rather than having that same dollar distributed among the others. Second, despite the asymmetry of information, it can still be common knowledge who is 'friends' and who is 'foes.' For instance, if $\delta_{k\ell}^{\max}, \delta_{\ell k}^{\max} < 0 < \delta_{ij}^{\min}, \delta_{ji}^{\min}$, then, in comparison, i and j are friends, whereas k and ℓ are foes. Likewise, it can be common knowledge that i likes j more than k, which is the case if $\delta_{ik}^{\max} < \delta_{ij}^{\min}$. And finally, while we assume that the variance of every δ_{ij} is strictly positive, it is allowed to be arbitrarily small. Reciprocal social preferences can thus be captured by letting $\Delta_{ij} = \Delta_{ji}$ and $\delta_{ij}^{\min} \approx \delta_{ij}^{\max}$.

The agents' problem is to choose a social alternative k and transfers $(t_i)_{i\in\mathcal{I}}$ such that the resulting *allocation*, i.e., the collection of private payoffs, is ex-post Pareto-efficient. We require that agents must do so without having access to an outside source of money, such that transfers must be *weakly budget-balanced*: $\sum_{i\in\mathcal{I}} t_i \leq 0$.

2.2 Revelation Mechanisms

A *direct* revelation mechanism involves the agents in a strategic game of incomplete information in which they are asked to report their types truthfully. Types are reported *simultaneously*. Based on their reports, a social alternative is chosen and transfers are made. As the *revelation principle* applies to the present setup (Myerson, 1979), there is no loss of generality in considering only *direct* mechanisms. Formally, a direct mechanism is given by a pair $\langle k, T\rangle$ with *allocation function* $k : \Theta \times \Delta \to K$ and *transfer scheme* $T = (t_i)_{i\in\mathcal{I}} : \Theta \times \Delta \to \mathbb{R}^n$. Notice that transfers may take arbitrary negative values.

Denote by $U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i)$ agent i's interim-expected utility from reporting $(\hat{\theta}_i, \hat{\delta}_i)$ if her true type is (θ_i, δ_i) while all the other agents report their types truthfully: $U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) = \sum_{j\in\mathcal{I}} \delta_{ij}\big[\bar{\pi}_{ij}(\hat{\theta}_i, \hat{\delta}_i) + \bar{t}_{ij}(\hat{\theta}_i, \hat{\delta}_i)\big]$, where $\bar{\pi}_{ij}(\theta_i, \delta_i) = \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[\pi_j(k(\theta,\delta) \,|\, \theta_j)\big]$ and $\bar{t}_{ij}(\theta_i, \delta_i) = \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[t_j(\theta,\delta)\big]$. For convenience, $U_i(\theta_i, \delta_i) = U_i(\theta_i, \delta_i \,|\, \theta_i, \delta_i)$. Then the mechanism $\langle k, T\rangle$

is *Bayesian incentive-compatible* if, for all $i \in \mathcal{I}$ and all $(\theta_i, \delta_i) \in \Theta_i \times \Delta_i$, we have $U_i(\theta_i, \delta_i) = \max_{(\hat{\theta}_i, \hat{\delta}_i) \in \Theta_i \times \Delta_i} U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i)$.[5]

2.3 Efficiency and Participation

The following Lemma links material efficiency (the maximum surplus of private payoffs) to Pareto efficiency. It allows us to focus on allocation functions that are *ex-post materially efficient*, $k(\theta, \delta) = k^\star(\theta) \in \arg\max_{k \in K} \sum_{i \in \mathcal{I}} \pi_i(k \,|\, \theta_i)$, and transfers $(t_i)_{i \in \mathcal{I}}$ that are (strictly, or *ex-post*) *budget-balanced*, $\sum_{i \in \mathcal{I}} t_i = 0$.

Lemma 1 *A mechanism is ex-post Pareto-efficient only if transfers are ex-post budget-balanced. If $|\delta_{ij}| < 1/(2n-3)$ for all i and all $j \neq i$, then an ex-post materially efficient allocation function is also ex-post Pareto-efficient; moreover, no ex-post budget-balanced transfer scheme ex-post Pareto-dominates another.*

Proof. See Appendix A.1. ■

The intuition behind Lemma 1 is this: If agents switch from a social alternative that is materially efficient to one that is not, or from one budget-balanced transfer scheme to another, then at least one agent must incur a material loss. Now consider the agent whose material loss is largest; if this agent i is sufficiently selfish, $|\delta_{ij}| < 1/(2n-3)$ for all $j \neq i$, then she would also incur a loss utility-wise. In contrast, the Pareto frontier can be indefinite for combinations of social types satisfying $|\delta_{ij}| \geq 1/(2n-3)$, in which case a subgroup of agents might be willing to transfer arbitrary amounts of money to their joint favorite agent.[6]

Finally, $\langle k, T \rangle$ is *interim individually rational* if it gains all agents' approval at the interim stage (i.e., unanimous approval constitutes a Bayes-Nash equilibrium at the stage where agents' types are private information). Following Segal and Whinston (2016), we represent *reservation utilities* by the interim-expected utilities that agents' derive from a Bayesian mechanism $\langle k^\circ, T^\circ \rangle$, with $k^\circ : \Theta \times \Delta \to K$ specifying "property rights" and $T^\circ = (t_i^\circ)_{i \in \mathcal{I}} : \Theta \times \Delta \to \mathbb{R}^n$ specifying "liability rules."

[5] Bayesian implementation has been criticized for assuming that the distribution of agents' types is common knowledge. Bergemann and Morris (2005) have proposed *ex-post implementation* for environments with interdependent utilities, requiring that truthful revelation of types constitutes a Nash equilibrium. However, Jehiel et al. (2006) show that ex-post implementation is 'generically' not feasible in the presence of informational externalities, a finding extended by Zik (2021) to our present context.

[6] An example is the group of three agents with $\delta_{13} = \delta_{23} > 1/3$, $\delta_{12} = \delta_{21} = -1/3$, and $\delta_{31} = \delta_{32} = 0$, in which agents 1 and 2 are willing to *jointly* transfer arbitrary individual amounts $t > 0$ to agent 3.

3 A Possibility Result

We establish our main result with the help of two concepts, *preference-separating mechanisms* and *participation-stimulating transfers*:

Definition 1 (Preference Separation and Participation Stimulation)
A preference-separating mechanism $\langle k^\star, T^\star \rangle$ *consists of the ex-post materially efficient allocation function* $k^\star : \Theta \to K$*, with* $k^\star(\theta) \in \arg\max_{k \in K} \sum_{i \in \mathcal{I}} \pi_i(k \,|\, \theta_i)$*, and an ex-post budget-balanced transfer scheme* $T^\star = (t_i^\star)_{i \in \mathcal{I}} : \Theta \times \Delta \to \mathbb{R}^n$ *defined by*

$$t_i^\star(\hat{\theta}, \hat{\delta}) \;=\; \underbrace{\sum_{j \neq i} \Big[\mathbb{E}_{\theta_{-i}} \big[\pi_j(k^\star(\hat{\theta}_i, \theta_{-i}) \,|\, \theta_j) \big] - \mathbb{E}_{\theta_{-j}} \big[\pi_i(k^\star(\hat{\theta}_j, \theta_{-j}) \,|\, \theta_i) \big] \Big]}_{\text{the terms of trade}} \;+\; \underbrace{s_i^\star(\hat{\delta})}_{\substack{\text{participation-} \\ \text{stimulating} \\ \text{transfers}}} \;,$$

where participation-stimulating (PS) transfers $s^\star = (s_i^\star)_{i \in \mathcal{I}} : \Delta \to \mathbb{R}^n$ *are defined by jointly satisfying the following conditions:*

(i) $s^\star$ *is strategy-proof: For all* $i \in \mathcal{I}$*, all* $\delta \in \Delta$*, and all* $\hat{\delta}_i \in \Delta_i$*,*

$$\sum_{j \in \mathcal{I}} \delta_{ij} s_j^\star(\delta) \;\geq\; \sum_{j \in \mathcal{I}} \delta_{ij} s_j^\star(\hat{\delta}_i, \delta_{-i}).$$

(ii) $s^\star$ *is ex-post budget-balanced: For all* $\delta \in \Delta$*,*

$$\sum_{j \in \mathcal{I}} s_j^\star(\delta) \;=\; 0.$$

(iii) From unanimous participation in $s^\star$*, each agent derives a strictly positive interim-expected utility gain: For all* $i \in \mathcal{I}$ *and all* $\delta_i \in \Delta_i$*,*

$$\sum_{j \in \mathcal{I}} \delta_{ij} \, \mathbb{E}_{\delta_{-i}} \big[s_j^\star(\delta) \big] \;>\; 0.$$

Theorem 1 (Efficient Implementation With At Least Three Agents)
If $n \geq 3$*, then there exists a* preference-separating mechanism $\langle k^\star, T^\star \rangle$ *that is Bayesian incentive-compatible, interim individually rational, ex-post budget-balanced, and ex-post materially efficient. If* $|\delta_{ij}| < 1/(2n-3)$ *for all* i *and all* $j \neq i$*, then* $\langle k^\star, T^\star \rangle$ *is necessarily ex-post Pareto-efficient.*

Before we prove Theorem 1, we shall discuss the inner logic of our mechanism. Notice first that, despite the decoupling of incentives to reveal payoff types and social types, our mechanism asks agents to report these types *simultaneously*.

Consider the *terms of trade*. Those operate on agents' payoff types and, as we will see, are *social-preference robust* in that they leave agents' social preferences strategically irrelevant. This is achieved by applying the mutual-concessions principle of the dyadical AGV-mechanism (Arrow, 1979; d'Aspremont and Gérard-Varet, 1979) to each and every single dyad: For the materially efficient social alternative $k^\star(\hat{\theta})$, the transfer of agent i to every other j equals j's expectation of i's material payoff when j reports payoff type $\hat{\theta}_j$; that is, i transfers $\mathbb{E}_{\theta_{-j}}\big[\pi_i(k^\star(\hat{\theta}_j, \theta_{-j}) \,|\, \theta_i)\big]$ to j and receives $\mathbb{E}_{\theta_{-i}}\big[\pi_j(k^\star(\hat{\theta}_i, \theta_{-i}) \,|\, \theta_j)\big]$ from j.

For *two* other-regarding agents, Bierbrauer and Netzer (2016) show that the AGV-mechanism is *social-preference robust*. Agents are incentivized to behave *as if* they are selfish: If $-i$ reports her payoff type truthfully, then $\mathbb{E}_{\theta_{-i}}\big[\pi_{-i}(k^\star(\hat{\theta}_i, \theta_{-i}) \,|\, \theta_{-i}) + t^\star_{-i}(\hat{\theta}_i, \theta_{-i})\big] = \mathbb{E}_\theta\big[\pi_i(k^\star(\theta) \,|\, \theta_i)\big]$; thereby, i's degree of altruism is rendered strategically irrelevant. Bierbrauer and Netzer (2016, p. 570) also show that, in their framework, the *conventional* n-agents AGV (see Mas-Colell, Whinston, and Green, 1995, pp. 886) is social-preference robust only under an additional symmetry condition. In our framework, social-preference robustness can be established for groups of arbitrary size without any symmetry requirements.

Consequently, the *terms of trade* preserve agents' privately known social preferences as a strategic degree of freedom, which is utilized by *participation-stimulating transfers*. Those are independent of the actual allocation problem and serve the purpose of stimulating agents' participation in the terms of trade. While being *ex-post budget balanced*, PS transfers yield agents an *interim-expected Pareto improvement upon the terms of trade*, by Definition 1(iii). If this interim-expected Pareto improvement is amplified sufficiently through uniformly scaling up the PS transfers, then agents' interim-expected utilities from unanimous participation will outweigh their reservation utilities. Notice that the scaling-up is only possible if we allow transfers to take arbitrary negative values.

Finally, we note that our participation-stimulation approach cannot succeed in dyads:

Proposition 1 *Participation-stimulating transfers do not exist if* $n = 2$.

Proof. Suppose the opposite is true. Then Definition 1(iii) requires that $0 < \mathbb{E}_{\delta_{-i}}\big[s^\star_i(\delta)\big] + \delta_i\, \mathbb{E}_{\delta_{-i}}\big[s^\star_{-i}(\delta)\big]$ for both $i \in \{1, 2\}$ and all $\delta_i \in \Delta_i \subset (-1, 1)$, while $s^\star_{-i}(\delta) = -s^\star_i(\delta)$ due to

ex-post budget balance. Hence, $0 < (1-\delta_i)\,\mathbb{E}_{\delta_{-i}}\big[s_i^\star(\delta)\big]$, implying that $0 < \mathbb{E}_{\delta_{-i}}\big[s_i^\star(\delta)\big]$ for all i, δ_i. But then, $0 < \mathbb{E}_{\delta}\big[s_i^\star(\delta)\big]$ for both i, contradicting ex-post budget balance. ■

The intuition behind Proposition 1 is straightforward: Unanimous participation requires each social type to interim-expect a utility gain, but as budgets must be balanced ex post while each agent values her own material well-being more than the other's, this requires each social type to interim-expect a material benefit. These interim expectations cannot be mutually consistent for all social types, regardless the specification of transfers $s^\star$; otherwise, both agents would benefit materially *ex ante*, contradicting budget balance.

4 Proof of Theorem 1

The proof of Theorem 1 proceeds in a series of Lemmas. Throughout, $n \geq 3$.

Lemma 2 Preference-separating mechanisms *are Bayesian incentive-compatible and ex-post materially efficient. If* $|\delta_{ij}| < 1/(2n-3)$ *for all* i *and all* $j \neq i$*, they are also ex-post Pareto-efficient.*

Proof. *Incentive compatibility*: Suppose the agents other than i reveal their types truthfully. Then the transfers that i interim-expects for herself and every other j read:

$$\begin{aligned}
\bar{t}_{ii}(\hat{\theta}_i, \hat{\delta}_i) &= \sum_{\ell \neq i} \mathbb{E}_{\theta_{-i}}\big[\pi_\ell(k^\star(\hat{\theta}_i, \theta_{-i}) \,|\, \theta_\ell)\big] - (n-1)\,\mathbb{E}_{\theta}\big[\pi_i(k^\star(\theta) \,|\, \theta_i)\big] + \mathbb{E}_{\delta_{-i}}\big[s_i^\star(\hat{\delta}_i, \delta_{-i})\big], \\
\bar{t}_{ij}(\hat{\theta}_i, \hat{\delta}_i) &\overset{j\neq i}{=} \sum_{\ell \neq j} \mathbb{E}_{\theta_{-i},\theta_{-j}}\big[\pi_\ell(k^\star(\theta) \,|\, \theta_\ell)\big] - \sum_{\ell \neq i,j} \mathbb{E}_{\theta_{-i},\theta_{-\ell}}\big[\pi_j(k^\star(\theta) \,|\, \theta_j)\big] \\
&\qquad -\mathbb{E}_{\theta_{-i}}\big[\pi_j(k^\star(\hat{\theta}_i, \theta_{-i}) \,|\, \theta_j)\big] + \mathbb{E}_{\delta_{-i}}\big[s_j^\star(\hat{\delta}_i, \delta_{-i})\big] \\
&= \sum_{\ell \in \mathcal{I}} \mathbb{E}_{\theta}\big[\pi_\ell(k^\star(\theta) \,|\, \theta_\ell)\big] - (n-1)\,\mathbb{E}_{\theta}\big[\pi_j(k^\star(\theta) \,|\, \theta_j)\big] \\
&\qquad - \mathbb{E}_{\theta_{-i}}\big[\pi_j(k^\star(\hat{\theta}_i, \theta_{-i}) \,|\, \theta_j)\big] + \mathbb{E}_{\delta_{-i}}\big[s_j^\star(\hat{\delta}_i, \delta_{-i})\big].
\end{aligned}$$

Agent i's interim-expected utility from reporting $(\hat{\theta}_i, \hat{\delta}_i)$ thus satisfies

$$
\begin{aligned}
(1) \quad U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) &= \sum_{j\in\mathcal{I}} \delta_{ij} \Big[\mathbb{E}_{\theta_{-i}}\big[\pi_j(k^\star(\hat{\theta}_i, \theta_{-i}) \,|\, \theta_j)\big] + \bar{t}_{ij}(\hat{\theta}_i, \hat{\delta}_i)\Big] \\
&= \mathbb{E}_{\theta_{-i}}\left[\sum_{\ell\in\mathcal{I}} \pi_\ell(k^\star(\hat{\theta}_i, \theta_{-i}) \,|\, \theta_\ell)\right] + \left(\sum_{j\neq i} \delta_{ij}\right) \mathbb{E}_\theta\left[\sum_{\ell\in\mathcal{I}} \pi_\ell(k^\star(\theta) \,|\, \theta_\ell)\right] \\
&\quad - (n-1)\,\mathbb{E}_\theta\left[\sum_{j\in\mathcal{I}} \delta_{ij}\pi_j(k^\star(\theta) \,|\, \theta_j)\right] + \sum_{j\in\mathcal{I}} \delta_{ij}\,\mathbb{E}_{\delta_{-i}}\big[s_j^\star(\hat{\delta}_i, \delta_{-i})\big].
\end{aligned}
$$

By equation (1), the incentives to reveal payoff types and social types are additively separated. As participation-stimulating transfers $s^\star$ are strategy-proof by Definition 1(i), preference-separating mechanisms are (dominant-strategy) incentive-compatible with respect to social types. On the other hand, if truthful revelation of her payoff type θ_i was inferior for some agent i, then there would exist $\hat{\theta}_i$ and θ_{-i} such that $\sum_{\ell\in\mathcal{I}} \pi_\ell(k^\star(\hat{\theta}_i, \theta_{-i}) \,|\, \theta_\ell) > \sum_{\ell\in\mathcal{I}} \pi_\ell(k^\star(\theta_i, \theta_{-i}) \,|\, \theta_\ell)$, implying that $\sum_{\ell\in\mathcal{I}} \pi_\ell(k \,|\, \theta_\ell) > \sum_{\ell\in\mathcal{I}} \pi_\ell(k^\star(\theta) \,|\, \theta_\ell)$ for some social alternative k, in contradiction to the definition of $k^\star$.

Efficiency: Preference-separating mechanisms are ex-post materially efficient by construction; hence, by Lemma 1, they are also ex-post Pareto-efficient if $|\delta_{ij}| < 1/(2n-3)$ for all i and all $j \neq i$. ■

By equation (1), the terms of trade are social-preference robust: Agents' social preferences are rendered strategically irrelevant when it comes to implementing the materially efficient allocation function $k^\star$. This opens up the possibility to operationalize the asymmetry of information about agents' social preferences to satisfy their interim participation constraints.

We construct participation-stimulating transfer schemes as follows. Let $M \in \mathcal{I}$ denote one (arbitrarily chosen) agent and define transfers $s^\star = (s_i^\star)_{i\in\mathcal{I}} : \Delta \to \mathbb{R}^n$ by

$$
\begin{aligned}
(2) \qquad s_M^\star(\delta) &= -\sum_{j\neq M} s_j^\star(\delta), \\
(3) \qquad s_j^\star(\delta) &= -C + g_j(\delta_j^\star) - \delta_j^\star g_j'(\delta_j^\star) + \sum_{\ell\neq j,M} g_\ell'(\delta_\ell^\star), \quad \text{for } j \neq M, \\
(4) \qquad g_j(\delta_j^\star) &= \mathrm{Var}_{\delta_j}[\delta_j^\star] + (\delta_j^\star - \mathbb{E}_{\delta_j}[\delta_j^\star])^2, \\
(5) \qquad \delta_j^\star &= \frac{\sum_{\ell\neq j,M}(\delta_{j\ell} - \delta_{jM})}{\delta_{jj} - \delta_{jM}},
\end{aligned}
$$

for some constant $C > 0$. In order to establish that this transfer scheme is participation-stimulating, we first consider the functions $(g_j)_{j \neq M}$ of equation (4):

Lemma 3 *Be $X_j : \Delta_j \to \mathbb{R}$ a continuous non-constant random variable. Then $\mathbb{E}_{\delta_j}[X_j]$ and $\mathrm{Var}_{\delta_j}[X_j]$ exist, and $g_j : \mathbb{R} \to \mathbb{R}$ defined by $g_j(X_j) = \mathrm{Var}_{\delta_j}[X_j] + \big(X_j - \mathbb{E}_{\delta_j}[X_j]\big)^2$ satisfies $g_j > 0$, $g_j'' > 0$, and $\mathbb{E}_{\delta_j}\big[g_j'(X_j)\big] = 0 = \mathbb{E}_{\delta_j}\big[g_j(X_j) - X_j g_j'(X_j)\big]$.*

Proof. $\mathbb{E}_{\delta_j}[X_j]$ and $\mathrm{Var}_{\delta_j}[X_j]$ exist, since Δ_j is compact and convex while X_j and the density of δ_j are continuous. Obviously, $g_j(X_j) > 0$, $g_j''(X_j) > 0$, and $\mathbb{E}_{\delta_j}\big[g_j'(X_j)\big] = 2\,\mathbb{E}_{\delta_j}\big[X_j - \mathbb{E}_{\delta_j}[X_j]\big] = 0$. On the other hand, as $\mathrm{Var}_{\delta_j}[X_j] = \mathbb{E}_{\delta_j}[X_j^2] - \mathbb{E}_{\delta_j}[X_j]^2$, one has $g_j(X_j) - X_j g_j'(X_j) = \mathbb{E}_{\delta_j}[X_j^2] - \mathbb{E}_{\delta_j}[X_j]^2 + X_j^2 - 2X_j\mathbb{E}_{\delta_j}[X_j] + \mathbb{E}_{\delta_j}[X_j]^2 - 2X_j\big(X_j - \mathbb{E}_{\delta_j}[X_j]\big) = \mathbb{E}_{\delta_j}[X_j^2] - X_j^2$; hence, $\mathbb{E}_{\delta_j}\big[g_j(X_j) - X_j g_j'(X_j)\big] = 0$. ■

We obtain that participation-stimulating transfer schemes do exist if $n \geq 3$:

Lemma 4 *The transfer scheme $s^\star$ defined by* (2)–(5) *is* participation-stimulating *in the manner of* Definition 1 *if $C > 0$ is chosen sufficiently small.*

Proof. *Strategy proofness*: Under $s^\star$, each agent $j \neq M$ reports a social type $\hat{\delta}_j$, which is strategically equivalent to reporting some signal $\hat{\delta}_j^\star \in \mathbb{R}$. Her ex-post utility is given by

$$\begin{aligned}\sum_{\ell \neq M}(\delta_{j\ell} - \delta_{jM}) s_\ell^\star(\hat{\delta}) \;&=\; (\delta_{jj} - \delta_{jM})\left[g_j(\hat{\delta}_j^\star) - \hat{\delta}_j^\star g_j'(\hat{\delta}_j^\star) + \sum_{\ell \neq j,M} g_\ell'(\hat{\delta}_\ell^\star)\right] \\ &\quad + \sum_{\ell \neq j,M}(\delta_{j\ell} - \delta_{jM})\left[g_\ell(\hat{\delta}_\ell^\star) - \hat{\delta}_\ell^\star g_\ell'(\hat{\delta}_\ell^\star) + \sum_{\ell' \neq \ell,j,M} g_{\ell'}'(\hat{\delta}_{\ell'}^\star)\right] \\ &\quad + \left[\sum_{\ell \neq j,M}(\delta_{j\ell} - \delta_{jM})\right] g_j'(\hat{\delta}_j^\star) - C\sum_{\ell \neq M}(\delta_{j\ell} - \delta_{jM}).\end{aligned}$$

Hence, when substituting for $\delta_j^\star = \sum_{\ell \neq j,M}(\delta_{j\ell} - \delta_{jM})/(\delta_{jj} - \delta_{jM})$, agent j maximizes $g_j(\hat{\delta}_j^\star) + (\delta_j^\star - \hat{\delta}_j^\star) g_j'(\hat{\delta}_j^\star)$ over the choice of $\hat{\delta}_j^\star$. As $g_j'' > 0$, each $j \neq M$ has the strictly dominant strategy to report $\hat{\delta}_j^\star = \delta_j^\star$. As agent M is not involved strategically, she has the weakly dominant strategy to report her true social type δ_M.

Ex-post budget balance: Immediate from equation (2).

Interim-expected Pareto improvement: When substituting for $\delta_j^\star$ and $\mathbb{E}_{\delta_\ell}\big[g'_\ell(\delta_\ell^\star)\big] = 0 = \mathbb{E}_{\delta_\ell}\big[g_\ell(\delta_\ell^\star) - \delta_\ell^\star g'_\ell(\delta_\ell^\star)\big]$, due to Lemma 3, then j's interim-expected utility from $s^\star$ is

$$\begin{aligned}\sum_{\ell\neq M}(\delta_{j\ell}-\delta_{jM})\,\mathbb{E}_{\delta_{-j}}[s_\ell^\star(\delta)] &= (\delta_{jj}-\delta_{jM})g_j(\delta_j^\star) - C\sum_{\ell\neq M}(\delta_{j\ell}-\delta_{jM}) \\ &= (\delta_{jj}-\delta_{jM})g_j(\delta_j^\star) - C(\delta_{jj}-\delta_{jM}) - C\sum_{\ell\neq j,M}(\delta_{j\ell}-\delta_{jM}) \\ &= (\delta_{jj}-\delta_{jM})\left[g_j(\delta_j^\star) - C(1+\delta_j^\star)\right].\end{aligned}$$

Recall that $\delta_{jj} = 1 > \delta_{jM}$ and $g_j(\delta_j^\star) \geq \mathrm{Var}_{\delta_j}[\delta_j^\star] > 0$. Notice that $\delta_j^\star < n-2$, since $\delta_{jj} - \delta_{jM} > \delta_{j\ell} - \delta_{jM}$ for all $\ell \neq j, M$. Hence, each agent $j \neq M$ derives positive interim-expected utility from unanimous participation if $C \leq \min_{j\neq M}\mathrm{Var}_{\delta_j}[\delta_j^\star]/(n-1)$. Due to Lemma 3 again, also M's interim-expected utility is positive if all agents participate: $\sum_{i\in\mathcal{I}}\delta_{Mi}\,\mathbb{E}_{\delta_{-M}}[s_i^\star(\delta)] = \sum_{j\neq M}(\delta_{Mj}-1)\,\mathbb{E}_\delta[s_j^\star(\delta)] = C\sum_{j\neq M}(1-\delta_{Mj}) > 0$. ■

Several remarks on the PS scheme (2)–(5) are in order. First, $s^\star$ is independent of agent M's social type, $(\delta_{Mj})_{j\neq M}$, such that M has no strategic role to play under $s^\star$. This feature is not a prerequisite for preference-separating implementation. Second, each agent $i \neq M$ has the strictly dominant strategy to report $\delta_i^\star = \sum_{\ell\neq i,M}(\delta_{i\ell}-\delta_{iM})/(\delta_{ii}-\delta_{iM})$ which is thus a one-dimensional sufficient statistic for i's social type. This fact allows for implementing the PS scheme by having agents reveal the necessary information about their social types via the choice of one-dimensional strategic variables, such as efforts. We discuss this in detail in Section 6.2. Third, the PS scheme implicitly assumes that the mean and variance of every $\delta_j^\star$ are commonly known. This assumption is sufficient but not necessary. As $s^\star$ is strategy-proof while the resulting interim-expected Pareto improvement is strict, it suffices that agents (and the designer) have sufficiently good estimates of those means and variances. Finally, notice that Lemma 4 is equally valid if a social-type set contains mass points.

With Lemmas 1 to 4 at hand, we can establish Theorem 1:

Proof of Theorem 1. Consider the preference-separating mechanism $\langle k^\star, T^\star\rangle$ with

$$t_i^\star(\hat\theta,\hat\delta) = \underbrace{\sum_{j\neq i}\Big[\mathbb{E}_{\theta_{-i}}\big[\pi_j(k^\star(\hat\theta_i,\theta_{-i})\,|\,\theta_j)\big] - \mathbb{E}_{\theta_{-j}}\big[\pi_i(k^\star(\hat\theta_j,\theta_{-j})\,|\,\theta_i)\big]\Big]}_{\text{the terms of trade}} + \underbrace{\alpha^\star\cdot s_i^\star(\hat\delta)}_{\text{PS transfers}},$$

where $(s_i^\star)_{i\in\mathcal{I}}$ is defined by equations (2) to (5) while $\alpha^\star > 0$. Notice that the conditions of Definition 1 are invariant under scaling all the components $s_i^\star$ with the same factor.

By Lemmas 2 and 4, this mechanism is Bayesian incentive-compatible. It is ex-post budget-balanced and ex-post materially efficient by construction. By Lemma 1, it is ex-post Pareto-efficient if $|\delta_{ij}| < 1/(2n-3)$ for all i and all $j \neq i$.

By equation (1), and since $\langle k^\star, T^\star\rangle$ is Bayesian incentive-compatible, agent i's interim-expected utility from unanimous participation in $\langle k^\star, T^\star\rangle$ is given by

$$\begin{aligned} U_i(\theta_i,\delta_i) \;=\; & \mathbb{E}_{\theta_{-i}}\left[\sum_{\ell\in\mathcal{I}}\pi_\ell(k^\star(\theta)\,|\,\theta_\ell)\right] + \left(\sum_{j\neq i}\delta_{ij}\right)\mathbb{E}_\theta\left[\sum_{\ell\in\mathcal{I}}\pi_\ell(k^\star(\theta)\,|\,\theta_\ell)\right] \\ & - (n-1)\,\mathbb{E}_\theta\left[\sum_{j\in\mathcal{I}}\delta_{ij}\pi_j(k^\star(\theta)\,|\,\theta_j)\right] + \alpha^\star\cdot\sum_{j\in\mathcal{I}}\delta_{ij}\,\mathbb{E}_{\delta_{-i}}\big[s_j^\star(\delta)\big], \end{aligned}$$

where $\sum_{j\in\mathcal{I}}\delta_{ij}\,\mathbb{E}_{\delta_{-i}}\big[s_j^\star(\delta)\big] > 0$ due to Lemma 4. Hence, if $\alpha^\star$ is chosen sufficiently large, agents' interim participation constraints are satisfied for any given collection of reservation utilities, specified in Section 2.3. ■

5 The Intuition Behind Participation Stimulation

In this Section, we outline the intuition behind our participation-stimulating transfers. We focus on the simplest possible case of three agents and dedicate to one of those a strategically inoperative (or, mediating) role. We refer to this agent as M and to the others as agents 1 and 2. With all else equal, we assume here that it is common knowledge that $\delta_{1M} = 0 = \delta_{2M}$, so we can write $\delta_1 = \delta_{12}$ and $\delta_2 = \delta_{21}$.

We construct PS transfers by first looking for a transfer scheme $s^\star$ that varies only in the social types of agents 1 and 2, is strategy-proof, and yields 1 and 2 each an *ex-ante* transfer of zero. Ex-post transfers to (from) 1 and 2 are paid (received) by M. Then transfers are *ex-post budget-balanced* among $\{1,2,M\}$, and interim-expected utility to M is zero. If participation in $s^\star$ yields 1 and 2 an interim-expected utility gain, then M can extract a monetary rent by demanding a (sufficiently small) uniform participation fee from 1 and 2. Thereby, also M obtains an interim-expected utility gain.

We first consider a discrete distribution with two equally likely social types; we thereby justify and generalize our money-pump example at the beginning. Then we show how the PS transfers of Lemma 4 can be constructed for arbitrary distributions.

	$\hat{\delta}_2 = \underline{\delta}$		$\hat{\delta}_2 = \overline{\delta}$	
$\hat{\delta}_1 = \underline{\delta}$	a	a	b	c
$\hat{\delta}_1 = \overline{\delta}$	c	b	d	d

(a) General case

	$\hat{\delta}_2 = \underline{\delta}$		$\hat{\delta}_2 = \overline{\delta}$	
$\hat{\delta}_1 = \underline{\delta}$	$-y$	$-y$	x	$-x$
$\hat{\delta}_1 = \overline{\delta}$	$-x$	x	y	y

(b) A minimum viable example

Figure 1: Participation-stimulating transfers for two social types.

5.1 A Simple Discrete Setup

Let $\Delta_{12} = \Delta_{21} = \left\{\underline{\delta}, \overline{\delta}\right\}$, with $-1/2 < \underline{\delta} < \overline{\delta} < 1/2$, and suppose both social types are equally likely. We refer to an agent of type $\underline{\delta}$ as a *relative egoist* and to an agent of type $\overline{\delta}$ as a *relative altruist*. In this scenario, PS transfers can be represented by a 2×2–*payoff matrix* specifying individual transfers for the feasible combinations of reported types $\hat{\delta}_i$. This matrix is depicted in Figure 1(a). For parameters to be determined, we denote the transfer scheme by $s = s[a, b, c, d]$. We gradually construct a transfer scheme that is strategy-proof and yields agents 1 and 2 strictly positive interim-expected utility while the sum of transfers is negative *ex ante* (such that M agrees to balance the budget).

We start out from the benchmark scheme $s_0 = s[0, 0, 0, 0]$ and consider the off-diagonal cells of the payoff matrix. Suppose we change off-diagonal payoffs according to $s_1 = s[0, x, -x, 0]$, for some $x > 0$, such that a reported relative egoist receives the amount x from a reported relative altruist. If agents are truthful, this change increases the *sum* of the agents' ex-post utilities in each off-diagonal cell from zero to $\left(\overline{\delta} - \underline{\delta}\right) x > 0$ and, thus, increases each agents' *ex-ante* expected utility by $\left(\overline{\delta} - \underline{\delta}\right) x/4$. We thus obtain an ex-ante expected utility gain by appropriately transferring money between agents, and this gain stems from the (potential) difference in agents' social preferences. In principle, this utility gain is the source for participation stimulation.

However, under $s_1 = s[0, x, -x, 0]$, truthful reporting is only incentive-compatible for a relative egoist. To incentivize a relative altruist, consider an additional change given by an increase of transfers for the meeting of two reported relative altruists (i.e., an increase of d) and a corresponding decrease for the meeting of two reported relative egoists (i.e., a decrease of a). Denoting this change $y > 0$, we obtain the scheme $s_2 = s[-y, x, -x, y]$

depicted in Figure 1(b). Switching from s_1 to s_2 further increases ex-ante expected utility by $\left(\overline{\delta} - \underline{\delta}\right) y/4 > 0$.[7]

If $(1 - \underline{\delta})\, x > (1 + \underline{\delta})\, y$, then a relative egoist reveals her type truthfully while interim-expecting a utility gain from unanimous participation. Similarly, a relative altruist is incentivized to be truthful and to participate if $\left(1 - \overline{\delta}\right) x < \left(1 + \overline{\delta}\right) y$. These two conditions can be satisfied simultaneously if $(1-\overline{\delta})/(1+\overline{\delta}) < (1-\underline{\delta})/(1+\underline{\delta})$ or, equivalently, if $\underline{\delta} < \overline{\delta}$. Hence, a positive variance of the social-type distribution, regardless how small, is already sufficient to allow for participation stimulation. As the interim-expected utility gains are strictly positive, each agent can be required to pay a participation fee $F > 0$ to M. Specifically, the transfer scheme $s_3 = s[-y-F, x-F, -x-F, y-F]$ is strategy-proof and induces unanimous participation if $2F < \min\left\{x(1 - \underline{\delta})/(1 + \underline{\delta}) - y \;;\; y - x(1 - \overline{\delta})/(1 + \overline{\delta})\right\}$. Notice that F can be arbitrarily large when increasing x while letting $y = x(1 - \underline{\delta}\ \overline{\delta})/[(1 + \underline{\delta})\,(1 + \overline{\delta})]$.

In conclusion, the (potential) differences in agents' social preferences can be utilized to generate an ex-ante expected utility gain. If this gain is distributed appropriately among the agents (including M), then they participate willingly in the respective ex-post budget-balanced game. As each agent values a dollar to herself more than a dollar to any other agent, distribution must take place *across* the different states of Nature (i.e., type realizations) and can be to everyone's advantage only in expectation.

5.2 Arbitrary Social-Type Distributions

We construct PS transfers for arbitrary social-type distributions by first looking for a (smooth) transfer scheme $s^\star = (s_1^\star, s_2^\star)$ that is *strategy-proof*,

$$\frac{\partial s_i^\star(\delta)}{\partial \delta_i} + \delta_i \frac{\partial s_{-i}^\star(\delta)}{\partial \delta_i} \;=\; 0, \tag{6}$$

and yields each agent $i \in \{1, 2\}$ an *ex-ante transfer of zero*,

$$\mathbb{E}_\delta\left[s_i^\star(\delta)\right] \;=\; 0, \tag{7}$$

[7]Notice that $s[-y, 0, 0, y]$, though yielding an ex-ante expected utility gain, too, is not strategy-proof for the relative egoist. In fact, we need to deploy all four cells of the payoff matrix.

as well as a *strictly positive interim-expected utility gain* from unanimous participation,

$$\mathbb{E}_{\delta_{-i}}\left[s_i^\star(\delta)\right] + \delta_i\,\mathbb{E}_{\delta_{-i}}\left[s_{-i}^\star(\delta)\right] \;=\; g_i(\delta_i) \tag{8}$$

for some function $g_i : \Delta_i \to (0,\infty)$. We can derive $s^\star$ from appropriate functions $(g_i)_i$:[8]

Proposition 2 *For smooth functions $g_i : \Delta_i \to (0,\infty)$ satisfying $g_i'' > 0$ and*

$$\mathbb{E}_{\delta_i}\left[g_i'(\delta_i)\right] \;=\; 0 \;=\; \mathbb{E}_{\delta_i}\left[g_i(\delta_i) - \delta_i g_i'(\delta_i)\right] \tag{9}$$

define transfers by $s_i^\star(\delta) = g_i(\delta_i) - \delta_i g_i'(\delta_i) + g_{-i}'(\delta_{-i})$. Then $s^\star = (s_1^\star, s_2^\star)$ satisfies conditions (6)–(8). From unanimous participation in $s^\star$, agent i derives an interim-expected utility gain of $g_i(\delta_i) > 0$ while interim-expecting a transfer of $\mathbb{E}_{\delta_{-i}}\left[s_i^\star(\delta)\right] = g_i(\delta_i) - \delta_i g_i'(\delta_i)$ to herself and a transfer of $\mathbb{E}_{\delta_{-i}}\left[s_{-i}^\star(\delta)\right] = g_i'(\delta_i)$ to agent $-i$.

Proof. We have $d[s_i^\star(\hat\delta_i, \delta_{-i}) + \delta_i s_{-i}^\star(\hat\delta_i, \delta_{-i})]/d\hat\delta_i = (\delta_i - \hat\delta_i) g_i''(\hat\delta_i)$; hence, $\hat\delta_i = \delta_i$. By (9), $\mathbb{E}_\delta\left[s_i^\star(\delta)\right] = 0$. By (9) again, $\mathbb{E}_{\delta_{-i}}\left[s_i^\star(\delta)\right] + \delta_i\,\mathbb{E}_{\delta_{-i}}\left[s_{-i}^\star(\delta)\right] = [g_i(\delta_i) - \delta_i g_i'(\delta_i)] + \delta_i\,[g_i'(\delta_i)] = g_i(\delta_i) > 0$. Hence, $s^\star$ satisfies (6)–(8). All else is obvious. ■

Under $s^\star$ of Proposition 2, the transfer that an agent interim-expects for herself is maximal (and positive) if that agent is a pure-payoff maximizer ($\delta_i = 0$), as $d\,\mathbb{E}_{\delta_{-i}}\left[s_i^\star(\delta)\right]/d\delta_i = -\delta_i g_i''(\delta_i)$ while $g_i'' > 0$. Money is thus redistributed *ex interim* to those agents who 'care least' about others. On the other hand, the transfer that an agent interim-expects for her opponent increases in her own social type, since $d\,\mathbb{E}_{\delta_{-i}}\left[s_{-i}^\star(\delta)\right]/d\delta_i = g_i''(\delta_i) > 0$, and is zero ex ante, since $\mathbb{E}_{\delta_i}\left[g_i'(\delta_i)\right] = 0$. Hence, least (most) altruistic types interim-expect to impose a negative (positive) externality on their opponent. This interim-expected externality, weighted with an agent's social type, overcompensates for interim-expected monetary losses: $\mathbb{E}_{\delta_{-i}}\left[s_i^\star(\delta)\right] + \delta_i\,\mathbb{E}_{\delta_{-i}}\left[s_{-i}^\star(\delta)\right] = g_i(\delta_i) > 0$.

It is easy to see that the functions

$$g_i(\delta_i) = \mathrm{Var}_{\delta_i}[\delta_i] + \left(\delta_i - \mathbb{E}_{\delta_i}[\delta_i]\right)^2 \tag{10}$$

[8]The sufficient conditions of Proposition 2 can be obtained as follows: By differentiating (6) with respect to δ_{-i} one obtains that $\partial^2 s_i^\star/\partial\delta_1\partial\delta_2 = 0$, implying that $s_i^\star$ is additively separable: $s_i^\star(\delta) = a_i(\delta_i) + b_i(\delta_{-i})$ for appropriate functions $a_i : \Delta_i \to \mathbb{R}$ and $b_i : \Delta_{-i} \to \mathbb{R}$. Hence, by condition (6) again, $a_i'(\delta_i) + \delta_i b_{-i}'(\delta_i) = 0$, such that partial integration yields $a_i(\delta_i) = -\delta_i b_{-i}(\delta_i) + \int_{\delta_i^{\min}}^{\delta_i} b_{-i}(x)dx + C$, for a constant C. Write $g_i(\delta_i) = \int_{\delta_i^{\min}}^{\delta_i} b_{-i}(x)dx + C$. Then, $a_i(\delta_i) = g_i(\delta_i) - \delta_i g_i'(\delta_i)$ and $b_i(\delta_{-i}) = g_{-i}'(\delta_{-i})$, yielding $s^\star$ of Proposition 2. Impose $g_i'' > 0$ to satisfy the SOC, and impose (9) to satisfy (7) and (8).

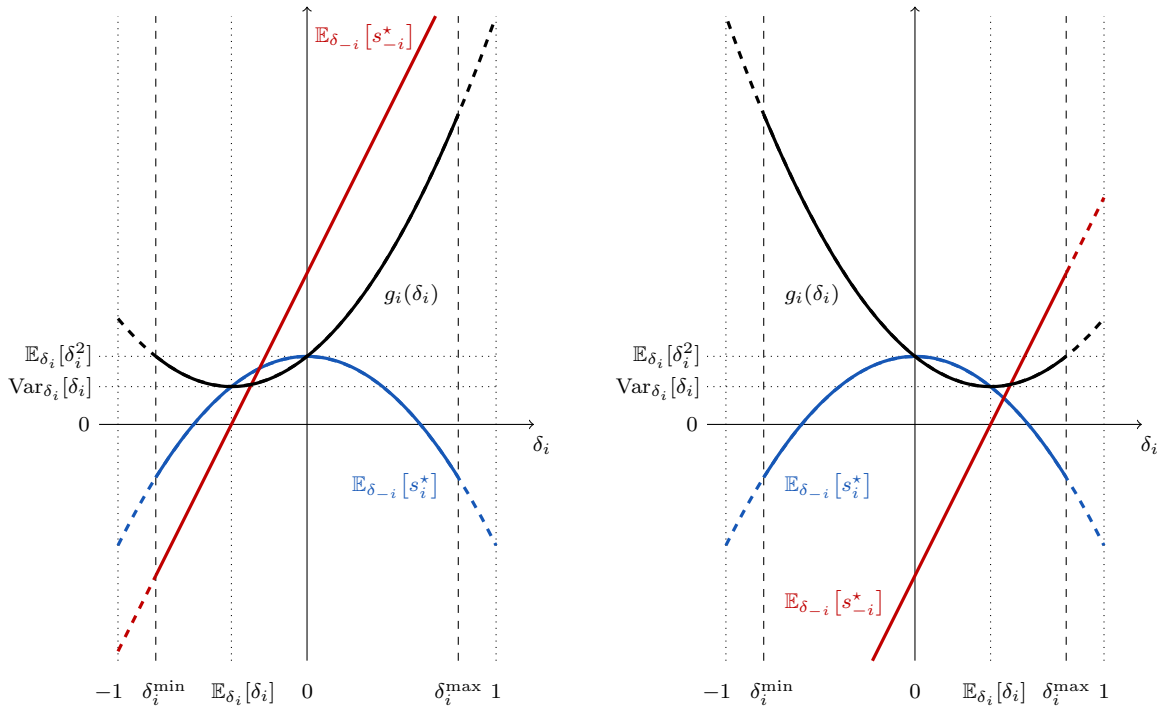

Figure 2: *The utility gain* $g_i(\delta_i) = \mathbb{E}_{\delta_{-i}}\big[s_i^\star(\delta)\big] + \delta_i\,\mathbb{E}_{\delta_{-i}}\big[s^\star_{-i}(\delta)\big] > 0$ *that a social type* δ_i *interim-expects under the transfer scheme* $s^\star$ *of equation* (11)*, for two different type distributions:* $\delta_i \in [\delta_i^{\min}, \delta_i^{\max}] = [-4/5, 4/5]$*,* $\mathbb{E}_{\delta_i}[\delta_i] = \mp 2/5$*, and* $\mathrm{Var}_{\delta_i}[\delta_i] = 1/5$*, such that* $\mathbb{E}_{\delta_{-i}}\big[s_i^\star(\delta)\big] = 9/25 - \delta_i^2$*,* $\mathbb{E}_{\delta_{-i}}\big[s^\star_{-i}(\delta)\big] = 2\delta_i \pm 4/5$*, and* $g_i(\delta_i) = (\delta_i \pm 2/5)^2 + 1/5$*.*

satisfy the conditions of Proposition 2; see also Lemma 3. This result has several desirable implications: It emphasizes the role of asymmetric information. With $\mathrm{Var}_{\delta_i}[\delta_i] = 0$ (and, thus, $\delta_i \equiv \mathbb{E}_{\delta_i}[\delta_i]$), the interim-expected utility gain of agent i is zero. Moreover, it shows that common-knowledge assumptions about social-type distributions can be weak. In fact, it suffices to assume common knowledge about their means and variances.

The transfer scheme corresponding to (10) reads

$$s_i^\star(\delta) = \mathbb{E}_{\delta_i}[\delta_i^2] - \delta_i^2 + 2\left(\delta_{-i} - \mathbb{E}_{\delta_{-i}}[\delta_{-i}]\right), \tag{11}$$

and Figure 2 depicts its interim-expected distributive effects: Social types satisfying $|\delta_i| > \sqrt{\mathbb{E}_{\delta_i}[\delta_i^2]}$ incur interim-expected monetary losses (blue), $\mathbb{E}_{\delta_{-i}}\big[s_i^\star(\delta)\big] < 0$, for which they are overcompensated through sufficiently strong interim-expected externalities (red), $\mathbb{E}_{\delta_{-i}}\big[s^\star_{-i}(\delta)\big] = 2\delta_i - 2\,\mathbb{E}_{\delta_i}[\delta_i]$. These interim-expected monetary losses of relatively strong social types are the source for attracting relatively selfish agents with interim-expected monetary gains (blue): $\mathbb{E}_{\delta_{-i}}\big[s_i^\star(\delta)\big] > 0$ for social types $|\delta_i| < \sqrt{\mathbb{E}_{\delta_i}[\delta_i^2]}$.

From here, we obtain our participation-stimulating transfers (2)–(5) as follows: The interim-expected distributive effects of $s_i^\star(\delta) = g_i(\delta_i) - \delta_i g_i'(\delta_i) + g'_{-i}(\delta_{-i})$, discussed

above, suggest that participation stimulation is driven by the externality that i imposes on $-i$ through the term $g'_{-i}(\delta_{-i})$. Hence, for the n-agents case, we let $s_i^\star(\hat{\delta}) = -C + g_i(\hat{\delta}_i^\star) - \hat{\delta}_i^\star g_i'(\hat{\delta}_i^\star) + \sum_{\ell \neq i,M} g_\ell'(\hat{\delta}_\ell^\star)$ for each $i \neq M$, with C a uniform participation fee given to M. Under this scheme, now re-accounting for the privately known social preferences toward M, each agent $i \neq M$ has the dominant strategy to report $\hat{\delta}_i^\star = \delta_i^\star$ of equation (5). Finally, the functions g_i of (10) must now be chosen with respect to the random variables $\delta_i^\star$. We thus obtain equation (4).

The term $\delta_i^\star = \sum_{\ell \neq i,M}(\delta_{i\ell} - \delta_{iM})/(\delta_{ii} - \delta_{iM})$ of equation (5) gives i's relative marginal utility from a redistribution of M's money either to the others, who obtain equal shares, or to i herself. It can be referred to as i's *relative spite* towards M, since $\delta_i^\star$ decreases in δ_{iM} and increases in i's prosociality toward the others, given by $\sum_{\ell \neq i,M} \delta_{i\ell}$. Notice from equation (3) that the transfer i interim-expects for herself is maximal if $\delta_i^\star = 0$, in which case i cares about M just as much as about the rest of the group. Money is thus redistributed ex interim to those agents who are (nearly) indifferent about any form of redistribution between M and the others. On the other hand, an agent who strongly cares more (less) about M than about the others interim-expects to invoke a redistribution from the others to M (from M to the others), which overcompensates her emotionally for interim-expected monetary losses.

6 Discussion

6.1 What If Social Types Are Common Knowledge?

Asymmetric information about agents' social preferences is a key assumption in the above analysis. We can easily rule out that participation stimulation in the manner of Definition 1 would work for *commonly known* social types: Under common knowledge, Definition 1(iii) would transform into the requirement that participation-stimulating transfers *ex-post* Pareto-dominate the transfer scheme $(s_i = 0)_{i \in \mathcal{I}}$ of *ex-post* budget-balanced zero-transfers (i.e., $\sum_{j \in \mathcal{I}} \delta_{ij} s_j^\star(\delta) > 0$ for all i and all δ), which is impossible due to Lemma 1.

We shall also discuss what *is* feasible if social types are common knowledge. Plausibly, if agents are sufficiently altruistic (i.e., $\delta_{ij} \to 1$ for all $i, j \neq i$), then individual rationality is satisfied for materially efficient allocation functions and budget-balanced transfers; see also Kucuksenel (2012). Seeking solutions that work for arbitrary social types, let us consider the following example which we owe to an Anonymous Referee:

Example. Suppose there are three agents and it is commonly known that $\delta_{12} = \delta_{23} = \delta_{31} = 1/10 < \delta_{13} = \delta_{21} = \delta_{32} = 1/5$. Now consider the following *liability rule*: If agent 1 refuses to participate while the other agents agree, then agent 3 must pay $x > 0$ to agent 2; if 2 refuses while the others agree, then 1 must pay x to 3; and if 3 refuses while the others agree, then 2 must pay x to 1. Under this liability rule, assuming the respective other agents participate, an agent who refuses incurs a utility loss of $x/10$. Letting x sufficiently large, *every* mechanism becomes individually rational in Bayes-Nash equilibrium. ■

In the Example, an agent who refuses to participate is (emotionally) penalized by forcing the agent she likes more to subsidize the agent she likes less. Obviously, this strategy works for every group in which each agent i prefers some agent j_i over some other agent ℓ_i. Commonly known social preferences can thus be exploited to push, rather than pull, agents into participation. The Example relates to the branch of literature that considers more general property rights and liability rules, allowing for redistribution even if some agents refuse to participate (Segal and Whinston, 2016) or allowing the designer to impose other threats against non-participation (Jehiel and Moldovanu, 2006, p. 108).

A caveat to such participation-enforcement strategies is that they presume substantial bargaining power for the designer. Moreover, the concept has a flavor of redundancy: Would the corresponding property rights and liability rules not require agents' approval in advance, potentially ruling out participation in the overall mechanism by backward induction? In contrast, our participation-stimulation approach works for any specification of property rights and liability rules. It thereby accounts for both the designer's limited bargaining power and the agents' free will.

6.2 Practical Implementation

An important question regarding possibility results concerns their practical relevance; whether they show how efficient design is attainable in practice, or whether they serve to point out practical difficulties in the manner of a "reductio ad absurdum critique." We shall therefore discuss the possibilities for and limitations to practically implementing our participation-stimulating transfers.

We argue that, in an abstract way, participation stimulation can be seen in practice. Observe that our PS transfers only require agents to report a one-dimensional sufficient statistic for their social type. Thus, reporting social types translates into agents selecting

one-dimensional strategies in a strategic game. It is this strategic game that renders participation attractive. In what follows, we illustrate how participation may be stimulated through various game forms.

The idea is to exploit the agents' social preferences by having them choose among different levels of a one-dimensional strategic variable and thereby impose positive or negative externalities on each other. For this purpose, define for each agent $i \neq M$ a set of *dedicated supporters* $\mathcal{S}_i \subseteq \mathcal{I} \setminus \{i, M\}$ and denote by $\mathcal{S}_{-i} = \mathcal{I} \setminus (\{i, M\} \cup \mathcal{S}_i)$ the set of i's *dedicated opponents*. For instance, if we let $\mathcal{S}_i = \mathcal{I} \setminus \{i, M\}$ for all $i \neq M$, then our formalism shall capture a *public-good game* among $\mathcal{I} \setminus \{M\}$, whereas $\mathcal{S}_i = \emptyset$ for each $i \neq M$ shall capture a *competition* between the agents other than M. For given sets $(\mathcal{S}_i)_{i \neq M}$, participation stimulation can be implemented as follows:

Proposition 3 *Participation stimulation can be implemented with an indirect mechanism under which agents $i \neq M$ invest $x_i \geq 0$ to receive* net returns $\hat{s}_i((x_j)_{j \neq M}) = -x_i + c_i + 2\mu\sqrt{x_i} + 2\sum_{j \in \mathcal{S}_i} \sqrt{x_j} - 2\sum_{j \in \mathcal{S}_{-i}} \sqrt{x_j}$, *for appropriate constants μ and $(c_i)_{i \neq M}$, while $\hat{s}_M = -\sum_{i \neq M} \hat{s}_i$.*

Proof. See Appendix A.2. ■

In the game of Proposition 3, agents' investments may take the form of monetary investments, labor effort, or physical exertion. An agent's investment imposes a positive (negative) payoff externality on those other agents for whom she is a dedicated supporter (opponent). If $\mathcal{S}_i = \mathcal{I} \setminus \{i, M\}$ for each $i \neq M$, agents are involved in a situation of *team-performance pay*, effectively a game of private contributions to a public good for $\mathcal{I} \setminus \{M\}$. Conversely, letting $\mathcal{S}_i = \emptyset$ for each $i \neq M$ yields a contest-like situation with *relative-performance pay*. Participation stimulation thus becomes a principal-agent scenario with M taking the role of the principal. Mixtures of relative- and team-performance pay are feasible, too. For instance, the partition $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2 \cup \{M\}$ with $\mathcal{S}_i = \mathcal{I}_\ell \setminus \{i\}$ for all $i \in \mathcal{I}_\ell$ and $\ell \in \{1, 2\}$ leads to a team competition between teams $\mathcal{I}_1$ and $\mathcal{I}_2$. In all those cases, each $i \neq M$ has the dominant strategy to invest $x_i = (\mu + \delta_i^{\mathcal{S}})^2$, where

$$\delta_i^{\mathcal{S}} = \frac{\sum_{\ell \neq i:\, i \in \mathcal{S}_\ell} (\delta_{i\ell} - \delta_{iM}) \;-\; \sum_{\ell \neq i:\, i \in \mathcal{S}_{-\ell}} (\delta_{i\ell} - \delta_{iM})}{\delta_{ii} - \delta_{iM}}, \tag{12}$$

while letting $\mu = \max_{j \neq M, \delta \in \Delta} |\delta_j^{\mathcal{S}}|$ ensures that the mechanism is well-defined.

Agent i's investment is strictly increasing in $\delta_i^{\mathcal{S}}$, and it increases (decreases) in i's relative pro-sociality towards those agents for whom i is a dedicated supporter (opponent).

Hence, whether a dedicated supporter (opponent) turns out to be an actual supporter (opponent) depends on that agent's social preferences. Moreover, i's investment increases (decreases) in i's preference for M if there are more (less) agents for whom i is a dedicated opponent rather than supporter. The transfer that i interim-expects for herself is maximal if $\delta_i^{\mathcal{S}} = 0$ (see equation (14) in Appendix A.2). Money is thus redistributed ex interim to those agents who are (nearly) indifferent about any form of redistribution between three parties: those they are meant to support, those they are meant to oppose, and finally M. On the other hand, an agent who has strong concerns about the distributive effects for and between these three parties will obey (if $\delta_i^{\mathcal{S}} \gg 0$) or disobey (if $\delta_i^{\mathcal{S}} \ll 0$) her dedicated roles; this results in an interim-expected monetary loss, overcompensated emotionally.

We accompany Proposition 3 with a real-world example. Think of a community organizing a fundraiser in support of their elementary school (e.g., to fund a new basketball court). The hard-core allocation problem underlying this event is obviously one of public-good provision, and the mechanism to resolve it, if only second-best, is actually quite simple, realistically speaking: 'Once you're in, you have to give,' as a matter of social norm. Events of this sort are often complemented with some soft-core incentive device, like awarding the best-dressed guest. The major purpose of such an add-on contest is not to make guests dress well, but rather to suppress free-riding-at-the-doorstep by compensating participants for their monetary losses (the lost returns from free-riding) with the social utility they derive from engaging in the contest. Awarding the best-dressed guest provides participants with a platform to live out their propensities to compete, and it is this attraction that helps pull them over the doorstep.[9]

6.3 Model Limitations

From the other angle, though, we must scrutinize the assumptions that render participation stimulation possible. As is standard in mechanism-design theory, we assume that transfers may take arbitrary negative values. This presumes that agents are endowed to pay these transfers. As interpersonal transfers play an important role in our study beyond standard theory, it is worthwhile discussing the impact of budget constraints. As is evident from equation (5), an agent's payment (i.e., negative transfer) increases with that agent's altruism toward that special agent M who is designated for balancing the budget; and as $\delta_{iM} \to 1$ (all else fixed), agent i's payment would exceed all limits. Hence,

[9] A similar point is frequently made in conceptual research on how to organize fundraisers; see, e.g., Webber (2004) and Peloza and Hassay (2007).

introducing budget constraints would conflict with allowing for arbitrary social-type sets. This raises the question whether participation stimulation would still work for bounded transfers if we confined social-type sets appropriately. In fact, this is not the case, for any type-set confinement: As is obvious from our derivation of PS transfers in Section 5.2, and from Figure 2 in particular, individual payments are *likely* largest for the extreme social types; on the other hand, narrowing the support tends to decrease the variance (which is the minimum value that the interim-expected utility gain from participation stimulation can take), so PS transfers must be amplified even further through the factor $\alpha^\star$ in the proof of Theorem 1. Similar arguments hold for our various versions of PS transfers. So we must conclude that budget constraints limit the scope of participation stimulation (as we constructed it). A way to resolve this problem would be to meet budget constraints with constraints on agents' reservation utilities.

Our assumption that agents' social preferences extend to each others' transfers is critical to our main result, and it distinguishes ours from other papers on mechanism design with social preferences. It implies that agents care about the overall distributive effects of the mechanism, but it requires that agents learn all other agents' full private payoffs ex post. As outlined by Sobel (2005, pp. 400), the domain of social preferences is critical in models with interdependent preferences. Yet, the literature provides little guidance in this regard. Very recent experimental studies suggest that some subjects sometimes apply their social preferences narrowly, but they conclude that more work is needed to explore the extent and drivers of narrow distributive concerns (see Ellis and Freedman, 2023; Exley and Kessler, 2023).

Our assumption that private payoffs are quasi-linear while utility is linear in private payoffs is crucial for both preference separation and participation stimulation. It implies that agents are *risk-neutral* with respect to transfers. We know from Section 6.1 that participation stimulation relies on agents accepting a *gamble* over the composition of social types at play. Plausibly, then, *risk-averse* agents are less susceptible to participation stimulation. We contend that, when relaxing these assumptions, participation stimulation, now generally understood as complementing a mechanism with an unrelated strategic game, may still prove helpful in attaining *individually rational second-best* implementation. We leave this for future work.

6.4 Relation to Mezzetti (2004)

Our bundling of two mechanisms resembles the approach of Mezzetti (2004); henceforth, Mezzetti. The key differences between his study and ours are the following.

In our model, agents' social preferences, and thus the allocational and informational externalities associated with them, extend to all agents' transfers. Mezzetti's agents can be other-regarding with respect to social alternatives but must disregard other agents' transfers; that is, they do not account for the *overall* distributive effects of a mechanism. As we will see, Mezzetti's mechanism is thus not incentive-compatible in our model.

While we consider a specific framework of *one*-dimensional allocational and informational externalities, Mezzetti considers a more general framework in which these externalities can be *multi*-dimensional. Jehiel and Moldovanu (2001) had shown that, with *multi*-dimensional externalities, there exists no mechanism that is both incentive-compatible and efficient, but they restricted attention to one-round-of-reporting mechanisms. Mezzetti shows that the conclusion changes when considering a two-stage mechanism: In the first round of reporting, each agent signals her preference type regarding a set of social alternatives; based on these reports, the designer ultimately chooses an alternative that maximizes aggregate utility. In the second round of reporting, each agent signals the payoff she realizes under this alternative, and interpersonal transfers are determined based on these reports. Specifically, the second-stage transfer scheme utilizes the principle of the VCG-mechanism (Vickrey, 1961; Clarke, 1971; Groves, 1973): Each agent is transferred the sum of all other agents' reported outcome-decision payoffs; as this transfer is independent of one's own report, each agent has the weakly dominant strategy to report her outcome-decision payoff truthfully. By backward induction, this mechanism makes each agent a residual claimant of the full surplus and thereby incentivizes truth-telling in the first reporting stage.

Having sketched Mezzetti's mechanism, we can rule out that it would be incentive-compatible in our model. It is appropriate to consider two versions of his mechanism. The first is a one-to-one adaption to our framework. In the first stage, agents report both their payoff types and social types; based on these reports, the designer chooses the alternative k that maximizes *aggregate utility* (which is a weighted sum of all agents' *private payoffs* under k). In the second stage, each agent reports the *utility* she derives under k; based on these reports, she receives a transfer that equals the sum of all the other agents' reported utility levels. This mechanism is clearly not incentive-compatible in the

second reporting stage: An agent's reported utility level affects every other's transfer, which she values according to her social type; she is indifferent only if the sum of her degrees of altruism toward the others equals zero and would otherwise under- or overstate her outcome-decision utility level. The second version shall account for our focusing on social alternatives that condition on payoff types. As we observed that our *terms of trade* implement the materially efficient alternative, it is natural to ask whether a version of Mezzetti's mechanism that merely operates on payoff types would achieve the same. But here, too, the second-stage transfer scheme is not incentive-compatible: Transferring to each agent the sum of the others' reported outcome-decision *payoffs* gives *almost all* social types the incentive to under- or overstate these payoffs.

Finally, Mezzetti's and our mechanism differ in the way they attract participation and allow the designer to extract the resulting surplus. (These issues are not discussed in Mezzetti, 2004, but in Mezzetti, 2003, 2007.) When applied to settings in which the surplus from the mechanism is strictly positive for any realization of types, Mezzetti's mechanism can be rendered individually rational through appropriate lump-sum transfers (see Mezzetti, 2003, Proposition 3). Deploying *side bets* that leverage the correlation in agents' second-stage payoff reports (similar to those in Crémer and McLean, 1985, 1988), the designer may extract *nearly* the full surplus (see Mezzetti, 2007, Theorem 4). In our model, by contrast, participation can be attracted whenever social-type distributions have strictly positive variance while transfers may take arbitrary negative values. By leveraging the differences in agents' other-regarding concerns, the designer can generate a *money pump* and extract far more than the gains from trade.

A Appendix

A.1 Proof of Lemma 1

Having required weak budget balance, Pareto efficiency implies strict budget balance: Suppose $\sum_{i\in\mathcal{I}} t_i = -\epsilon$ for some $\epsilon > 0$. Then a Pareto improvement can be achieved through transfers $(t_i + \epsilon/n)_{i\in\mathcal{I}}$, since $\sum_{j\in\mathcal{I}} \delta_{ij} > 0$ by assumption.

In the following, let $|\delta_{ij}| < 1/(2n-3)$ for all i and all $j \neq i$. Suppose that, for any fixed transfers $(t_i)_{i\in\mathcal{I}}$, there exists a social alternative $k^\circ(\theta)$ that Pareto-dominates the alternative $k^\star(\theta) \in \arg\max_{k\in K} \sum_{i\in\mathcal{I}} \pi_i(k \,|\, \theta_i)$ while $\sum_{i\in\mathcal{I}} \pi_i(k^\circ \,|\, \theta_i) < \sum_{i\in\mathcal{I}} \pi_i(k^\star \,|\, \theta_i)$.

Then there must exist agents i who make strict material losses when switching from $k^\star$ to k°; that is, $\pi_i(k^\circ \,|\, \theta_i) - \pi_i(k^\star \,|\, \theta_i) = -\epsilon_i < 0$. Be $i^\star$ one of the agents for whom this material loss is largest. Agent $i^\star$ is *not worse off* utility-wise under k° than under $k^\star$ if and only if she is 'emotionally' compensated through the distributive effects on the others: $\sum_{j\neq i^\star} \delta_{i^\star j}\big[\pi_j(k^\circ \,|\, \theta_j) - \pi_j(k^\star \,|\, \theta_j)\big] \geq \epsilon_{i^\star}$. We show that this is impossible.

First suppose $\delta_{i^\star j} \leq 0$ for all $j \neq i^\star$. Then $i^\star$ obtains the maximum 'emotional' compensation feasible if also each $j \neq i^\star$ realizes the maximum material loss of $-\epsilon_{i^\star}$ when switching from $k^\star$ to k°; that is, if $\pi_j(k^\circ \,|\, \theta_j) - \pi_j(k^\star \,|\, \theta_j) = -\epsilon_{i^\star} < 0$. But even then, $\sum_{j\neq i^\star} \delta_{i^\star j}\big[\pi_j(k^\circ \,|\, \theta_j) - \pi_j(k^\star \,|\, \theta_j)\big] = \sum_{j\neq i^\star} \delta_{i^\star j}(-\epsilon_{i^\star}) < \epsilon_{i^\star}$, since $0 \geq \delta_{i^\star j} > -1/(2n-3) \geq -1/(n-1)$.

Now suppose $\max_{j\neq i^\star} \delta_{i^\star j} > 0$, and let $j^\star \in \arg\max_{j\neq i^\star} \delta_{i^\star j}$ be the favorite agent of $i^\star$. Then $i^\star$ obtains the maximum 'emotional' compensation feasible if $j^\star$ realizes a maximum material gain when switching from $k^\star$ to k°, under the constraint that $\sum_{j\in\mathcal{I}} \pi_j(k^\circ \,|\, \theta_j) < \sum_{j\in\mathcal{I}} \pi_j(k^\star \,|\, \theta_j)$. This is the case if each $j \neq i^\star, j^\star$ also realizes the maximum material loss of $-\epsilon_{i^\star}$ while aggregate losses, amounting to $(n-1)\epsilon_{i^\star}$, serve as a subsidy to agent $j^\star$; that is, if $\pi_j(k^\circ \,|\, \theta_j) - \pi_j(k^\star \,|\, \theta_j) = -\epsilon_{i^\star} < 0$ for all $j \neq i^\star, j^\star$ while $\pi_{j^\star}(k^\circ \,|\, \theta_{j^\star}) - \pi_{j^\star}(k^\star \,|\, \theta_{j^\star}) = (n-1)\,\epsilon_{i^\star}$. But even then, $\sum_{j\neq i^\star} \delta_{i^\star j}\big[\pi_j(k^\circ \,|\, \theta_j) - \pi_j(k^\star \,|\, \theta_j)\big] = \sum_{j\neq i^\star, j^\star} \delta_{i^\star j}(-\epsilon_{i^\star}) + \delta_{i^\star j^\star}(n-1)\,\epsilon_{i^\star} < \epsilon_{i^\star}(n-2)/(2n-3) + \epsilon_{i^\star}(n-1)/(2n-3) = \epsilon_{i^\star}$, since $|\delta_{i^\star j}| < 1/(2n-3)$ for all $j \neq i^\star$.

Hence, agent $i^\star$ is worse off under k° than under $k^\star$, implying $k^\star$ is Pareto-efficient.

It remains to show that, for any fixed social alternative k, no ex-post budget-balanced transfer scheme ex-post Pareto-dominates another if $|\delta_{ij}| < 1/(2n-3)$ for all i and all $j \neq i$: Suppose the opposite is true, and transfers $(t_i^\circ)_{i\in\mathcal{I}}$ ex-post Pareto-dominate transfers $(t_i^\star)_{i\in\mathcal{I}}$, while both are ex-post budget-balanced. Then there is an agent $i^\star$ who

suffers the maximum monetary loss when switching from $(t_i^\star)_{i\in\mathcal{I}}$ to $(t_i^\circ)_{i\in\mathcal{I}}$. From here, the proof proceeds exactly as above. ■

A.2 Proof of Proposition 3

For any given sets $(\mathcal{S}_i)_{i\neq M}$, we obtain participation-stimulating transfers by modifying the transfer scheme (2)–(5) as follows:

$$
\begin{aligned}
(13)\quad s_M^\star(\delta) &= -\sum_{j\neq M} s_j^\star(\delta),\\
(14)\quad s_j^\star(\delta) &= -C + g_j(\delta_j^{\mathcal{S}}) - \delta_j^{\mathcal{S}} g_j'(\delta_j^{\mathcal{S}}) + \sum_{\ell\neq j,M} (-1)^{\mathbb{1}_{\mathcal{S}_{-j}}(\ell)}\cdot g_\ell'(\delta_\ell^{\mathcal{S}}), \quad \text{for } j\neq M,\\
(15)\quad g_j(\delta_j^{\mathcal{S}}) &= \mathrm{Var}_{\delta_j}[\delta_j^{\mathcal{S}}] + (\delta_j^{\mathcal{S}} - \mathbb{E}_{\delta_j}[\delta_j^{\mathcal{S}}])^2,\\
(16)\quad \delta_j^{\mathcal{S}} &= \frac{\sum_{\ell\neq j,M}(-1)^{\mathbb{1}_{\mathcal{S}_{-\ell}}(j)}\cdot(\delta_{j\ell}-\delta_{jM})}{\delta_{jj}-\delta_{jM}},
\end{aligned}
$$

for some constant $C>0$, where $\mathbb{1}_A(x)$ is the indicator function (i.e., $\mathbb{1}_A(x)=1$ if $x\in A$ and $\mathbb{1}_A(x)=0$ if $x\notin A$) . To see this, we follow the proof of Lemma 4:

Strategy proofness: Under $s^\star$, each agent $j\neq M$ reports a social type $\hat{\delta}_j$, which is strategically equivalent to reporting some signal $\hat{\delta}_j^{\mathcal{S}}\in\mathbb{R}$. Her ex-post utility is given by

$$
\begin{aligned}
\sum_{\ell\neq M}(\delta_{j\ell}-\delta_{jM})s_\ell^\star(\hat{\delta}) &= (\delta_{jj}-\delta_{jM})\left[g_j(\hat{\delta}_j^{\mathcal{S}}) - \hat{\delta}_j^{\mathcal{S}} g_j'(\hat{\delta}_j^{\mathcal{S}}) + \sum_{\ell\neq j,M}(-1)^{\mathbb{1}_{\mathcal{S}_{-j}}(\ell)}\cdot g_\ell'(\hat{\delta}_\ell^{\mathcal{S}})\right]\\
&\quad + \sum_{\ell\neq j,M}(\delta_{j\ell}-\delta_{jM})\left[g_\ell(\hat{\delta}_\ell^{\mathcal{S}}) - \hat{\delta}_\ell^{\mathcal{S}} g_\ell'(\hat{\delta}_\ell^{\mathcal{S}}) + \sum_{\ell'\neq\ell,j,M}(-1)^{\mathbb{1}_{\mathcal{S}_{-\ell}}(\ell')}\cdot g_{\ell'}'(\hat{\delta}_{\ell'}^{\mathcal{S}})\right]\\
&\quad + \left[\sum_{\ell\neq j,M}(-1)^{\mathbb{1}_{\mathcal{S}_{-\ell}}(j)}\cdot(\delta_{j\ell}-\delta_{jM})\right] g_j'(\hat{\delta}_j^{\mathcal{S}}) - C\sum_{\ell\neq M}(\delta_{j\ell}-\delta_{jM}).
\end{aligned}
$$

Hence, when substituting for $\delta_j^{\mathcal{S}} = \sum_{\ell\neq j,M}(-1)^{\mathbb{1}_{\mathcal{S}_{-\ell}}(j)}\cdot(\delta_{j\ell}-\delta_{jM})/(\delta_{jj}-\delta_{jM})$, agent j maximizes $g_j(\hat{\delta}_j^{\mathcal{S}}) + (\delta_j^{\mathcal{S}}-\hat{\delta}_j^{\mathcal{S}})g_j'(\hat{\delta}_j^{\mathcal{S}})$ over the choice of $\hat{\delta}_j^{\mathcal{S}}$. As $g_j''>0$, each $j\neq M$ has the strictly dominant strategy to report $\hat{\delta}_j^{\mathcal{S}}=\delta_j^{\mathcal{S}}$. As agent M is not involved strategically, she has the weakly dominant strategy to report her true social type δ_M.

Ex-post budget balance: Immediate from equation (13).

Interim-expected Pareto improvement: When substituting for $\delta_j^{\mathcal{S}}$ and $\mathbb{E}_{\delta_\ell}\left[g'_\ell(\delta_\ell^{\mathcal{S}})\right] = 0 = \mathbb{E}_{\delta_\ell}\left[g_\ell(\delta_\ell^{\mathcal{S}}) - \delta_\ell^{\mathcal{S}} g'_\ell(\delta_\ell^{\mathcal{S}})\right]$, due to Lemma 3, then j's interim-expected utility from $s^\star$ is

$$\begin{aligned}\sum_{\ell\neq M}(\delta_{j\ell}-\delta_{jM})\,\mathbb{E}_{\delta_{-j}}[s_\ell^\star(\delta)] &= (\delta_{jj}-\delta_{jM})g_j(\delta_j^{\mathcal{S}}) - C\sum_{\ell\neq M}(\delta_{j\ell}-\delta_{jM})\\ &= (\delta_{jj}-\delta_{jM})g_j(\delta_j^{\mathcal{S}}) - C(\delta_{jj}-\delta_{jM}) - C\sum_{\ell\neq j,M}(\delta_{j\ell}-\delta_{jM})\\ &= (\delta_{jj}-\delta_{jM})\left[g_j(\delta_j^{\mathcal{S}}) - C(1+\delta_j^\star)\right],\end{aligned}$$

for $\delta_i^\star = \sum_{\ell\neq i,M}(\delta_{i\ell}-\delta_{iM})/(\delta_{ii}-\delta_{iM})$. Recall that $\delta_{jj}=1>\delta_{jM}$ and $g_j(\delta_j^{\mathcal{S}}) \geq \mathrm{Var}_{\delta_j}[\delta_j^{\mathcal{S}}] > 0$ and that $\delta_j^\star < n-2$, since $\delta_{jj}-\delta_{jM} > \delta_{j\ell}-\delta_{jM}$ for all $\ell \neq j, M$. We thus obtain that each $j \neq M$ derives positive interim-expected utility from unanimous participation if we let $C \leq \min_{j\neq M}\mathrm{Var}_{\delta_j}[\delta_j^{\mathcal{S}}]/(n-1)$. Finally, due to Lemma 3 again, also M's interim-expected utility is positive if all agents participate: $\sum_{i\in\mathcal{I}}\delta_{Mi}\,\mathbb{E}_{\delta_{-M}}[s_i^\star(\delta)] = \sum_{j\neq M}(\delta_{Mj}-1)\,\mathbb{E}_\delta[s_j^\star(\delta)] = C\sum_{j\neq M}(1-\delta_{Mj}) > 0$.

To implement $s^\star:\Delta\to\mathbb{R}$ with an indirect mechanism $\hat{s}:[0,\infty)^n\to\mathbb{R}$, we observe that

$$\begin{aligned}s_j^\star(\delta) &= 2\sum_{\ell\in\mathcal{S}_j}\left(\delta_\ell^{\mathcal{S}} - \mathbb{E}_{\delta_\ell}[\delta_\ell^{\mathcal{S}}]\right) - 2\sum_{\ell\in\mathcal{S}_{-j}}\left(\delta_\ell^{\mathcal{S}} - \mathbb{E}_{\delta_\ell}[\delta_\ell^{\mathcal{S}}]\right) + \mathbb{E}_{\delta_j}[(\delta_j^{\mathcal{S}})^2] - (\delta_j^{\mathcal{S}})^2 - C\\ &= 2\hat{c}_j - (\mu+\delta_j^{\mathcal{S}})^2 + 2\mu(\mu+\delta_j^{\mathcal{S}}) + 2\sum_{\ell\in\mathcal{S}_j}(\mu+\delta_\ell^{\mathcal{S}}) - 2\sum_{\ell\in\mathcal{S}_{-j}}(\mu+\delta_\ell^{\mathcal{S}})\\ &= c_j - x_j + 2\mu\sqrt{x_j} + 2\sum_{\ell\in\mathcal{S}_j}\sqrt{x_\ell} - 2\sum_{\ell\in\mathcal{S}_{-j}}\sqrt{x_\ell}\\ &= \hat{s}_j((x_\ell)_{\ell\neq M})\end{aligned}$$

when letting $\sqrt{x_\ell} = \mu + \delta_\ell^{\mathcal{S}}$ for $\mu = \max_{j\neq M,\delta\in\Delta}|\delta_j^{\mathcal{S}}|$ while letting $c_j = 2\hat{c}_j$ for

$$\hat{c}_j = \mu\cdot|\mathcal{S}_{-j}| - \mu\cdot|\mathcal{S}_j| - \frac{1}{2}\mu^2 + \frac{1}{2}\mathbb{E}_{\delta_j}[(\delta_j^{\mathcal{S}})^2] - \sum_{\ell\in\mathcal{S}_j}\mathbb{E}_{\delta_\ell}[\delta_\ell^{\mathcal{S}}] + \sum_{\ell\in\mathcal{S}_{-j}}\mathbb{E}_{\delta_\ell}[\delta_\ell^{\mathcal{S}}] - \frac{1}{2}C.$$

Since agent $j\neq M$ has the strictly dominant strategy to report $\delta_j^{\mathcal{S}}$ under $s^\star$, and since $dx_j/d\delta_j^{\mathcal{S}} > 0$, she also has the dominant strategy to invest $x_j = (\mu+\delta_j^{\mathcal{S}})^2$ under $\hat{s}$. ■

References

Andreoni, James and John Miller. 2002. "Giving according to GARP: An experimental test of the consistency of preferences for altruism." *Econometrica* 70 (2):737–753.

Antler, Yair. 2015. "Two-sided matching with endogenous preferences." *American Economic Journal: Microeconomics* 7 (3):241–258.

———. 2023. "Multilevel marketing: Pyramid-shaped schemes or exploitative scams?" *Theoretical Economics* 18 (2):633–668.

Arrow, Kenneth. 1979. "The property rights doctrine and demand revelation under incomplete information." In *Economics and Human Welfare*, edited by M. J. Boskin. New York, NY: Academic Press.

Bergemann, Dirk and Stephen Morris. 2005. "Robust mechanism design." *Econometrica* 73 (6):1771–1813.

Bierbrauer, Felix and Nick Netzer. 2016. "Mechanism design and intentions." *Journal of Economic Theory* 163:557–603.

Border, Kim and Uzi Segal. 1994. "Dutch books and conditional probability." *Economic Journal* 104 (422):71–75.

Bruhin, Adrian, Ernst Fehr, and Daniel Schunk. 2019. "The many faces of human sociality: Uncovering the distribution and stability of social preferences." *Journal of the European Economic Association* 17 (4):1025–1069.

Charness, Gary and Matthew Rabin. 2002. "Understanding social preferences with simple tests." *Quarterly Journal of Economics* 117 (3):817–869.

Chen, Jing, Silvio Micali, and Rafael Pass. 2015. "Tight revenue bounds with possibilistic beliefs and level-k rationality." *Econometrica* 83 (4):1619–1639.

Clarke, Edward. 1971. "Multipart pricing of public goods." *Public Choice* 11 (1):17–33.

Crémer, Jacques and Richard McLean. 1985. "Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent." *Econometrica* 53 (2):345–361.

———. 1988. "Full extraction of surplus in Bayesian and dominant strategy auctions." *Econometrica* 56 (6):1247–1257.

d'Aspremont, Claude and Louis-André Gérard-Varet. 1979. "Incentives and incomplete information." *Journal of Public Economics* 11 (1):25–45.

Desiraju, Ramarao and David Sappington. 2007. "Equity and adverse selection." *Journal of Economics and Management Strategy* 16 (2):285–318.

Eliaz, Kfir and Ran Spiegler. 2007. "A mechanism design approach to speculative trade." *Econometrica* 75 (3):875–884.

———. 2009. "Bargaining over bets." *Games and Economic Behavior* 66 (1):78–97.

Ellis, Andrew and David Freedman. 2023. "Revealing choice bracketing." arXiv:2006.14869.

Exley, Christine and Judd Kessler. 2023. "Equity concerns are narrowly framed." *American Economic Journal: Microeconomics* forthcoming.

Fehr, Ernst, Karla Hoff, and Mayuresh Kshetramade. 2008. "Spite and development." *American Economic Review: Papers & Proceedings* 98 (2):494–99.

Groves, Theodore. 1973. "Incentives in teams." *Econometrica* 41 (4):617–631.

Jehiel, Philippe, Moritz Meyer-ter Vehn, Benny Moldovanu, and William Zame. 2006. "The limits of ex post implementation." *Econometrica* 74 (3):585–610.

Jehiel, Philippe and Benny Moldovanu. 2001. "Efficient design with interdependent valuations." *Econometrica* 69 (5):1237–1259.

———. 2006. "Allocative and informational externalities in auctions and related mechanisms." In *Advances in Economics and Econometrics. Theory and Applications, Ninth World Congress*, vol. 1, chap. 3. Cambridge University Press, 102–135.

Kosenok, Grigory and Sergei Severinov. 2008. "Individually rational, budget-balanced mechanisms and allocation of surplus." *Journal of Economic Theory* 140 (1):126–161.

Kozlovskaya, Maria and Antonio Nicoló. 2019. "Public good provision mechanisms and reciprocity." *Journal of Economic Behavior & Organization* 167 (3):235–244.

Kucuksenel, Serkan. 2012. "Behavioral mechanism design." *Journal of Public Economic Theory* 14 (5):767–789.

Mailath, George and Andrew Postlewaite. 1990. "Asymmetric information bargaining problems with many agents." *Review of Economic Studies* 57 (3):351–367.

Mas-Colell, Andreu, Michael Whinston, and Jerry Green. 1995. *Microeconomic Theory*. Oxford University Press.

McAfee, Preston and Philip Reny. 1992. "Correlated information and mechanism design." *Econometrica* 60 (2):395–421.

McLean, Richard and Andrew Postlewaite. 2004. "Informational size and efficient auctions." *Review of Economic Studies* 71 (3):809–827.

Mezzetti, Claudio. 2003. "Auction design with interdependent valuations: The generalized revelation principle, efficiency, full surplus extraction and information acquisition." mimeo. Available under: https://ideas.repec.org/p/fem/femwpa/2003.21.html.

———. 2004. "Mechanism design with interdependent valuations: Efficiency." *Econometrica* 72 (5):1617–1626.

———. 2007. "Mechanism design with interdependent valuations: Surplus extraction." *Economic Theory* 31 (3):473–488.

Myerson, Roger. 1979. "Incentive compatibility and the bargaining problem." *Econometrica* 47 (1):61–73.

Myerson, Roger and Mark Satterthwaite. 1983. "Efficient mechanisms for bilateral trading." *Journal of Economic Theory* 29 (2):265–281.

Nau, Robert. 1992. "Joint coherence in games of incomplete information." *Management Science* 38 (3):374–387.

Peloza, John and Derek Hassay. 2007. "A typology of charity support behaviors: Toward a holistic view of helping." *Journal of Nonprofit & Public Sector Marketing* 17 (1-2):135–151.

Prediger, Sebastian, Björn Vollan, and Benedikt Herrmann. 2014. "Resource scarcity and antisocial behavior." *Journal of Public Economics* 119:1–9.

Rubinstein, Ariel and Ran Spiegler. 2008. "Money pumps in the market." *Journal of the European Economic Association* 6 (1):237–253.

Saijo, Tatsuyoshi and Hideki Nakamura. 1995. "The spite dilemma in voluntary contribution mechanism experiments." *Journal of Conflict Resolution* 39 (3):535–560.

Segal, Ilya and Michael Whinston. 2016. "Property rights and the efficiency of bargaining." *Journal of the European Economic Association* 14 (6):1287–1328.

Sobel, Joel. 2005. "Interdependent preferences and reciprocity." *Journal of Economic Literature* 43 (2):392–436.

Tang, Pingzhong and Tuomas Sandholm. 2012. "Optimal auctions for spiteful bidders." In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. 1457–1463.

Vickrey, William. 1961. "Counterspeculation, auctions, and competitive sealed tenders." *Journal of Finance* 16 (1):8–37.

Webber, Daniel. 2004. "Understanding charity fundraising events." *International Journal of Nonprofit and Voluntary Sector Marketing* 9 (2):122–134.

Werner, Jan. 2022. "Speculative trade under ambiguity." *Journal of Economic Theory* 199:105200.

Williams, Steven. 1999. "A characterization of efficient, Bayesian incentive compatible mechanisms." *Economic Theory* 14 (1):155–180.

Yaari, Menahem. 1998. "On the Role of 'Dutch Books' in the Theory of Choice Under Risk (1985)." In *Frontiers of Research in Economic Theory. The Nancy L. Schwartz Memorial Lectures*, edited by Donald P. Jacobs, Ehud Kalai, Morton I. Kamien, and Nancy L. Schwartz, chap. 3. Cambridge University Press, 33–46.

Zik, Boaz. 2021. "Ex-post implementation with social preferences." *Social Choice and Welfare* 56 (3):467–485.