ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Daske, Thomas

Working Paper

The Incentive Costs of Welfare Judgments

Suggested Citation: Daske, Thomas (2021) : The Incentive Costs of Welfare Judgments, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/230318

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

The Incentive Costs of Welfare Judgments

Thomas Daske*

February 11, 2021†

This paper draws an incentive-theoretical perspective on the concept of social welfare. In a simple mechanism-design framework, agents' interpersonal preferences and private payoffs are all subject to asymmetric information. Under reasonable normative assumptions, the following result is established: A policy can be implemented with a budget-balanced mechanism if and only if it is consistent with *materialistic utilitarianism*, which seeks to maximize aggregate material wealth, not utility. Any other policy, to be implementable, must violate budget balance and therefore comes at incentive costs. The corresponding mechanism is virtually unique, which allows for conclusions upon distributive and procedural justice.

JEL classification: C78; D60; D82

Keywords: Mechanism design; social welfare; distributive justice; procedural justice; utilitarianism; dictatorship

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

For their helpful comments and critical remarks, I want to thank Christian Feilcke, Julian Hackinger, Bård Harstad, Michael Kurschilgen, Christoph March, David Miller, Marco Sahm, Johannes Schneider, and Robert von Weizsäcker. I am also grateful for helpful comments by participants of the Asian Meeting of the Econometric Society in Hong Kong, the European Meeting on Game Theory in Paris, the Annual Congress of the Association for Public Economic Theory in Paris, the Annual Congress of the International Institute of Public Finance in Tokyo, the European Winter Meeting of the Econometric Society in Barcelona, the Conference on the Political Economy of Democracy and Dictatorship in Münster, the Meeting of the European Public Choice Society in Rome, les Journées Gérard-Varet à Aix-en-Provence, the European Meeting of the Econometric Society in Cologne, and a seminar at the University of Bamberg.

†An earlier version circulated under the title "Externality Assessments, Welfare Judgments, and Mechanism Design" and is available under http://hdl.handle.net/10419/172494.

1 Introduction

Imagine a judge who seeks to balance interests in a divorce battle. In the process, the judge receives a very valuable, or perhaps not at all valuable, piece of information: *She* despises *him*, but *he* does still love *her*. Her contempt is entirely unrelated to the couple's economic standing; she rather cannot stand his manners any longer and feels entitled to see him suffer for that waste of her time. How should our judge process this piece of information?

Following Harsanyi (1977), one might argue that the judge must ignore *her* feelings:

> "Some preferences ... must be altogether excluded from our social-utility function. In particular we must exclude all clearly antisocial preferences as sadism, envy, resentment, and malice. ... 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."

Following Blanchet and Fleurbaey (2006), one might argue that the judge must also ignore *his* feelings; otherwise, *she* would benefit from *his* greater willingness to concede (which would be the effective outcome if the judge was guided by, for instance, Bentham utilitarianism).

The shortcoming of these arguments, from an economic perspective, is that they are entirely moral. One could also argue that *he* should be privileged for his greater willingness to concede; or, that *she* should be disadvantaged for seeking awful revenge; or, to have the Amazons their way, that *he* must go to hell!—All these welfare judgments seem equally justifiable; they all seem consistent with the Pareto principle.

What is there to lose when abiding, or not abiding, by one of these judgments? Economically speaking, what are the *costs* of abiding by one welfare judgment and not by another (assuming there is at least one good that is valuable to all of the conflicting parties, in entities of which these costs could be measured)?

Mechanism-design theory, the normative counterpart of positive game theory, can be an ideal device for studying the matters of *distributive* and *procedural justice* on positive, incentive-theoretical grounds. What it takes is a plausible framework that allows for results in between arbitrariness and impossibility. This paper analyses such a framework.

In a simple quasi-linear environment, agents are privately informed about their preferences for consumption and their interpersonal, more or less altruistic or spiteful, preferences. (It can well be common knowledge who likes or dislikes whom; what is supposed to be private information is the *intensity* of these interpersonal feelings.) That is, next to their allocative preferences, agents are privately informed about their distributive preferences, which specify how they internalize the overall distributive effects of a mechanism.

The idea is to endogenize the social goal and search for all those policies (whether they prescribe the provision level of a public good or how to divide a given resource) that can be implemented with an ex-post budget-balanced mechanism.[1]

While the incentive-compatibility constraint is beyond dispute, it should be stressed that budget (im-)balance constitutes a reasonable qualitative measure for the *incentive costs* of the welfare judgment inherent to a policy: Budget (im-)balance is conceptually attractive because it is measured by tangible entities, not utility, and is thus objective in this sense. It is a neutral measure in that it is independent of conceptions of distributive and procedural justice. Moreover, ex-post budget balance is in the interest of the group of agents: if the mechanism does not run a deficit, then the amount of distributable material wealth is not diminished; and if the mechanism need not be subsidized from the outside, then group autonomy is preserved.

The analysis focuses on policies that can be implemented irrespective of the distribution of agents' preference types. Such policies will be referred to as being *definitely* implementable.[2]

The main result, Theorem 1, states that a non-constant[3] policy is *definitely* implementable with an *ex-post budget-balanced* mechanism if and only if it is consistent with *materialistic utilitarianism*, seeking to maximize aggregate material wealth, not utility; neither may such policy account for the agents' interpersonal concerns, nor may it reflect a social planner's conception of distributive justice. Any other policy, to be *definitely* implementable, must violate budget balance and therefore comes at incentive costs. Moreover, regarding policy choice, materialistic utilitarianism is not indifferent about its distributive effects; its zero-cost implementation requires a virtually unique incentive scheme, a finding that allows for conclusions upon distributive and procedural justice.

[1] As the model is one of informational externalities, it must rely on Bayesian implementation; see the discussion in Section 2.2.

[2] This restriction will be substantiated on normative grounds in Section 2.3.

[3] A constant policy is one that selects the same social alternative irrespective of the agents' preferences.

Two other results clarify the roles of budget balance and agents' interpersonal concerns. Theorem 2 shows that nearly any policy is *definitely* implementable if budget balance is not imposed.[4] Likewise, by Theorem 3, nearly any policy is *definitely* implementable at zero incentive costs if the agents' social preferences are common knowledge.[5] This latter result shows that Theorem 1 does not per se rely on agents exhibiting social preferences, it rather relies on the asymmetry of information about these preferences.

Theorem 1 contributes to the untiring debate about the appropriate evaluation of overall economic outcomes.[6,7] Several prominent studies have argued in favor of (weighted) Benthamite utilitarianism (seeking to maximize a possibly weighted sum of individual *utilities*); these studies argue on purely normative grounds, disregarding incentive compatibility (e.g., Harsanyi, 1955; Arrow, 1973; Maskin, 1978). Others have argued, again normatively, that welfare judgments should be based on the resources or prospects that individuals are assigned, not their perceived utilities (e.g., Rawls, 1971; Sen, 1992); in this manner, Piacquadio (2017) has laid out a normative foundation of "opportunity-equivalent" utilitarianism. Theorem 1 supports these latter conceptions by showing that conflicts of interests must be *objectified* (reduced to consumption-wise well-being) to be resolvable at zero incentive costs; in particular, it rationalizes the moral convictions of Harsanyi (1977) and Blanchet and Fleurbaey (2006), quoted in the beginning, that agents' interpersonal concerns must be irrelevant, both formally and even strategically.

From a different angle, this study adds to the bargaining literature. Theorem 1 puts several prominent bargaining solutions (Nash, 1950; Kalai and Smorodinsky, 1975; Kalai, 1977) into incentive-theoretical perspective and proposes a different solution.

Another finding, Theorem 4, contributes to the literature on the economic performance of *dictatorship* (e.g., Olson, 1993; Wintrobe, 2000; Acemoglu and Robinson, 2006) by showing that dictatorial policies come almost always at incentive costs.

[4]This finding substantiates budget (im-)balance as a qualitative measure of the incentive costs of a welfare judgment. It shows that the uniqueness result of Theorem 1 does not rely on the incentive-compatibility constraint alone but on imposing budget balance besides.

[5]Theorem 3 holds in particular under the conventional assumption of *commonly known* selfish agents.

[6]The standard, normative approach to distributive justice takes an *ex-ante* perspective on the socially desirable allocation of resources among risk-*averse* agents whose preferences are common knowledge (e.g., Harsanyi, 1955; Rawls, 1971; Arrow, 1973; Fleurbaey, 2010; Grant et al., 2010; Eden, 2020). The present result is based on strategic decision-making at the *interim* stage, with privately known individual preferences and risk-*neutral* agents. While risk neutrality is restrictive, it is interesting that a clear-cut aggregation result can be obtained without assuming decreasing marginal utility.

[7]Consistent with the mechanism-design literature on interdependent preferences, the here conceived model assumes that agents' utilities are cardinal and interpersonally comparable. For a general discussion of the informational bases of different approaches to welfare measurement, see Sen (1974).

While this paper considers a simple framework, its main insights might carry over to second-best implementation in more complex model economies.[8] Anyhow, inherent in the debate about moral principles is a notion of proclaimed universality of these principles. If there are moral principles that can be deemed universal, those can be uncovered by considering a simple framework. If, however, these very principles cannot be retrieved in more complex environments, then universality is merely impossible from an incentive-theoretical point of view, which would support the conviction of Yaari and Bar-Hillel (1984) that "[s]weeping solutions and world-embracing theories are not likely to be adequate for dealing with the intricacies inherent in the problem of How to Distribute." This paper takes a step in uncovering the incentive-theoretical possibility of moral principles that are, in fact, universal.

The paper proceeds as follows. Section 2 outlines the model framework (for bilateral bargaining problems). Section 3 establishes the main result. Section 4 scrutinizes favoritism and dictatorship. Section 5 concludes with an interpretation in terms of distributive and procedural justice. (Appendix A contains omitted proofs. A supplement establishes the central results for the n-agents case.)

2 The Analytical Framework

2.1 Allocations, Utility, and Information

There is a continuum $K \subset \mathbb{R}$ of social alternatives, bounded or unbounded, and there are two agents, indexed by $i \in \{1, 2\}$. (The n-agents case is discussed in the supplement to this paper.) From alternative $k \in K$ and a monetary transfer $t_i \in \mathbb{R}$, agent i gains a *private payoff* $\Pi_i(k, t_i \,|\, \theta_i) = \theta_i v_i(k) + w_i(k) + t_i$, where the functions $v_i : K \to (0, \infty)$ and $w_i : K \to \mathbb{R}$ are twice continuously differentiable; furthermore, $dv_i/dk \neq 0$. Agent i's *payoff type* θ_i belongs to a closed (proper) interval $\Theta_i = [\theta_i^{\min}, \theta_i^{\max}]$.[9]

[8] The here conceived model is one of *one*-dimensional informational externalities. As shown by Jehiel and Moldovanu (2001), higher-dimensional informational externalities render first-best efficient implementation (seeking to maximize aggregate material wealth) impossible, irrespective of further constraints such as budget balance.

[9] An example is a bargaining problem in which agents must come to an agreement upon the division of a given resource; e.g., $K = [0, 1]$, $v_1 = \sqrt{k}$, $v_2 = \sqrt{1 - k}$, $w_i = 0$. Another example is the provision of a public good the costs of which, before transfers, are shared equally among agents; e.g., $K = [0, 1]$, $v_i = k$, $w_i = -k^2/2$, for production costs k^2 at provision level k.

Agents exhibit interpersonal preferences in the form of altruism or spite:[10] From the allocation of payoffs, agent i derives ex-post *utility*

$$u_i(k, t_i, t_{-i}, \theta_{-i} \,|\, \theta_i, \delta_i) \;\; = \;\; \Pi_i(k, t_i \,|\, \theta_i) + \delta_i \cdot \Pi_{-i}(k, t_{-i} \,|\, \theta_{-i}),$$

where the degree δ_i of i's altruism or spite towards $-i$ belongs to $\Delta_i = [\delta_i^{\min}, \delta_i^{\max}] \subset (-1, 1)$. Refer to δ_i as i's *social type*. Notice that $(-1, 1)$ is the maximum range of interpersonal altruism, or spite, for which agents care about overall material efficiency while still being selfish to the extent that each prefers a dollar to be her own rather than having it given to the other.

Refer to the pair (θ_i, δ_i) as i's *type*, and denote payoff types by $\theta = (\theta_1, \theta_2)$ and social types by $\delta = (\delta_1, \delta_2)$. For Cartesian products $\Theta = \Theta_1 \times \Theta_2$ and $\Delta = \Delta_1 \times \Delta_2$, denote the *type space* by $\Theta \times \Delta$. For convenience, define also $\pi_i(k \,|\, \theta_i) = \theta_i v_i(k) + w_i(k)$ and $u_i(k, \theta_{-i} \,|\, \theta_i, \delta_i) = \pi_i(k \,|\, \theta_i) + \delta_i \pi_{-i}(k \,|\, \theta_{-i})$.

Each agent is privately informed about her payoff type and social type, which realize independently according to continuous densities. At the interpersonal level, agents' types are independent, too. In particular, the variance of each δ_i, while assumed strictly positive, is allowed to be arbitrarily small. Consequently, although social types are assumed independent, *reciprocal* interpersonal preferences can be captured by letting $\Delta_1 = \Delta_2$ and $\delta_i^{\min} \approx \delta_i^{\max}$.

The agents' (or, a social planner's) problem is to choose a social alternative k and transfers (t_1, t_2) such that the resulting allocation is 'desirable.' What this can or should mean is the very topic of this paper.

2.2 Incentives

As agents are free to adapt to the economic environment they are exposed to, a welfare judgment upon the choice of k, to become meaningful, must be *incentive-compatible* in that agents' behavior is consistent with the specific social goal.

A *direct revelation mechanism*, or *social contract*, involves the agents in a strategic game of incomplete information. In this game, agents are asked to report their types

[10] For evidence on altruism, 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).

truthfully.[11] Based on their reports, a social alternative will be implemented and transfers will be made. Specifically, a mechanism is defined by a *policy* $k : \Theta \times \Delta \to K$ and a *transfer scheme* $T = (t_1, t_2) : \Theta \times \Delta \to \mathbb{R}^2$. Throughout, attention is restricted to transfer schemes that are continuous on the social-type space Δ.

A mechanism $\langle k, T \rangle$ is Bayesian *incentive-compatible* if, under the rules of the resulting game, the strategy to truthfully reveal her type maximizes each agent's interim-expected utility, such that truthful revelation by both agents is mutually consistent:

$$(\theta_i, \delta_i) \quad \in \quad \arg\max_{(\hat{\theta}_i, \hat{\delta}_i)} \quad \mathbb{E}_{\theta_{-i}, \delta_{-i}} \Big[u_i \big(k, t_i, t_{-i}, \theta_{-i} \,\big|\, \theta_i, \delta_i \big) \Big],$$

on $\Theta_i \times \Delta_i$ for both i, where k, t_i, and t_{-i} are functions of $(\hat{\theta}_i, \hat{\delta}_i, \theta_{-i}, \delta_{-i})$. In this case, the policy k is said to be Bayesian *implementable*.[12] The mechanism is ex-post *budget-balanced* if $t_1(\theta, \delta) + t_2(\theta, \delta) = 0$ on $\Theta \times \Delta$.

2.3 Welfare Judgments

For the present purpose, there is no need for starting out from an axiomatization of a welfare function. Instead, two definitions specify the welfare judgments that are to be considered.[13]

Definition 1 (Generic Policies)

A policy $k^\star : \Theta \times \Delta \to K$ *is* generic *if it is partially differentiable and satisfies* $\partial k^\star / \partial \theta_i \neq 0$*, for all* (θ, δ) *and both agents* i*.*

[11] By the *revelation principle*, which applies to the present setup (Myerson, 1979), there is no loss of generality in identifying message sets, from which agents draw their reports, with agents' type sets.

[12] Bayesian implementation has been criticized upon assuming that type distributions are common knowledge. In order to cope with Wilson's (1987) call for avoiding such strong assumption, Bergemann and Morris (2005) have proposed *ex-post implementation* for model economies with interdependent utilities (in which dominant-strategy implementation is not feasible; Williams and Radner, 1988), requiring that truthful revelation of types constitutes a Nash equilibrium under the respective mechanism. However, as shown by Jehiel et al. (2006) and Zik (2020), ex-post implementation is not feasible in the presence of informational externalities.

[13] The results apply in particular to any welfare judgment of the form $k^\star = \arg\max_{k \in K} W$, for some arbitrary welfare function $W : (\pi_i)_{i \in \{1,2\}} \times \Theta \times \Delta \times K \to \mathbb{R}$, as long as the choice of W is consistent with the Definitions 1 and 2 below; agents' utilities, if made the primitive, would enter W in the form of $\pi_i + \delta_i \pi_{-i}$. Welfare may condition separately on agents' types and social alternatives k, allowing a social planner to pursue objectives next to satisfying the agents' needs, such as punishing spite, or steering agents towards a specific alternative k. Examples of admissable welfare functions are the materialistic-utilitarian ($W = \pi_1 + \pi_2$), the Bentham-utilitarian ($W = u_1 + u_2$), the Nash (1950) product, Atkinson's (1970) isoelastic welfare, and Arrow's (1973) CES welfare. A further admissable *policy* is the bargaining solution of Kalai and Smorodinsky (1975). All these functions may depend either on the agents' utilities or private payoffs, and they may or may not account for the distributive effects of interpersonal transfers.

The welfare judgment inherent to Definition 1 is the following: As each agent is first of all concerned with her material well-being, the choice between social alternatives should be responsive to changes in each agent's payoff type. (As it will turn out, this restriction is basically without loss of generality; see Corollary 1 in Section 3.) Partial differentiability can be regarded as a postulate of *local consistency*.

The second definition states that the implementability of a policy should not depend on what agents' preferences *could be* or *could have been*, but that a 'just' policy is implementable for whatever the agents' preferences *actually are*. The implementability of a policy should not depend on the statistical distribution of agents' types.

Definition 2 (Definitely Implementable Policies)
A policy $k^\star : \Theta \times \Delta \to K$ *is* definitely *implementable if it is implementable for arbitrary type distributions.*

This requirement can be interpreted as a variant of the often considered axiom of *independence of irrelevant alternatives*, which has been given different meanings in different contexts. In the present context, it means that social choice is independent of those individual characteristics that, in the process of preference revelation, have ultimately been eliminated and can thus be excluded from the type space $\Theta \times \Delta$ when looking at it from an ex-post perspective. The welfare comparison of social alternatives should only be concerned with the world's actual state, irrespective of its likelihood. Likewise, changes to the type distribution should not change a spectator's moral judgment upon the conflict of interests between those very agents she ultimately observes.[14]

Another way to argue is in terms of *procedural justice*: The parties to a conflict of interests should be judged in their own rights, not in their relative standing to the rest of society.

Notice that definite implementability does not require an incentive-compatible mechanism as a whole to be independent of type distributions, which is impossible in the presence of informational externalities.[15] The concept rather draws a normative distinction between means (the transfer scheme) and ends (the policy).

[14] An established, incentive-theoretically founded welfare judgment that does condition on type distributions is the generalized Nash product of Harsanyi and Selten (1972).

[15] For this reason, the concept should not be confused with *robust*, or *ex-post implementation* in the manner of Bergemann and Morris (2005); see Jehiel et al. (2006) and Zik (2020).

Contrary to most welfare-economic studies, the present one does not impose the Pareto principle from the outset. However, next to being attractive on normative grounds, the Pareto principle is essential from an incentive-theoretical point of view, too. It implies that implementable policies are renegotiation-proof; otherwise, agents would have an incentive to seek other, Pareto-efficient, 'bargaining solutions' and, thereby, undermine the practical relevance of the welfare judgment upon the mechanism under consideration. Therefore, mechanisms that are desirable in terms of the above definitions must be evaluated in terms of Pareto efficiency.

Special attention will be paid to those policies that are implementable through *budget-balanced* transfers. Such mechanisms need not be subsidized from the outside, which guarantees group autonomy, and they do not run a deficit, such that the resolution of the agents' conflict of interests is not complicated by diminishing overall material wealth. If a policy is implementable, but not implementable through budget-balanced transfers, then this policy will be said to come at *incentive costs*.[16]

In this manner, a policy is considered desirable if and only if it is *generic* and *definitely* implementable at *zero incentive costs*.

3 Justifying Materialistic Utilitarianism

This Section proves the following Theorem and discusses it from various angles.

Theorem 1 *A generic policy* $k^\star : \Theta \times \Delta \to K$ *is* definitely *implementable through ex-post budget-balanced transfers if and only if it is consistent with* materialistic utilitarianism: $k^\star(\theta) = \arg\max_{k \in K} \pi_1(k \,|\, \theta_1) + \pi_2(k \,|\, \theta_2)$. *The corresponding transfers* $T^\star = (t_i^\star)_{i=1,2}$ *are necessarily of* AGV-type: *For reported types* $(\hat{\theta}, \hat{\delta}) \in \Theta \times \Delta$, *transfers are given by*

$$(1) \qquad t_i^\star(\hat{\theta}, \hat{\delta}) \;=\; \mathbb{E}_{\theta_{-i}}\big[\pi_{-i}(k^\star(\hat{\theta}_i, \theta_{-i}) \,|\, \theta_{-i})\big] - \mathbb{E}_{\theta_i}\big[\pi_i(k^\star(\theta_i, \hat{\theta}_{-i}) \,|\, \theta_i)\big] + s_i(\hat{\theta}, \hat{\delta}),$$

where $s_i : \Theta \times \Delta \to \mathbb{R}$ *must be chosen such that* $s_1 + s_2 = 0$ *on* $\Theta \times \Delta$, *while* $\mathbb{E}_{\theta_j, \delta_j}[s_i(\theta, \delta)]$ *is constant for all* $i, j \in \{1, 2\}$. *The resulting allocations are ex-post Pareto-efficient.*

[16] For the present purpose, there is no value in *quantifying* these costs; but if one seeks to do so, then a reasonable measure might be $C(k^\star) = \inf_{(t_1,t_2) \in \mathcal{T}(k^\star)} \mathbb{E}_{\theta,\delta}\big[\,|t_1 + t_2|\,\big]$, with $\mathcal{T}(k^\star)$ denoting the set of all those transfer schemes that (definitely) implement the policy $k^\star$. Searching for those policies $k^\star$ that minimize $C(k^\star)$ could then constitute the objective when evaluating welfare judgments on the grounds of second-best implementation in more complex economic environments.

In what follows, I refer to the mechanisms of Theorem 1 as *AGV-type mechanisms*, as they coincide with the *expected-externality mechanism* of Arrow (1979) and d'Aspremont and Gérard-Varet (1979) for bilateral bargaining problems. Theorem 1 states in particular that the mechanisms that implement 'desirable' policies are basically unique: The transfer components $(s_i)_i$ must be rendered strategically inoperative and are thus irrelevant when it comes to the choice between social alternatives k; they capture the possibility of interpersonal transfers that are (strategically) unrelated to the actual allocation problem. From the interim perspective, the stage of decision-making, these components represent a (potential) redistribution bias towards one agent: by Theorem 1, there must exist constants S_i such that $S_1 + S_2 = 0$ and $S_i = \mathbb{E}_{\theta_1,\delta_1}[s_i(\theta,\delta)] = \mathbb{E}_{\theta_2,\delta_2}[s_i(\theta,\delta)]$ for both i, implying that agents have identical perceptions of this redistribution bias.[17]

Under AGV-type mechanisms, when leaving the $(s_i)_i$ aside, each agent i pays to agent $-i$ the money equivalent of what $-i$ believes to contribute to i's material well-being when reporting her payoff type θ_{-i}. As i herself receives $\mathbb{E}_{\theta_{-i}}\big[\pi_{-i}(k^\star(\hat{\theta}_i,\theta_{-i})\,|\,\theta_{-i})\big]$ from $-i$, agent i's social type becomes strategically irrelevant: $\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[\Pi_{-i}(k^\star,t^\star_{-i}\,|\,\theta_{-i})\big] = \mathbb{E}_{\theta}\big[\pi_i(k^\star(\theta)\,|\,\theta_i)\big]$. In this respect, AGV-type mechanisms are *social-preference robust*.

That Bayesian implementation *can be* social-preference robust was shown first by Bierbrauer and Netzer (2016).[18] The present study provides conditions under which social-preference robustness is even necessary.[19]

Interestingly, generic policies that are *definitely* implementable at zero incentive costs are automatically ex-post Pareto-efficient. However, this implication requires that agents are moderately altruistic or spiteful; for two agents, this means that $|\delta_i| < 1$.[20]

[17] **Proof:** If $\mathbb{E}_{\theta_j,\delta_j}[s_i(\theta,\delta)]$ is constant for all i,j, then there exists a constant $\alpha_i = \sum_j \mathbb{E}_{\theta_j,\delta_j}[s_i(\theta,\delta)]$. Taking expectations over $(\theta_\ell,\delta_\ell)$ yields $\alpha_i = \mathbb{E}_{\theta_\ell,\delta_\ell}[s_i(\theta,\delta)] + \mathbb{E}_{\theta,\delta}[s_i(\theta,\delta)]$. Hence, $\mathbb{E}_{\theta_\ell,\delta_\ell}[s_i(\theta,\delta)] = S_i$ for all $i,\ell \in \{1,2\}$, where $S_i = \alpha_i - \mathbb{E}_{\theta,\delta}[s_i(\theta,\delta)]$ is constant.

[18] Although these authors are concerned with intention-based social preferences in the manner of Rabin (1993), they observe that their result does also hold for unconditional pro- or anti-social preferences, such as altruism and spite. These and other authors (e.g., Bartling and Netzer, 2016, and Bierbrauer et al., 2017) deem social-preference robustness a desirable property because it allows for avoiding unrealistic common-knowledge assumptions (here: about social-type distributions), as urged for by Wilson (1987).

[19] The analysis of the n-agents case reveals that, regarding the choice between social alternatives k, social-preference robustness is always necessary. However, if $n \geq 3$, then the transfer scheme might leave agents' social preferences strategically operative; and I show in Daske (2020) how the asymmetry of information about them can be operationalized to satisfy agents' participation constraints if those are considered.

[20] Otherwise, the Pareto frontier might be indefinite; either an agent is willing to transfer arbitrary amounts of money to a grateful recipient ($\delta_1 > 1 > \delta_2$), or both agents are willing to give up arbitrary amounts ($\delta_1,\delta_2 < -1$). For groups of more than two agents, the constraint on interpersonal degrees of altruism must be sharper, but the implication prevails qualitatively.

The following Propositions give proof of Theorem 1. The results also show that if the economic environment does not allow for an interior solution to $\max_{k\in K}\ \pi_1(k\,|\,\theta_1) + \pi_2(k\,|\,\theta_2)$, then no generic policy is *definitely* implementable; in this case, only *constant* policies (i.e., $\partial k^\star/\partial\theta_i = 0$ and $\partial k^\star/\partial\delta_i = 0$ for both i) are *definitely* implementable.

The sufficiency part is to be addressed first.[21]

Proposition 1 *The policy* $k^\star(\theta) = \arg\max_{k\in K}\ \pi_1(k\,|\,\theta_1) + \pi_2(k\,|\,\theta_2)$ *is* definitely *implementable through AGV-type transfers. The mechanism* $\langle k^\star, T^\star\rangle$ *is ex-post Pareto-efficient.*

Proof. *Definite implementability*: Suppose agent $-i$ reveals her type truthfully. Then agent i reports $(\hat\theta_i, \hat\delta_i)$ so as to maximize her interim-expected utility. Without loss of generality, normalize $s(\hat\theta, \hat\delta) = 0$. By equation (1),

$$\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[u_i\big] \;=\; \mathbb{E}_{\theta_{-i}}\big[\pi_i(k^\star(\hat\theta_i,\theta_{-i})\,|\,\theta_i) + \pi_{-i}(k^\star(\hat\theta_i,\theta_{-i})\,|\,\theta_{-i})\big] - (1-\delta_i)\,\mathbb{E}_\theta\big[\pi_i(k^\star(\theta)\,|\,\theta_i)\big],$$

where the second term on the right-hand side is independent of $\hat\theta_i$. If truthfully reporting θ_i was inferior for i, then there would exist some θ_{-i} such that $\pi_i(k^\star(\hat\theta_i,\theta_{-i})\,|\,\theta_i) + \pi_{-i}(k^\star(\hat\theta_i,\theta_{-i})\,|\,\theta_{-i}) > \pi_i(k^\star(\theta)\,|\,\theta_i) + \pi_{-i}(k^\star(\theta)\,|\,\theta_{-i})$, which contradicts the definition of $k^\star$. Hence, agent i has no incentive to misreport her payoff type and, obviously, she has no incentive to misreport her social type. As this argument holds for any set of type distributions, AGV-type transfers *definitely* implement $k^\star$.

Ex-post Pareto efficiency: Suppose there exists an allocation k' satisfying $\pi_1(k'\,|\,\theta_1) + \pi_2(k'\,|\,\theta_2) < \pi_1(k^\star\,|\,\theta_1) + \pi_2(k^\star\,|\,\theta_2)$ that, for some types (θ,δ), Pareto-improves upon $k^\star(\theta)$. As not both agents can be materially better off under k' than under $k^\star$, suppose that agent 1 suffers a loss and that this loss is (weakly) greater than the (potential) loss of agent 2. Then the differences $d_i = \pi_i(k^\star\,|\,\theta_i) - \pi_i(k'\,|\,\theta_i)$ satisfy $d_1 > 0$ and $d_1 \geq d_2 > -d_1$. Consequently, $u_1(k',\theta_2\,|\,\theta_1,\delta_1) - u_1(k^\star,\theta_2\,|\,\theta_1,\delta_1) = -(d_1 + \delta_1 d_2) < 0$, since $\delta_1 \in (-1,1)$. Hence, agent 1 is worse of under k' than under $k^\star$. By the same line of reasoning, one obtains that no ex-post budget-balanced transfer scheme Pareto-dominates another. Hence, AGV-type mechanisms are ex-post Pareto-efficient. ■

[21] That AGV-type mechanisms are incentive-compatible for *two* other-regarding agents and arbitrary type distributions has been shown earlier by Bierbrauer and Netzer (2016) and Bartling and Netzer (2016). The proof is restated in the present notation for the sake of completeness. Bierbrauer and Netzer (2016) also observe that the standard AGV fails to be incentive-compatible for more than two other-regarding agents. The supplement to this paper shows why: The mutual-concessions principle of the bilateral AGV must be applied to each and every bilateral relationship if there are $n \geq 3$ agents.

The next two Propositions give proof of the necessity part of Theorem 1. A Lemma eases the exposition.[22]

Lemma 1 *A partially differentiable policy $k^\star : \Theta \times \Delta \to K$ is* definitely *implementable through budget-balanced transfers only if it satisfies*

$$\left[\sum_{j=1,2} \frac{d\pi_j(k^\star \,|\, \theta_j)}{dk}\right] \frac{\partial k^\star}{\partial \theta_i} = (1-\delta_i)\frac{\partial v_i(k^\star)}{\partial \delta_i}, \tag{2}$$

$$\begin{aligned}\left[\sum_{j=1,2} \frac{d\pi_j(k^\star \,|\, \theta_j)}{dk}\right] \frac{\partial k^\star}{\partial \delta_i} = {} & (1-\delta_i)\frac{\partial^2}{\partial \delta_i^2}\int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s,\theta_{-i},\delta))\,ds \\ & - \frac{\partial}{\partial \delta_i}\int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s,\theta_{-i},\delta))\,ds \\ & + (1-\delta_i)\frac{d^2 p_i(\delta_i)}{d\delta_i^2} - \frac{dp_i(\delta_i)}{d\delta_i}\end{aligned} \tag{3}$$

on $\Theta \times \Delta$ for each $i \in \{1,2\}$, for twice differentiable functions $p_i : \Delta_i \to \mathbb{R}$.

If $k^\star$ is independent of social types, then $k^\star$ is implementable through budget-balanced transfers only if the transfer to each i satisfies $\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[t_i^\star(\theta,\delta)\big] = c_i + \mathbb{E}_{\theta_{-i}}\big[\pi_{-i}(k^\star(\theta) \,|\, \theta_{-i})\big]$ for all (θ_i,δ_i) and some constant c_i.

Proof. See Appendix A.1. ■

Lemma 1 shows that the asymmetry of information about how agents perceive the distributive effects of a mechanism constrains desirable policies to satisfy a system of intertwined partial differential equations. The Lemma already suggests that materialistic utilitarianism cannot easily be thrust aside, for the left-hand sides of identities (2) and (3) entail the derivative of aggregate private payoffs with respect to social alternatives k, and become the first-order condition of $\max_{k\in K} \sum_j \pi_j(k^\star \,|\, \theta_j)$ if the policy is social-type independent. The central question is, thus, whether desirable policies may depend on social types. The second part of the Lemma shows that if a desirable policy is social-type independent, then the mechanisms to implement it are necessarily social-preference robust.

The following result rationalizes the moral convictions of Harsanyi (1977) and Blanchet and Fleurbaey (2006), as outlined in the beginning, that policies shall not depend on agents' interpersonal concerns:

[22] Technically speaking, the focus on *definitely* implementable policies serves the purpose of eliminating the expectations operator from the agents' first-order conditions, resulting in the necessary conditions stated by Lemma 1.

Proposition 2 *A partially differentiable policy $k^\star : \Theta \times \Delta \to K$ is* definitely *implementable through budget-balanced transfers only if it is social-preference independent.*

Proof. By Lemma 1, the policy $k^\star$ must satisfy conditions (2) and (3). Integrating (2) with respect to θ_i, while integrating $d\pi_i(k^\star \,|\, \theta_i)/dk$ by parts,[23] yields

$$C_i + \sum_{j=1,2} \pi_j(k^\star \,|\, \theta_j) - \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s, \theta_{-i}, \delta))\, ds \;\; = \;\; (1-\delta_i)\frac{\partial}{\partial \delta_i} \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s, \theta_{-i}, \delta))\, ds,$$

for some function $C_i : \Theta_{-i} \times \Delta \to \mathbb{R}$. Differentiating this with respect to δ_i yields

$$\begin{aligned} &\frac{\partial C_i(\theta_{-i}, \delta)}{\partial \delta_i} + \left[\sum_{j=1,2} \frac{d\pi_j(k^\star \,|\, \theta_j)}{dk} \right] \frac{\partial k^\star}{\partial \delta_i} - \frac{\partial}{\partial \delta_i} \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s, \theta_{-i}, \delta))\, ds \\ = \;\; &(1-\delta_i)\frac{\partial^2}{\partial \delta_i^2} \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s, \theta_{-i}, \delta))\, ds - \frac{\partial}{\partial \delta_i} \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s, \theta_{-i}, \delta))\, ds. \end{aligned}$$

Substituting for identity (3) yields

$$\frac{\partial C_i(\theta_{-i}, \delta)}{\partial \delta_i} + (1-\delta_i)\frac{d^2 p_i(\delta_i)}{d\delta_i^2} \;\; - \;\; \frac{dp_i(\delta_i)}{d\delta_i} \;\; = \;\; \frac{\partial}{\partial \delta_i} \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s, \theta_{-i}, \delta))\, ds.$$

Differentiating the latter with respect to θ_i implies that $0 = \partial v_i(k^\star)/\partial \delta_i$. As $dv_i/dk \neq 0$ by assumption, $k^\star$ must satisfy $\partial k^\star(\theta, \delta)/\partial \delta_i = 0$, for all (θ, δ) and both agents i. ■

Reconsidering Lemma (1), the next result draws the rather obvious conclusion: As agents' social preferences must be ignored, any costless welfare judgment must be consistent with materialistic utilitarianism, and the mechanisms to implement it must be of AGV-type:

Proposition 3 *A generic, social-preference independent policy $k^\star : \Theta \to K$ is* definitely *implementable through budget-balanced transfers only if $k^\star(\theta) = \arg\max_{k \in K} \pi_1(k \,|\, \theta_1) + \pi_2(k \,|\, \theta_2)$. The mechanisms that implement $k^\star$ are necessarily of AGV-type.*

Proof. As $k^\star : \Theta \to K$ is generic, $\partial k^\star/\partial \theta_i \neq 0$. By assumption, $\partial k^\star/\partial \delta_i = 0$. Hence, condition (2) of Lemma 1 implies that $k^\star$ must satisfy $0 = \sum_i d\pi_i(k^\star \,|\, \theta_i)/dk$. (If this equation has no solution, then condition (2) and Proposition 2 imply that $k^\star$ must be constant: $\partial k^\star/\partial \theta_i = 0 = \partial k^\star/\partial \delta_i$.) Hence, for each θ, the policy $k^\star(\theta)$ is either a minimizer, saddle point, or maximizer of $\sum_i \pi_i(k \,|\, \theta_i)$, which we keep in mind for a moment.

[23] $\int \frac{d\pi_i(k^\star \,|\, \theta_i)}{dk} \frac{\partial k^\star}{\partial \theta_i} d\theta_i = \int \theta_i \frac{\partial v_i(k^\star)}{\partial \theta_i} d\theta_i + \int \frac{\partial w_i(k^\star)}{\partial \theta_i} d\theta_i = \theta_i v_i(k^\star) - \int v_i(k^\star) d\theta_i + w_i(k^\star) + C_i.$

Suppose $T^\star$ is a budget-balanced transfer scheme that implements $k^\star(\theta)$. Notice that one can always write $t_i^\star(\hat\theta,\hat\delta) = \mathbb{E}_{\theta_{-i}}\big[\pi_{-i}(k^\star(\hat\theta_i,\theta_{-i})\,|\,\theta_{-i})\big] - \mathbb{E}_{\theta_i}\big[\pi_i(k^\star(\theta_i,\hat\theta_{-i})\,|\,\theta_i)\big] + s_i(\hat\theta,\hat\delta)$, for appropriate functions $s_i : \Theta\times\Delta \to \mathbb{R}$ satisfying $s_1 + s_2 = 0$ on $\Theta\times\Delta$. But then, $\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[t_i^\star(\theta,\delta)\big] = \mathbb{E}_{\theta_{-i}}\big[\pi_{-i}(k^\star(\theta)\,|\,\theta_{-i})\big] - \mathbb{E}_{\theta}\big[\pi_i(k^\star(\theta)\,|\,\theta_i)\big] + \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[s_i(\theta,\delta)\big]$. When substituting for $\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[t_i^\star(\theta,\delta)\big] = c_i + \mathbb{E}_{\theta_{-i}}\big[\pi_{-i}(k^\star(\theta)\,|\,\theta_{-i})\big]$ from Lemma 1, one observes that each function s_i must satisfy $\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[s_i(\theta,\delta)\big] = c_i + \mathbb{E}_{\theta}\big[\pi_i(k^\star(\theta)\,|\,\theta_i)\big]$, which is constant; as $s_1 + s_2 = 0$, also $\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[s_{-i}(\theta,\delta)\big]$ is constant. Hence, *transfers* must be of AGV-type.

Now reconsider the possible nature of $k^\star$. Under AGV-type transfers, the mechanism $\langle k^\star, T^\star\rangle$ yields agent i an interim-expected utility level of

$$\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[u_i\big] \;=\; \mathbb{E}_{\theta_{-i}}\big[\pi_i(k^\star(\hat\theta_i,\theta_{-i})\,|\,\theta_i) + \pi_{-i}(k^\star(\hat\theta_i,\theta_{-i})\,|\,\theta_{-i})\big] - (1-\delta_i)\,\mathbb{E}_{\theta}\big[\pi_i(k^\star(\theta)\,|\,\theta_i)\big].$$

As $k^\star$ is supposed to be *definitely* implementable, one must have $\pi_i(k^\star(\hat\theta_i,\theta_{-i})\,|\,\theta_i) + \pi_{-i}(k^\star(\hat\theta_i,\theta_{-i})\,|\,\theta_{-i}) \le \pi_i(k^\star(\theta)\,|\,\theta_i) + \pi_{-i}(k^\star(\theta)\,|\,\theta_{-i})$ for all $\hat\theta_i,\theta$. This is obviously impossible if $k^\star$ is a minimizer or saddle point. Hence, $k^\star(\theta) = \arg\max_{k\in K}\ \pi_1(k\,|\,\theta_1)+\pi_2(k\,|\,\theta_2)$, and $\langle k^\star, T^\star\rangle$ is of AGV-type. ■

Theorem 1 is thus established.

A closer look at condition (2) reveals that Propositions 1 to 3 even establish a more general result:

Corollary 1 *A partially differentiable policy $k^\star : \Theta\times\Delta \to K$ is* definitely *implementable through budget-balanced transfers if and only if it is either constant (i.e., selects the same social alternative irrespective of the agents' preferences) or consistent with materialistic utilitarianism. (In the first case, transfers can be zero, in the second, they must be of AGV-type.)*

Corollary 1 has the following qualitative meaning: If 'costless' policies should at all depend on the agents' preferences, then they must vary with agents' preferences for consumption (and only those), and no agent must be ignored. The latter implication calls for a closer look at favoritism and dictatorship; in Section 4. The former implication shows that focusing on generic, payoff-type dependent policies is without loss of generality

if the social-choice problem, how to *aggregate* individual preferences, is supposed to be meaningful.

Another implication is that a social planner, if at all concerned with the agents' needs, must not attach value to social alternatives as such. For instance, at most for specific type distributions could a policy of the form $k^\star = \arg\max_{k\in K} k + \pi_1(k \,|\, \theta_1) + \pi_2(k \,|\, \theta_2)$ (e.g., 'as much environmental protection as possible while keeping an eye on the economy') be implemented at a balanced budget.

The next Theorem shows that various generic policies are *definitely* implementable if one waives budget balance. Hence, many welfare judgments other than the materialistic utilitarian are feasible, but all of those involve incentive costs.

Theorem 2 *If budget balance is not imposed, then a twice continuously partially differentiable policy* $k^\star : \Theta \times \Delta \to K$ *is* definitely *implementable if it satisfies the following conditions: (i)* $\inf_{\Theta\times\Delta} \partial v_i(k^\star)/\partial\theta_i > 0$*; (ii)* $\partial v_i(k^\star)/\partial\delta_i$ *is bounded; (iii)* $\partial^2 v_i(k^\star)/\partial\delta_i^2$ *is bounded below.*

Proof. The proof is constructive. See Appendix A.2. ■

Another Theorem clarifies the critical role of *asymmetric information* about agents' interpersonal concerns: It is not social preferences per se that constrict the set of costless welfare judgments but rather the asymmetry of information about them.

Theorem 3 *If social preferences are common knowledge, then any partially differentiable policy* $k^\star : \Theta \times \Delta \to K$ *satisfying* $\partial v_i(k^\star)/\partial\theta_i \geq 0$ *is* definitely *implementable through budget-balanced transfers.*

Proof. The proof is constructive. See Appendix A.3. ■

Notice that Theorems 2 and 3 explicitly allow for policies that depend on agents' interpersonal concerns. Moreover, the mechanism in the proof of Theorem 3 illustrates that other than AGV-type transfer schemes would implement the materialistic-utilitarian policy if the agents' social preferences were common knowledge.

4 Favoritism and Dictatorship

By Theorem 1, a generic policy is *definitely* implementable at zero incentive costs only if it is unbiased. Both agents' private payoffs, and only those, must enter the welfare judgment with equal weight. Any form of *favoritism*, e.g., $k^\star = \arg\max_{k\in K}(\alpha_1 u_1 + \alpha_2 u_2)$ or $k^\star = \arg\max_{k\in K}(\alpha_1 \pi_1 + \alpha_2 \pi_2)$, with $\alpha_1 \neq \alpha_2$, would come at incentive costs.[24]

The extreme form of favoritism is *dictatorship*: $k^\star = \arg\max_{k\in K} u_i$ for one agent i. Dictatorship plays a prominent role in social-choice theory; it constitutes a simple benchmark for how to derive a 'social' preference from the preferences of society's members, while preserving rationality and guaranteeing Pareto efficiency (e.g., Arrow, 1950). Concepts of *random dictatorship* have thus attracted some interest (e.g., Gibbard, 1977).

The results of Section 3 allow us to assess *the incentive costs of dictatorship*. Notice that the *definite* implementability of her policy is in the dictator's own best interest.

Theorem 4 *A dictatorial policy, if it is interior solution to* $\max_{k\in K} u_i(k, \theta_{-i} \,|\, \theta_i, \delta_i)$*, is* definitely *implementable through budget-balanced transfers if and only if the dictator is either perfectly selfish,* $\delta_i = 0$*, or perfectly benevolent,* $\delta_i = 1$*. (This equivalence also holds if the dictator's preferences are common knowledge.)*

Proof. The sufficiency part is trivial for $\delta_i = 0$ (transfers can even be zero), and it is immediate from Proposition 1 for $\delta_i = 1$ (in which case the dictator is indifferent between any interpersonal transfers).

For the necessity part, let $\delta_i \notin \{0, 1\}$. If dictator i's problem has always an interior solution $k^\star$, then $0 = du_i(k^\star, \theta_{-i} \,|\, \theta_i, \delta_i)/dk$ and $0 > d^2u_i(k^\star, \theta_{-i} \,|\, \theta_i, \delta_i)/dk^2$, implying that $\partial k^\star/\partial \theta_{-i} = -\delta_i \cdot v_{-i}/(d^2u_i/dk^2) \neq 0$. As also $\partial k^\star/\partial \delta_{-i} = 0$, condition (2) of Lemma 1, when applied to $-i$, implies that $k^\star$ must satisfy $0 = \sum_j d\pi_j(k^\star \,|\, \theta_j)/dk$, while $0 = du_i(k^\star, \theta_{-i} \,|\, \theta_i, \delta_i)/dk = d\pi_i(k^\star \,|\, \theta_i)/dk + \delta_i \cdot d\pi_{-i}(k^\star \,|\, \theta_{-i})/dk$. These conditions on $k^\star$ hold simultaneously only in the *null event* that $d\pi_i(k^\star \,|\, \theta_i)/dk = 0 = d\pi_{-i}(k^\star \,|\, \theta_{-i})/dk$, at most for some specific θ, requiring that i and $-i$ prefer the same social alternative $k^\star$.

This line of reasoning, including the derivation of condition (2) for $-i$, builds solely on the assumption that the preferences of $-i$ are private information.[25] ■

[24] This result, which carries over to the n-agents case, puts into perspective the often conducted social-utility weights approach in the theory of optimal taxation; see, e.g, Saez and Stantcheva (2016).

[25] Common knowledge about the dictator's preferences simply means that the expectations operator $\mathbb{E}_{\theta_i,\delta_i}[\,\cdot\,]$ in the proof of Lemma 1 vanishes when taking the perspective of subordinate $-i$. In this case, also the restriction to *definitely* implementable policies becomes superfluous.

Theorem 4 entails a somewhat surprising non-monotonicity: Contrary to self-serving and perfectly benevolent dictatorship, malevolent as well as imperfectly benevolent dictatorship ($\delta_i < 0$, or $0 < \delta_i < 1$) must come at incentive costs. By contrast, Theorems 2 and 3 apply at least to imperfectly benevolent dictatorship, such that the dictator's will could be implemented if budget balance was not a constraint or even at zero incentive costs if the subordinate's feelings for the dictator were common knowledge.

Theorem 4 builds on the realistic assumption that the dictator's knowledge about her subordinate's allocative preferences is imperfect. This imperfection of dictatorial knowledge is problematic if the dictator is imperfectly selfish. If she cares about her subordinate's material well-being, for the better or worse, then incentive compatibility with respect to θ_{-i} becomes a constraint. As she is also imperfectly informed about the subordinate's feelings for her, incentive compatibility becomes even more demanding because the subordinate might misrepresent θ_{-i} due to its effect on the dictator.

The generalization of the theory to the n-agents case implies that costless dictatorship treats all subordinates as equals, with the dictator either ignoring all of them or regarding them as equal to herself. In particular, dictatorship favoring one subgroup over another always comes at incentive costs.

Theorem 4 contributes a simple rationale to the debate on the economic performance and stability of dictatorship, and it highlights the critical role of the *character* of dictatorial policies: An unbalanced budget means either economic mal-performance, triggering regime instability, or dependence on outside subsidies, implying non-autonomy. Notice that a dictator's perfect selfishness is a null event, not only theoretically but also practically, because any regime relies to some extent on the goodwill of an elite, be it the leader's party, the military, the aristocracy, or the bureaucracy; in these cases, the dictator's social preference might merely be instrumental, not intrinsic. On the other hand, perfect benevolence, which is equally unlikely for the very same reasons, would render the dictator's regime effectively non-dictatorial.

While the prominent theories of (or against) dictatorship all take a dynamic, political-economy perspective (e.g., Olson, 1993; Wintrobe, 2000; Acemoglu and Robinson, 2006), the present result is in the spirit of Hayek's (1945) argument against *centrally planned* economies, stressing the critical role of information asymmetries.

5 Distributive and Procedural Justice

Thus far, I have focused on the welfare judgment inherent to the choice between social alternatives. I conclude the paper with a general interpretation of Theorem 1 in terms of distributive and procedural justice by taking interpersonal transfers into account.

I have argued, in Section 2.3, that a policy is socially desirable if it is *generic* and *definitely* implementable through budget-balanced transfers. Based on the assumption that agents' allocative and distributive preferences are all subject to asymmetric information, I have shown that desirable policies must maximize aggregate *material* wealth. The corresponding mechanism must oblige agents to make mutual concessions amounting to the interim-expected *material* externalities they impose on each other under this policy, which is precisely the renowned AGV-mechanism for bilateral bargaining problems.

These implications carry over to the n-agents case, as shown in the supplement to this paper. Interpersonal transfers must then be made by applying the same mutual-concessions principle to every single bilateral relationship, which differs from the AGV if $n \geq 3$. While the AGV subsidizes or sanctions the *average* externalities that an agent imposes on the rest of the group, 'desirable' mechanisms treat interpersonal externalities on the *bilateral* level.

The results imply that distributive and procedural justice are conceptually not distinct. Consider, for instance, Rawls' (1971) conception of *perfect procedural justice*: It calls for a fundamental concept of allocative fairness (independent of the means to implement it), and, only second, a procedure that always yields the fair outcome by incentivizing agents to act accordingly; it takes distributive justice as the objective and requires the incentive scheme to serve on its behalf. This conception of justice suggests that allocative fairness could be specified at will.[26]

The present study shows that, under reasonable assumptions, there is but one form of perfect procedural justice that preserves a balanced budget: The choice between social alternatives must be consistent with *materialistic utilitarianism*, and agents must com-

[26] At this point, I should comment on the Rawlsian (1971) maximin-welfare criterion and, likewise, the bargaining solution of Kalai (1977): When applied to the choice between social alternatives, whether with respect to private payoffs ($k^\star = \arg\max_{k\in K} \min_i \pi_i(k \,|\, \theta_i)$) or utilities ($k^\star = \arg\max_{k\in K} \min_i u_i(k, \theta_{-i} \,|\, \theta_i, \delta_i)$), then it is Theorem 4 that rejects definite implementability at zero incentive costs; because, locally, the worse-off agent becomes dictator, whose perfect selfishness or perfect benevolence are null events. When taking the distributive effects of budget-balanced transfers into account, then the maximin principle is *egalitarian*, seeking to equalize either private payoffs or utilities; it is easy to see that in neither case is the resulting transfer scheme of AGV-type, in violation of Theorem 1.

pensate each other for the material externalities they expect to impose on each other under this policy. As interpersonal transfers thus vary with payoff-type distributions, the evaluation of ex-post allocations, too, must be consistent with materialistic utilitarianism. In particular, perfect procedural justice *objectifies* a conflict of interests by rendering the agents' interpersonal concerns entirely irrelevant, both formally and strategically.

Referring to incomplete information and transaction costs, Rawls' (1971, pp.74–75) concluded that "[p]retty clearly, perfect procedural justice is rare, if not impossible, in cases of much practical interest." He thus invoked *pure procedural justice*, meaning that "there is no independent criterion for the right result: instead there is a correct or fair procedure such that the outcome is likewise correct or fair, whatever it is, provided that the procedure has been properly followed." As the properties imposed on 'desirable' mechanisms in Section 2.3 are concerned with the procedure, not with allocative fairness, the results obtained in this paper are equally decisive when interpreting them in terms of pure procedural justice. Indeed, when imposing budget balance, perfect and pure procedural justice turn out to coincide.

Concerning the choice between social alternatives, the general features of pure procedural justice are no favoritism, no dictatorship; independence of agents' interpersonal concerns; and a materialistic-utilitarian policy seeking to maximize aggregate material wealth (which can well be interpreted as maximizing the aggregate money equivalents of individuals' prospects and opportunities). The more specific feature is, effectively, a unique conception of allocative fairness: Redistribution of material wealth on a *bilateral* basis (also in the n-agents case) determined by the net externalities that individuals expect to impose on each other when acting freely under the materialistic-utilitarian policy.—Any other procedure would come at incentive costs, either by diminishing distributable material wealth or by relying on outside subsidies, thus undermining group autonomy.

From a different angle, pure procedural justice involves a unique bargaining solution. The conflicting parties settle on a policy that maximizes 'what is in for all,' and based on their 'promises' on how to act under this policy, they each request compensation for how keeping promises would probably favor every single other party. In this respect, pure procedural justice has an exciting appeal to decentralization (once more echoing

Hayek, 1945): Incentivizing individuals to act in favor of the *social* goal requires them to compensate each other, at least effectively, through *bilateral* bargains.

While some of these features of procedural justice might critically depend on the assumptions made, the more general ones might carry over to second-best implementation in more complex, perhaps more realistic, economic environments. Nonetheless, most of these features seem practically impossible to implement, but the principal purpose of moral philosophy, even from an incentive-theoretical perspective, is not to give practical advice but orientation.

A Omitted Proofs

A.1 Proof of Lemma 1

Suppose $k^\star : \Theta \times \Delta \to \mathbb{R}$ is partially differentiable and *definitely* implementable through budget-balanced transfers $T = (t_1, t_2) : \Theta \times \Delta \to \mathbb{R}^2$. Define

$$\bar{v}_i(\hat{\theta}_i, \hat{\delta}_i) = \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^\star(\hat{\theta}_i, \hat{\delta}_i, \theta_{-i}, \delta_{-i}))\big], \tag{4}$$

$$\bar{w}_i(\hat{\theta}_i, \hat{\delta}_i) = \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[w_i(k^\star(\hat{\theta}_i, \hat{\delta}_i, \theta_{-i}, \delta_{-i}))\big], \tag{5}$$

$$\bar{\pi}_i(\hat{\theta}_i, \hat{\delta}_i) = \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[\pi_i(k^\star(\hat{\theta}_i, \hat{\delta}_i, \theta_{-i}, \delta_{-i}) \,|\, \hat{\theta}_i)\big], \tag{6}$$

$$\bar{\pi}_{-i}(\hat{\theta}_i, \hat{\delta}_i) = \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[\pi_{-i}(k^\star(\hat{\theta}_i, \hat{\delta}_i, \theta_{-i}, \delta_{-i}) \,|\, \theta_{-i})\big], \tag{7}$$

$$\bar{t}_i(\hat{\theta}_i, \hat{\delta}_i) = \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[t_i(\hat{\theta}_i, \hat{\delta}_i, \theta_{-i}, \delta_{-i})\big], \tag{8}$$

$$\bar{t}_{-i}(\hat{\theta}_i, \hat{\delta}_i) = \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[t_{-i}(\hat{\theta}_i, \hat{\delta}_i, \theta_{-i}, \delta_{-i})\big]. \tag{9}$$

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) and if agent $-i$ reports her type truthfully:

$$U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) = \theta_i \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i) + \bar{w}_i(\hat{\theta}_i, \hat{\delta}_i) + \bar{t}_i(\hat{\theta}_i, \hat{\delta}_i) + \delta_i \bar{\pi}_{-i}(\hat{\theta}_i, \hat{\delta}_i) + \delta_i \bar{t}_{-i}(\hat{\theta}_i, \hat{\delta}_i).$$

Ease notation by also defining $U_i(\theta_i, \delta_i) = U_i(\theta_i, \delta_i \,|\, \theta_i, \delta_i)$. Then the following must hold for all $\theta_i, \hat{\theta}_i \in \Theta_i$ and all $\delta_i, \hat{\delta}_i \in \Delta_i$:

$$U_i(\theta_i, \delta_i) \geq U_i(\hat{\theta}_i, \delta_i \,|\, \theta_i, \delta_i) = U_i(\hat{\theta}_i, \delta_i) + (\theta_i - \hat{\theta}_i)\bar{v}_i(\hat{\theta}_i, \delta_i), \tag{10}$$

$$U_i(\hat{\theta}_i, \delta_i) \geq U_i(\theta_i, \delta_i \,|\, \hat{\theta}_i, \delta_i) = U_i(\theta_i, \delta_i) + (\hat{\theta}_i - \theta_i)\bar{v}_i(\theta_i, \delta_i), \tag{11}$$

$$U_i(\theta_i, \delta_i) \geq U_i(\theta_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) = U_i(\theta_i, \hat{\delta}_i) + (\delta_i - \hat{\delta}_i)\Big[\bar{\pi}_{-i}(\theta_i, \hat{\delta}_i) + \bar{t}_{-i}(\theta_i, \hat{\delta}_i)\Big], \tag{12}$$

$$U_i(\theta_i, \hat{\delta}_i) \geq U_i(\theta_i, \delta_i \,|\, \theta_i, \hat{\delta}_i) = U_i(\theta_i, \delta_i) + (\hat{\delta}_i - \delta_i)\Big[\bar{\pi}_{-i}(\theta_i, \delta_i) + \bar{t}_{-i}(\theta_i, \delta_i)\Big]. \tag{13}$$

Without loss of generality, suppose $\hat{\theta}_i > \theta_i$. Then (10) and (11) imply that

$$\bar{v}_i(\hat{\theta}_i, \delta_i) \geq \frac{U_i(\hat{\theta}_i, \delta_i) - U_i(\theta_i, \delta_i)}{\hat{\theta}_i - \theta_i} \geq \bar{v}_i(\theta_i, \delta_i). \tag{14}$$

As $\bar{v}_i$ is continuous on Θ_i, letting $\hat{\theta}_i$ approach θ_i yields $\partial U_i(\theta_i, \delta_i)/\partial\theta_i = \bar{v}_i(\theta_i, \delta_i)$. Integrating the latter with respect to θ_i yields the identity

$$U_i(\theta_i, \delta_i) \;=\; p_i(\delta_i) + \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\, ds, \tag{15}$$

with some function $p_i : \Delta_i \to \mathbb{R}$. Similarly, suppose $\hat{\delta}_i > \delta_i$. Then (12) and (13) yield

$$\bar{\pi}_{-i}(\theta_i, \hat{\delta}_i) + \bar{t}_{-i}(\theta_i, \hat{\delta}_i) \;\geq\; \frac{U_i(\theta_i, \hat{\delta}_i) - U_i(\theta_i, \delta_i)}{\hat{\delta}_i - \delta_i} \;\geq\; \bar{\pi}_{-i}(\theta_i, \delta_i) + \bar{t}_{-i}(\theta_i, \delta_i).$$

As $\bar{\pi}_{-i}$ and $\bar{t}_{-i}$ are continuous on Δ_i by assumption, $U_i(\theta_i, \delta_i)$ is differentiable in δ_i; letting $\hat{\delta}_i$ approach δ_i yields $\partial U_i(\theta_i, \delta_i)/\partial\delta_i = \bar{\pi}_{-i}(\theta_i, \delta_i) + \bar{t}_{-i}(\theta_i, \delta_i)$. Integrating the latter with respect to δ_i yields

$$U_i(\theta_i, \delta_i) \;=\; q_i(\theta_i) + \int_{\delta_i^{\min}}^{\delta_i} \bar{\pi}_{-i}(\theta_i, r)\, dr + \int_{\delta_i^{\min}}^{\delta_i} \bar{t}_{-i}(\theta_i, r)\, dr, \tag{16}$$

with some function $q_i : \Theta_i \to \mathbb{R}$. Jointly, (15) and (16) imply that

$$\int_{\delta_i^{\min}}^{\delta_i} \bar{t}_{-i}(\theta_i, r)\, dr \;=\; p_i(\delta_i) - q_i(\theta_i) + \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\, ds - \int_{\delta_i^{\min}}^{\delta_i} \bar{\pi}_{-i}(\theta_i, r)\, dr. \tag{17}$$

As $\bar{v}_i$ is differentiable in δ_i, identity (17) implies that also p_i is differentiable in δ_i. As $k^\star$ and T are assumed partially differentiable on Δ, while each π_i is differentiable in k, the functions p_i are twice differentiable. Differentiating (17) with respect to δ_i yields

$$\bar{t}_{-i}(\theta_i, \delta_i) \;=\; \frac{dp_i(\delta_i)}{d\delta_i} - \bar{\pi}_{-i}(\theta_i, \delta_i) + \frac{\partial}{\partial\delta_i} \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\, ds. \tag{18}$$

Budget balance requires in particular that $\bar{t}_i(\theta_i, \delta_i) = -\bar{t}_{-i}(\theta_i, \delta_i)$ on $\Theta_i \times \Delta_i$, such that truthful revelation, $(\hat{\theta}_i, \hat{\delta}_i) = (\theta_i, \delta_i)$, is incentive-compatible for agent i only if the following partial first-order condition is satisfied:

$$\begin{aligned} 0 \;&=\; \frac{\partial}{\partial\hat{\theta}_i} \Big[\theta_i \bar{v}_i(\hat{\theta}_i, \delta_i) + \bar{w}_i(\hat{\theta}_i, \delta_i) + \delta_i \bar{\pi}_{-i}(\hat{\theta}_i, \delta_i) - (1 - \delta_i)\bar{t}_{-i}(\hat{\theta}_i, \delta_i)\Big]_{\hat{\theta}_i = \theta_i} \\ &=\; \mathbb{E}_{\theta_{-i}, \delta_{-i}} \left[\sum_{j=1,2} \frac{d\pi_j(k^\star(\theta, \delta) \,|\, \theta_j)}{dk} \frac{\partial k^\star}{\partial\theta_i} - (1 - \delta_i) \frac{\partial v_i(k^\star(\theta, \delta))}{\partial\delta_i} \right], \end{aligned} \tag{19}$$

where the second equality is implied by (18). As $k^\star$ is supposed to be *definitely* implementable, (19) must hold for arbitrary type distributions. As the argument of $\mathbb{E}_{\theta_{-i},\delta_{-i}}[\,\cdot\,]$ is continuous in $(\theta_{-i},\delta_{-i})$, the policy $k^\star$ must satisfy condition (2). Similarly, the first-order condition with respect to $\hat{\delta}_i$ requires that

$$\begin{aligned} 0 &= \frac{\partial}{\partial\hat{\delta}_i}\Big[\theta_i\bar{v}_i(\hat{\theta}_i,\delta_i)+\bar{w}_i(\hat{\theta}_i,\delta_i)+\delta_i\bar{\pi}_{-i}(\hat{\theta}_i,\delta_i)-(1-\delta_i)\bar{t}_{-i}(\hat{\theta}_i,\delta_i)\Big]_{\hat{\delta}_i=\delta_i} \\ &= \mathbb{E}_{\theta_{-i},\delta_{-i}}\Bigg[\sum_{j=1,2}\frac{d\pi_j(k^\star(\theta,\delta)\,|\,\theta_j)}{dk}\frac{\partial k^\star}{\partial\delta_i} - (1-\delta_i)\frac{d^2p_i(\delta_i)}{d\delta_i^2} + \frac{dp_i(\delta_i)}{d\delta_i} \\ &\qquad - (1-\delta_i)\frac{\partial^2}{\partial\delta_i^2}\int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s,\theta_{-i},\delta))\,ds \\ &\qquad + \frac{\partial}{\partial\delta_i}\int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s,\theta_{-i},\delta))\,ds\Bigg], \end{aligned}$$

where the second equality is implied by (18). As $k^\star$ is supposed to be *definitely* implementable, $k^\star$ must satisfy condition (3). This proves the first part of the Lemma.

For the second part, reconsider identities (15) and (18). Jointly, they imply that

$$\begin{aligned} p_i(\delta_i)+\int_{\theta_i^{\min}}^{\theta_i}\bar{v}_i(s,\delta_i)\,ds &= U_i(\theta_i,\delta_i) \\ &= \theta_i\bar{v}_i(\theta_i,\delta_i)+\bar{w}_i(\theta_i,\delta_i)+\bar{t}_i(\theta_i,\delta_i) \\ &\quad +\delta_i\frac{dp_i(\delta_i)}{d\delta_i}+\delta_i\frac{\partial}{\partial\delta_i}\int_{\theta_i^{\min}}^{\theta_i}\bar{v}_i(s,\delta_i)\,ds. \end{aligned} \tag{20}$$

Now suppose $\partial k^\star/\partial\delta_i=0$ for both i. Due to (18) and (20), $\bar{t}_i$ and $\bar{t}_{-i}$ then satisfy

$$\bar{t}_i(\theta_i,\delta_i) = p_i(\delta_i)-\delta_i\frac{dp_i(\delta_i)}{d\delta_i}-\theta_i\bar{v}_i(\theta_i,\delta_i)-\bar{w}_i(\theta_i,\delta_i)+\int_{\theta_i^{\min}}^{\theta_i}\bar{v}_i(s,\delta_i)\,ds, \tag{21}$$

$$\bar{t}_{-i}(\theta_i,\delta_i) = \frac{dp_i(\delta_i)}{d\delta_i}-\bar{\pi}_{-i}(\theta_i,\delta_i), \tag{22}$$

where, now, only the terms containing p_i effectively depend on δ_i. Due to budget balance, identities (21) and (22) imply that p_i must solve

$$a_i = p_i(\delta_i)+(1-\delta_i)\frac{dp_i(\delta_i)}{d\delta_i}, \tag{23}$$

where a_i is a constant. Differentiating (23) with respect to δ_i yields $d^2p_i(\delta_i)/d\delta_i^2=0$, such that $dp_i(\delta_i)/d\delta_i=-c_i$ for some constant c_i. Hence, identity (22) reads $\bar{t}_{-i}(\theta_i,\delta_i)=$

$-c_i - \bar{\pi}_{-i}(\theta_i, \delta_i)$, implying that $\bar{t}_i(\theta_i, \delta_i) = c_i + \bar{\pi}_{-i}(\theta_i, \delta_i) = c_i + \mathbb{E}_{\theta_{-i}}\big[\pi_{-i}(k^\star(\hat{\theta}) \,|\, \theta_{-i})\big]$, due to budget balance and $\partial k^\star / \partial \delta_i = 0$. ■

A.2 On Theorem 2

A.2.1 Proof of Theorem 2

Be $k^\star$ as described. For twice differentiable functions $p_i : \Delta_i \to \mathbb{R}$, define $T^\star$ by

$$\begin{aligned} t_i^\star(\hat{\theta}, \hat{\delta}) \;=\; & p_i(\hat{\delta}_i) - \hat{\delta}_i \frac{dp_i(\hat{\delta}_i)}{d\hat{\delta}_i} - \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[\pi_i(k^\star(\hat{\theta}_i, \hat{\delta}_i, \theta_{-i}, \delta_{-i}) \,|\, \hat{\theta}_i)\big] \\ & + \int_{\theta_i^{\min}}^{\hat{\theta}_i} \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^\star(s, \hat{\delta}_i, \theta_{-i}, \delta_{-i}))\big] ds \\ & - \hat{\delta}_i \frac{\partial}{\partial \hat{\delta}_i} \int_{\theta_i^{\min}}^{\hat{\theta}_i} \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^\star(s, \hat{\delta}_i, \theta_{-i}, \delta_{-i}))\big] ds \\ & + \frac{\partial p_{-i}(\hat{\delta}_{-i})}{\partial \hat{\delta}_{-i}} - \mathbb{E}_{\theta_i,\delta_i}\big[\pi_i(k^\star(\theta_i, \delta_i, \hat{\theta}_{-i}, \hat{\delta}_{-i}) \,|\, \theta_i)\big] \\ & + \frac{\partial}{\partial \hat{\delta}_{-i}} \int_{\theta_{-i}^{\min}}^{\hat{\theta}_{-i}} \mathbb{E}_{\theta_i,\delta_i}\big[v_{-i}(k^\star(\theta_i, \delta_i, s, \hat{\delta}_{-i}))\big] ds. \end{aligned}$$

Then $T^\star$ *definitely* implements $k^\star$ if the functions p_i are chosen such that

$$\frac{\Big[\frac{\partial}{\partial \delta_i}\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^\star(\theta,\delta))\big]\Big]^2}{\frac{\partial}{\partial \theta_i}\mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^\star(\theta,\delta))\big]} \;<\; \frac{\partial^2}{\partial \delta_i^2}\left[p_i(\delta_i) + \int_{\theta_i^{\min}}^{\theta_i} \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k^\star(s,\theta_{-i},\delta))\big] ds\right] \tag{24}$$

for all (θ_i, δ_i). For example, one can choose $p_i(\delta_i) = \frac{1}{2} c_i \delta_i^2$, with

$$c_i \;>\; \frac{\sup_{\Theta\times\Delta}\left[\frac{\partial v_i(k^\star)}{\partial \delta_i}\right]^2}{\inf_{\Theta\times\Delta}\frac{\partial v_i(k^\star)}{\partial \theta_i}} \;-\; (\theta_i^{\max} - \theta_i^{\min}) \cdot \inf_{\Theta\times\Delta} \frac{\partial^2 v_i(k^\star)}{\partial \delta_i^2}. \tag{25}$$

Such numbers c_i exist due to conditions (i) to (iii).

To see this, suppose agent $-i$ reports her type truthfully. With notation adopted from equations (4) to (9), $T^\star$ satisfies

$$\begin{aligned}
\bar{t}_i(\hat{\theta}_i, \hat{\delta}_i) &= a_i + p_i(\hat{\delta}_i) - \hat{\delta}_i \frac{dp_i(\hat{\delta}_i)}{d\hat{\delta}_i} - \bar{\pi}_i(\hat{\theta}_i, \hat{\delta}_i) \\
&\quad + \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds - \hat{\delta}_i \frac{\partial}{\partial \hat{\delta}_i} \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds, \\
\bar{t}_{-i}(\hat{\theta}_i, \hat{\delta}_i) &= b_i + \frac{dp_i(\hat{\delta}_i)}{d\hat{\delta}_i} - \bar{\pi}_{-i}(\hat{\theta}_i, \hat{\delta}_i) + \frac{\partial}{\partial \hat{\delta}_i} \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds,
\end{aligned}$$

with appropriate constants $a_i, b_i \in \mathbb{R}$. From reporting some type $(\hat{\theta}_i, \hat{\delta}_i)$, agent i of true type (θ_i, δ_i) gains interim-expected utility

$$\begin{aligned}
U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) &= \theta_i \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i) + \bar{w}_i(\hat{\theta}_i, \hat{\delta}_i) + \bar{t}_i(\hat{\theta}_i, \hat{\delta}_i) \\
&\quad + \delta_i \bar{\pi}_{-i}(\hat{\theta}_i, \hat{\delta}_i) + \delta_i \bar{t}_{-i}(\hat{\theta}_i, \hat{\delta}_i) \\
&= (\theta_i - \hat{\theta}_i) \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i) + a_i + p_i(\hat{\delta}_i) - \hat{\delta}_i \frac{dp_i(\hat{\delta}_i)}{d\hat{\delta}_i} \\
&\quad + \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds - \hat{\delta}_i \frac{\partial}{\partial \hat{\delta}_i} \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds \\
&\quad + \delta_i b_i + \delta_i \frac{dp_i(\hat{\delta}_i)}{d\hat{\delta}_i} + \delta_i \frac{\partial}{\partial \hat{\delta}_i} \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds.
\end{aligned}$$

Partial derivatives thus satisfy

$$\frac{\partial}{\partial \hat{\theta}_i} U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) = (\theta_i - \hat{\theta}_i) \frac{\partial}{\partial \hat{\theta}_i} \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i) + (\delta_i - \hat{\delta}_i) \frac{\partial}{\partial \hat{\delta}_i} \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i), \tag{26}$$

$$\begin{aligned}
\frac{\partial}{\partial \hat{\delta}_i} U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) &= (\theta_i - \hat{\theta}_i) \frac{\partial}{\partial \hat{\delta}_i} \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i) \\
&\quad + (\delta_i - \hat{\delta}_i) \frac{\partial^2}{\partial \hat{\delta}_i^2} \left[p_i(\hat{\delta}_i) + \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds \right].
\end{aligned} \tag{27}$$

Ease notation by defining $A_i = \frac{\partial}{\partial \hat{\delta}_i} \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i)$, $B_i = \frac{\partial}{\partial \hat{\theta}_i} \bar{v}_i(\hat{\theta}_i, \hat{\delta}_i)$, and

$$C_i = \frac{\partial^2}{\partial \hat{\delta}_i^2} \left[p_i(\hat{\delta}_i) + \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(s, \hat{\delta}_i)\, ds \right].$$

Then (26) and (27) read

$$\frac{\partial}{\partial\hat{\theta}_i}U_i(\hat{\theta}_i,\hat{\delta}_i\,|\,\theta_i,\delta_i) \;=\; (\theta_i-\hat{\theta}_i)B_i+(\delta_i-\hat{\delta}_i)A_i, \tag{28}$$

$$\frac{\partial}{\partial\hat{\delta}_i}U_i(\hat{\theta}_i,\hat{\delta}_i\,|\,\theta_i,\delta_i) \;=\; (\theta_i-\hat{\theta}_i)A_i+(\delta_i-\hat{\delta}_i)C_i. \tag{29}$$

By condition (i), $B_i > 0$. Choose $p_i(\delta_i) = \frac{1}{2}c_i\delta_i^2$, with c_i as defined in condition (25). Then $C_i > 0$; and condition (24) is satisfied: $A_i^2 < B_iC_i$.

Notice first that $(\hat{\theta}_i,\hat{\delta}_i) = (\theta_i,\delta_i)$ is the unique stationary point of $U_i(\hat{\theta}_i,\hat{\delta}_i\,|\,\theta_i,\delta_i)$, since $\frac{\partial}{\partial\hat{\theta}_i}U_i(\hat{\theta}_i,\hat{\delta}_i\,|\,\theta_i,\delta_i) = 0 = \frac{\partial}{\partial\hat{\delta}_i}U_i(\hat{\theta}_i,\hat{\delta}_i\,|\,\theta_i,\delta_i)$ implies that $(\theta_i-\hat{\theta}_i) = -(\delta_i-\hat{\delta}_i)\frac{A_i}{B_i}$ and, thus, $0 = (\delta_i-\hat{\delta}_i)\frac{1}{B_i}(B_iC_i-A_i^2)$, where $B_i > 0$ and $B_iC_i - A_i^2 > 0$. Evaluating the Hessian $\mathcal{H}_i$ of $U_i(\hat{\theta}_i,\hat{\delta}_i\,|\,\theta_i,\delta_i)$ at $(\hat{\theta}_i,\hat{\delta}_i) = (\theta_i,\delta_i)$ yields

$$\mathcal{H}_i \;=\; \begin{pmatrix} -B_i & -A_i \\ -A_i & -C_i \end{pmatrix}.$$

The principal minors of $\mathcal{H}_i$, namely $-B_i < 0$ and $\det(\mathcal{H}_i) = B_iC_i - A_i^2 > 0$, are alternating in sign, with the first-order principal minor being negative. Hence, $\mathcal{H}_i$ is negative definite at (θ_i,δ_i), such that (θ_i,δ_i) is a local maximizer of $U_i(\hat{\theta}_i,\hat{\delta}_i\,|\,\theta_i,\delta_i)$.

It remains to show that truth-telling is indeed the (unique) global expected-utility maximizer of agent i. It suffices to show that $U_i(\hat{\theta}_i,\hat{\delta}_i\,|\,\theta_i,\delta_i)$ has no local maximizer on the boundary of $\Theta_i\times\Delta_i$.

Suppose a local maximizer is located on $(\theta_i^{\min},\theta_i^{\max})\times\{\delta_i^{\min}\}$ or $(\theta_i^{\min},\theta_i^{\max})\times\{\delta_i^{\max}\}$. As $U_i(\hat{\theta}_i,\hat{\delta}_i\,|\,\theta_i,\delta_i)$ is twice continuously partially differentiable, this maximizer, $(\hat{\theta}_i,\hat{\delta}_i)$, must satisfy $0 = \frac{\partial}{\partial\hat{\theta}_i}U_i(\hat{\theta}_i,\hat{\delta}_i\,|\,\theta_i,\delta_i)$ and, thus, $(\theta_i-\hat{\theta}_i) = -(\delta_i-\hat{\delta}_i)\frac{A_i}{B_i}$. Substituting the latter into (29) yields $\frac{\partial}{\partial\hat{\delta}_i}U_i(\hat{\theta}_i,\hat{\delta}_i\,|\,\theta_i,\delta_i) = (\delta_i-\hat{\delta}_i)\frac{1}{B_i}(B_iC_i-A_i^2)$. As $\frac{1}{B_i}(B_iC_i-A_i^2) > 0$, the reporting of $\hat{\delta}_i\in\{\delta_i^{\min},\delta_i^{\max}\}$ is not optimal, which contradicts the assumption. By a similar argument one can show that no local maximizer is located on $\{\theta_i^{\min}\}\times(\delta_i^{\min},\delta_i^{\max})$ or $\{\theta_i^{\max}\}\times(\delta_i^{\min},\delta_i^{\max})$. Hence, only the corners of $\Theta_i\times\Delta_i$ qualify as further maximizers.

First suppose $(\theta_i^{\max},\delta_i^{\max})$ is a local maximizer. Then $0\le\frac{\partial}{\partial\hat{\theta}_i}U_i(\theta_i^{\max},\delta_i^{\max}\,|\,\theta_i,\delta_i)$ and $0\le\frac{\partial}{\partial\hat{\delta}_i}U_i(\theta_i^{\max},\delta_i^{\max}\,|\,\theta_i,\delta_i)$ must hold. As $(\theta_i-\theta_i^{\max}),(\delta_i-\delta_i^{\max}) < 0$, while $B_i,C_i > 0$, this implies that $A_i < 0$. However, by (28) and (29), $(\delta_i-\delta_i^{\max}) \ge -(\theta_i-\theta_i^{\max})\frac{A_i}{C_i}$; thus,

$$0 \;\le\; (\theta_i-\theta_i^{\max})B_i+(\delta_i-\delta_i^{\max})A_i \;\le\; (\theta_i-\theta_i^{\max})\frac{1}{C_i}(B_iC_i-A_i^2) \;<\; 0.$$

Next suppose $(\theta_i^{\max}, \delta_i^{\min})$ is a local maximizer. Then $0 \leq \frac{\partial}{\partial \hat{\theta}_i} U_i(\theta_i^{\max}, \delta_i^{\min} \,|\, \theta_i, \delta_i)$ and $0 \geq \frac{\partial}{\partial \hat{\delta}_i} U_i(\theta_i^{\max}, \delta_i^{\min} \,|\, \theta_i, \delta_i)$ must hold. As $(\theta_i - \theta_i^{\max}) < 0$, while $(\delta_i - \delta_i^{\min}), B_i, C_i > 0$, this implies that $A_i > 0$. However, by (28) and (29), $(\theta_i - \theta_i^{\max}) \geq -(\delta_i - \delta_i^{\min})\frac{A_i}{B_i}$; thus,

$$0 \;\geq\; (\theta_i - \theta_i^{\max})A_i + (\delta_i - \delta_i^{\min})C_i \;\geq\; (\delta_i - \delta_i^{\min})\frac{1}{B_i}(B_iC_i - A_i^2) \;>\; 0.$$

Now suppose $(\theta_i^{\min}, \delta_i^{\min})$ is a local maximizer. Then $0 \geq \frac{\partial}{\partial \hat{\theta}_i} U_i(\theta_i^{\min}, \delta_i^{\min} \,|\, \theta_i, \delta_i)$ and $0 \geq \frac{\partial}{\partial \hat{\delta}_i} U_i(\theta_i^{\min}, \delta_i^{\min} \,|\, \theta_i, \delta_i)$ must hold. As $(\theta_i - \theta^{\min}), (\delta_i - \delta^{\min}), B_i, C_i > 0$, this implies that $A_i < 0$. However, by (28) and (29), $(\delta_i - \delta^{\min}) \leq -(\theta_i - \theta^{\min})\frac{A_i}{C_i}$; thus,

$$0 \;\geq\; (\theta_i - \theta_i^{\min})B_i + (\delta_i - \delta_i^{\min})A_i \;\geq\; (\theta_i - \theta_i^{\min})\frac{1}{C_i}(B_iC_i - A_i^2) \;>\; 0.$$

Finally, suppose $(\theta_i^{\min}, \delta_i^{\max})$ is a local maximizer. Then $0 \geq \frac{\partial}{\partial \hat{\theta}_i} U_i(\theta_i^{\min}, \delta_i^{\max} \,|\, \theta_i, \delta_i)$ and $0 \leq \frac{\partial}{\partial \hat{\delta}_i} U_i(\theta_i^{\min}, \delta_i^{\max} \,|\, \theta_i, \delta_i)$ must hold. As $(\delta_i - \delta_i^{\max}) < 0$ and $(\theta_i - \theta_i^{\min}), B_i, C_i > 0$, this implies that $A_i > 0$. However, by (28) and (29), $(\theta_i - \theta_i^{\min}) \leq -(\delta_i - \delta_i^{\max})\frac{A_i}{B_i}$; thus,

$$0 \;\leq\; (\theta_i - \theta_i^{\min})A_i + (\delta_i - \delta_i^{\max})C_i \;\leq\; (\delta_i - \delta_i^{\max})\frac{1}{B_i}(B_iC_i - A_i^2) \;<\; 0.$$

Altogether, (θ_i, δ_i) is the unique global maximizer of $U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i)$. As this is true for any set of type distributions, $T^\star$ *definitely* implements $k^\star$. ■

A.2.2 Derivation of $T^\star$ in the Proof of Theorem 2

Condition (18) in the proof of Lemma 1 states that transfers must satisfy

$$\bar{t}_{-i}(\theta_i, \delta_i) \;=\; \frac{dp_i(\delta_i)}{d\delta_i} - \bar{\pi}_{-i}(\theta_i, \delta_i) + \frac{\partial}{\partial \delta_i}\int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\, ds, \tag{30}$$

where $p_i : \Delta_i \to \mathbb{R}$ is some differentiable function. By conditions (15) and (30),

$$\begin{aligned} p_i(\delta_i) + \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\, ds \;&=\; U_i(\theta_i, \delta_i) \\ &=\; \bar{\pi}_i(\theta_i, \delta_i) + \bar{t}_i(\theta_i, \delta_i) + \delta_i\bar{\pi}_{-i}(\theta_i, \delta_i) + \delta_i\bar{t}_{-i}(\theta_i, \delta_i) \\ &=\; \bar{\pi}_i(\theta_i, \delta_i) + \bar{t}_i(\theta_i, \delta_i) + \delta_i\frac{dp_i(\delta_i)}{d\delta_i} + \delta_i\frac{\partial}{\partial \delta_i}\int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\, ds. \end{aligned}$$

Hence, transfers must also satisfy

$$\begin{aligned}(31)\qquad \bar{t}_i(\theta_i,\delta_i) &= p_i(\delta_i) - \delta_i \frac{dp_i(\delta_i)}{d\delta_i} - \bar{\pi}_i(\theta_i,\delta_i) \\ &\quad + \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s,\delta_i)\,ds - \delta_i \frac{\partial}{\partial \delta_i} \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s,\delta_i)\,ds.\end{aligned}$$

From identities (30) and (31), $T^\star$ can be guessed. The specific choice of p_i ensures that truth-telling is the unique optimal strategy for agent i. ■

A.3 On Theorem 3

A.3.1 Proof of Theorem 3

Be $k^\star$ as described. For agents $i \in \{1,2\}$ of commonly known social types $\delta = (\delta_1, \delta_2)$, define the functions $S_i : \Theta \times \Delta \to \mathbb{R}$ by

$$S_i(\hat{\theta},\delta) = \int_{\theta_i^{\min}}^{\hat{\theta}_i} v_i(k^\star(s,\hat{\theta}_{-i},\delta))\,ds - \pi_i(k^\star(\hat{\theta},\delta)\,|\,\hat{\theta}_i) - \delta_i \pi_{-i}(k^\star(\hat{\theta},\delta)\,|\,\hat{\theta}_{-i}).$$

Then the budget-balanced transfer scheme $T^\star$ defined by

$$\begin{aligned}t_i^\star(\hat{\theta},\delta) &= \frac{1}{1-\delta_i}\Big[S_i(\hat{\theta},\delta) - \mathbb{E}_{\theta_i}\big[S_i(\theta_i,\hat{\theta}_{-i},\delta)\big]\Big] \\ &\quad + \frac{1}{1-\delta_{-i}}\Big[-S_{-i}(\hat{\theta},\delta) + \mathbb{E}_{\theta_{-i}}\big[S_{-i}(\theta_{-i},\hat{\theta}_i,\delta)\big]\Big]\end{aligned}$$

definitely implements $k^\star$. To see this, notice first that the functions S_i and $T^\star$ satisfy

$$(32)\quad \mathbb{E}_{\theta_{-i}}\big[t_i^\star(\hat{\theta}_i,\theta_{-i},\delta) + \delta_i t_{-i}^\star(\hat{\theta}_i,\theta_{-i},\delta)\big] = \mathbb{E}_{\theta_{-i}}\big[S_i(\hat{\theta}_i,\theta_{-i},\delta)\big] - \mathbb{E}_{\theta_i,\theta_{-i}}\left[S_i(\theta_i,\theta_{-i},\delta)\right],$$

$$\begin{aligned}(33)\qquad \mathbb{E}_{\theta_{-i}}\big[S_i(\hat{\theta}_i,\theta_{-i},\delta)\big] &= \int_{\theta_i^{\min}}^{\hat{\theta}_i} \mathbb{E}_{\theta_{-i}}\big[v_i(k^\star(s,\theta_{-i},\delta))\big]\,ds \\ &\quad - \mathbb{E}_{\theta_{-i}}\big[\pi_i(k^\star(\hat{\theta}_i,\theta_{-i},\delta)\,|\,\hat{\theta}_i)\big] \\ &\quad - \delta_i \cdot \mathbb{E}_{\theta_{-i}}\big[\pi_{-i}(k^\star(\hat{\theta}_i,\theta_{-i},\delta)\,|\,\theta_{-i})\big].\end{aligned}$$

Suppose $-i$ reports her payoff type truthfully. By (32) and (33), agent i's interim-expected utility from reporting $\hat{\theta}_i$ satisfies

$$\begin{aligned}\mathbb{E}_{\theta_{-i}}\big[u_i(\,\cdot\,)\big] &= \mathbb{E}_{\theta_{-i}}\big[\pi_i(k^\star(\hat{\theta}_i,\theta_{-i},\delta)\,|\,\theta_i)\big] - \mathbb{E}_{\theta_{-i}}\big[\pi_i(k^\star(\hat{\theta}_i,\theta_{-i},\delta)\,|\,\hat{\theta}_i)\big] \\ &\quad -\mathbb{E}_{\theta_i,\theta_{-i}}\left[S_i(\theta_i,\theta_{-i},\delta)\right] + \int_{\theta_i^{\min}}^{\hat{\theta}_i} \mathbb{E}_{\theta_{-i}}\big[v_i(k^\star(s,\theta_{-i},\delta))\big]\,ds.\end{aligned}$$

Her marginal utility is given by $\frac{\partial}{\partial\hat{\theta}_i}\mathbb{E}_{\theta_{-i}}\big[u_i(\,\cdot\,)\big] = (\theta_i - \hat{\theta}_i)\cdot\frac{\partial}{\partial\hat{\theta}_i}\mathbb{E}_{\theta_{-i}}\Big[v_i(k^\star(\hat{\theta}_i,\theta_{-i},\delta))\Big]$, where the last factor is non-negative, as $v_i \geq 0$. Hence, truthful revelation is optimal for i. As this is true for any set of type distributions, $T^\star$ *definitely* implements $k^\star$. ■

A.3.2 Derivation of $T^\star$ in the Proof of Theorem 3

By condition (18) in the proof Lemma 1, transfers must be such that

$$U_i(\theta_i\,|\,\theta_i,\delta) = p_i(\delta) + \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s,\delta)\,ds \tag{34}$$

for some function $p_i : \Delta \to \mathbb{R}$. Thus,

$$\mathbb{E}_{\theta_2}\big[t_1\big] + \delta_1\mathbb{E}_{\theta_2}\big[t_2\big] = p_1(\delta) + \int_{\theta_1^{\min}}^{\theta_1} \bar{v}_1(s,\delta)\,ds - \mathbb{E}_{\theta_2}\big[\pi_1\big] - \delta_1\mathbb{E}_{\theta_2}\big[\pi_2\big], \tag{35}$$

$$\mathbb{E}_{\theta_1}\big[t_2\big] + \delta_2\mathbb{E}_{\theta_1}\big[t_1\big] = p_2(\delta) + \int_{\theta_2^{\min}}^{\theta_2} \bar{v}_2(s,\delta)\,ds - \mathbb{E}_{\theta_1}\big[\pi_2\big] - \delta_2\mathbb{E}_{\theta_1}\big[\pi_1\big]. \tag{36}$$

Due to budget balance, (35) and (36) imply that

$$\begin{aligned}(1-\delta_1)\mathbb{E}_{\theta_2}\big[t_1\big] &= p_1(\delta) + \int_{\theta_1^{\min}}^{\theta_1} \bar{v}_1(s,\delta)\,ds - \mathbb{E}_{\theta_2}\big[\pi_1\big] - \delta_1\mathbb{E}_{\theta_2}\big[\pi_2\big], \\ -(1-\delta_2)\mathbb{E}_{\theta_1}\big[t_1\big] &= p_2(\delta) + \int_{\theta_2^{\min}}^{\theta_2} \bar{v}_2(s,\delta)\,ds - \mathbb{E}_{\theta_1}\big[\pi_2\big] - \delta_2\mathbb{E}_{\theta_1}\big[\pi_1\big].\end{aligned}$$

From these conditions, $T^\star$ can be guessed. ■

References

Acemoglu, Daron and James A Robinson. 2006. *Economic origins of dictatorship and democracy*. Cambridge, UK: Cambridge University Press.

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.

Arrow, Kenneth J. 1950. "A difficulty in the concept of social welfare." *Journal of Political Economy* 58 (4):328–346.

———. 1973. "Some ordinalist-utilitarian notes on Rawls's theory of justice." *Journal of Philosophy* 70 (9):245–263.

———. 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.

Atkinson, Anthony B. 1970. "On the measurement of inequality." *Journal of Economic Theory* 2 (3):244–263.

Bartling, Björn and Nick Netzer. 2016. "An externality-robust auction: Theory and experimental evidence." *Games and Economic Behavior* 97:186–204.

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.

Bierbrauer, Felix, Axel Ockenfels, Andreas Pollak, and Désirée Rückert. 2017. "Robust mechanism design and social preferences." *Journal of Public Economics* 149:59–80.

Blanchet, Didier and Marc Fleurbaey. 2006. "Selfishness, altruism and normative principles in the economic analysis of social transfers." *Handbook of the Economics of Giving, Altruism and Reciprocity* 2:1465–1503.

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.

Daske, Thomas. 2020. "Efficient incentives in social networks: Gamification and the Coase theorem." *Mimeo*. Available under: http://hdl.handle.net/10419/222527.

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

Eden, Maya. 2020. "Welfare Analysis with Heterogeneous Risk Preferences." *Journal of Political Economy* 128 (12):4574–4613.

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

Fleurbaey, Marc. 2010. "Assessing risky social situations." *Journal of Political Economy* 118 (4):649–680.

Gibbard, Allan. 1977. "Manipulation of schemes that mix voting with chance." *Econometrica* 45 (3):665–681.

Grant, Simon, Atsushi Kajii, Ben Polak, and Zvi Safra. 2010. "Generalized utilitarianism and Harsanyi's impartial observer theorem." *Econometrica* 78 (6):1939–1971.

Harsanyi, John C. 1955. "Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility." *Journal of Political Economy* 63 (4):309–321.

———. 1977. "Morality and the theory of rational behavior." *Social Research* 44 (4):623–656.

Harsanyi, John C and Reinhard Selten. 1972. "A generalized Nash solution for two-person bargaining games with incomplete information." *Management Science* 18 (5-part-2):80–106.

Hayek, Friedrich A. 1945. "The use of knowledge in society." *American Economic Review* 35 (4):519–530.

Jehiel, Philippe, Moritz Meyer-ter Vehn, Benny Moldovanu, and William R 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.

Kalai, Ehud. 1977. "Proportional solutions to bargaining situations: Interpersonal utility comparisons." *Econometrica* 45 (7):1623–1630.

Kalai, Ehud and Meir Smorodinsky. 1975. "Other solutions to Nash's bargaining problem." *Econometrica* 43 (3):513–518.

Maskin, Eric. 1978. "A theorem on utilitarianism." *Review of Economic Studies* 45 (1):93–96.

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

Nash, John F. 1950. "The bargaining problem." *Econometrica* 18 (2):155–162.

Olson, Mancur. 1993. "Dictatorship, democracy, and development." *American Political Science Review* 87 (3):567–576.

Piacquadio, Paolo Giovanni. 2017. "A fairness justification of utilitarianism." *Econometrica* 85 (4):1261–1276.

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

Rabin, Matthew. 1993. "Incorporating fairness into game theory and economics." *American Economic Review* 83 (5):1281–1302.

Rawls, John B. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.

Saez, Emmanuel and Stefanie Stantcheva. 2016. "Generalized social marginal welfare weights for optimal tax theory." *American Economic Review* 106 (1):24–45.

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

Sen, Amartya. 1974. "Informational bases of alternative welfare approaches: Aggregation and income distribution." *Journal of Public Economics* 3 (4):387–403.

———. 1992. *Inequality reexamined*. Cambridge, UK: Oxford University Press.

Williams, Steven R and Roy Radner. 1988. "Informational externalities and the scope of efficient dominant strategy mechanisms." Discussion paper 761, Northwestern University. Available under: https://ideas.repec.org/p/nwu/cmsems/761.html.

Wilson, Robert. 1987. "Game-theoretic analyses of trading processes." In *Advances in Economic Theory: Fifth World Congress*, edited by T. Bewley, chap. 2. Cambridge U.K.: Cambridge University Press, 33–70.

Wintrobe, Ronald. 2000. *The political economy of dictatorship*. Cambridge, UK: Cambridge University Press.

Yaari, Menahem E and Maya Bar-Hillel. 1984. "On dividing justly." *Social Choice and Welfare* 1 (1):1–24.

Zik, Boaz. 2020. "Ex-post implementation with social preferences." *Social Choice and Welfare*. https://doi.org/10.1007/s00355-020-01291-x.

Supplement to "The Incentive Costs of Welfare Judgments": The n-Agents Case

Thomas Daske*

February 11, 2021

This supplement generalizes the main results, Theorems 1 and 4, obtained in "The Incentive Costs of Welfare Judgments" to groups of arbitrary size.

1 The Analytical Framework

There is a continuum $K \subset \mathbb{R}$ of social alternatives, bounded or unbounded, and there is a group $\mathcal{I} = \{1, \ldots, n\}$ of $n \geq 2$ agents. From alternative $k \in K$ and a monetary transfer $t_i \in \mathbb{R}$, agent i gains a *private payoff* $\Pi_i(k, t_i \,|\, \theta_i) = \theta_i v_i(k) + w_i(k) + t_i$, where the functions $v_i : K \to [0, \infty)$ and $w_i : K \to \mathbb{R}$ are twice continuously differentiable; furthermore, $dv_i/dk \neq 0$. Agent i's *payoff type* θ_i belongs to a closed (proper) interval $\Theta_i = [\theta_i^{\min}, \theta_i^{\max}]$. The collection of agents' payoff types is denoted by $\theta = (\theta_i, \theta_{-i})$, with $\theta_{-i} = (\theta_j)_{j \neq i}$. Agents exhibit interpersonal preferences in the form of altruism or spite: From the allocation of payoffs, 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, for $j \neq i$, belongs to $\Delta_{ij} = [\delta_{ij}^{\min}, \delta_{ij}^{\max}] \subset \left(\frac{-1}{n-1}, 1\right)$, while $\delta_{ii} = 1$. Refer to δ_{ij} as i's *degree of altruism towards* j, to the collection $\delta_i = (\delta_{ij})_{j \neq i}$ as i's *social type*, and to the pair (θ_i, δ_i) as i's *type*. Denote the collection of social types by $\delta = (\delta_i, \delta_{-i})$, with $\delta_{-i} = (\delta_j)_{j \neq i}$, and Cartesian products of type sets by $\Theta = \prod_i \Theta_i$, $\Delta_i = \prod_{j \neq i} \Delta_{ij}$, and $\Delta = \prod_i \Delta_i$. For convenience, define also $\pi_i(k \,|\, \theta_i) = \theta_i v_i(k) + w_i(k)$ and $u_i(k, \theta_{-i} \,|\, \theta_i, \delta_i) = \sum_{j \in \mathcal{I}} \delta_{ij} \, \pi_j(k \,|\, \theta_j)$.

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

Agents are privately informed about their payoff types and social types, all of which realize independently according to continuous, strictly positive densities.

A *direct revelation mechanism* $\langle k, T\rangle$ is defined by a *policy* $k : \Theta\times\Delta \to K$ and a *transfer scheme* $T = (t_i)_{i\in\mathcal{I}} : \Theta \times \Delta \to \mathbb{R}^n$.[1] Throughout, attention is restricted to transfer schemes that are partially differentiable on the social-type space Δ. The mechanism $\langle k, T\rangle$ is Bayesian *incentive-compatible* if $(\theta_i, \delta_i) \in \arg\max_{(\hat\theta_i,\hat\delta_i)} \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[u_i\big(k, t_i, t_{-i}, \theta_{-i} \,\big|\, \theta_i, \delta_i\big)\big]$ on $\Theta_i \times \Delta_i$ for all i, where k and $(t_j)_{j\in\mathcal{I}}$ are functions of $(\hat\theta_i, \hat\delta_i, \theta_{-i}, \delta_{-i})$. In this case, the policy k is said to be Bayesian *implementable*. The mechanism is *budget-balanced* if $\sum_{j\in\mathcal{I}} t_j = 0$ on $\Theta \times \Delta$.

As outlined in detail in "The Incentive Costs of Welfare Judgments," the analysis will focus on policies that conform with the following definitions:

Definition 1. (Generic Policies)
A policy $k^\star : \Theta\times\Delta \to K$ *is* generic *if it is partially differentiable and satisfies* $\partial k^\star/\partial\theta_i \neq 0$ *for all* (θ, δ) *and all* $i \in \mathcal{I}$.

Definition 2. (Definitely Implementable Policies)
A policy $k^\star : \Theta \times \Delta \to K$ *is* definitely implementable *if it is implementable for any set of type distributions.*

2 Incentive Compatibility

Characterizing the incentives costs of welfare judgments requires a clear understanding of what incentive compatibility means in the present framework. For this purpose, define

$$\begin{aligned}
\bar\pi_i(\hat\theta_i, \hat\delta_i \,|\, \theta_i) &= \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[\pi_i(k(\hat\theta_i, \hat\delta_i, \theta_{-i}, \delta_{-i}) \,|\, \theta_i)\big],\\
\bar\pi_{ij}(\theta_i, \delta_i) &= \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[\pi_j(k(\theta, \delta) \,|\, \theta_j)\big],\\
\bar t_{ij}(\theta_i, \delta_i) &= \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[t_j(\theta, \delta)\big],\\
\bar v_i(\theta_i, \delta_i) &= \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[v_i(k(\theta, \delta))\big],
\end{aligned}$$

for all $i, j \in \mathcal{I}$. 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) and if all the other agents report their types truthfully:

[1] By the *revelation principle*, which applies to the present setup (Myerson, 1979), there is no loss of generality in identifying message sets, from which agents draw their reports, with agents' type sets.

$$U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) \;=\; \bar{\pi}_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i) + \bar{t}_{ii}(\hat{\theta}_i, \hat{\delta}_i) + \sum_{j \neq i} \delta_{ij} \left[\bar{\pi}_{ij}(\hat{\theta}_i, \hat{\delta}_i) + \bar{t}_{ij}(\hat{\theta}_i, \hat{\delta}_i) \right].$$

For convenience, define also $U_i(\theta_i, \delta_i) = U_i(\theta_i, \delta_i \,|\, \theta_i, \delta_i)$.

Lemma 1. *A partially differentiable policy $k : \Theta \times \Delta \to K$ is Bayesian implementable only if it satisfies the following conditions for all $i \in \mathcal{I}$ and all $j \in \mathcal{I} \setminus \{i\}$:*

(i) *$\bar{v}_i(\theta_i, \delta_i)$ is non-decreasing in θ_i.*

(ii) *$\bar{\pi}_{ij}(\theta_i, \delta_i) + \bar{t}_{ij}(\theta_i, \delta_i)$ is non-decreasing in δ_{ij}.*

(iii) *There exist a twice partially differentiable function $p_i : \Delta_i \to \mathbb{R}$ and functions $q_{ij} : \Theta_i \times \prod_{\ell \neq i,j} \Delta_{i\ell} \to \mathbb{R}$, partially differentiable in θ_i, such that, for $\delta_i^{-j} = (\delta_{i\ell})_{\ell \neq i,j}$,*

$$U_i(\theta_i, \delta_i) \;=\; p_i(\delta_i) + \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(r, \delta_i)\, dr \tag{1}$$

$$\;=\; q_{ij}(\theta_i, \delta_i^{-j}) + \int_{\delta_{ij}^{\min}}^{\delta_{ij}} \left[\bar{\pi}_{ij}(\theta_i, r, \delta_i^{-j}) + \bar{t}_{ij}(\theta_i, r, \delta_i^{-j}) \right] dr. \tag{2}$$

Proof. Suppose the mechanism $\langle k, T \rangle$ is incentive-compatible. Then the following must hold for all $\theta_i, \hat{\theta}_i \in \Theta_i$ and all $\delta_i, \hat{\delta}_i \in \Delta_i$:

$$U_i(\theta_i, \delta_i) \;\geq\; U_i(\hat{\theta}_i, \delta_i \,|\, \theta_i, \delta_i) \;=\; U_i(\hat{\theta}_i, \delta_i) + (\theta_i - \hat{\theta}_i)\bar{v}_i(\hat{\theta}_i, \delta_i), \tag{3}$$

$$U_i(\hat{\theta}_i, \delta_i) \;\geq\; U_i(\theta_i, \delta_i \,|\, \hat{\theta}_i, \delta_i) \;=\; U_i(\theta_i, \delta_i) + (\hat{\theta}_i - \theta_i)\bar{v}_i(\theta_i, \delta_i), \tag{4}$$

$$\begin{aligned} U_i(\theta_i, \delta_i) \;&\geq\; U_i(\theta_i, \hat{\delta}_{ij}, \delta_i^{-j} \,|\, \theta_i, \delta_i) \\ &=\; U_i(\theta_i, \hat{\delta}_{ij}, \delta_i^{-j}) + (\delta_{ij} - \hat{\delta}_{ij}) \left[\bar{\pi}_{ij}(\theta_i, \hat{\delta}_i, \delta_i^{-j}) + \bar{t}_{ij}(\theta_i, \hat{\delta}_i, \delta_i^{-j}) \right], \end{aligned} \tag{5}$$

$$\begin{aligned} U_i(\theta_i, \hat{\delta}_{ij}, \delta_i^{-j}) \;&\geq\; U_i(\theta_i, \delta_i \,|\, \theta_i, \hat{\delta}_i, \delta_i^{-j}) \\ &=\; U_i(\theta_i, \delta_i) + (\hat{\delta}_{ij} - \delta_{ij}) \left[\bar{\pi}_{ij}(\theta_i, \delta_i) + \bar{t}_{ij}(\theta_i, \delta_i) \right]. \end{aligned} \tag{6}$$

Without loss of generality, let $\hat{\theta}_i > \theta_i$. Then (3) and (4) imply that

$$\bar{v}_i(\hat{\theta}_i, \delta_i) \;\geq\; \frac{U_i(\hat{\theta}_i, \delta_i) - U_i(\theta_i, \delta_i)}{\hat{\theta}_i - \theta_i} \;\geq\; \bar{v}_i(\theta_i, \delta_i). \tag{7}$$

Hence, $\bar{v}_i(\theta_i, \delta_i)$ is non-decreasing in θ_i. As $\bar{v}_i$ is continuous on Θ_i, letting $\hat{\theta}_i$ approach θ_i yields $\partial U_i(\theta_i, \delta_i)/\partial\theta_i = \bar{v}_i(\theta_i, \delta_i)$. Integrating the latter with respect to θ_i yields the condition (1) for some function $p_i : \Delta_i \to \mathbb{R}$. Similarly, let $\hat{\delta}_i > \delta_i$. By (5) and (6),

$$(8) \qquad \begin{aligned} \bar{\pi}_{ij}(\theta_i, \hat{\delta}_{ij}, \delta_i^{-j}) + \bar{t}_{ij}(\theta_i, \hat{\delta}_{ij}, \delta_i^{-j}) &\geq \frac{U_i(\theta_i, \hat{\delta}_{ij}, \delta_i^{-j}) - U_i(\theta_i, \delta_i)}{\hat{\delta}_{ij} - \delta_{ij}} \\ &\geq \bar{\pi}_{ij}(\theta_i, \delta_i) + \bar{t}_{ij}(\theta_i, \delta_i). \end{aligned}$$

Hence, $\bar{\pi}_{ij}(\theta_i, \delta_i) + \bar{t}_{ij}(\theta_i, \delta_i)$ is non-decreasing in δ_{ij} for each $j \neq i$. Letting $\hat{\delta}_{ij}$ approach δ_{ij} in (8) implies that (2) must hold for some function $q_{ij} : \Theta_i \times \prod_{\ell \neq i,j} \Delta_{i\ell} \to \mathbb{R}$. By comparison of (1) and (2), p_i and q_{ij} are partially differentiable in δ_i and θ_i, respectively. As k and T are assumed partially differentiable on Δ, while each π_i is differentiable in k, the functions p_i are twice partially differentiable. ■

Conditions (i) and (1) are the well-known results for Bayesian incentive compatibility in linear settings with independent valuations (Myerson, 1981), and conditions (ii) and (2) are their social-preference equivalents. Condition (2) implies in particular that an incentive-compatible mechanism must take into account that agents internalize its distributive effects; interim-expected transfers are thus linked.

It will prove useful to fully characterize incentive compatible mechanisms for policies that are social-type independent: $k : \Theta \to K$. Contrary to linear settings with independent valuations, where conditions (i) and (1) are even sufficient (Myerson, 1981), the conditions of Lemma 1 do not yet ensure incentive compatibility:

Lemma 2. *A partially differentiable policy $k : \Theta \to K$ is Bayesian implementable if and only if it satisfies the following conditions for all $i \in \mathcal{I}$ and all $j \in \mathcal{I} \setminus \{i\}$:*

(i) $\bar{v}_i(\theta_i)$ is non-decreasing in θ_i.

(ii) There exists a twice partially differentiable function $p_i : \Delta_i \to \mathbb{R}$ such that, for all $\theta_i \in \Theta_i$ and all $\delta_i, \hat{\delta}_i \in \Delta_i$,

$$(9) \qquad \bar{t}_{ij}(\theta_i, \delta_i) = \frac{\partial p_i(\delta_i)}{\partial \delta_{ij}} - \bar{\pi}_{ij}(\theta_i), \quad \text{for } j \neq i,$$

$$(10) \qquad \bar{t}_{ii}(\theta_i, \delta_i) = p_i(\delta_i) - \nabla p_i(\delta_i) \cdot \delta_i - \bar{\pi}_{ii}(\theta_i) + \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(r)\, dr,$$

$$(11) \qquad p_i(\delta_i) - p_i(\hat{\delta}_i) \geq \nabla p_i(\hat{\delta}_i) \cdot (\delta_i - \hat{\delta}_i).$$

Proof. Suppose the mechanism $\langle k, T\rangle$, with $k : \Theta \to K$, is incentive-compatible. Condition (i) simply restates Lemma 1(i). By Lemma 1(iii), when differentiating with respect to δ_{ij}, there exists a partially differentiable function $p_i : \Delta_i \to \mathbb{R}$ such that $\partial p_i(\delta_i)/\partial\delta_{ij} = \bar{\pi}_{ij}(\theta_i) + \bar{t}_{ij}(\theta_i, \delta_i)$, which yields condition (9). Due to (1) and (9),

$$\begin{aligned} p_i(\delta_i) + \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(r)\, dr &= \bar{\pi}_{ii}(\theta_i) + \bar{t}_{ii}(\theta_i, \delta_i) + \sum_{j\neq i} \delta_{ij} \left[\bar{\pi}_{ij}(\theta_i) + \bar{t}_{ij}(\theta_i, \delta_i)\right] \\ &= \bar{\pi}_{ii}(\theta_i) + \bar{t}_{ii}(\theta_i, \delta_i) + \nabla p_i(\delta_i) \cdot \delta_i, \end{aligned}$$

which yields condition (10). If conditions (9) and (10) hold, then

$$\begin{aligned} (12) \qquad U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) &= \bar{\pi}_i(\hat{\theta}_i \,|\, \theta_i) + \bar{t}_{ii}(\hat{\theta}_i, \hat{\delta}_i) + \sum_{j\neq i} \delta_{ij} \left[\bar{\pi}_{ij}(\hat{\theta}_i) + \bar{t}_{ij}(\hat{\theta}_i, \hat{\delta}_i)\right] \\ &= (\theta_i - \hat{\theta}_i)\bar{v}_i(\hat{\theta}_i) + \int_{\theta_i^{\min}}^{\hat{\theta}_i} \bar{v}_i(r)\, dr + p_i(\hat{\delta}_i) + \nabla p_i(\hat{\delta}_i) \cdot (\delta_i - \hat{\delta}_i), \end{aligned}$$

implying in particular that i's interim-expected utility is additively separable in her payoff type and social type. Hence, if (12) holds, then $U_i(\hat{\theta}_i, \delta_i \,|\, \theta_i, \delta_i) \geq U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i)$ holds for all $\theta_i, \hat{\theta}_i, \delta_i, \hat{\delta}_i$ if and only if $p_i(\hat{\delta}_i) + \nabla p_i(\hat{\delta}_i) \cdot (\delta_i - \hat{\delta}_i) \leq p_i(\delta_i)$ holds for all $\delta_i, \hat{\delta}_i$. That is, with respect to *social* types, condition (ii) is necessary and sufficient for incentive compatibility. On the other hand, if (12) holds while $\bar{v}_i(\hat{\theta}_i)$ is non-decreasing, then $U_i(\theta_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) \geq U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i)$ does hold for all $\theta_i, \hat{\theta}_i, \delta_i, \hat{\delta}_i$. Hence, with respect to *payoff* types, conditions (i), (9), and (10) are necessary and sufficient for incentive compatibility. ■

By Lemma 2, incentive compatibility requires interim-expected transfers to be additively separable in an agent's payoff type and social type. Conditions (i), (9), and (10) follow from Lemma 1; they ensure incentive compatibility with respect to payoff types. The distinctive feature of Lemma 2 is condition (11), which ensures incentive compatibility with respect to social types. When choosing p_i affine, such that interim-expected transfers are social-type independent, the necessary conditions of Lemma 1 are already sufficient.

By Condition (9), the private payoff that an agent i interim expects for every other j must be independent of i's payoff type: $\mathbb{E}_{\theta_{-i}}\big[\,\Pi_j(k(\theta), t_j(\theta) \,|\, \theta_j)\,\big] = \partial p_i(\delta_i)/\partial\delta_{ij}$, such that $U_i(\hat{\theta}_i, \hat{\delta}_i \,|\, \theta_i, \delta_i) = \bar{\pi}_i(\hat{\theta}_i \,|\, \theta_i) + \bar{t}_{ii}(\hat{\theta}_i, \hat{\delta}_i) + \nabla p_i(\hat{\delta}_i) \cdot \delta_i$ for all $(\theta_i, \delta_i), (\hat{\theta}_i, \hat{\delta}_i)$. That is,

with respect to the revelation of their *payoff types*, agents must be incentivized to behave *as if* they were selfish. In other words, the mechanism must be *social-preference robust*.

Bierbrauer and Netzer (2016) coined this property the 'insurance property,' as it ensures agents against the other-regarding concerns of other agents. They discuss the 'insurance property' in the range of asymmetric information about agents' intention-based social preferences, but they also observe that (some of) their results apply to models of outcome-based social preferences, such as altruism and spite. For arbitrary network size, they prove that a mechanism is incentive-compatible if it both has the insurance property and is incentive-compatible among selfish agents; this result is obtained from Lemma 2 above when choosing p_i affine. For dyads, they provide conditions under which the insurance property is also necessary; in the present model, for policies $k : \Theta \to K$, the insurance property is always necessary.

3 Justifying Materialistic Utilitarianism

This Section generalizes Theorem 1 in "The Incentive Costs of Welfare Judgments" to groups of arbitrary size.

The conditions of Lemma 2(ii), when confronting them with the requirement of budget balance, suggest to focusing on the following class of mechanisms. These mechanisms coincide with the "AGV-type" mechanisms in "The Incentive Costs of Welfare Judgments" if $n = 2$. (The concept builds on an early draft of Daske, 2020.)

Definition 3. (Social-Preference Compatible Mechanisms; SPC-Mechanisms) *An* SPC-mechanism $\langle k^\star, T^\star \rangle$ *is given by the ex-post materially efficient policy* $k^\star(\theta) \in \arg\max_{k \in K} \sum_{i \in \mathcal{I}} \pi_i(k \,|\, \theta_i)$ *and a budget-balanced transfer scheme* $T^\star = (t_i^\star)_{i \in \mathcal{I}}$ *of the following form: For reported types* $(\hat{\theta}, \hat{\delta})$*,*

$$t_i^\star(\hat{\theta}, \hat{\delta}) \;=\; \sum_{\ell \neq i} \Big[\mathbb{E}_{\theta_{-i}} \big[\pi_\ell(k^\star(\hat{\theta}_i, \theta_{-i}) \,|\, \theta_\ell) \big] - \mathbb{E}_{\theta_{-\ell}} \big[\pi_i(k^\star(\hat{\theta}_\ell, \theta_{-\ell}) \,|\, \theta_i) \big] \Big] + s_i(\hat{\theta}, \hat{\delta}),$$

where the components $s_i : \Theta \times \Delta \to \mathbb{R}$ *satisfy the following conditions for all* $i, j \in \mathcal{I}$*:*

(i) $\sum_{i \in \mathcal{I}} s_i(\theta, \delta) = 0$ *on* $\Theta \times \Delta$*.*

(ii) $\mathbb{E}_{\theta_{-i}, \delta_{-i}}[s_j(\theta, \delta)]$ *is independent of* θ_i *and, thus, independent of* $k^\star$*.*

(iii) $\left(\mathbb{E}_{\theta_{-i},\delta_{-i}}[s_j(\theta,\delta)]\right)_{j\neq i} = \nabla p_i(\delta_i)$ *on* Δ_i*, for some partially differentiable function* $p_i : \Delta_i \to \mathbb{R}$ *satisfying* $p_i(\delta_i) + \nabla p_i(\delta_i)\cdot(\mathbf{1}-\delta_i) = 0$ *and* $p_i(\delta_i) - p_i(\hat{\delta}_i) \geq \nabla p_i(\hat{\delta}_i)\cdot(\delta_i - \hat{\delta}_i)$ *for all* $\delta_i, \hat{\delta}_i \in \Delta_i$.[2]

SPC-mechanisms apply the principle of the dyadical AGV-mechanism of Arrow (1979), and d'Aspremont and Gérard-Varet (1979) to groups of arbitrary size:[3] Leaving the $(s_i)_i$ aside, each agent i pays to every other j the money equivalent of what j believes to contribute to i's material well-being when reporting, or acting according to, her payoff type θ_j. As i herself thus receives $\mathbb{E}_{\theta_{-i}}\left[\pi_{-i}(k^\star(\hat{\theta}_i,\theta_{-i})\,|\,\theta_{-i})\right]$ from j, SPC-mechanisms render i's degree of altruism towards j strategically irrelevant, since

$$\mathbb{E}_{\theta_{-i},\delta_{-i}}\left[\pi_{-i}(k^\star(\hat{\theta}_i,\theta_{-i})\,|\,\theta_{-i}) + t^\star_{-i}(\hat{\theta}_i,\theta_{-i})\right] \;=\; \mathbb{E}_{\theta}\left[\pi_i(k^\star(\theta)\,|\,\theta_i)\right].$$

That is, SPC-mechanisms can be social-preference robust.

Now consider the incentives to reveal social preferences, which are fully determined by the appropriate choice of partially differentiable functions $p_i : \Delta_i \to \mathbb{R}$. By Lemma 1, these functions determine agents' interim-expected utilities from the residual transfer scheme $(s_i)_i$. The inequality of Definition 3(iii) implies the convexity of p_i, and any convex *continuously* partially differentiable function satisfies this inequality. The first-order partial differential equation can be interpreted in terms of Euler's theorem for homogeneous functions: This condition states that the transform $\tilde{p}_i(\delta_i) = p_i(\mathbf{1}-\delta_i)$ satisfies $\tilde{p}_i(\delta_i) - \nabla\tilde{p}_i(\delta_i)\cdot\delta_i = 0$, which is equivalent to $\tilde{p}_i$ being homogeneous of degree one if p_i is differentiable.[4] In Daske (2020), I have shown how the residual transfer scheme $(s_i)_i$ can be chosen so as to satisfy agents' interim participation constraints, which are not of interest in the present study.

Theorem 1. *A generic policy* $k^\star : \Theta\times\Delta \to K$ *is* definitely *implementable at zero incentive costs if and only if it is consistent with materialistic utilitarianism:* $k^\star(\theta) =$

[2] **Notation:** $\mathbf{1} = (1,\ldots,1) \in \mathbb{R}^{n-1}$.

[3] SPC-mechanisms belong to the class of *expected-externality mechanisms*, defined by the materially efficient policy $k^\star$ and transfers $t_i(\hat{\theta}) = \sum_{\ell\neq i}\mathbb{E}_{\theta_{-i}}\left[\pi_\ell(k^\star(\hat{\theta}_i,\theta_{-i})\,|\,\theta_\ell)\right] + h_i(\hat{\theta}_{-i})$, where the $h_i : \Theta_{-i} \to \mathbb{R}$ are arbitrary functions. Notice that the AGV-mechanism, defined through $h_i(\hat{\theta}_{-i}) = \frac{-1}{n-1}\sum_{j\neq i}\sum_{\ell\neq j}\mathbb{E}_{\theta_{-j}}\left[\pi_\ell(k^\star(\hat{\theta}_j,\theta_{-j})\,|\,\theta_\ell)\right]$, is social-preference compatible *if and only if* $n = 2$. While the AGV subsidizes or sanctions the *average* externalities that an agent imposes on the rest of the group, SPC-mechanisms treat interpersonal externalities on the *bilateral* level.

[4] A non-trivial example of a function satisfying these two conditions is given by $p_i(\delta_i) = ||\mathbf{1}-\delta_i||$, where $||\cdot|| : \mathbb{R}^{n-1} \to [0,\infty)$ is a continuously partially differentiable norm.

$\arg\max_{k\in K}\sum_{i\in\mathcal{I}}\pi_i(k\,|\,\theta_i)$. *The respective mechanisms are necessarily of SPC-type. The resulting allocations are ex-post Pareto-efficient if agents are moderately altruistic or spiteful:* $|\delta_{ij}|<\frac{1}{2n-3}$ *for all* i *and all* $j\neq i$.[5]

The following three Propositions give proof of Theorem 1. The sufficiency part is to be addressed first.

Proposition 1. *The materially efficient policy* $k^\star(\theta)=\arg\max_{k\in K}\sum_{i\in\mathcal{I}}\pi_i(k\,|\,\theta_i)$ *is* definitely *implementable through SPC-transfers. SPC-mechanisms are ex-post Pareto-efficient if* $|\delta_{ij}|<\frac{1}{2n-3}$ *for all* i *and all* $j\neq i$.

Proof. The proof is the same as in Daske (2020, Lemma 1 and Proposition 1); it is restated here for the sake of completeness. Suppose the agents other than i reveal their types truthfully. Then, $\bar{t}_{ii}(\hat{\theta}_i,\hat{\delta}_i)=\sum_{\ell\neq i}\mathbb{E}_{\theta_{-i}}\big[\pi_\ell(k^\star(\hat{\theta}_i,\theta_{-i})\,|\,\theta_\ell)\big]-(n-1)\,\mathbb{E}_\theta\big[\pi_i(k^\star(\theta)\,|\,\theta_i)\big]-\nabla p_i(\hat{\delta}_i)\cdot\mathbf{1}$ and

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

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

$$\begin{aligned}(13)\;\;U_i(\hat{\theta}_i,\hat{\delta}_i\,|\,\theta_i,\delta_i)\;&=\;\mathbb{E}_{\theta_{-i}}\left[\sum_{\ell\in\mathcal{I}}\pi_\ell(k^\star(\hat{\theta}_i,\theta_{-i})\,|\,\theta_\ell)\right]+\left(\sum_{j\neq i}\delta_{ij}\right)\mathbb{E}_\theta\left[\sum_{\ell\in\mathcal{I}}\pi_\ell(k^\star(\theta)\,|\,\theta_\ell)\right]\\&\qquad-(n-1)\,\mathbb{E}_\theta\left[\sum_{j\in\mathcal{I}}\delta_{ij}\pi_j(k^\star(\theta)\,|\,\theta_j)\right]-\nabla p_i(\hat{\delta}_i)\cdot(\mathbf{1}-\delta_i).\end{aligned}$$

If truthful revelation of θ_i was inferior for i, then there would exist $\hat{\theta}_i$ and θ such that $\sum_{\ell\in\mathcal{I}}\pi_\ell(k^\star(\hat{\theta}_i,\theta_{-i})\,|\,\theta_\ell)>\sum_{\ell\in\mathcal{I}}\pi_\ell(k^\star(\theta_i,\theta_{-i})\,|\,\theta_\ell)$, which contradicts the definition of $k^\star$. If truthful revelation of δ_i was inferior for i, then there would exist $\hat{\delta}_i$ such that $\nabla p_i(\hat{\delta}_i)\cdot$

[5] Constraining the δ_{ij} guarantees that the Pareto frontier is definite; otherwise, for instance, a coalition of agents might be willing to transfer arbitrary amounts of money to their joint favorite agent.

$(\mathbf{1}-\delta_i) < \nabla p_i(\delta_i)\cdot(\mathbf{1}-\delta_i)$; but since p_i satisfies $p_i(\delta_i)+\nabla p_i(\delta_i)\cdot(\mathbf{1}-\delta_i)=0$, this would imply that $0 > \nabla p_i(\hat{\delta}_i)\cdot(\mathbf{1}-\delta_i)-\nabla p_i(\delta_i)\cdot(\mathbf{1}-\delta_i) = \nabla p_i(\hat{\delta}_i)\cdot(\mathbf{1}-\delta_i)+p_i(\delta_i)-p_i(\hat{\delta}_i)-\nabla p_i(\hat{\delta}_i)\cdot(\mathbf{1}-\hat{\delta}_i) = \nabla p_i(\hat{\delta}_i)\cdot(\hat{\delta}_i-\delta_i)+p_i(\delta_i)-p_i(\hat{\delta}_i)$ and, thus, $\nabla p_i(\hat{\delta}_i)\cdot(\delta_i-\hat{\delta}_i) > p_i(\delta_i)-p_i(\hat{\delta}_i)$; a contradiction to the assumption that $p_i(\delta_i)-p_i(\hat{\delta}_i) \geq \nabla p_i(\hat{\delta}_i)\cdot(\delta_i-\hat{\delta}_i)$.

For the result on Pareto efficiency, assume in the following that $|\delta_{ij}| < \frac{1}{2n-3}$ for all i and all $j \neq i$. Suppose that, for given transfers $(t_i)_{i\in\mathcal{I}}$, there exists a social alternative $k^\circ(\theta)$ that Pareto-dominates $k^\star(\theta) \in \arg\max_{k\in K} \sum_{i\in\mathcal{I}} \pi_i(k \,|\, \theta_i)$. Then there must exist agents i who make strict material losses when switching from $k^\star$ to k°: $\pi_i(k^\circ \,|\, \theta_i) - \pi_i(k^\star \,|\, \theta_i) = -\epsilon_i < 0$. Be $i^\star$ one of the agents for whom this material loss is largest. Then $i^\star$ is *not* worse off under k° than under $k^\star$ if and only if she is *mentally compensated* through the distributive effects on all the others: $\sum_{j\neq i^\star} \delta_{i^\star j}\big[\pi_j(k^\circ \,|\, \theta_j) - \pi_j(k^\star \,|\, \theta_j)\big] \geq \epsilon_{i^\star}$.

First suppose $\delta_{i^\star j} \leq 0$ for all $j \neq i^\star$. Then $i^\star$ obtains the *maximum mental compensation* feasible if each $j \neq i^\star$ also 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 $0 \geq \delta_{i^\star j} > \frac{-1}{2n-3} \geq \frac{-1}{n-1}$.

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 mental 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, $\sum_{j\neq i^\star} \delta_{i^\star j}\big[\pi_j(k^\circ \,|\, \theta_j) - \pi_j(k^\star \,|\, \theta_j)\big] = \sum_{j\neq i^\star, j^\star} \delta_{i^\star j}(-\epsilon_{i^\star}) + \delta_{i^\star j^\star}(n-1)\,\epsilon_{i^\star} < \frac{n-2}{2n-3}\,\epsilon_{i^\star} + \frac{n-1}{2n-3}\,\epsilon_{i^\star} = \epsilon_{i^\star}$, since $|\delta_{i^\star j}| < \frac{1}{2n-3}$ for all $j \neq i^\star$.

Hence, agent $i^\star$ is worse off under k° than under $k^\star$, implying that $k^\star$ is Pareto-efficient. The reasoning is exactly the same when showing that, for any fixed social alternative k, no ex-post budget-balanced transfer scheme Pareto-dominates another, implying that SPC-mechanisms, which are ex-post budget-balanced, are ex-post Pareto-efficient under the condition imposed on $(\delta_{ij})_{i,j\neq i}$. ■

The next two Propositions establish the necessity part. A Lemma eases the exposition.

Lemma 3. *A partially differentiable policy $k^\star : \Theta \times \Delta \to K$ is* definitely *implementable through budget-balanced transfers only if it satisfies the following conditions for all $i \in \mathcal{I}$, all $\ell \in \mathcal{I} \setminus \{i\}$, and all (θ, δ):*

$$\left[\sum_{j\in\mathcal{I}} \frac{d\pi_j(k^\star \,|\, \theta_j)}{dk}\right] \frac{\partial k^\star}{\partial \theta_i} = \sum_{j\in\mathcal{I}\setminus\{i\}} (1-\delta_{ij}) \frac{\partial v_i(k^\star)}{\partial \delta_{ij}}, \tag{14}$$

$$\begin{aligned}\left[\sum_{j\in\mathcal{I}} \frac{d\pi_j(k^\star \,|\, \theta_j)}{dk}\right] \frac{\partial k^\star}{\partial \delta_{i\ell}} &= \sum_{j\in\mathcal{I}\setminus\{i\}} (1-\delta_{ij}) \frac{\partial^2}{\partial\delta_{i\ell}\partial\delta_{ij}} \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s,\theta_{-i},\delta))\,ds \\ &\quad - \frac{\partial}{\partial\delta_{i\ell}} \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s,\theta_{-i},\delta))\,ds \\ &\quad - \frac{\partial p_i(\delta_i)}{\partial\delta_{i\ell}} + \sum_{j\in\mathcal{I}\setminus\{i\}} (1-\delta_{ij}) \frac{\partial^2 p_i(\delta_i)}{\partial\delta_{i\ell}\partial\delta_{ij}},\end{aligned} \tag{15}$$

for a twice partially differentiable function $p_i : \Delta_i \to \mathbb{R}$.

Proof. Differentiating equation (2) of Lemma 1 with respect to δ_{ij} yields

$$\bar{t}_{ij}(\theta_i, \delta_i) = \frac{\partial p_i(\delta_i)}{\partial \delta_{ij}} - \bar{\pi}_{ij}(\theta_i, \delta_i) + \frac{\partial}{\partial \delta_{ij}} \int_{\theta_i^{\min}}^{\theta_i} \bar{v}_i(s, \delta_i)\,ds. \tag{16}$$

Budget balance requires in particular that $\bar{t}_{ii}(\theta_i, \delta_i) = -\sum_{j\neq i} \bar{t}_{ij}(\theta_i, \delta_i)$ on $\Theta_i \times \Delta_i$, such that truthful revelation, $(\hat{\theta}_i, \hat{\delta}_i) = (\theta_i, \delta_i)$, is incentive-compatible for agent i only if the following first-order condition with respect to $\hat{\theta}_i$ is satisfied:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \hat{\theta}_i} \left[\bar{\pi}_i(\hat{\theta}_i, \delta_i \,|\, \theta_i) + \sum_{j\neq i} \delta_{ij} \bar{\pi}_{ij}(\hat{\theta}_i, \delta_i) - \sum_{j\neq i} (1-\delta_{ij}) \bar{t}_{ij}(\hat{\theta}_i, \delta_i) \right]_{\hat{\theta}_i = \theta_i} \\ &= \mathbb{E}_{\theta_{-i}, \delta_{-i}} \left[\sum_{j\in\mathcal{I}} \frac{d\pi_j(k^\star(\theta,\delta) \,|\, \theta_j)}{dk} \frac{\partial k^\star}{\partial \theta_i} - \sum_{j\in\mathcal{I}} (1-\delta_{ij}) \frac{\partial v_i(k^\star(\theta,\delta))}{\partial \delta_{ij}} \right], \end{aligned} \tag{17}$$

where the second equality is implied by (16). As $k^\star$ is supposed to be *definitely* implementable, (17) must hold for arbitrary type distributions. As the argument of $\mathbb{E}_{\theta_{-i},\delta_{-i}}[\,\cdot\,]$ is continuous in $(\theta_{-i}, \delta_{-i})$, the policy $k^\star$ must satisfy condition (14).

Similarly, the first-order conditions with respect to $\hat{\delta}_{i\ell}$ must be satisfied for every $\ell \neq i$:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \hat{\delta}_{i\ell}} \left[\bar{\pi}_i(\hat{\theta}_i, \delta_i \,|\, \theta_i) + \sum_{j \neq i} \delta_{ij} \bar{\pi}_{ij}(\hat{\theta}_i, \delta_i) - \sum_{j \neq i} (1 - \delta_{ij}) \bar{t}_{ij}(\hat{\theta}_i, \delta_i) \right]_{\hat{\delta}_{i\ell} = \delta_{i\ell}} \\ &= \mathbb{E}_{\theta_{-i}, \delta_{-i}} \Bigg[\sum_{j \in \mathcal{I}} \frac{d\pi_j(k^\star(\theta, \delta) \,|\, \theta_j)}{dk} \frac{\partial k^\star}{\partial \delta_{i\ell}} - \sum_{j \in \mathcal{I} \setminus \{i\}} (1 - \delta_{ij}) \frac{\partial^2 p_i(\delta_i)}{\partial \delta_{i\ell} \partial \delta_{ij}} + \frac{\partial p_i(\delta_i)}{\partial \delta_{i\ell}} \\ &\qquad - \sum_{j \in \mathcal{I} \setminus \{i\}} (1 - \delta_{ij}) \frac{\partial^2}{\partial \delta_{i\ell} \partial \delta_{ij}} \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s, \theta_{-i}, \delta))\, ds \\ &\qquad + \frac{\partial}{\partial \delta_{i\ell}} \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s, \theta_{-i}, \delta))\, ds \Bigg], \end{aligned}$$

where the second equality is implied by (16). As $k^\star$ is supposed to be *definitely* implementable, $k^\star$ must satisfy condition (15). ■

Proposition 2. *A partially differentiable policy $k^\star : \Theta \times \Delta \to K$ is* definitely *implementable through budget-balanced transfers only if it is social-preference independent.*

Proof. By Lemma 3, the policy $k^\star$ must satisfy conditions (14) and (15). Integrating (14) with respect to θ_i, while integrating $d\pi_i(k^\star \,|\, \theta_i)/dk$ by parts,[6] yields

$$\begin{aligned} &C_i(\theta_{-i}, \delta) + \sum_{j \in \mathcal{I}} \pi_j(k^\star \,|\, \theta_j) - \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s, \theta_{-i}, \delta))\, ds \\ &= \sum_{j \in \mathcal{I} \setminus \{i\}} (1 - \delta_{ij}) \frac{\partial}{\partial \delta_{ij}} \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s, \theta_{-i}, \delta))\, ds, \end{aligned} \tag{18}$$

for some function $C_i : \Theta_{-i} \times \Delta \to \mathbb{R}$. Differentiating (18) with respect to $\delta_{i\ell}$ yields

$$\begin{aligned} &\frac{\partial C_i(\theta_{-i}, \delta)}{\partial \delta_{i\ell}} + \left[\sum_{j \in \mathcal{I}} \frac{d\pi_j(k^\star \,|\, \theta_j)}{dk} \right] \frac{\partial k^\star}{\partial \delta_{i\ell}} - \frac{\partial}{\partial \delta_{i\ell}} \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s, \theta_{-i}, \delta))\, ds \\ &= \sum_{j \in \mathcal{I} \setminus \{i\}} (1 - \delta_{ij}) \frac{\partial^2}{\partial \delta_{i\ell} \partial \delta_{ij}} \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s, \theta_{-i}, \delta))\, ds - \frac{\partial}{\partial \delta_{i\ell}} \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s, \theta_{-i}, \delta))\, ds. \end{aligned} \tag{19}$$

Substituting for (15) in equation (19) yields

$$\frac{\partial C_i(\theta_{-i}, \delta)}{\partial \delta_{i\ell}} + \sum_{j \in \mathcal{I} \setminus \{i\}} (1 - \delta_{ij}) \frac{\partial^2 p_i(\delta_i)}{\partial \delta_{i\ell} \partial \delta_{ij}} - \frac{\partial p_i(\delta_i)}{\partial \delta_{i\ell}} = \frac{\partial}{\partial \delta_{i\ell}} \int_{\theta_i^{\min}}^{\theta_i} v_i(k^\star(s, \theta_{-i}, \delta))\, ds.$$

[6] $\int \frac{d\pi_i(k^\star \,|\, \theta_i)}{dk} \frac{\partial k^\star}{\partial \theta_i} d\theta_i = \int \theta_i \frac{\partial v_i(k^\star)}{\partial \theta_i} d\theta_i + \int \frac{\partial w_i(k^\star)}{\partial \theta_i} d\theta_i = \theta_i v_i(k^\star) - \int v_i(k^\star) d\theta_i + w_i(k^\star) + C_i.$

Differentiating the latter with respect to θ_i implies that $\partial v_i(k^\star)/\partial\delta_{i\ell} = 0$. As $dv_i/dk \neq 0$ by assumption, $k^\star$ must satisfy $\partial k^\star/\partial\delta_{i\ell} = 0$ for all i and all $\ell \neq i$. ■

Proposition 3. *A generic, social-type independent policy $k^\star : \Theta \to K$ is* definitely *implementable through budget-balanced transfers only if $k^\star(\theta) = \arg\max_{k\in K} \sum_{i\in\mathcal{I}} \pi_i(k \,|\, \theta_i)$. The mechanisms that implement $k^\star$ are necessarily of SPC-type.*

Proof. As $k^\star : \Theta \to K$ is generic, $\partial k^\star/\partial\theta_i \neq 0$. By assumption, $\partial k^\star/\partial\delta_i = 0$. Hence, condition (14) of Lemma 3 implies that $k^\star$ must satisfy $0 = \sum_i d\pi_i(k^\star \,|\, \theta_i)/dk$. (If this equation has no solution, then condition (14) and Proposition 2 imply that $k^\star$ must be constant: $\partial k^\star/\partial\theta_i = 0 = \partial k^\star/\partial\delta_i$.) Hence, for each θ, the policy $k^\star(\theta)$ is either a minimizer, saddle point, or maximizer of $\sum_i \pi_i(k \,|\, \theta_i)$, which we keep in mind for a moment.

Suppose $T^\star$ is a budget-balanced transfer scheme that implements $k^\star(\theta)$. Notice that one can always write $t_i^\star(\hat\theta, \hat\delta) = \sum_{\ell\neq i} \Big[\mathbb{E}_{\theta_{-i}}\big[\pi_\ell(k^\star(\hat\theta_i, \theta_{-i}) \,|\, \theta_\ell)\big] - \mathbb{E}_{\theta_{-\ell}}\big[\pi_i(k^\star(\hat\theta_\ell, \theta_{-\ell}) \,|\, \theta_i)\big]\Big] + s_i(\hat\theta, \hat\delta)$, for appropriate functions $s_i : \Theta \times \Delta \to \mathbb{R}$ satisfying $\sum_{i\in\mathcal{I}} s_i = 0$. Then $T^\star$ is budget-balanced by construction. It has to be shown that the $(s_i)_{i\in\mathcal{I}}$ must satisfy the conditions of Definition 3(ii),(iii). By the same reasoning as above, $\bar t_{ii}(\hat\theta_i, \hat\delta_i) = \sum_{\ell\neq i} \mathbb{E}_{\theta_{-i}}\big[\pi_\ell(k^\star(\hat\theta_i, \theta_{-i}) \,|\, \theta_\ell)\big] - (n-1)\,\mathbb{E}_\theta\big[\pi_i(k^\star(\theta) \,|\, \theta_i)\big] + \mathbb{E}_{\theta_{-i},\delta_{-i}}[s_i(\hat\theta_i, \theta_{-i}, \hat\delta_i, \delta_{-i})]$ and

$$
\begin{aligned}
(20) \qquad \bar t_{ij}(\hat\theta_i, \hat\delta_i) \;&\overset{j\neq i}{=}\; \sum_{\ell\in\mathcal{I}} \mathbb{E}_\theta\big[\pi_\ell(k^\star(\theta) \,|\, \theta_\ell)\big] - (n-1)\,\mathbb{E}_\theta\big[\pi_j(k^\star(\theta) \,|\, \theta_j)\big] \\
&\qquad - \mathbb{E}_{\theta_{-i}}\big[\pi_j(k^\star(\hat\theta_i, \theta_{-i}) \,|\, \theta_j)\big] + \mathbb{E}_{\theta_{-i},\delta_{-i}}\big[s_j(\hat\theta_i, \theta_{-i}, \hat\delta_i, \delta_{-i})\big] \\
(21) \qquad &= \; \frac{\partial p_i(\hat\delta_i)}{\partial\delta_{ij}} - \mathbb{E}_{\theta_{-i}}\big[\pi_j(k^\star(\hat\theta_i, \theta_{-i}) \,|\, \theta_j)\big],
\end{aligned}
$$

where equation (21) is implied by Lemma 2(ii), for some function $p_i : \Delta_i \to \mathbb{R}$. Hence, $\mathbb{E}_{\theta_{-i},\delta_{-i}}[s_\ell(\hat\theta_i, \theta_{-i}, \hat\delta_i, \delta_{-i})]$ must be independent of $\hat\theta_i$. As the $(s_i)_{i\in\mathcal{I}}$ are required to be budget-balanced, also $\mathbb{E}_{\theta_{-i},\delta_{-i}}[s_i(\hat\theta_i, \theta_{-i}, \hat\delta_i, \delta_{-i})]$ must be independent of $\hat\theta_i$. This establishes condition (ii). For condition (iii), define $p_i(\hat\delta_i) = -\sum_{\ell\neq i}(1-\delta_{i\ell})\bar s_{i\ell}(\hat\delta_i)$, where $\bar s_{i\ell}(\hat\theta_i, \hat\delta_i) = \mathbb{E}_{\theta_{-i},\delta_{-i}}[s_\ell(\hat\theta_i, \theta_{-i}, \hat\delta_i, \delta_{-i})]$. Notice that $p_i(\hat\delta_i)$ gives precisely agent i's interim-expected utility from $(s_i)_{i\in\mathcal{I}}$ when reporting $\hat\delta_i$. By Lemma 1(iii), incentive compatibility with respect to δ_i thus requires that $p_i(\delta_i) = q_{i\ell}(\delta_i^{-\ell}) + \int_{\delta_{i\ell}^{\min}}^{\delta_{i\ell}} \bar s_{i\ell}(r, \delta_i^{-\ell})\,dr$ for all $\ell \neq i$. Hence, $\bar s_{i\ell} = \partial p_i/\partial\delta_{i\ell}$ for all $\ell \neq i$, implying that $(\bar s_{i\ell})_{\ell\neq i} = \nabla p_i$ and $p_i(\delta_i) = -\sum_{\ell\neq i}(1-\delta_{i\ell})\bar s_{i\ell}(\delta_i) = -\nabla p_i(\delta_i)\cdot(\mathbf{1}-\delta_i)$. Incentive compatibility thus requires $\nabla p_i(\hat\delta_i)\cdot(\mathbf{1}-\delta_i) \geq \nabla p_i(\delta_i)\cdot(\mathbf{1}-\delta_i)$. By the same reasoning as in the proof of

Proposition 1, these conditions on p_i imply that $p_i(\delta_i) - p_i(\hat{\delta}_i) \geq \nabla p_i(\hat{\delta}_i) \cdot (\delta_i - \hat{\delta}_i)$. Hence, *transfers* must be of SPC-type.

Now reconsider the possible nature of $k^\star$. Under SPC-type transfers, the mechanism $\langle k^\star, T^\star \rangle$ yields agent i an interim-expected utility level of

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

As $k^\star$ is supposed to be *definitely* implementable, we must have $\sum_{\ell \in \mathcal{I}} \pi_\ell(k^\star(\hat{\theta}_i, \theta_{-i}) \,|\, \theta_\ell) \leq \sum_{\ell \in \mathcal{I}} \pi_\ell(k^\star(\theta_i, \theta_{-i}) \,|\, \theta_\ell)$ for all $\hat{\theta}_i, \theta$. This is obviously impossible if $k^\star$ is a minimizer or saddle point; it is only possible if $k^\star$ is a maximizer, $k^\star(\theta) = \arg\max_{k \in K} \sum_{\ell \in \mathcal{I}} \pi_\ell(k \,|\, \theta_\ell)$. Hence, $\langle k^\star, T^\star \rangle$ is of SPC-type. ■

Theorem 1 is thus established.

4 Scrutinizing Dictatorship

Through Lemma 3 and Proposition 1, also the result on dictatorship can be generalized. Assume for simplicity that $d^2\pi_j(k \,|\, \theta_j)/dk^2 < 0$ on $K \times \Theta_j$ for all $j \in \mathcal{I}$.

Theorem 4. *A dictatorial policy, if it is interior solution to* $\max_{k \in K} u_i(k, \theta_{-i} \,|\, \theta_i, \delta_i)$, *is* definitely *implementable through budget-balanced transfers if and only if dictator i is either perfectly selfish, $\delta_{ij} = 0$ for all $j \neq i$, or perfectly benevolent, $\delta_{ij} = 1$ for all $j \neq i$.* (*This equivalence also holds if the dictator's preferences are common knowledge.*)

Proof. The sufficiency part is trivial if $\delta_{ij} = 0$ for all $j \neq i$ (transfers can even be zero), and it is immediate from Proposition 1 if $\delta_i = 1$ for all $j \neq i$ (in which case the dictator is indifferent between any interpersonal transfers).

For the necessity part, suppose dictator i's problem has always an interior solution $k^\star$ and that $\delta_{ij} \neq 0$ for at least one $j \neq i$. Then, $0 = du_i(k^\star, \theta_{-i} \,|\, \theta_i, \delta_i)/dk$ and $0 > d^2u_i(k^\star, \theta_{-i} \,|\, \theta_i, \delta_i)/dk^2$, implying that $\partial k^\star/\partial \theta_j = -\delta_{ij} \cdot v_j/(d^2u_i/dk^2) \neq 0$ for all $j \neq i$. As $\partial k^\star/\partial \delta_{j\ell} = 0$ for all $\ell \neq j$, condition (14) of Lemma 3, when applied to j, implies

that $k^\star$ must satisfy $0 = \sum_{\ell\in\mathcal{I}} d\pi_\ell(k^\star \,|\, \theta_\ell)/dk$; hence, $\partial k^\star/\partial\theta_j = -v_j/(\sum_{\ell\in\mathcal{I}} d^2\pi_\ell/dk^2)$. Jointly, these conditions on $\partial k^\star/\partial\theta_j$ yield

$$\delta_{ij} \cdot \sum_{\ell\in\mathcal{I}} \frac{d^2\pi_\ell(k^\star \,|\, \theta_\ell)}{dk^2} \quad = \quad \frac{d^2u_i(k^\star, \theta_{-i} \,|\, \theta_i, \delta_i)}{dk^2} \quad = \quad \sum_{\ell\in\mathcal{I}} \delta_{i\ell} \cdot \frac{d^2\pi_\ell(k^\star \,|\, \theta_\ell)}{dk^2}, \tag{22}$$

for all $j \neq i$. Hence, $\delta_{ij} = \delta_{i\ell}$ for all $j, \ell \neq i$ and thus, again by (22), $\delta_{ij} = 1$ for all $j \neq i$.

This line of reasoning, including the derivation of condition (14) for $j \neq i$, builds solely on the assumption that the preferences of every $j \neq i$ are private information. ■

References

Arrow, Kenneth J. 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.

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

Daske, Thomas. 2020. "Efficient incentives in social networks: Gamification and the Coase theorem." *Mimeo*. Available under: http://hdl.handle.net/10419/222527.

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

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

———. 1981. "Optimal auction design." *Mathematics of Operations Research* 6 (1):58–73.