Robust Coalitional Implementation*

Huiyi Guo† Nicholas C. Yannelis‡

January 28, 2022

Abstract

The paper introduces coalition structures to study belief-free full implementation. When the mechanism designer does not know which coalitions are admissible, we provide necessary and almost sufficient conditions on when a social choice function is robustly coalitionally implementable, i.e., implementable regardless of the coalition pattern and the belief structure. Robust coalitional implementation is a strong requirement that imposes stringent conditions on implementable social choice functions. However, when the mechanism designer has additional information on which coalitions are admissible, we show that coalitional manipulations may help a mechanism designer to implement social choice functions that are not robustly implementable in the sense of Bergemann and Morris (2009, 2011). As different social choice functions are implementable under different coalition patterns, the paper provides insights on when agents should be allowed to play cooperatively.

Keywords: Belief-free implementation; Full implementation; Coalition.

JEL: C71; D82.

*Part of the results reported in the current paper is based on the authors' earlier project entitled *Robust Strong Nash Implementation*. We are grateful to two anonymous referees and an advisory editor whose comments helped us improve the paper significantly. We also thank Rabah Amir, Rodrigo Velez, and participants at various conferences for valuable discussions.

†Texas A&M University, Department of Economics, 4228 TAMU, College Station, TX 77843, huiyiguo@tamu.edu. Corresponding author.

‡The University of Iowa, Department of Economics, W380 John Pappajohn Business Building, Iowa City, IA 52242, nicholasyannelis@gmail.com.

1 Introduction

In Bayesian implementation problems (see, for example, Jackson (1991)), agents' private information is canonically modeled by a type space that is common knowledge between the mechanism designer (assumed to be female) and all agents (assumed to be male). Inspired by the Wilson doctrine, Bergemann and Morris (2009, 2011), among others, relax the common knowledge assumption and introduce a belief-free approach to study when a social choice function is fully implementable under all type spaces. This is the robust (full) implementation problem. The existing literature on robust implementation has been assuming that agents behave non-cooperatively without considering potential coalitional manipulations. However, the needs to make a mechanism robust to agents' belief structures and to make it immune from collusion may coexist. The current paper thus introduces coalition structures into the research program of robust implementation.

In our paper, the coalition pattern, i.e., the collection of admissible coalitions, is exogenously given by $\mathcal{S}$. When an admissible coalition has a profitable joint deviation, members of this coalition will coordinately deviate such that every member is better off. The equilibrium played by agents is called the interim $\mathcal{S}$ equilibrium, which is immune from deviations of any coalition in $\mathcal{S}$. In one extreme case where only singleton coalitions are permissible in $\mathcal{S}$, interim $\mathcal{S}$ equilibrium reduces to the interim equilibrium (also called the Bayesian equilibrium). In the other extreme case where $\mathcal{S}$ includes all coalitions, the corresponding interim $\mathcal{S}$ equilibrium is a variant of the strong equilibrium of Aumann (1959) in an incomplete information setting. Although the coalition pattern $\mathcal{S}$ is common knowledge among agents, the mechanism designer may or may not have access to this information. Depending on whether the mechanism designer knows the coalition pattern, we study two problems: robust coalitional implementation and robust $\mathcal{S}$ implementation.

The first problem we examine is robust coalitional implementation, in which the mechanism designer has no information on which coalitions are admissible. In this case, she wishes to construct a mechanism such that the social choice function coincides with the interim $\mathcal{S}$ equilibrium outcomes regardless of the coalition pattern $\mathcal{S}$ and the belief structure. We provide a group of sufficient conditions on robust coalitional implementation: a social choice

function is robustly coalitionally implementable if it satisfies the robust coalitional incentive compatibility condition, the robust coalitional monotonicity condition, and the interior coalitional reward property. Among these conditions, robust coalitional incentive compatibility and robust coalitional monotonicity are also necessary. The two necessary and almost sufficient conditions are stronger than those of Bergemann and Morris (2011), because we want to make sure that the mechanism is invulnerable to the additional uncertainty facing the designer, i.e., agents' coalition pattern, beyond her uncertainty on agents' belief structure. As the set of robustly coalitionally implementable social choice functions shrinks compared to the one under non-cooperative robust implementation, not knowing the coalition pattern is costly to the mechanism designer. Example 1 in the paper presents a social choice function that is robustly implementable in the sense of Bergemann and Morris (2009), but it is not robustly coalitionally implementable.

When the mechanism designer has information on the coalition pattern $\mathcal{S}$, she knows the set of possible equilibria played by agents. Our robust $\mathcal{S}$ implementation question requires the social choice function to coincide with the interim $\mathcal{S}$ equilibria outcomes regardless of agents' belief structures. We establish sufficient conditions for robust $\mathcal{S}$ implementation: robust $\mathcal{S}$ incentive compatibility, robust $\mathcal{S}$ monotonicity, and the interior $\mathcal{S}$ reward property. When only singleton coalitions are permissible, our sufficiency result implies that of Bergemann and Morris (2011) on robust implementation under the non-cooperative framework. When the coalition pattern is richer, our robust $\mathcal{S}$ incentive compatibility condition becomes more demanding, but the robust $\mathcal{S}$ monotonicity condition may be weaker. Hence, introducing non-trivial coalition structures may give the mechanism designer leeway to implement social choice functions that are not robustly implementable in the sense of Bergemann and Morris (2011). Intuitively, allowing for coalitional manipulations makes the existence of a good equilibrium more difficult, but can potentially make it easier to dissolve bad equilibria. When the second effect dominates, the mechanism designer can benefit from non-trivial coalitions. Example 2 in the paper presents a social choice function that violates the robust monotonicity condition and thus is not implementable in the sense of Bergemann and Morris (2011). However, we demonstrate its implementability under the richest coalition pattern, implying that robust monotonicity is not necessary for robust $\mathcal{S}$ implementation under a

non-trivial coalition pattern.

Our study of robust coalitional implementation and robust $\mathcal{S}$ implementation demonstrates the importance of mechanism designer's knowledge on coalition patterns in robust implementation problems. In addition, the comparison between robust $\mathcal{S}$ implementation under different coalition patterns highlights the value of having different coalition patterns for robust implementation problems. In particular, introducing non-trivial coalition patterns may help to implement social choice functions that are non-implementable under the non-cooperative framework.

Related Literature: The paper fits into the literature on robust full implementation. In a single crossing environment, Bergemann and Morris (2009) characterize social choice functions that are robustly fully implementable under direct mechanisms. In a general environment, Bergemann and Morris (2011) propose necessary and almost sufficient conditions for robust implementation under general mechanisms. Saijo et al. (2007) and Adachi (2014) focus on private value environments and establish necessary and sufficient conditions for secure implementation (in dominant strategy equilibrium and in Nash equilibrium) and for robust implementation. Oury and Tercieux (2012) propose a robust partial implementation concept called continuous implementation and explore its connection with full implementation in rationalizable strategies. Penta (2015) and Müller (2016) further extend the belief-free mechanisms to dynamic ones. Instead of assuming that the mechanism designer knows nothing about agents' belief structures, Ollár and Penta (2017) allow the mechanism designer to have partial information on the belief structure and to design transfers. All the above works have been assuming that agents behave non-cooperatively without considering coalitional manipulations. The current paper extends the literature on robust implementation by taking into account coalitional manipulations and exploring the value of cooperation to robust implementation problems.

Besides, the paper is closely related to the literature on full implementation with coalition structures. To the best of our knowledge, only two papers look into the problem of Bayesian full implementation with coalitions. One is Hahn and Yannelis (2001). In exchange economies with general preferences, they generalize the strong equilibrium concept to the incomplete information setting and provide conditions for full implementation under

this equilibrium. The other is Safronov (2018), where the expected externality mechanism is redesigned. Essentially, the newly designed mechanism can fully implement the set of efficient social choice functions under the independent private value environment regardless of the coalition pattern. The most important difference between the current paper and the above two is that we adopt a belief-free approach but their results rely on the mechanism designer's knowledge of the belief structure. Other than full implementation, there is a branch of literature studying partial implementation with coalitional manipulations with or without adopting a belief-free approach, including but not limited to Bennett and Conn (1977), Green and Laffont (1979), Chen and Micali (2012), Bierbrauer and Hellwig (2011, 2015, 2016), Guo (2020), Guo and Yannelis (2020). The partial implementation literature focuses on the existence of a good equilibrium that leads to outcomes consistent with the social choice function while the full implementation literature, including the current paper, also aims to ensure the non-existence of bad equilibria.

When agents do not possess private information, more papers have studied Nash implementation problems with coalitional manipulations. Maskin (1978) initiates the concept of fully implementing a social choice correspondence in strong equilibrium. Subsequently, Maskin (1979) studies when full implementation can be guaranteed under all coalition patterns, which he calls a double implementation problem.[1] Then, a few papers, including but not limited to Maskin et al. (1985), Dutta and Sen (1991), Suh (1996, 1997), Pasin (2009), and Korpela (2013), further explore the problem of implementation in strong equilibrium or the problem of double implementation, and provide various characterizations or sufficient conditions. The Maskin monotonicity condition, which is necessary for Nash implementation, is also necessary for implementation in strong equilibrium (and for double implementation). This contrasts with our finding that the robust monotonicity condition is not necessary for robust $\mathcal{S}$ implementation under non-trivial coalition patterns.

Recently, under the complete information setting, Koray and Yildiz (2018) and Korpela et al. (2020) bring to the literature the idea of designing a rights structure or a code of rights,

[1]The term double implementation has been used to refer to other implementation concepts unrelated to coalitions. To highlight our focus on coalition manipulations, we call the robust implementation problem under all coalition patterns the robust coalitional implementation problem.

which specifies the collection of coalitions having the right to act cooperatively. We differ in our incomplete information setup and in our exogenous coalition structure. Our finding that some social choice functions are robustly implementable under certain non-trivial coalition patterns shares a similar implication with theirs in that non-trivial coalitions can bring value to institution design.

The paper proceeds as follows. Section 2 presents the primitives of the environment. We then motivate the study of robust coalitional implementation and robust $\mathcal{S}$ implementation with two examples in Section 3. The main results of our paper, sufficient conditions for robust coalitional implementation and robust $\mathcal{S}$ implementation, and additional examples are introduced in Sections 4 and 5. We then conclude in Section 6.

2 Asymmetric Information Environment

We first consider an asymmetric information environment without any specification on beliefs, namely a **payoff environment**, given by $\mathcal{E} = \{I, A, (\Theta_i, u_i)_{i=1}^n\}$, where

- $I = \{1, ..., n\}$ is the set of **agents**;
- A is **the set of feasible outcomes**, i.e., the set of all lotteries over a deterministic feasible outcome set X;[2]
- $\Theta = \Theta_1 \times ... \times \Theta_n$ is a countable **payoff type set**, and $\theta_i \in \Theta_i$ is agent i's **payoff type**;
- $u_i : X \times \Theta \to \mathbb{R}$, agent i's **utility function**, represents agent i's utility of consuming a deterministic outcome $a \in X$, when the realized payoff type profile is $\theta = (\theta_i)_{i \in I}$; then extend the domain of u_i to $A \times \Theta$ so that for any $a \in A = \Delta(X)$ with measure μ, $u_i(a, \theta) = \int_{x \in X} u_i(x, \theta) d\mu$.[3]

A **social choice function** $f : \Theta \to A$ is an exogenous rule to assign feasible outcomes contingent on agents' payoff types. Notice that the outcome prescribed by a social choice function does not depend on agents' belief assessments of each other's private information, which will be introduced later.

[2]Lotteries are feasible in many papers in the literature, e.g., Bergemann and Morris (2011), Pram (2020), Jain (2021).

[3]The integral form of the utility function is used when we construct lotteries in Theorems 1 and 2.

Given a sequence of outcomes $(a^k \in A)_{k=1,2,...}$ and a sequence of weights $(\rho^k \geq 0)_{k=1,2,...}$ such that $\sum_{k=1,2,...} \rho^k = 1$, i.e., $(\rho^k \geq 0)_{k=1,2,...} \in \Delta$, we let $\sum_{k=1,2,...} \rho^k a^k$ denote a compound lottery whose realization is a^k with probability ρ^k. Similarly, for a sequence of social choice functions $(f^k : \Theta \to A)_{k=1,2,...}$, $\sum_{k=1,2,...} \rho^k f^k$ denotes a new social choice function so that at each $\theta \in \Theta$, $\sum_{k=1,2,...} \rho^k f^k(\theta)$ is the outcome.

In this paper, we assume that the payoff environment $\mathcal{E}$ is common knowledge between the mechanism designer and all agents. However, the following belief structure, including the type space and the belief revising rule, is not known to the mechanism designer.

Agents' beliefs are ex-post payoff-irrelevant, but they affect the strategic interaction between agents in the interim stage. A **type space** is a collection $\mathcal{T} = (T_i, \hat{\theta}_i, \pi_i)_{i \in I}$, where

- $t_i \in T_i$ is a **type** of agent i, which represents agent i's private information; the set of all type profiles is denoted by $T = \prod_{i \in I} T_i$ and a generic element is denoted by $t = (t_i)_{i \in I}$; to avoid technicality, we assume that each T_i is a countable set;
- agent i with type t_i has a payoff type $\hat{\theta}_i(t_i)$, which is defined by a surjective mapping $\hat{\theta}_i : T_i \to \Theta_i$; denote $\hat{\theta}(t) = (\hat{\theta}_i(t_i))_{i \in I}$;
- agent i with type t_i has a **belief type** $\pi_i(t_i)$, which is a probability distribution over $T_{-i} = \prod_{j \neq i} T_j$, assigning probability $\pi_i(t_i)[t_{-i}]$ to the event that others have type profile $t_{-i} = (t_j)_{j \neq i}$.

A key feature of this paper is that we consider profitable deviations of coalitions. A **coalition** is a non-empty subset of I and an agent in the subset is called a **member**. A **coalition pattern**, denoted by $\mathcal{S}$, is the set of all admissible coalitions and is exogenously given. We assume that all singletons are included in $\mathcal{S}$, i.e., agents can always choose to play non-cooperatively. One example of a coalition pattern is the **minimal** (or trivial) coalition pattern, which we denote by $\underline{\mathcal{S}} = \{\{i\} : i \in I\}$. The other extreme case is the **maximal** coalition pattern, denoted by $\bar{\mathcal{S}} = 2^I \backslash \{\emptyset\}$, which has the richest coalition structure. In applications, one may be interested in other patterns formed by partisanship, cultural differences, geographic isolation, etc.

When an agent acquires new information on other agents' types, the new information may

surprise him.[4] Hence, we have to consider how agents revise beliefs under zero probability events. For each distribution $\pi_i(t_i^*) \in \Delta(T_{-i})$ and non-singleton $S \subseteq I$ containing i, let the notation $\pi_i(t_i^*)[t_{S\backslash\{i\}}^*]$ represent the marginal probability that coalition $S\backslash\{i\}$ has type profile $t_{S\backslash\{i\}}^* = (t_j^*)_{j\in S\backslash\{i\}}$. Whenever $\pi_i(t_i^*)[t_{S\backslash\{i\}}^*] = 0$, a **belief revising rule** specifies a posterior belief $(\pi_i(t_i^*)[t_{-i}|t_{S\backslash\{i\}}^*])_{t_{-i}\in T_{-i}}$ over T_{-i} whose marginal probability on the event that coalition $S\backslash\{i\}$ has type profile $t_{S\backslash\{i\}}^*$ is 1. The posterior belief is defined by the Bayes rule whenever $\pi_i(t_i^*)[t_{S\backslash\{i\}}^*] > 0$.

A **mechanism** is a pair $(M, g) = (\prod_{i\in I} M_i, g)$, where M_i is the **message space** of agent i, i.e., the set of all messages that agent i can submit, and $g : M \to A$ is an **outcome function**, which assigns to each message profile $m = (m_i)_{i\in I}$ a feasible outcome. Agent i's **strategy** $\sigma_i : T_i \to M_i$ is a private information contingent plan of submitting messages. We focus on pure strategies in this paper for simplicity. Denote by σ_S the strategy profile $(\sigma_i)_{i\in S}$, by σ_{-S} the profile $(\sigma_i)_{i\notin S}$, and by σ the profile $(\sigma_i)_{i\in I}$. In the special case that $M_i = \Theta_i$ for all $i \in I$ and $g(\hat{\theta}(t)) = f(\hat{\theta}(t))$ for all $t \in T$, the mechanism elicits agents' payoff types and is called a **direct mechanism** f.

When the coalition pattern is $\mathcal{S}$, this paper assumes that agents play an interim $\mathcal{S}$ equilibrium. The equilibrium requires that there does not exist an admissible coalition as well as a type profile and a deviating strategy profile of the coalition, such that under coalition members' pooled information, deviation makes every member strictly better off.

Definition 1: *Given a type space and a belief revising rule, the strategy profile σ^* is an **interim $\mathcal{S}$ equilibrium** of the mechanism (M, g) if there does not exist $S \in \mathcal{S}$, $t_S^* \in T_S$, and strategy profile σ_S', such that for all $i \in S$,*

$$\sum_{t_{-i}\in T_{-i}} u_i\Big(g\big(\sigma_S'(t_S^*), \sigma_{-S}^*(t_{-S})\big), \hat{\theta}(t_S^*, t_{-S})\Big)\pi_i(t_i^*)[t_{-i}|t_{S\backslash\{i\}}^*]$$
$$> \sum_{t_{-i}\in T_{-i}} u_i\Big(g\big(\sigma^*(t_S^*, t_{-S})\big), \hat{\theta}(t_S^*, t_{-S})\Big)\pi_i(t_i^*)[t_{-i}|t_{S\backslash\{i\}}^*].$$

Under the maximal coalition pattern $\bar{\mathcal{S}}$, the interim $\bar{\mathcal{S}}$ equilibrium can be viewed as a variant of Aumann (1959)'s strong equilibrium under asymmetric information. Hence, we

[4] A related question shows up in dynamic environments, where Penta (2015) and Müller (2016) have explored how belief revising rule under zero probability events affects robust dynamic implementation.

also call an interim $\bar{\mathcal{S}}$ equilibrium an **interim strong equilibrium**. Similarly, under the minimal coalition pattern $\underline{\mathcal{S}}$, the interim $\underline{\mathcal{S}}$ equilibrium becomes the widely adopted **interim equilibrium** (or Bayesian equilibrium) in the mechanism design literature.

To interpret the interim $\mathcal{S}$ equilibrium, imagine that each $S \in \mathcal{S}$ has access to an outside intermediary that is benevolent to coalition S. We assume that members of coalition $S \in \mathcal{S}$ disclose their private information to the intermediary of S. Based on the pooled information of S, the intermediary computes each member's updated belief and determines if the coalition has a way to profitably deviate. If there exists a profitable joint deviation, the intermediary *will* coordinate one. Since each coalition in $\mathcal{S}$ containing agent i has its intermediary, the posterior beliefs of agent i are different in different coalitions. Our interim $\mathcal{S}$ equilibrium can be viewed as a refinement of interim equilibrium that survives after the attempt of every coalition $S \in \mathcal{S}$ to seek profitable coalitional deviation.

Notice it is assumed that with the assistance of a third-party intermediary, agents within an admissible coalition essentially act as a utility-maximizing pseudo agent without encountering within-coalition interactions. This assumption is consistent with the coalition-proofness notions of Bennett and Conn (1977), Green and Laffont (1979), Chen and Micali (2012), Safronov (2018), etc, in partial implementation problems (or the partial implementation direction of full implementation problems). The assumption simplifies the analysis by helping us focus on the interaction between a coalition and all others out of the coalition. Although there are alternative models considering within-coalition strategic interactions that potentially undermine the power of coalitional manipulations (see, e.g., Che and Kim (2006), Moreno-García and Torres-Martínez (2020), and Koutsougeras (2020)), our definition of interim $\mathcal{S}$ equilibrium imposes a strong stability requirement on the equilibrium notion and may serve as a benchmark to study implementation with coalition concerns.

The mechanism designer may or may not have knowledge about the coalition pattern, and thus may or may not know which coalitions have access to such intermediaries to coordinate profitable deviations. Hence, we consider two implementation concepts: robust coalitional implementation and robust $\mathcal{S}$ implementation.

Our first implementation concept, robust coalitional implementation, addresses two robustness concerns. First, following the robust mechanism design literature, it assumes that

the mechanism designer has no information on agents' belief structure. Second, the mechanism designer has no information on the true coalition pattern, i.e., she does not know which coalitions have access to third-party intermediaries to coordinate profitable coalitional deviations. Robust coalitional implementation requires the existence of a mechanism (M, g), such that under each belief structure and coalition pattern $\mathcal{S}$, there exists a good interim $\mathcal{S}$ equilibrium (i.e., one leading to outcomes consistent with f), and there does not exist any bad interim $\mathcal{S}$ equilibrium (i.e., one leading to outcomes inconsistent with f). Notice that given any belief structure, since the set of all interim strong equilibria is equal to the intersection of interim $\mathcal{S}$ equilibria across all coalition patterns, the existence of a good interim $\mathcal{S}$ equilibrium under all coalition patterns is equivalent to the existence of a good interim strong equilibrium. Similarly, since the set of all interim equilibria is equal to the union of interim $\mathcal{S}$ equilibria across all coalition patterns, the non-existence of a bad interim $\mathcal{S}$ equilibrium under any coalition pattern $\mathcal{S}$ is equivalent to the non-existence of any bad interim equilibrium. We thus adopt the following definition of robust coalitional implementation.

Definition 2: *A social choice function f is said to be* ***robustly coalitionally implementable*** *if there is a mechanism (M, g) such that under all type spaces and all belief revising rules,*

(*i*) *there exists an interim strong equilibrium σ of the mechanism (M, g) such that $g\big(\sigma(t)\big) = f\big(\hat{\theta}(t)\big)$ for all $t \in T$;*

(*ii*) *if σ is an interim equilibrium of the mechanism (M, g), then $g\big(\sigma(t)\big) = f\big(\hat{\theta}(t)\big)$ for all $t \in T$.*

Our second implementation concept, robust $\mathcal{S}$ implementation, assumes that the mechanism designer knows that coalitions in $\mathcal{S}$ will deviate whenever there is room for profit. It requires the set of interim $\mathcal{S}$ equilibria to coincide with the social choice function under all belief structures.

Definition 3: *A social choice function f is said to be* ***robustly $\mathcal{S}$ implementable*** *if there is a mechanism (M, g) such that under all type spaces and all belief revising rules,*

(*i*) *there exists an interim* $\mathcal{S}$ *equilibrium* σ *of the mechanism* (M, g) *such that* $g\big(\sigma(t)\big) = f\big(\hat{\theta}(t)\big)$ *for all* $t \in T$*;*

(*ii*) *if* σ *is an interim* $\mathcal{S}$ *equilibrium of the mechanism* (M, g)*, then* $g\big(\sigma(t)\big) = f\big(\hat{\theta}(t)\big)$ *for all* $t \in T$*.*

Specifically, under the minimal coalition pattern, Definition 3 becomes the robust implementation notion of Bergemann and Morris (2011). To differentiate all implementation concepts mentioned in the current paper, the term robust implementation refers to the one of Bergemann and Morris (2011) exclusively henceforth. Having a larger coalition pattern makes the existence of a good equilibrium more difficult, but may help to dissolve bad equilibria. Due to the two conflicting forces, for any $\mathcal{S}$ larger than $\underline{\mathcal{S}}$, robust $\mathcal{S}$ implementation does not imply robust implementation, and vice versa. Also, since robust $\mathcal{S}$ implementation relies on the assumption that profitable coalitional deviations *will* happen, rather than *may* happen, the robust $\mathcal{S}$ implementation concept should not be viewed as a more robust or less robust notion than robust implementation.

To see the relationship between our two implementation concepts, robust coalitional implementation implies robust $\mathcal{S}$ implementation for all coalition pattern $\mathcal{S}$. However, robust $\mathcal{S}$ implementation does not imply robust coalitional implementation, regardless of $\mathcal{S}$. One may wonder if robust $\bar{\mathcal{S}}$ implementation implies robust coalitional implementation, but this is not true because all interim strong equilibria are good does not warrant that all interim equilibria are good. Nevertheless, if there is a mechanism robustly $\bar{\mathcal{S}}$ implementing f and robustly $\underline{\mathcal{S}}$ implementing f simultaneously, it robustly coalitionally implements f.

If we require the two conditions in Definition 2 (resp. Definition 3) to hold under a given type space and belief revising rule only, we say the social choice function f is **interim coalitionally implementable** (resp. **interim $\mathcal{S}$ implementable**).

3 Motivating Examples

We present two examples to motivate the study of coalitional manipulations in implementation problems. The first one is a variant of the public good example of Bergemann and Morris

(2009): we have discrete types and allow the use of indirect mechanisms. The example shows that robustly implementable social choice functions may be vulnerable to coalitional manipulations. Thus they may not be robustly $\mathcal{S}$ implementable for some coalition pattern $\mathcal{S} \neq \underline{\mathcal{S}}$, and a fortiori may not be robustly coalitionally implementable. The example also shows that robustly coalitionally implementable social choice functions that are non-dictatorial exist, although the requirement of robust coalitional implementation is demanding.

Example 1: *Consider an environment with two agents, where each agent's payoff type set* Θ_i *is* $\{0, 0.5, 1\}$. *The mechanism designer can construct public goods and charge both agents. For simplicity, the private value utility of agent* i *is denoted by* $u_i(x, \theta_i) = \theta_i x_0 + x_i$ *rather than* $u_i(x, \theta)$, *when* x_0 *units of public good are provided and* i *receives a monetary transfer of* x_i *(equivalently,* i *is charged a payment of* $-x_i$*). A deterministic social choice function* f *is given by* $f(\theta) = (f_0(\theta), f_1(\theta), f_2(\theta))$ *for all* $\theta \in \Theta$, *where the public good provision level is* $f_0(\theta) = \theta_1 + \theta_2$ *and the transfer is* $f_i(\theta) = -0.5{\theta_i}^2$ *for all* $i \in I$. *We assume for simplicity that the set of deterministic feasible outcomes is given by* $X = f(\Theta)$.

Bergemann and Morris (2009) have shown that f *is robustly implementable (or equivalently, robustly* $\underline{\mathcal{S}}$ *implementable) by the direct mechanism. However, the truth-telling interim equilibrium in the direct mechanism is vulnerable to group manipulations. For example, when the grand coalition has payoff type profile* $\theta^* = (0.5, 0.5)$, *the group can jointly misreport payoff type profile* $\theta' = (1, 1)$ *so that each agent* $i \in I$ *earns a payoff of* $u_i(f(\theta'), \theta_i^*) = 0.5 > u_i(f(\theta^*), \theta_i^*) = 0.375$.

In fact, f *is neither robustly* $\bar{\mathcal{S}}$ *implementable nor robustly coalitionally implementable by any mechanism because there may not exist a good interim strong equilibrium under all belief structures.*[5] *To see this, we can fix any type space* $\mathcal{T}$ *with type set* T *and any belief revising rule. Suppose by way of contradiction that there is a mechanism* (M, g) *admitting a good interim strong equilibrium* σ. *Then* $g(\sigma(t)) = f(\hat{\theta}(t))$ *for all* $t \in T$. *Now, fix any type profile* $t^* \in T$ *with payoff types* $\theta^* = (0.5, 0.5)$ *and another type profile* $t' \in T$ *with payoff types* $\theta' = (1, 1)$. *By jointly deviating from playing* σ *to the alternative strategy profile* σ' *defined*

[5]Essentially, this is because f violates the robust coalitional incentive compatibility condition which will be introduced later.

by $\sigma_i'(t_i) = \sigma_i(t_i')$ for all $i \in I$ and $t_i \in T_i$, each type-t_i^* agent $i \in I$ in the grand coalition earns a payoff of $u_i(g(\sigma'(t^*)), \theta_i^*) = u_i(g(\sigma(t')), \theta_i^*) = u_i(f(\theta'), \theta_i^*) = 0.5 > u_i(f(\theta^*), \theta_i^*) = 0.375$. This contradicts the supposition that σ is an interim strong equilibrium. Hence, f is not robustly $\bar{\mathcal{S}}$ implementable, and a fortiori not robustly coalitionally implementable.

We remark that if the payoff type set is reduced to $\Theta_i = \{0, 1\}$ for all $i \in I$, then it is easy to see that the direct mechanism robustly coalitionally implements f (and thus robustly $\bar{\mathcal{S}}$ implements f). In particular, no coalition can profitably deviate from truthfully reporting. Besides, every bad interim equilibrium can be dissolved by a singleton's deviation. Hence, robustly coalitionally implementable social choice functions that are non-dictatorial exist although stringent conditions are imposed on them.

Example 2 presents a social choice function that is only robustly implementable under the maximal coalition pattern. It shows that having a non-trivial coalition pattern may help a mechanism designer to implement social choice functions that are non-implementable under the non-cooperative framework. It also implicitly shows that the robust monotonicity condition (defined and proved to be necessary for robust implementation in Bergemann and Morris (2011)) is not necessary for robust $\mathcal{S}$ implementation in general.

Example 2: *Consider a public good environment similar to Example 1 except that (i) Θ_i is $\{0, 1\}$ for both agents; (ii) agents have common value utility functions: the utility of agent i is given by $u_i(x, \theta) = (\theta_1 + \theta_2)x_0 + x_i$ when agents have payoff types θ_1 and θ_2, x_0 units of public good are provided, and agent i receives a monetary transfer of x_i; (iii) the social choice function $f = (f_0, f_1, f_2)$ is given by $f_0(\theta) = 2(\theta_1 + \theta_2)$, $f_i(\theta) = -(\theta_1 + \theta_2)^2$ for each $i \in \{1, 2\}$. Again, the set of deterministic feasible outcomes is given by $X = f(\Theta)$.*

We claim that f is not robustly implementable in the sense of Bergemann and Morris (2011).[6] *Suppose by way of contradiction that a mechanism (M, g) robustly implements f. Then there exists an interim equilibrium σ such that $g(\sigma(t)) = f(\hat{\theta}(t))$ for all $t \in T$ in the common prior type space defined below. For each $i \in \{1, 2\}$, the type set of agent i is given by $T_i = \{t_i^0, t_i^1\}$, where type t_i^0 has payoff type 0, and type t_i^1 has payoff type 1. Agents' beliefs are updated from the prior in the table below, where ϵ is a sufficiently small positive number.*

[6] This is essentially because f violates the robust monotonicity condition.

	t_2^0	t_2^1
t_1^0	ϵ^2	ϵ
t_1^1	$1-\epsilon-2\epsilon^2$	ϵ^2

Table 3.1: Common Prior

Consider the strategy profile σ' *defined by* $\sigma'_1(t_1) = \sigma_1(t_1^0)$ *for all* $t_1 \in T_1$ *and* $\sigma'_2(t_2) = \sigma_2(t_2^1)$ *for all* $t_2 \in T_2$. *The strategy profile* σ' *leads to unwanted outcomes: for example,* $g(\sigma'(t_1^1, t_2^1)) = g(\sigma(t_1^0, t_2^1)) = f(\hat{\theta}(t_1^0, t_2^1)) = f(0,1) \neq f(1,1) = f(\hat{\theta}(t_1^1, t_2^1))$. *We now show that* σ' *is an interim equilibrium for* $\epsilon > 0$ *sufficiently small, contradicting the supposition that* (M, g) *robustly implements* f. *By the definition of strategy profile* σ', *the interim payoff for type-*t_1^0 *agent* 1 *under* σ' *is equal to*

$$\begin{aligned}
&\frac{\epsilon}{1+\epsilon} u_1(g(\sigma'_1(t_1^0), \sigma'_2(t_2^0)), (0,0)) + \frac{1}{1+\epsilon} u_1(g(\sigma'_1(t_1^0), \sigma'_2(t_2^1)), (0,1)) \\
=&\frac{\epsilon}{1+\epsilon} u_1(g(\sigma_1(t_1^0), \sigma_2(t_2^1)), (0,0)) + \frac{1}{1+\epsilon} u_1(g(\sigma_1(t_1^0), \sigma_2(t_2^1)), (0,1)) \\
=&\frac{\epsilon}{1+\epsilon} u_1(f(0,1), (0,0)) + \frac{1}{1+\epsilon} u_1(f(0,1), (0,1)),
\end{aligned}$$

which is close to $u_1(f(0,1), (0,1))$ *when* ϵ *is sufficiently small. Since* $f(0,1)$ *maximizes the above expression among all feasible outcomes, type-*t_1^0 *agent* 1 *is playing best response under* σ'. *Similarly, we can verify that all other types and agents play best responses under* σ', *which implies that* σ' *is an interim equilibrium. This contradicts the supposition that* (M, g) *robustly implements* f.

However, f *is robustly* $\bar{\mathcal{S}}$ *implemented under all type spaces and all belief revising rules by the direct mechanism. Truthfully reporting of both agents constitutes a good interim strong equilibrium. To see this, notice that under each* $t \in T$, $f(\hat{\theta}(t))$ *assigns the optimal feasible outcome to both agents, and thus neither unilateral deviation nor coalitional deviation is profitable. Also, the mechanism does not admit any bad interim strong equilibrium. To see this, whenever a strategy profile* σ'' *and a type profile* $t \in T$ *are such that* $g(\sigma''(t)) \neq f(\hat{\theta}(t))$, *both agents receive a sub-optimal feasible outcome at* $t \in T$. *When the type-*t *grand coalition coordinately deviates from* σ'' *to truthfully reporting members' payoff types, both players will strictly improve their utility levels at* t. *Hence,* σ'' *cannot be an interim strong equilibrium.*

In fact, one can generalize the environment to admit more agents ($n \geq 2$) and larger payoff type sets ($|\Theta_i| \geq 3$ for all $i \in I$). Suppose each $i \in I$ has common value utility $u_i(x, \theta) = \sum_{j \in I} \theta_j x_0 + x_i$. Define a social choice function $f = (f_0, (f_i)_{i \in I})$ where $f_0(\theta) = n \sum_{j \in I} \theta_j$ and $f_i(\theta) = -0.5n(\sum_{j \in I} \theta_j)^2$ for all $\theta \in \Theta$ and $i \in I$. It is easy to see that the direct mechanism robustly $\bar{\mathcal{S}}$ implements f.

4 Robust Coalitional Implementation

We begin by assuming that the mechanism designer does not know the coalition pattern. This section introduces conditions for robust coalitional implementation and then constructs a mechanism for it.

4.1 Conditions

4.1.1 Necessary Conditions

The first condition we introduce is the robust coalitional incentive compatibility condition.

Definition 4: *A social choice function f satisfies the **robust coalitional incentive compatibility** condition if for any coalition $S \subseteq I$ and payoff type profiles $\theta'_S \neq \theta^*_S$, there exists $i \in S$ such that*

$$u_i\big(f(\theta^*_S, \theta_{-S}), (\theta^*_S, \theta_{-S})\big) \geq u_i\big(f(\theta'_S, \theta_{-S}), (\theta^*_S, \theta_{-S})\big) \text{ for all } \theta_{-S} \in \Theta_{-S}.$$

The condition guarantees the existence of a good interim strong equilibrium under all type spaces and belief revising rules. Similar to the coalition-proofness notions of Bennett and Conn (1977), Green and Laffont (1979), Chen and Micali (2012), and Safronov (2018), our condition disincentivizes any coalition from jointly misreporting members' payoff type profile in a direct mechanism. A difference is that within a coalition, our model neither assumes transferable utility nor common belief towards agents out of the coalition. Notice that when coalition $S = I$, robust coalitional incentive compatibility excludes the existence of $\theta^*, \theta' \in \Theta$ such that $f(\theta')$ is preferred to $f(\theta^*)$ for all agents under true payoff types θ^*. Namely, f should be ex-post weakly Pareto efficient within $f(\Theta)$. A global version

of ex-post weak Pareto efficiency is unnecessary: for example, a constant inefficient social choice function is robustly coalitionally implementable. We remark that there are papers in the literature imposing a surjectivity assumption on the social choice function (e.g, Maskin (1977) and Williams (2001)). One reason for this is to establish the almost sufficiency of Maskin monotonicity in the Nash implementation problem. However, the current paper adopts sufficient conditions and mechanisms without relying on the surjectivity assumption to dissolve bad equilibria, which allows us to look at social choice functions with $f(\Theta) \subsetneq A$.

The robust coalitional incentive compatibility condition is in general a strong condition. Allowing all coalitions to be admissible imposes a stronger stability requirement than the familiar ex-post incentive compatibility condition. In addition, we do not introduce strategic interactions within a coalition that potentially undermine the power of coalitions. However, there are environments in which robust coalitional incentive compatibility is implied by familiar conditions. For example, in private value environments, if a social choice function is obviously strategy-proof (see Li (2017)), then it satisfies robust coalitional incentive compatibility. Besides, in two-agent environments, robust coalitional incentive compatibility can be guaranteed by ex-post incentive compatibility and ex-post weak Pareto efficiency.

The following proposition shows that the robust coalitional incentive compatibility condition is necessary for robust coalitional implementation. We leave the proof to the Appendix.

Proposition 1: *If a social choice function f is robustly coalitionally implementable, then f satisfies the robust coalitional incentive compatibility condition.*

To prevent the existence of bad interim equilibria, we introduce the robust coalitional monotonicity condition. Define a deception of agent i's payoff type as a set-valued mapping $\beta_i : \Theta_i \to 2^{\Theta_i} \backslash \{\emptyset\}$. The symbol $\beta = (\beta_i)_{i \in I}$ denotes a profile of deceptions. For any coalition $S \subseteq I$ and payoff type profile θ_S, denote $\beta_S(\theta_S) = (\beta_i(\theta_i))_{i \in S}$. We adopt the notation $\theta'_S \in \beta_S(\theta_S)$ when $\theta'_i \in \beta_i(\theta_i)$ for each $i \in S$. The deception profile is **acceptable** if $f(\theta) = f(\theta')$ for all $\theta \in \Theta$ and $\theta' \in \beta(\theta)$. Otherwise, we say the deception profile is **unacceptable**. The **coalitional reward set** of agent i, denoted by Y_i, is the collection of **coalitional reward functions** $y : \Theta_{-i} \to A$ satisfying the following conditions: for each

$S \subseteq I$ containing i, $\theta''_S \in \Theta_S$, and $\theta'_{S\setminus\{i\}} \in \Theta_{S\setminus\{i\}}$, there exists $j \in S$ such that

$$u_j\big(f(\theta''_S, \theta_{-S}), (\theta''_S, \theta_{-S})\big) \geq u_j\big(y(\theta'_{S\setminus\{i\}}, \theta_{-S}), (\theta''_S, \theta_{-S})\big),\ \forall \theta_{-S} \in \Theta_{-S}.$$

We remark that when f satisfies the robust coalitional incentive compatibility condition, the set Y_i is non-empty. This is because we can fix any θ_i and let $y : \Theta_{-i} \to A$ be defined by $y(\theta_{-i}) = f(\theta)$ for all $\theta_{-i} \in \Theta_{-i}$. By the robust coalitional incentive compatibility condition, when some coalition $S \ni i$ with payoff type profile θ''_S misreports $(\theta_i, \theta'_{S\setminus\{i\}})$, there should exist $j \in S$ such that j is weakly worse-off under all θ_{-S}.

Definition 5: *A social choice function f is said to satisfy the* ***robust coalitional monotonicity condition*** *if whenever a deception profile β is unacceptable, there exists $i \in I$, $\theta_i \in \Theta_i$, and $\theta'_i \in \beta_i(\theta_i)$ such that for any conjecture $\psi_i \in \Delta(\{(\theta_{-i}, \theta'_{-i}) | \theta_{-i} \in \Theta_{-i}, \theta'_{-i} \in \beta_{-i}(\theta_{-i})\})$, there exists $y \in Y_i$ such that*

$$\sum_{\theta_{-i} \in \Theta_{-i}, \theta'_{-i} \in \beta_{-i}(\theta_{-i})} u_i\big(y(\theta'_{-i}), (\theta_i, \theta_{-i})\big)\psi_i(\theta_{-i}, \theta'_{-i}) > \sum_{\theta_{-i} \in \Theta_{-i}, \theta'_{-i} \in \beta_{-i}(\theta_{-i})} u_i\big(f(\theta'_i, \theta'_{-i}), (\theta_i, \theta_{-i})\big)\psi_i(\theta_{-i}, \theta'_{-i}).$$

We call the agent i above a whistle-blower and the function $y \in Y_i$ a successful coalitional reward function. The robust coalitional monotonicity condition conveys the following meaning: when agents are assigned f but follow an unacceptable deception profile β, there exists a whistle-blower i who can signal that a bad equilibrium is reached, so that regardless of his conjecture of other agents' true and reported payoff types, i can profitably deviate by proposing a successful coalitional reward function y.

Our robust coalitional monotonicity condition is stronger than the robust monotonicity condition of Bergemann and Morris (2011). We will define the latter condition as a special case of Definition 8 and explain this relationship in Section 5.

The proposition below shows that the robust coalitional monotonicity condition is necessary for robust coalitional implementation. We relegate its proof to the Appendix.

Proposition 2: *If a social choice function f is robustly coalitionally implementable, then f satisfies the robust coalitional monotonicity condition.*

We provide below a group of easy-to-check (and stronger) sufficient conditions for robust coalitional monotonicity and robust coalitional implementation. This proposition may be useful in applications because the mechanism used for robust coalitional implementation is the direct mechanism, which is simpler than the general indirect mechanism in Theorem 1. We remark that this proposition can be applied to demonstrate the robust coalitional implementability of social choice functions in Example 1 (when $\Theta_i = \{0, 1\}$ for all $i \in I$) and in Examples 3 and 4.

Proposition 3: *If a social choice function f:* (1) *satisfies the robust coalitional incentive compatibility condition, and* (2) *is such that whenever a deception profile β is unacceptable, there exists an agent $i \in I$ with $\theta_i \in \Theta_i$ and $\theta_i' \in \beta_i(\theta_i)$ such that $u_i(f(\theta_i, \theta_{-i}'), (\theta_i, \theta_{-i})) > u_i(f(\theta_i', \theta_{-i}'), (\theta_i, \theta_{-i}))$ for all $\theta_{-i} \in \Theta_{-i}$ and $\theta_{-i}' \in \beta_{-i}(\theta_{-i})$, then f: (i) satisfies the robust coalitional monotonicity condition and (ii) is robustly coalitionally implemented by the direct mechanism f.*

If conditions (1) and (2) hold, then for each unacceptable deception profile β, there exists an agent i with payoff type θ_i misreporting $\theta_i' \in \beta_i(\theta_i)$ who can be a whistle-blower. The function $y : \Theta_{-i} \to A$ defined by $y(\theta_{-i}) = f(\theta_i, \theta_{-i})$ for all $\theta_{-i} \in \Theta_{-i}$ is a successful coalitional reward function. Intuitively, this is because this agent has a strict incentive to correct his own misreport regardless of the true and reported payoff types of other agents as long as they misreport according to β_{-i}. Hence, there is no bad interim equilibrium. The complete proof is in the Appendix.

4.1.2 Sufficient Conditions

The two conditions introduced in Section 4.1.1 are part of the sufficient conditions for robust coalitional implementation. We then introduce the interior coalitional reward property to complete the group of sufficient conditions. The property is not necessary since Proposition 3 has provided a sufficiency result for robust coalitional implementation in direct mechanisms without relying on it.

Definition 6: *A social choice function f satisfies the* ***interior coalitional reward prop-***

erty*, if for any agent $i \in I$, there exists a countable set $\hat{Y}_i \subseteq Y_i$, such that:*

(i) for all $\theta_i \in \Theta_i$ and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$, there exists $\underline{y}, \bar{y} \in \hat{Y}_i$ such that

$$\sum_{\theta_{-i} \in \Theta_{-i}, \theta'_{-i} \in \beta_{-i}(\theta_{-i})} u_i(\bar{y}(\theta'_{-i}), \theta)\psi_i(\theta_{-i}, \theta'_{-i}) > \sum_{\theta_{-i} \in \Theta_{-i}, \theta'_{-i} \in \beta_{-i}(\theta_{-i})} u_i(\underline{y}(\theta'_{-i}), \theta)\psi_i(\theta_{-i}, \theta'_{-i});$$

(ii) for any function $y \in Y_i$, sequence $(y^k \in \hat{Y}_i)_{k=1,2,\ldots}$, and vector $(\rho^k)_{k=0,1,2,\ldots} \in \Delta$, the function $\rho^0 y + \sum_{k=1,2,\ldots} \rho^k y^k \in Y_i$.

The above property implies that for each agent i, there is a countable subset $\hat{Y}_i \subseteq Y_i$, such that for each type of him, there always exist at least two rankable functions in $\hat{Y}_i$. Besides, it is required that for every $y \in Y_i$, a lottery over $\hat{Y}_i \cup \{y\}$ is still a coalitional reward function.

The interior coalitional reward property holds trivially when agents have quasilinear utility functions, the non-linear utility parts are bounded, and the monetary transfers can be sufficiently low. For example, for agent $i \in I$, consider two functions $\bar{y}, \underline{y} \in Y_i$, which always offer sufficiently low transfer to every agent. Let agents' transfers under $\bar{y}$ to be slightly higher than those under $\underline{y}$ and every other component under $\bar{y}$ and $\underline{y}$ be the same. Then the set $\hat{Y}_i \equiv \{\bar{y}, \underline{y}\}$ can satisfy the two requirements in the interior coalitional reward property.

The fact that there are two rankable functions in $\hat{Y}_i$ is used to dissolve bad equilibria in our mechanism in the next section. We use the rankable functions to create an open set of outcomes from which a consumer cannot find an optimal one. The idea is similar to the conditional no total indifference property of Bergemann and Morris (2011).

4.2 Mechanism

To establish the following sufficiency theorem on robust coalitional implementation, we will construct a new mechanism explicitly. Then we will explain why the existing mechanism of Bergemann and Morris (2011) cannot fulfill the goal of robust coalitional implementation.

Theorem 1: *If a social choice function f satisfies the robust coalitional incentive compatibility condition, the robust coalitional monotonicity condition, and the interior coalitional reward property, then f is robustly coalitionally implementable.*

Consider a mechanism where each agent i reports a message $m_i = (m_i^1, m_i^2, m_i^3, m_i^4) \in M_i^1 \times M_i^2 \times M_i^3 \times M_i^4$. The first component $m_i^1 \in M_i^1 \equiv \Theta_i$ reports a payoff type, the second one $m_i^2 \in M_i^2 \equiv \{B, NB\}$ means to blow a whistle or not, $m_i^3 \in M_i^3 \equiv Y_i$ proposes a coalitional reward function, and $m_i^4 \in M_i^4 \equiv \mathbb{N}_+$ is a non-negative integer.

We partition the message space into subsets $\bar{M}$ and $\hat{M}$ as follows:

$\bar{M} = \{m | m_i = (\cdot, NB, \cdot, \cdot) \, \forall i \in I\}$,

$\hat{M}(S) = \{m | m_i = (\cdot, B, \cdot, \cdot) \, \forall i \in S; m_j = (\cdot, NB, \cdot, \cdot) \, \forall j \notin S\}$,

$\hat{M} = \bigcup_{S \in 2^I \setminus \{\emptyset\}} \hat{M}(S)$.

Rule 1. If $m \in \bar{M}$, let the outcome allocation be $g(m) = f(m^1)$.

Rule 2. If $m \in \hat{M}$, there exists a unique coalition $S \subseteq I$ such that $m \in \hat{M}(S)$. We define $i^* \equiv \min S$, i.e., i^* is the agent with the smallest index among those who blow a whistle. By the interior coalitional reward property, there exists a countable set $\hat{Y}_{i^*} \subseteq Y_{i^*}$. List the elements of $\hat{Y}_{i^*}$ by y^1, y^2,... When the cardinality of $\hat{Y}_{i^*}$, denoted by K, is finite, then we define $y^k \equiv y^K$ for all $k > K$. Let the outcome $g(m)$ be a lottery of realization $m_{i^*}^3(m_{-i^*}^1)$ with probability $\frac{m_{i^*}^4}{m_{i^*}^4+1}$ and of realization $y^k(m_{-i^*}^1)$ with probability $\frac{0.5^k}{m_{i^*}^4+1}$ for $k = 1, 2, ...$

In the Appendix, we prove that (M, g) robustly coalitionally implements f. Now we provide a sketch of the proof. For convenience of notation, for each $i \in I$, we decompose $\sigma_i : T_i \to M_i$ into $\sigma_i = (\sigma_i^1, \sigma_i^2, \sigma_i^3, \sigma_i^4)$ so that $\sigma_i^k(t_i) \in M_i^k$ for each $k = 1, ..., 4$ and $t_i \in T_i$.

Claim 1 in the Appendix establishes that regardless of the type space and belief revising rule, it is an interim strong equilibrium for each agent to truthfully report his payoff type without blowing a whistle. This strategy profile always triggers Rule 1. By robust coalitional incentive compatibility, no coalition can profit from staying with Rule 1 but misreporting payoff types. In addition, no coalition can benefit from triggering Rule 2 because deviating to a coalitional reward function is not profitable.

Claim 2 demonstrates that in any interim equilibrium under any type space and belief revising rule, agents do not blow a whistle. Suppose that there is a type space, a belief revising rule, and an interim equilibrium σ in which some agent blows a whistle. Then, we can find an agent-type pair denoted by j and t_j^* such that regardless of $t_{-j} \in T_{-j}$, type-t_j^* agent j is always the agent with the smallest index who blows a whistle under $\sigma(t_j^*, t_{-j})$. From t_j^*'s point of view, by playing $\sigma_j(t_j^*)$, the outcome is assigned according to the coalitional

reward function $y \equiv \sigma_j^3(t_j^*)$ with probability $\frac{\sigma_j^4(t_j^*)}{1+\sigma_j^4(t_j^*)}$ and according to a full-support lottery over $\hat{Y}_j$ with probability $\frac{1}{1+\sigma_j^4(t_j^*)}$. However, we show that t_j^* can be better off by proposing a better coalitional reward function than y or decreasing the probability that the full-support lottery is assigned.

Claim 3 further shows that in any interim equilibrium, agents follow an acceptable deception profile to report payoff types. Otherwise, there exists a whistle-blower who can profitably deviate by proposing a successful coalitional reward function and submitting a large integer so that the outcome approximates the one under the coalitional reward function.

The three claims jointly establish that (M, g) robustly coalitionally implements f.

We remark that our mechanism mainly differs from the one of Bergemann and Morris (2011) in the allocation when $m \in \hat{M}(S)$ for some non-singleton S. In our mechanism, we let the agent with the smallest index among those who blow a whistle propose a coalitional reward function. However, their mechanism lets each agent propose an unrestricted outcome of his choice, and the outcome proposed by each agent is realized with positive probability. As the unrestricted outcome might lead to a profitable coalitional deviation from the good strategy profile described in our Claim 1, we cannot follow their mechanism for robust coalitional implementation.

4.3 Examples

In the following example, we present an environment with $n \geq 3$ and $|\Theta_i| \geq 3$ for all $i \in I$, as well as a social choice function on public good provision based on Bierbrauer and Hellwig (2016). The social choice function is robustly coalitionally implementable.

Example 3: *There is a finite set $V = \{v^1, v^2, ..., v^L\}$ such that $\Theta_i = V$ for all $i \in I$. Suppose it is feasible to produce $x_0 \in \mathbb{R}_+$ units of public good at the total cost $ncx_0 \geq 0$. Each agent i has a private value utility $u_i(x, \theta_i) = \theta_i x_0 + x_i$, where x_i is the transfer received by agent i. Suppose it is not feasible for the mechanism designer to run into budget deficit and the set of deterministic feasible outcomes is given by $X = \{(x_0, (x_i)_{i\in I}) : x_0 \in \mathbb{R}_+, x_i \in \mathbb{R}, \forall i \in I, ncx_0 + \sum_{i\in I} x_i \leq 0\}$. Assume $v^1 < c < v^L$ for the problem to be interesting and $c \notin V$ to avoid indifference.*

Define $s_1(\theta) \equiv |\{i \in I : \theta_i > c\}|$, *which represents the number of agents whose private evaluations are higher than* c. *Let* $q : \{0, 1, 2, ..., n\} \to \mathbb{R}_+$ *be a strictly increasing function. Consider a social choice function* $f = (f_0, (f_i)_{i \in I})$ *defined by* $f_0(\theta) = q(s_1(\theta))$ *and* $f_i(\theta) = -cf_0(\theta)$ *for all* $\theta \in \Theta$. *Namely, the public good provision level is decided by the number of agents whose evaluations are higher than* c *through a strictly increasing* q *function; all agents share the cost of production.*

The social choice function f *satisfies robust coalitional incentive compatibility. Suppose a coalition* $S \subseteq I$ *with payoff types* θ_S *can benefit from misreporting* θ'_S. *Then there exists* θ_{-S} *such that* $f(\theta) \neq f(\theta'_S, \theta_{-S})$, *which implies that* $s_1(\theta) \neq s_1(\theta'_S, \theta_{-S})$. *Notice that in the special case* $S = I$, *this should read as* $s_1(\theta) \neq s_1(\theta')$. *If* $s_1(\theta) < s_1(\theta'_S, \theta_{-S})$, *then a higher level of public good is provided, which is profitable for coalition* S *with payoff types* θ_S *only if* $\theta_i > c$ *for all* $i \in S$. *However, under* f, *coalition* S *with payoff types* θ_S *cannot misreport to further increase the level of public good provision. The case that* $s_1(\theta) > s_1(\theta'_S, \theta_{-S})$ *can be analyzed similarly. Hence,* f *satisfies robust coalitional incentive compatibility.*

The social choice function f *also satisfies condition* (2) *required in Proposition 3. To see this, whenever a deception profile* β *is unacceptable, there exists an agent* i *with a true evaluation* $\theta_i < c$ *misreporting some evaluation* $\theta'_i > c$ *or the other way around. For this agent* i, $u_i(f(\theta_i, \theta'_{-i}), \theta_i) > u_i(f(\theta'_i, \theta'_{-i}), \theta_i)$ *for all* $\theta'_{-i} \in \Theta_{-i}$. *Namely, he has an incentive to correct his misreport regardless of the strategy taken by other agents. Hence, condition* (2) *required in Proposition 3 is satisfied.*

By Proposition 3, f *is robustly coalitionally implemented by the direct mechanism* f. *In fact, a voting mechanism that elicits if each* θ_i *is higher than* c *or not robustly coalitionally implements this social choice function.*

In the following example, we present a robustly coalitionally implementable social choice function f on private good allocation based on Bergemann and Morris (2009).

Example 4: *The mechanism designer has one unit of indivisible private good to allocate. Normalize* Θ_i *so that* $\Theta_i \subseteq [0, 1]$ *for all* i. *Agent* i *has private value utility* $u_i(x, \theta_i) = \theta_i x_i^0 + x_i^1$ *where* x_i^0 *and* x_i^1 *represent the allocation and transfer to agent* i *respectively. The set of deterministic feasible outcomes is* $X = \{(x_i^0, x_i^1)_{i \in I} : x_i^0 \in \{0, 1\}, x_i^1 \in \mathbb{R}, \sum_{i \in I} x_i^0 \in$

$\{0,1\}, \sum_{i\in I} x_i^1 \leq 0\}$.

We define a social choice function f by defining $\bar{f}$ and $\hat{f}^j$ for each $j \in I$ first. Let $\bar{f}$ be a second-price auction with the tie-breaking rule in favor of the agent with the smallest index. Formally, $\bar{f} = (\bar{f}_i^0, \bar{f}_i^1)_{i\in I}$, where $(\bar{f}_i^0(\theta), \bar{f}_i^1(\theta)) = (1, -\max_{j\neq i}\{\theta_j\})$ if i is the agent with the smallest index among those reporting the highest payoff type, and $(\bar{f}_i^0(\theta), \bar{f}_i^1(\theta)) = (0,0)$ otherwise. Notice that $\bar{f}$ is ex-post incentive compatible, but there may not always be a strict incentive to truthfully report.

Then, for each $j \in I$, define a social choice function $\hat{f}^j = (\hat{f}_i^{j0}, \hat{f}_i^{j1})_{i\in I}$ where

$$\hat{f}_j^{j0}(\theta) = \theta_j, \hat{f}_j^{j1}(\theta) = -0.5\theta_j^2, \hat{f}_i^{j0}(\theta) = \hat{f}_i^{j1}(\theta) = 0 \text{ for all } i \neq j \text{ and } \theta \in \Theta.$$

Agent j's allocation (which is stochastic) and transfer depend on his own report only. Other agents receive neither the good nor any transfer. Notice that this inefficient social choice function can strictly elicit agent j's payoff type without affecting others' incentives.

Fix a constant $\epsilon \in (0,1)$ that is sufficiently close to zero. Define $f \equiv \epsilon \sum_{j\in I} \hat{f}^j/n + (1-\epsilon)\bar{f}$, which can be viewed as a stochastic mechanism equal to $\bar{f}$ with probability $1-\epsilon$ and equal to $\hat{f}^j$ with probability ϵ/n for each $j \in I$. Since ϵ is small, f is approximately equal to the second-price auction. Notice that the only difference between f and the one proposed by Bergemann and Morris (2009) is that we replace their uniform tie-breaking rule with a biased one to avoid profitable coalitional deviations.

We claim that f satisfies robust coalitional incentive compatibility. Since f is ex-post incentive compatible, no individual will unilaterally misreport. Whenever a coalition $S \subseteq I$ with $|S| \geq 2$ and payoff types θ_S misreports θ'_S, at most one member in S can strictly benefit due to the biased tie-breaking rule, and thus there is no profitable coalitional deviation either.

Now fix any unacceptable deception profile β. There must exist $i \in I$ with payoff type θ_i misreporting $\theta'_i \neq \theta_i$. For this agent, $u_i(f(\theta_i, \theta'_{-i}), \theta_i) > u_i(f(\theta'_i, \theta'_{-i}), \theta_i)$ for all $\theta'_{-i} \in \Theta_{-i}$, since $\hat{f}^i$ gives him a strict incentive to truthfully report. This inequality implies condition (2) *required by Proposition 3.*

By Proposition 3, f is robustly coalitionally implemented by the direct mechanism f.

5 Robust $\mathcal{S}$ Implementation

In this section, we assume that the mechanism designer knows the coalition pattern $\mathcal{S}$ and thus knows that agents play an interim $\mathcal{S}$ equilibrium. We will provide sufficient conditions for robust $\mathcal{S}$ implementation and construct a mechanism explicitly. When a coalition pattern $\mathcal{S}$ is richer than $\underline{\mathcal{S}}$, the sufficient conditions for robust $\mathcal{S}$ implementation do not imply those for robust implementation, and vice versa. This leaves leeway to robustly $\mathcal{S}$ implement some social choice functions that are not robustly implementable in the non-cooperative framework.

5.1 Conditions

The first condition is the robust $\mathcal{S}$ incentive compatibility condition, which prevents any admissible coalition from misreporting.

Definition 7: *A social choice function f is said to satisfy the* ***robust $\mathcal{S}$ incentive compatibility*** *condition if for all $S \in \mathcal{S}$ and $\theta'_S \neq \theta^*_S$, there exists $i \in S$ such that*

$$u_i\big(f(\theta^*_S, \theta_{-S}), (\theta^*_S, \theta_{-S})\big) \geq u_i\big(f(\theta'_S, \theta_{-S}), (\theta^*_S, \theta_{-S})\big) \text{ for all } \theta_{-S} \in \Theta_{-S}.$$

The smaller the coalition pattern is, the weaker the robust $\mathcal{S}$ incentive compatibility condition is. In particular, robust $\underline{\mathcal{S}}$ incentive compatibility is equivalent to the **ex-post incentive compatibility** condition in the literature. Robust coalitional incentive compatibility implies robust $\mathcal{S}$ incentive compatibility for all coalition pattern $\mathcal{S}$ and is equivalent to robust $\bar{\mathcal{S}}$ incentive compatibility.

Similar to Proposition 1, it is easy to show that robust $\mathcal{S}$ incentive compatibility is necessary for robust $\mathcal{S}$ implementation.

Then we define the $\mathcal{S}$ reward set and the robust $\mathcal{S}$ monotonicity condition. For each $S \subseteq I$, the $\mathcal{S}$ **reward set**, $Y_S[\mathcal{S}]$, is the collection of all $\mathcal{S}$ **reward functions** $y : \Theta_{-S} \to A$ subject to the following restriction: for each $\bar{S}$ such that $S \subseteq \bar{S} \in \mathcal{S}$, payoff type profile $\theta'_{\bar{S}\setminus S} \in \Theta_{\bar{S}\setminus S}$, and payoff type profile $\theta''_{\bar{S}} \in \Theta_{\bar{S}}$, there exists $i \in \bar{S}$ such that

$$u_i\big(f(\theta''_{\bar{S}}, \theta_{-\bar{S}}), (\theta''_{\bar{S}}, \theta_{-\bar{S}})\big) \geq u_i\big(y\big(\theta'_{\bar{S}\setminus S}, \theta_{-\bar{S}}\big), (\theta''_{\bar{S}}, \theta_{-\bar{S}})\big), \forall \theta_{-\bar{S}} \in \Theta_{-\bar{S}}.$$

To unify the notation, in the special case $S = I$, the set Θ_{-S} degenerates and each $y : \Theta_{-S} \to A$ should be viewed as a constant function with the range in A. When f satisfies robust $\mathcal{S}$ incentive compatibility, the set $Y_S[\mathcal{S}]$ is non-empty for any coalition $S \subseteq I$. To see this, when there does not exist $\bar{S}$ such that $S \subseteq \bar{S} \in \mathcal{S}$, $Y_S[\mathcal{S}]$ is the collection of all mappings from Θ_{-S} to A and thus is non-empty. When there exists $\bar{S}$ such that $S \subseteq \bar{S} \in \mathcal{S}$, we can fix any $\theta_S \in \Theta_S$ and define $y(\theta_{-S}) = f(\theta)$ for all $\theta_{-S} \in \Theta_{-S}$. By robust $\mathcal{S}$ incentive compatibility, for all $\bar{S}$ satisfying $S \subseteq \bar{S} \in \mathcal{S}$, $\theta'_{\bar{S}\setminus S} \in \Theta_{\bar{S}\setminus S}$, and $\theta''_{\bar{S}} \in \Theta_{\bar{S}}$, there exists $i \in \bar{S}$ such that

$$u_i\big(f(\theta''_{\bar{S}}, \theta_{-\bar{S}}), (\theta''_{\bar{S}}, \theta_{-\bar{S}})\big) \geq u_i\big(f(\theta_S, \theta'_{\bar{S}\setminus S}, \theta_{-\bar{S}}), (\theta''_{\bar{S}}, \theta_{-\bar{S}})\big) = u_i\big(y\big(\theta'_{\bar{S}\setminus S}, \theta_{-\bar{S}}\big), (\theta''_{\bar{S}}, \theta_{-\bar{S}})\big)$$

for all $\theta_{-\bar{S}} \in \Theta_{-\bar{S}}$. Hence, $y \in Y_S[\mathcal{S}]$, and moreover, $Y_S[\mathcal{S}]$ is non-empty again.

We remark that the coalitional reward set $Y_i = Y_i[\bar{\mathcal{S}}] \subseteq Y_i[\mathcal{S}]$ for all $i \in I$ and coalition pattern $\mathcal{S}$.

Definition 8: *A social choice function f satisfies the **robust $\mathcal{S}$ monotonicity** condition if whenever a deception profile β is unacceptable, there exists $S \in \mathcal{S}$, $\theta_S \in \Theta_S$, and $\theta'_S \in \beta_S(\theta_S)$ such that for any conjectures $(\psi_i \in \Delta(\{(\theta_{-S}, \theta'_{-S}) | \theta_{-S} \in \Theta_{-S}, \theta'_{-S} \in \beta_{-S}(\theta_{-S})\}))_{i \in S}$, there exists $y \in Y_S[\mathcal{S}]$ such that for all $i \in S$,*

$$\sum_{\theta_{-S} \in \Theta_{-S}, \theta'_{-S} \in \beta_{-S}(\theta_{-S})} u_i\big(y(\theta'_{-S}), (\theta_S, \theta_{-S})\big)\psi_i(\theta_{-S}, \theta'_{-S}) > \sum_{\theta_{-S} \in \Theta_{-S}, \theta'_{-S} \in \beta_{-S}(\theta_{-S})} u_i\big(f(\theta'_S, \theta'_{-S}), (\theta_S, \theta_{-S})\big)\psi_i(\theta_{-S}, \theta'_{-S}).$$

The robust $\mathcal{S}$ monotonicity condition allows a coalition $S \in \mathcal{S}$ to dissolve a bad equilibrium by proposing a function in the $\mathcal{S}$ reward set. Briefly speaking, in various monotonicity conditions under non-cooperative frameworks, when a deception profile is unacceptable, one agent reverses his ranking between two outcomes: one reward outcome and one social choice outcome, under two states. In our robust $\mathcal{S}$ monotonicity condition, one coalition switches its ranking rather than one agent. In the literature, Hahn and Yannelis (2001)'s coalitional Bayesian monotonicity condition under a given type space and Pasin (2009)'s coalitional monotonicity condition under complete information have a similar feature.

The robust coalitional monotonicity condition introduced in Section 4.1.1 is stronger than the robust $\mathcal{S}$ monotonicity condition for any coalition pattern $\mathcal{S}$.

It is easy to see that the robust $\underline{\mathcal{S}}$ monotonicity condition is equivalent to the **robust monotonicity** condition of Bergemann and Morris (2011). One may wonder if robust monotonicity is equivalent to robust coalitional monotonicity and the answer is no in general. This is because a singleton whistle-blower only needs to propose an element in $Y_i[\underline{S}]$ rather than Y_i in the robust monotonicity condition. However, in quasilinear environments where the transfers can be sufficiently low, the robust monotonicity condition is equivalent to robust coalitional monotonicity. To see this, suppose agents follow an unacceptable deception profile. When the robust monotonicity condition is satisfied, there exists an agent i who can benefit from proposing some $y \in Y_i[\underline{\mathcal{S}}]$. By sufficiently decreasing the transfer of each $j \neq i$ in y to construct $\hat{y}$, one can see that $\hat{y}$ is a successful coalitional reward function and thus the robust coalitional monotonicity condition holds.

The fact that robust monotonicity may be equivalent to robust coalitional monotonicity, which further implies robust $\mathcal{S}$ monotonicity for all $\mathcal{S}$, gives us leeway to implement some social choice functions that are not robustly implementable. For instance, the one in Example 2 does not satisfy robust monotonicity and fails to be robustly implementable, but it satisfies robust $\bar{\mathcal{S}}$ monotonicity and is robustly $\bar{\mathcal{S}}$ implementable. This observation may be surprising because it implies that robust monotonicity is not necessary for robust $\mathcal{S}$ implementation in general (e.g., when $\mathcal{S} = \bar{\mathcal{S}}$), although the Maskin monotonicity condition is necessary for implementation in strong equilibrium in complete information setting (see Maskin (1978)). Under the robust monotonicity condition, given any bad deception profile, there should be a whistle-blower to dissolve the deception profile regardless of his conjecture about the true and reported payoff types of all other agents. However, under the robust $\mathcal{S}$ monotonicity condition, it suffices to have an admissible coalition of agents who can dissolve the bad deception profile regardless of their conjectures about agents out of the coalition. Notice that the latter condition imposes no restriction on coalition members' conjectures with respect to each other, which is why there might exist a coalition of whistle-blowers to dissolve a bad deception profile though no singleton can play this role.

We remark that the robust $\mathcal{S}$ monotonicity condition is in general unnecessary for robust $\mathcal{S}$ implementation, although it is necessary under some special cases, for example, when $\mathcal{S} = \underline{\mathcal{S}}$ or $n = 2$. Robust $\mathcal{S}$ monotonicity assumes that given any unacceptable deception

profile β, there should be a coalition $S \in \mathcal{S}$ that can act as whistle-blowers regardless of the coalition's conjecture of those out of the coalition. In a sense, this requires the existence of the same whistle-blowing coalition under all belief structures. However, to dissolve bad interim $\mathcal{S}$ equilibria, the coalition of deviators may depend on the belief structure: when agents hold some belief structure, a coalition $S \in \mathcal{S}$ can be deviators, and under some other belief structure, a coalition $S' \in \mathcal{S}$ ($S \neq S'$ but S and S' can overlap) can be deviators. To support this, Example 6 in the Appendix provides a social choice function that does not satisfy robust $\mathcal{S}$ monotonicity, but is robustly $\mathcal{S}$ implementable.

We define the interim $\mathcal{S}$ monotonicity condition (Definition 12) in the Appendix, and by modifying the argument of Proposition 2, one can identify a necessary condition for robust $\mathcal{S}$ implementation: the interim $\mathcal{S}$ monotonicity condition should hold under all type spaces and all belief revising rules. However, whether the necessary condition holds or not is much more difficult to check than robust $\mathcal{S}$ monotonicity because one needs to consider all type spaces and belief revising rules. Because of this, we focus on the sufficient but unnecessary robust $\mathcal{S}$ monotonicity condition in the paper, which has already allowed us to robustly $\mathcal{S}$ implement some social choice functions that are not robustly implementable.

Similar to Proposition 3, we provide below a relatively easy-to-check sufficient condition for robust $\mathcal{S}$ monotonicity and robust $\mathcal{S}$ implementation.

Proposition 4: *If a social choice function f:* (1) *satisfies the robust $\mathcal{S}$ incentive compatibility condition, and* (2) *is such that whenever a deception profile β is unacceptable, there exists a coalition $S \in \mathcal{S}$ with $\theta_S \in \Theta_S$ and $\theta'_S \in \beta_S(\theta_S)$ such that $u_i(f(\theta_S, \theta'_{-S}), (\theta_S, \theta_{-S})) > u_i(f(\theta'_S, \theta'_{-S}), (\theta_S, \theta_{-S}))$ for all $i \in S$, $\theta_{-S} \in \Theta_{-S}$ and $\theta'_{-S} \in \beta_{-S}(\theta_{-S})$, then f: (i) satisfies the robust $\mathcal{S}$ monotonicity condition and (ii) is robustly $\mathcal{S}$ implemented by the direct mechanism.*

Condition (2) of the proposition requires that whenever β is unacceptable, an admissible coalition S with payoff type profile θ_S misreporting $\theta'_S \in \beta_S(\theta_S)$ has a strict incentive to revert to truthfully report regardless of the true and reported payoff types of other agents as long as they misreport according to β_{-S}. This coalition can be whistle-blowers and propose the $\mathcal{S}$ reward function $y : \Theta_{-S} \to A$ given by $y(\theta_{-S}) = f(\theta_S, \theta_{-S})$ for all $\theta_{-S} \in \Theta_{-S}$. The

rest of the proof is similar to Proposition 3 and thus is omitted.

Proposition 4 can be used to demonstrate the robust $\bar{\mathcal{S}}$ implementability of the social choice function in Example 2. In that example, whenever a deception profile is unacceptable, the grand coalition with some true and misreported payoff type profiles has the incentive to revert to truthfully report. Hence, the robust $\bar{\mathcal{S}}$ monotonicity condition holds. Recall that the belief structure is such that no individual has the incentive to unilaterally correct his own misreport.

At last, we introduce a weak condition, the interior $\mathcal{S}$ reward property, to complete the group of sufficient conditions. This property is also not necessary for robust $\mathcal{S}$ implementation.

Definition 9: *A social choice function f satisfies the **interior $\mathcal{S}$ reward property**, if for any coalition $S \subseteq I$ such that there exists $\bar{S}$ satisfying $S \subseteq \bar{S} \in \mathcal{S}$, there exists a countable set $\hat{Y}_S[\mathcal{S}] \subseteq Y_S[\mathcal{S}]$ such that:*

(i) for all $i \in S, \theta_i \in \Theta_i$, and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-S})$, there exists $\underline{y}, \bar{y} \in \hat{Y}_S[\mathcal{S}]$ such that

$$\sum_{\theta_{-i} \in \Theta_{-i}, \theta'_{-S} \in \Theta_{-S}} u_i(\bar{y}(\theta'_{-S}), \theta)\psi_i(\theta_{-i}, \theta'_{-S}) > \sum_{\theta_{-i} \in \Theta_{-i}, \theta'_{-S} \in \Theta_{-S}} u_i(\underline{y}(\theta'_{-S}), \theta)\psi_i(\theta_{-i}, \theta'_{-S});$$

(ii) for any function $y \in Y_S[\mathcal{S}]$, sequence $(y^k \in \hat{Y}_S[\mathcal{S}])_{k=1,2,\ldots}$, and vector $(\rho^k)_{k=0,1,2,\ldots} \in \Delta$, the function $\rho^0 y + \sum_{k=1,2,\ldots} \rho^k y^k \in Y_S[\mathcal{S}]$.

According to this property, whenever there exists $\bar{S}$ such that $S \subseteq \bar{S} \in \mathcal{S}$, there always exists a countable set $\hat{Y}_S[\mathcal{S}] \subseteq Y_S[\mathcal{S}]$, such that for each $i \in S$, there are rankable functions in $\hat{Y}_S[\mathcal{S}]$. Furthermore, for each $y \in Y_S[\mathcal{S}]$, every lottery over $\hat{Y}_S[\mathcal{S}] \cup \{y\}$ should still fall in the $\mathcal{S}$ reward set $Y_S[\mathcal{S}]$. Similar to the interior coalitional reward property, when agents have quasilinear utility functions where the non-linear parts are bounded and the monetary transfers can be sufficiently low, the interior $\mathcal{S}$ reward property holds trivially.

It is easy to see that the interior $\bar{\mathcal{S}}$ reward property implies the interior coalitional reward property, and the latter implies the interior $\underline{\mathcal{S}}$ reward property.[7] Moreover, when X is

[7] The fact that the interior coalitional reward property may not imply the interior $\mathcal{S}$ reward property does not contradict the fact that robust coalitional implementation implies robust $\mathcal{S}$ implementation, because the interior $\mathcal{S}$ reward property is not necessary for robust $\mathcal{S}$ implementation.

countable, f satisfies the interior $\underline{\mathcal{S}}$ reward property if and only if it satisfies the conditional no total indifference property introduced by Bergemann and Morris (2009) (provided in Definition 13 in the Appendix). The "if" direction is formally established in the proof of our Corollary 1 and the "only if" direction is trivial.

5.2 Mechanism

Under the sufficient conditions in Section 5.1, the mechanism constructed for Theorem 1 cannot robustly $\mathcal{S}$ implement f: when all agents follow an unacceptable deception profile and do not blow a whistle, the bad interim $\mathcal{S}$ equilibrium may not be dissolved even if $S \in \mathcal{S}$ is a coalition of whistle-blowers in the robust $\mathcal{S}$ monotonicity condition. This is because a successful $\mathcal{S}$ reward function $y \in Y_S[\mathcal{S}]$ may not be in Y_i for any $i \in S$. Hence, we propose a new mechanism to fulfill the goal of robust $\mathcal{S}$ implementation. The difference between this mechanism and the one in Theorem 1 is that we allow each i to propose an element of $Y_S[\mathcal{S}]$ contingent on different $S \subseteq I$ that he is a member of.

Theorem 2: *If a social choice function f satisfies the robust $\mathcal{S}$ incentive compatibility condition, the robust $\mathcal{S}$ monotonicity condition, and the interior $\mathcal{S}$ reward property, then f is robustly $\mathcal{S}$ implementable.*

In the mechanism (M, g), each agent i reports a message $m_i = (m_i^1, m_i^2, m_i^3, m_i^4) \in M_i^1 \times M_i^2 \times M_i^3 \times M_i^4$. The M_i^1, M_i^2, and M_i^4 components of the message space are identical to those in Theorem 1. The third component $m_i^3 \in M_i^3 \equiv \prod_{S \ni i, S \subseteq I} Y_S[\mathcal{S}]$ is a vector of $\mathcal{S}$ reward functions corresponding to different coalitions containing i. The partition of message space is identical to that in Theorem 1.

Rule 1. If $m \in \bar{M}$, let the outcome of the mechanism be $g(m) = f(m^1)$.

Rule 2. If $m \in \hat{M}$, there exists a unique coalition $S \subseteq I$ such that $m \in \hat{M}(S)$. Define $i^*[S] = \min S$, i.e., the agent with the smallest index who blows a whistle. If there exists $\bar{S} \in \mathcal{S}$ such that $S \subseteq \bar{S}$, we define $S^*[S] = S$; otherwise, let $S^*[S] = \{i^*[S]\}$. In the remainder of this paragraph, we adopt notations i^* and S^* rather than $i^*[S]$ and $S^*[S]$ for simplicity. Denote the component of $m_{i^*}^3$ that is in $Y_{S^*}[\mathcal{S}]$ by y. By the interior $\mathcal{S}$ reward property,

there exists a countable subset of $Y_{S^*}[\mathcal{S}]$, denoted by $\hat{Y}_{S^*}[\mathcal{S}] = \{y^1, y^2, ...\}$. When $\hat{Y}_{S^*}[\mathcal{S}]$ has cardinality $K < \infty$, define $y^k \equiv y^K$ for all $k > K$. Then let the outcome $g(m)$ be a lottery of realization $y(m^1_{-S^*})$ with probability $\frac{m^4_{i^*}}{m^4_{i^*}+1}$ and of realization $y^k(m^1_{-S^*})$ with probability $\frac{0.5^k}{m^4_{i^*}+1}$ for each $k = 1, 2, ...$

To prove that the mechanism (M, g) robustly $\mathcal{S}$ implements f, we relegate the analysis to the Appendix and only provide a sketch here.

Claim 4 in the Appendix shows that under all belief structures, it is an interim $\mathcal{S}$ equilibrium for agents to truthfully report payoff types without blowing a whistle. This follows from the robust $\mathcal{S}$ incentive compatibility condition and the definition of the $\mathcal{S}$ reward set.

Claim 5 shows that under all belief structures, it is never an interim $\mathcal{S}$ equilibrium for some agent to blow a whistle. Otherwise, we can find an agent j and a type t^*_j for whom there is a profitable unilateral deviation.

Claim 6 shows that under every belief structure and in every interim $\mathcal{S}$ equilibrium, agents should report according to an acceptable deception profile. Otherwise, some $S \in \mathcal{S}$ can deviate by blowing whistles and proposing a successful $\mathcal{S}$ reward function in $Y_S[\mathcal{S}]$.

By setting $\mathcal{S} = \bar{\mathcal{S}}$, the above mechanism can also be used to prove Theorem 1. However, we choose to present the simpler mechanism over there which can also be compared with the one of Bergemann and Morris (2011) more easily.

When $\mathcal{S} = \underline{\mathcal{S}}$, Theorem 2 provides sufficient conditions for robust implementation. By focusing on a countable set of deterministic feasible outcomes X, Theorem 2 of Bergemann and Morris (2011) proves that if f satisfies the robust monotonicity condition and the conditional no total indifference property, then f is robustly implementable. In the Appendix, we present the conditional no total indifference property and show that the sufficient conditions in their Theorem 2 imply ours. Hence, their Theorem 2 can be viewed as a special case of our Theorem 2.

Corollary 1 (Theorem 2, Bergemann and Morris (2011))**:** *Suppose the set of deterministic feasible outcomes X is countable. If a social choice function f satisfies the robust monotonicity condition and the conditional no total indifference property, then f is robustly implementable under all type spaces.*

5.3 Example

Based on Bierbrauer and Hellwig (2016), we present a social choice function on public good provision that is robustly $\bar{\mathcal{S}}$ implementable but not robustly implementable.

Example 5: *Consider the same setup as in Example 3, except for two modifications:* (1) *let V be the set of all rational numbers in the interval $[0,1]$,* (2) *let q be a weakly increasing function: $q(0) = q(1) = ... = q(n-1) = 0$ and $q(n) = 1$. Thus, the public good provision level is 1 if all agents agree to pay their share of cost $c \in [0,1] \backslash V$ and is 0 otherwise. We remark that these changes are crucial so that the social choice function f is robustly $\bar{\mathcal{S}}$ implementable but not robustly implementable.*

Similar to Example 3, it is easy to see that f satisfies robust $\bar{\mathcal{S}}$ incentive compatibility.

To establish robust $\bar{\mathcal{S}}$ monotonicity, first, recall that the outcome assigned by the social choice function f only depends on the number of agents whose reported private evaluations are higher than c. Thus, if a deception profile β is unacceptable, then there must be a true payoff type profile θ and a reported payoff type profile $\theta' \in \beta(\theta)$ such that $s_1(\theta) \neq s_1(\theta')$. As a result, an unacceptable deception profile β belongs to one of the following two cases.

Case 1, under β, there exists some agent with a payoff type higher than c misreporting a payoff type lower than c, i.e., there exists $i \in I$, $\theta_i > c$, and $\theta'_i \in \beta_i(\theta_i)$ with $\theta'_i < c$. We now identify a true payoff type profile θ and a misreported payoff type profile $\theta' \in \beta(\theta)$ such that grand coalition with true payoff type profile θ misreporting θ' can propose a successful $\bar{\mathcal{S}}$ reward function. Fix any θ_{-i} with $\theta_j > c$ for all $j \neq i$ and $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$. The grand coalition I with payoff type profile θ misreporting θ' can be the whistle-blowers proposing a constant successful $\bar{\mathcal{S}}$ reward function defined by $y = f(\theta)$: it is easy to see that $y \in Y_I[\bar{\mathcal{S}}]$ and $u_j(y, \theta_j) = u_j(f(\theta), \theta_j) > u_j(f(\theta'), \theta_j)$ for all $j \in I$.

Case 2, under β, there exists some agent with a payoff type lower than c misreporting a payoff type higher than c, but no agent misreports in the other way around. Intuitively, we want to find payoff type profiles $\theta \in \Theta$ and $\theta' \in \beta(\theta)$ such that truthful reporting and misreporting lead to zero and negative aggregate utilities of the grand coalition I respectively. Then, instead of consuming $f(\theta')$, by modifying $f(\theta)$ and transferring money within the grand coalition, the grand coalition with payoff type profile θ can propose a successful $\bar{\mathcal{S}}$ reward

function to benefit every agent. In particular, we let S denote the set of all agent i for whom there exist payoff types $\theta_i \in \Theta_i$ and $\theta'_i \in \beta_i(\theta_i)$ such that $\theta_i < c$ and $\theta'_i > c$. If $S \neq I$, we fix any θ_{-S} and $\theta'_{-S} \in \beta_{-S}(\theta_{-S})$ satisfying $\theta_j > c$ for all $j \notin S$ and $\sum_{j\in I} \theta_j < nc$. The existence of such a type profile θ_{-S} follows from the fact that V is dense and that $\sum_{i\in S} \theta_i < |S|c$. Since Case 2 assumes that no agent with a payoff type higher than c misreports a payoff type lower than c, in the payoff type profile θ'_{-S}, it must be true that $\theta'_j > c$ for all $j \notin S$. We summarize a few key observations as follows: $\sum_{j\in I} \theta_j < nc$, $f(\theta)$ involves not producing the public good, and $f(\theta')$ involves producing the public good. Then, we have $\sum_{j\in I} u_j(f(\theta), \theta_j) = 0 > \sum_{j\in I} u_j(f(\theta'), \theta_j)$. Thus, there exists a vector $(\tau_j)_{j\in I} \in \mathbb{R}^n$ such that $\sum_{j\in I} \tau_j = 0$ and $u_j(f(\theta), \theta_j) + \tau_j > u_j(f(\theta'), \theta_j)$ for all $j \in I$. It is easy to check that the constant function $y \equiv (f_0(\theta), (f_j(\theta) + \tau_j)_{j\in I})$ is in the set $Y_I[\bar{\mathcal{S}}]$ and is a successful $\bar{\mathcal{S}}$ reward function when the grand coalition with payoff type profile θ misreporting θ' serves as whistle-blowers.

To this end, we have verified the robust $\bar{\mathcal{S}}$ monotonicity condition.

The interior $\bar{\mathcal{S}}$ reward property is trivial, because of the quasilinearity setup. We omit the details.

By Theorem 2, f is robustly $\bar{\mathcal{S}}$ implementable.

At last, we remark that robust $\underline{\mathcal{S}}$ monotonicity fails and thus f is not robustly implementable. Consider an unacceptable deception profile where all agents always report payoff type 0. Suppose by way of contradiction that some agent i with payoff type θ_i can act as a whistle-blower and propose a successful $\underline{\mathcal{S}}$ reward function $y : \Theta_{-i} \to A$. Then the following inequalities must hold simultaneously: $u_i(y(\theta'_{-i}), \theta_i) > u_i(f(\theta'), \theta_i)$ and $u_i(f(\theta_i, \theta'_{-i}), \theta_i) \geq u_i(y(\theta'_{-i}), \theta_i)$, where $\theta'_j = 0$ for all $j \neq i$. This is a contradiction since $u_i(f(\theta'), \theta_i) = u_i(f(\theta_i, \theta'_{-i}), \theta_i) = 0$. Hence, f does not satisfy the robust $\underline{\mathcal{S}}$ monotonicity condition.

6 Concluding Remarks

This paper introduces coalition structures to study belief-free implementation. When the mechanism designer does not know what the coalition pattern is, we provide sufficient conditions to robustly coalitionally implement a social choice function under all type spaces and belief revising rules. When she knows that agents play an interim $\mathcal{S}$ equilibrium, we present

sufficient conditions for robust $\mathcal{S}$ implementation. Robust $\mathcal{S}$ implementation provides new insights on implementing some social choice functions that are not robustly implementable under the non-cooperative framework.

In our paper, coalition patterns are exogenously given. Since there are social choice functions that are not implementable under the non-cooperative framework but implementable under a cooperative framework, the mechanism designer may benefit from endogenously engineering coalitions. Koray and Yildiz (2018) and Korpela et al. (2020) have introduced the idea of designing rights structure or code of rights to Nash implementation problems. One may consider extending their approach to benefit the mechanism designer in Bayesian implementation or robust implementation problems. We leave this exercise for future study.

The recent literature has also explored the effects of various behavioral assumptions on implementation problems. For example, Hayashi et al. (2020) study strong implementation under complete information in a setting where players' choices need not be rational. Also under the complete information setting, Dutta and Sen (2012) and Lombardi and Yoshihara (2018, 2020) extend the Nash implementation literature by assuming that agents only misreport when they can strictly profit from doing so. Velez and Brown (2020) follow a behavioral approach to refine Nash equilibrium and to characterize implementable social choice functions under the alternative equilibrium concept. Under an incomplete information setting, there are papers studying how the assumption of ambiguity aversion affects the partial or full implementation of efficient and individually rational social choice functions. See, e.g., Liu (2016), de Castro et al. (2017a,b), Guo (2019), Guo and Yannelis (2021), Liu and Yannelis (2021). In future studies, it may be of interest to further explore how these behavioral components affect coalition manipulations in implementation problems.

A Appendix

Definition 10: *Given a type space and a belief revising rule, a social choice function f satisfies the* ***interim coalitional incentive compatibility*** *condition if there is no coalition $S \subseteq I$ and type profiles $t_S^* \neq t_S' \in T_S$ such that for all $i \in S$,*

$$\sum_{t_{-i}\in T_{-i}} u_i\Big(f\big(\hat{\theta}(t'_S,t_{-S})\big),\hat{\theta}(t^*_S,t_{-S})\Big)\pi_i(t^*_i)[t_{-i}|t^*_{S\setminus\{i\}}]$$
$$> \sum_{t_{-i}\in T_{-i}} u_i\Big(f\big(\hat{\theta}(t^*_S,t_{-S})\big),\hat{\theta}(t^*_S,t_{-S})\Big)\pi_i(t^*_i)[t_{-i}|t^*_{S\setminus\{i\}}].$$

Lemma 1: *If a social choice function f satisfies the interim coalitional incentive compatibility condition under all type spaces and all belief revising rules, then it satisfies the robust coalitional incentive compatibility condition.*

Proof. We prove by contrapositive. Suppose that f does not satisfy the robust coalitional incentive compatibility condition, i.e., there exists a coalition S and payoff type profiles $\theta^*_S \neq \theta'_S \in \Theta_S$ such that for all $i \in S$, there exists $\theta^i_{-S} \in \Theta_{-S}$ such that $u_i\big(f(\theta'_S,\theta^i_{-S}),(\theta^*_S,\theta^i_{-S})\big) > u_i\big(f(\theta^*_S,\theta^i_{-S}),(\theta^*_S,\theta^i_{-S})\big)$. Consider any payoff type space (a type space where for all $i \in I$, there is a bijection between T_i and Θ_i) satisfying the following restriction: for all $i \in S$ and $t^*_i \in T_i$ with $\hat{\theta}_i(t^*_i) = \theta^*_i$, $\pi_i(t^*_i)[t_{-i}] = 1$ for the type profile t_{-i} with payoff type profile $(\theta^*_{S\setminus\{i\}},\theta^i_{-S})$. For each $i \in S$, let t'_i denote the type with payoff type θ'_i. It is easy to see that type-t^*_S coalition S has the incentive to misreport t'_S. Therefore, f does not satisfy interim coalitional incentive compatibility. This is so under every belief revising rule. □

Proof of Proposition 1. Suppose f is robustly coalitionally implemented by (M,g), but does not satisfy robust coalitional incentive compatibility. By Lemma 1, there exists a type space and a belief revising rule under which there exists a coalition $S \subseteq I$ and type profiles $t^*_S \neq t'_S$ such that for all $i \in S$,

$$\sum_{t_{-i}\in T_{-i}} u_i\Big(f\big(\hat{\theta}(t'_S,t_{-S})\big),\hat{\theta}(t^*_S,t_{-S})\Big)\pi_i(t^*_i)[t_{-i}|t^*_{S\setminus\{i\}}]$$
$$> \sum_{t_{-i}\in T_{-i}} u_i\Big(f\big(\hat{\theta}(t^*_S,t_{-S})\big),\hat{\theta}(t^*_S,t_{-S})\Big)\pi_i(t^*_i)[t_{-i}|t^*_{S\setminus\{i\}}].$$

As f is robustly coalitionally implemented by (M,g), under the type space and the belief revising rule there exists an interim strong equilibrium σ such that $g(\sigma(t)) = f(\hat{\theta}(t))$ for all $t \in T$. Define a constant strategy σ'_i by $\sigma'_i(t_i) = \sigma_i(t'_i)$ for all $t_i \in T_i$ and $i \in S$. The strategy profile $(\sigma'_i)_{i\in S}$ makes type-t^*_S coalition S strictly better off, a contradiction. □

Definition 11: *Given a type space and a belief revising rule, a social choice function f satisfies the* ***interim coalitional monotonicity*** *condition if whenever a profile of mappings*

$(\alpha_i : T_i \to T_i)_{i\in I}$ is such that $f\big(\hat{\theta}(\bar{t})\big) \neq f\big(\hat{\theta}(\alpha(\bar{t}))\big)$ for some $\bar{t} \in T$, there exists an agent $i \in I$, a type $t_i^* \in T_i$, and a function $h : T \to A$ such that

$$(i)\ \sum_{t_{-i}\in T_{-i}} u_i\Big(h(\alpha(t_i^*, t_{-i})), \hat{\theta}(t_i^*, t_{-i})\Big)\pi_i(t_i^*)[t_{-i}] > \sum_{t_{-i}\in T_{-i}} u_i\Big(f\big(\hat{\theta}\big(\alpha(t_i^*, t_{-i})\big)\big), \hat{\theta}(t_i^*, t_{-i})\Big)\pi_i(t_i^*)[t_{-i}];$$

(ii) for any $S \subseteq I$ containing i and type profiles $t'_S, t''_S \in T_S$, there exists $j \in S$ such that

$$\sum_{t_{-j}\in T_{-j}} u_j\Big(f\big(\hat{\theta}(t''_S, t_{-S})\big), \hat{\theta}(t''_S, t_{-S})\Big)\pi_j(t''_j)[t_{-j}|t''_{S\setminus\{j\}}] \geq \sum_{t_{-j}\in T_{-j}} u_j\big(h(t'_S, t_{-S}), \hat{\theta}(t''_S, t_{-S})\big)\pi_j(t''_j)[t_{-j}|t''_{S\setminus\{j\}}].$$

Lemma 2: *If a social choice function f satisfies the interim coalitional monotonicity condition under all type spaces and all belief revising rules, then it satisfies the robust coalitional monotonicity condition.*

Proof. Suppose f satisfies interim coalitional monotonicity under all type spaces and all belief revising rules, but robust coalitional monotonicity fails. Then, there exists an unacceptable deception profile β, such that for all $i \in I$, $\theta_i \in \Theta_i$, and $\theta'_i \in \beta_i(\theta_i)$, there exists $\psi_i \in \Delta(\{(\theta_{-i}, \theta'_{-i})|\theta_{-i} \in \Theta_{-i}, \theta'_{-i} \in \beta_{-i}(\theta_{-i})\})$ such that for all $y \in Y_i$, it holds that

$$\sum_{\theta_{-i}\in\Theta_{-i}, \theta'_{-i}\in\beta_{-i}(\theta_{-i})} u_i\big(y(\theta'_{-i}), \theta\big)\psi_i(\theta_{-i}, \theta'_{-i}) \leq \sum_{\theta_{-i}\in\Theta_{-i}, \theta'_{-i}\in\beta_{-i}(\theta_{-i})} u_i\big(f(\theta'), \theta\big)\psi_i(\theta_{-i}, \theta'_{-i}). \quad (1)$$

It is without loss of generality to assume that β in the previous paragraph satisfies $\theta_i \in \beta_i(\theta_i)$ for all $i \in I$ and $\theta_i \in \Theta_i$. To see this, we show case by case that the unacceptable deception profile $\bar{\beta}$ defined by $\bar{\beta}_i(\theta_i) = \beta_i(\theta_i) \cup \{\theta_i\}$ for all $i \in I$ and $\theta_i \in \Theta_i$ can replace β in the previous paragraph. Case 1: for $\theta_i \in \Theta_i$, $\theta'_i \in \beta_i(\theta_i) \subseteq \bar{\beta}_i(\theta_i)$, there exists $\psi_i \in \Delta(\{(\theta_{-i}, \theta'_{-i})|\theta_{-i} \in \Theta_{-i}, \theta'_{-i} \in \beta_{-i}(\theta_{-i}) \subseteq \bar{\beta}_{-i}(\theta_{-i})\})$ such that whenever $y \in Y_i$, expression (1) is satisfied. Case 2: for each $i \in I$, $\theta_i \in \Theta_i$, and $\theta'_i \in \bar{\beta}_i(\theta_i)\setminus\beta_i(\theta_i)$, θ'_i has to be equal to θ_i. We can arbitrarily pick $\theta^*_{-i} \in \Theta_{-i}$ and let ψ_i be a distribution such that $\psi_i(\theta^*_{-i}, \theta^*_{-i}) = 1$. Then for any $y \in Y_i$, by the definition of Y_i, the following inequality holds:

$$\sum_{\theta_{-i}\in\Theta_{-i}, \theta'_{-i}\in\bar{\beta}_{-i}(\theta_{-i})} u_i\big(y(\theta'_{-i}), \theta\big)\psi_i(\theta_{-i}, \theta'_{-i}) = u_i\big(y(\theta^*_{-i}), (\theta_i, \theta^*_{-i})\big) \leq u_i\big(f(\theta_i, \theta^*_{-i}), (\theta_i, \theta^*_{-i})\big) = \sum_{\theta_{-i}\in\Theta_{-i}, \theta'_{-i}\in\bar{\beta}_{-i}(\theta_{-i})} u_i\big(f(\theta'), \theta\big)\psi_i(\theta_{-i}, \theta'_{-i}).$$

Since it is without loss of generality to assume that $\theta_i \in \beta_i(\theta_i)$ for all $i \in I$ and $\theta_i \in \Theta_i$, we can construct a type set $T_i = T_i^1 \cup T_i^2$ for each $i \in I$ by following Bergemann and Morris (2008). Then we will specify a belief revising rule.

Step 1. Define T_i^1. For each $i \in I$, define a bijection $\xi_i^1 : T_i^1 \to \{(\theta_i, \theta_i') : \theta_i \in \Theta_i, \theta_i' \in \beta_i(\theta_i)\}$ so that type t_i with $\xi_i^1(t_i) = (\theta_i, \theta_i')$ has a payoff type θ_i and belief type:

$$\pi_i(t_i)[t_{-i}] = \begin{cases} \psi_i(\theta_{-i}, \theta_{-i}') & \text{if } t_{-i} = \left([\xi_j^1]^{-1}(\theta_j, \theta_j')\right)_{j \neq i} \in T_{-i}^1; \\ 0 & \text{elsewhere.} \end{cases}$$

Step 2. Define T_i^2. Let the set T_i^2 be a bijection to Θ under $\xi_i^2 : T_i^2 \to \Theta$. Specifically, for type $t_i \in T_i^2$ with $\xi_i^2(t_i) = \theta$, let $\hat{\theta}_i(t_i) = \theta_i$ and the belief of t_i be

$$\pi_i(t_i)[t_{-i}] = \begin{cases} 1 & \text{if } t_{-i} = \left([\xi_j^1]^{-1}(\theta_j, \theta_j)\right)_{j \neq i} \in T_{-i}^1; \\ 0 & \text{elsewhere.} \end{cases}$$

Step 3. For each $i \in I$, define a mapping $\alpha_i : T_i \to T_i$ by:

$$\alpha_i(t_i) = \begin{cases} [\xi_i^1]^{-1}(\theta_i', \theta_i') & \text{if } t_i = [\xi_i^1]^{-1}(\theta_i, \theta_i') \in T_i^1; \\ t_i & \text{elsewhere.} \end{cases}$$

Step 4. Define the belief revising rule. For each $i \in I$, $t_i \in T_i$, $S \subseteq I$ containing i, and $t_{S\setminus\{i\}}$ happening with zero probability under distribution $\pi_i(t_i)$, we specify the following belief revising rule: let the revised belief $\pi_i(t_i)[(t_{S\setminus\{i\}}, t_{-S})|t_{S\setminus\{i\}}] = \pi_i(t_i)[t_{-S}]$ for all $t_{-S} \in T_{-S}$. Meanwhile, let $\pi_i(t_i)[(t'_{S\setminus\{i\}}, t_{-S})|t_{S\setminus\{i\}}] = 0$ for all $t'_{S\setminus\{i\}} \neq t_{S\setminus\{i\}}$ and $t_{-S} \in T_{-S}$.

Step 5. Yield a contradiction. As it is not true that $f\big(\hat{\theta}(t)\big) = f\big(\hat{\theta}(\alpha(t))\big)$ for all $t \in T$, by the interim coalitional monotonicity condition, there exists $i \in I$, $t_i^* \in T_i$, and $h : T \to A$ such that conditions (i) and (ii) in Definition 11 are satisfied. Define a function $y : \Theta_{-i} \to A$ by $y(\theta_{-i}) = h\Big(\alpha_i(t_i^*), \left([\xi_j^1]^{-1}(\theta_j, \theta_j)\right)_{j \neq i}\Big)$ for all $\theta_{-i} \in \Theta_{-i}$. According to condition (ii), by having $S \subseteq I$ go over every set containing i and t_S'' go over every type profile in T_S^2, it is easy to verify that $y \in Y_i$. We also know that $t_i^* \notin T_i^2$. Otherwise, condition (i) in Definition 11 would contradict with (ii) if we set $S = \{i\}$, given types in T_i^2 expect other agents to truthfully report under α_{-i} (i.e., for all $t_i \in T_i^2$ and $t_{-i} \in T_{-i}$ such that $\pi_i(t_i)[t_{-i}] > 0$, it holds that $\alpha_{-i}(t_{-i}) = t_{-i}$). Thus, $t_i^* \in T_i^1$ and condition (i) implies that

$$\sum_{\theta_{-i} \in \Theta_{-i}, \theta_{-i}' \in \beta_{-i}(\theta_{-i})} u_i\big(y(\theta_{-i}'), \theta\big)\psi_i(\theta_{-i}, \theta_{-i}') > \sum_{\theta_{-i} \in \Theta_{-i}, \theta_{-i}' \in \beta_{-i}(\theta_{-i})} u_i\big(f(\theta'), \theta\big)\psi_i(\theta_{-i}, \theta_{-i}'),$$

a contradiction with expression (1). Hence, robust coalitional monotonicity holds. □

Proof of Proposition 2. Suppose f is robustly coalitionally implemented by (M, g), but fails to satisfy robust coalitional monotonicity. From Lemma 2, there exists some type space and belief revising rule under which f does not satisfy interim coalitional monotonicity although it is interim coalitionally implementable. Under this type space and belief revising rule, let σ^* be an interim strong equilibrium such that $g\big(\sigma^*(t)\big) = f\big(\hat{\theta}(t)\big)$ for all $t \in T$. If under a profile of mappings $(\alpha_i : T_i \to T_i)_{i\in I}$, there exists $\bar{t} \in T$ such that $f\big(\hat{\theta}(\bar{t})\big) \neq f\big(\hat{\theta}(\alpha(\bar{t}))\big)$, then $\sigma^* \circ \alpha \equiv (\sigma_i^* \circ \alpha_i)_{i\in I}$ is not an interim equilibrium by Definition 2. Hence, there exists $i \in I$, $t_i^* \in T_i$, and $\sigma_i' : T_i \to M_i$ such that

$$\sum_{t_{-i}\in T_{-i}} u_i\Big(g\Big(\sigma_i'(t_i^*), \sigma_{-i}^*\big(\alpha_{-i}(t_{-i})\big)\Big), \hat{\theta}(t_i^*, t_{-i})\Big)\pi_i(t_i^*)[t_{-i}] \\ > \sum_{t_{-i}\in T_{-i}} u_i\Big(g\Big(\sigma^*\big(\alpha(t_i^*, t_{-i})\big)\Big), \hat{\theta}(t_i^*, t_{-i})\Big)\pi_i(t_i^*)[t_{-i}].$$

By defining $h : T \to A$ by $h(t) = g\big(\sigma_i'(t_i^*), \sigma_{-i}^*(t_{-i})\big)$ for all $t \in T$, we have

$$\sum_{t_{-i}\in T_{-i}} u_i\Big(h\big(\alpha(t_i^*, t_{-i})\big), \hat{\theta}(t_i^*, t_{-i})\Big)\pi_i(t_i^*)[t_{-i}] > \sum_{t_{-i}\in T_{-i}} u_i\Big(f\big(\hat{\theta}\big(\alpha(t_i^*, t_{-i})\big)\big), \hat{\theta}(t_i^*, t_{-i})\Big)\pi_i(t_i^*)[t_{-i}]. \tag{2}$$

Since σ^* is an interim strong equilibrium, for any coalition $S \subseteq I$ containing i with types $t_S'' \in T_S$, deviating to $(\sigma_i'(t_i^*), \sigma_{S\setminus\{i\}}^*(t'_{S\setminus\{i\}}))$ is never profitable regardless of $t'_{S\setminus\{i\}}$. Therefore, there exists an agent $j \in S$ such that

$$\sum_{t_{-j}\in T_{-j}} u_j\Big(g\big(\sigma^*(t_S'', t_{-S})\big), \hat{\theta}(t_S'', t_{-S})\Big)\pi_j(t_j'')[t_{-j}|t''_{S\setminus\{j\}}] \\ \geq \sum_{t_{-j}\in T_{-j}} u_j\Big(g\big(\sigma_i'(t_i^*), \sigma_{-i}^*(t'_{S\setminus\{i\}}, t_{-S})\big), \hat{\theta}(t_S'', t_{-S})\Big)\pi_j(t_j'')[t_{-j}|t''_{S\setminus\{j\}}].$$

Since the outcome assigned by h is independent of i's type, for all t_i' and $t'_{S\setminus\{i\}}$, we have

$$\sum_{t_{-j}\in T_{-j}} u_j\Big(f\big(\hat{\theta}(t_S'', t_{-S})\big), \hat{\theta}(t_S'', t_{-S})\Big)\pi_j(t_j'')[t_{-j}|t''_{S\setminus\{j\}}] \\ \geq \sum_{t_{-j}\in T_{-j}} u_j\big(h(t'_S, t_{-S}), \hat{\theta}(t_S'', t_{-S})\big)\pi_j(t_j'')[t_{-j}|t''_{S\setminus\{j\}}]. \tag{3}$$

Expressions (2) and (3) establish interim coalitional monotonicity, a contradiction. □

Proof of Proposition 3. To establish (i), suppose conditions (1) and (2) hold. Then for any unacceptable deception profile β, the agent i with payoff type θ_i misreporting $\theta_i' \in \beta_i(\theta_i)$ in condition (2) can be a whistle-blower. To see this, the function $y : \Theta_{-i} \to A$ defined by $y(\theta_{-i}) = f(\theta_i, \theta_{-i})$ for all $\theta_{-i} \in \Theta_{-i}$ is a coalitional reward function, because f satisfies robust coalitional incentive compatibility. The function y is also a successful coalitional reward function because of the strict inequality in condition (2). Hence, f satisfies the robust coalitional monotonicity condition.

To establish (ii), let f be the direct mechanism. Since robust coalitional incentive compatibility holds, $\sigma_i^*(t_i) = \hat{\theta}_i(t_i)$ for all $i \in I$ and $t_i \in T_i$ constitutes an interim strong equilibrium. Suppose by way of contradiction that there is an interim equilibrium σ, such that $f(\sigma(\bar{t})) \neq f(\hat{\theta}(\bar{t}))$ for some $\bar{t} \in T$. Define a deception β_i by $\beta_i(\theta_i) = \bigcup_{\{t_i \in T_i | \hat{\theta}_i(t_i) = \theta_i\}} \{\sigma_i(t_i)\}$ for all $i \in I$ and $\theta_i \in \Theta_i$. The deception profile β is unacceptable. By condition (2), there exists $i \in I$, $\theta_i \in \Theta_i$, and $\theta_i' \in \beta_i(\theta_i)$ such that $u_i(f(\theta_i, \theta_{-i}'), (\theta_i, \theta_{-i})) > u_i(f(\theta_i', \theta_{-i}'), (\theta_i, \theta_{-i}))$ for all $\theta_{-i} \in \Theta_{-i}$ and $\theta_{-i}' \in \beta_{-i}(\theta_{-i})$. Thus, for any type t_i such that $\hat{\theta}_i(t_i) = \theta_i$ and $\sigma_i(t_i) = \theta_i'$,

$$\sum_{t_{-i} \in T_{-i}} u_i(f(\sigma_i^*(t_i), \sigma_{-i}(t_{-i})), \hat{\theta}(t_i, t_{-i}))\pi_i(t_i)[t_{-i}] > \sum_{t_{-i} \in T_{-i}} u_i(f(\sigma(t_i, t_{-i})), \hat{\theta}(t_i, t_{-i}))\pi_i(t_i)[t_{-i}],$$

i.e., reverting to truthfully report is profitable. This contradicts the supposition that σ is an interim equilibrium. Hence, the direct mechanism robustly coalitionally implements f. $\square$

Proof of Theorem 1. We prove that (M, g) robustly coalitionally implements f.

Claim 1: *Under any type space and any belief revising rule, $\sigma_i^*(t_i) = (\hat{\theta}_i(t_i), NB, \cdot, \cdot)$ for all $i \in I$ and $t_i \in T_i$ constitutes an interim strong equilibrium of (M, g).*

Proof: We want to show that for any coalition $S \subseteq I$, realized type profile $t_S \in T_S$, and strategy profile σ_S, σ_S is not a profitable deviation from σ_S^*.

Case 1. Suppose $\sigma_i(t_i) = (\cdot, NB, \cdot, \cdot)$ for all $i \in S$. By robust coalitional incentive compatibility, σ_S is not profitable.

Case 2. Suppose there exists a non-empty subset $\underline{S} \subseteq S$ such that $\sigma_i(t_i) = (\cdot, B, \cdot, \cdot)$ for all $i \in \underline{S}$ and $\sigma_i(t_i) = (\cdot, NB, \cdot, \cdot)$ for all $i \in S \backslash \underline{S}$. Define $j \equiv \min \underline{S}$. For each $t_{-S} \in T_{-S}$, $g(\sigma_S(t_S), \sigma_{-S}^*(t_{-S})) \in \hat{M}(\underline{S})$ and thus the outcome is a compound lottery of

$y(\sigma^1_{S\setminus\{j\}}(t_{S\setminus\{j\}}), \hat{\theta}_{-S}(t_{-S}))$ and $\sum_{k=1,2,...} 0.5^k y^k(\sigma^1_{S\setminus\{j\}}(t_{S\setminus\{j\}}), \hat{\theta}_{-S}(t_{-S}))$, where $y \equiv \sigma^3_j(t_j) \in Y_j$ and $\sum_{k=1,2,...} 0.5^k y^k \in \hat{Y}_j$. By condition (ii) of the interior coalitional reward property, σ_S is not profitable for S.

Claim 2: *Under any type space and any belief revising rule, if σ is an interim equilibrium of the mechanism (M, g), then $\sigma(t) \in \bar{M}$ for all $t \in T$.*

Proof: We prove by contrapositive. Suppose there exists $\bar{t} \in T$ such that $\sigma(\bar{t}) \notin \bar{M}$. Let j be the agent with the smallest index for whom there exists $t^*_j \in T_j$ such that $\sigma^2_j(t^*_j) = B$. Notice that agent j is uniquely defined. We fix one type t^*_j with $\sigma^2_j(t^*_j) = B$ and will show below that t^*_j has a profitable deviation. Let θ^*_j denote $\hat{\theta}_j(t^*_j)$.

Denote $\hat{Y}_j = \{y^1, y^2, ...\}$ and $y = \sigma^3_j(t^*_j)$. For each $t_{-j} \in T_{-j}$, $g(\sigma(t^*_j, t_{-j}))$ is a lottery of realization $y(\sigma^1_{-j}(t_{-j}))$ with probability $\frac{\sigma^4_j(t^*_j)}{1+\sigma^4_j(t^*_j)}$ and of realization $y^k(\sigma^1_{-j}(t_{-j}))$ with probability $\frac{0.5^k}{1+\sigma^4_j(t^*_j)} > 0$ for $k = 1, 2, ...$ The distribution $\psi_j \in \Delta(\{(\theta_{-j}, \theta'_{-j}) | \theta_{-j} \in \Theta_{-j}, \theta'_{-j} \in \beta_{-j}(\theta_{-j})\})$ defined by

$$\psi_j(\theta_{-j}, \theta'_{-j}) \equiv \sum_{\hat{\theta}_{-j}(t_{-j})=\theta_{-j}, \sigma^1_{-j}(t_{-j})=\theta'_{-j}} \pi_j(t^*_j)[t_{-j}]$$

is the probability that t_{-j} has payoff type profile θ_{-j} and misreports θ'_{-j}. Thus, type-t^*_j agent j's expected utility is equal to

$$\frac{\sigma^4_j(t^*_j)}{1+\sigma^4_j(t^*_j)} \sum_{\theta_{-j}\in\Theta_{-j}, \theta'_{-j}\in\beta_{-j}(\theta_{-j})} u_j\big(y(\theta'_{-j}), (\theta^*_j, \theta_{-j})\big)\psi_j(\theta_{-j}, \theta'_{-j})$$
$$+ \frac{1}{1+\sigma^4_j(t^*_j)} \sum_{\theta_{-j}\in\Theta_{-j}, \theta'_{-j}\in\beta_{-j}(\theta_{-j})} \sum_{k=1,2,...} 0.5^k u_j\big(y^k(\theta'_{-j}), (\theta^*_j, \theta_{-j})\big)\psi_j(\theta_{-j}, \theta'_{-j}).$$

We now define a deviating strategy σ'_j based on two cases.

Case 1: suppose

$$\sum_{\theta_{-j}\in\Theta_{-j}, \theta'_{-j}\in\beta_{-j}(\theta_{-j})} u_j\big(y(\theta'_{-j}), (\theta^*_j, \theta_{-j})\big)\psi_j(\theta_{-j}, \theta'_{-j})$$
$$\leq \sum_{\theta_{-j}\in\Theta_{-j}, \theta'_{-j}\in\beta_{-j}(\theta_{-j})} \sum_{k=1,2,...} 0.5^k u_j(y^k(\theta'_{-j}), (\theta^*_j, \theta_{-j}))\psi_j(\theta_{-j}, \theta'_{-j}). \quad (4)$$

By the interior coalitional reward property, there exist integers $k' \neq k''$ such that

$$\sum_{\theta_{-j}\in\Theta_{-j},\theta'_{-j}\in\beta_{-j}(\theta_{-j})} u_j(y^{k'}(\theta'_{-j}),(\theta^*_j,\theta_{-j}))\psi_j(\theta_{-j},\theta'_{-j}) > \sum_{\theta_{-j}\in\Theta_{-j},\theta'_{-j}\in\beta_{-j}(\theta_{-j})} u_j(y^{k''}(\theta'_{-j}),(\theta^*_j,\theta_{-j}))\psi_j(\theta_{-j},\theta'_{-j}).$$

Thus, there must exist some $k \geq 1$ such that

$$\sum_{\theta_{-j}\in\Theta_{-j},\theta'_{-j}\in\beta_{-j}(\theta_{-j})} u_j\big(y(\theta'_{-j}),(\theta^*_j,\theta_{-j})\big)\psi_j(\theta_{-j},\theta'_{-j}) < \sum_{\theta_{-j}\in\Theta_{-j},\theta'_{-j}\in\beta_{-j}(\theta_{-j})} u_j\big(y^k(\theta'_{-j}),(\theta^*_j,\theta_{-j})\big)\psi_j(\theta_{-j},\theta'_{-j}).$$

Type-t^*_j agent j will be better off by deviating to σ'_j defined by $\sigma'_j(t^*_j) = (\sigma^1_j(t^*_j), \sigma^2_j(t^*_j), y^k, \sigma^4_j(t^*_j))$ and $\sigma'_j(t_j) = \sigma_j(t_j)$ for $t_j \neq t^*_j$.

Case 2: suppose expression (4) does not hold. Then t^*_j is better off by deviating to σ'_j defined by $\sigma'_j(t^*_j) = (\sigma^1_j(t^*_j), \sigma^2_j(t^*_j), \sigma^3_j(t^*_j), \sigma^4_j(t^*_j)+1)$ and $\sigma'_j(t_j) = \sigma_j(t_j)$ for $t_j \neq t^*_j$.

In both cases, σ is not an interim equilibrium.

Claim 3: *Under any type space and any belief revising rule, if σ is an interim equilibrium of (M,g), then $g(\sigma(t)) = f(\hat{\theta}(t))$ for all $t \in T$.*

Proof: From Claim 2, $g(\sigma(t)) = f(\sigma^1(t))$ for all $t \in T$. Suppose by way of contradiction that there exists $\bar{t} \in T$ such that $g(\sigma(\bar{t})) \neq f(\hat{\theta}(\bar{t}))$. Define a deception β_i by $\beta_i(\theta_i) = \bigcup_{\{t_i\in T_i|\hat{\theta}_i(t_i)=\theta_i\}}\{\sigma^1_i(t_i)\}$ for all $i \in I$ and $\theta_i \in \Theta_i$. The deception profile β is unacceptable.

By robust coalitional monotonicity, there exists $i \in I$, $\theta^*_i \in \Theta_i$, and $\theta'_i \in \beta_i(\theta^*_i)$ such that for any $\psi_i \in \Delta(\{(\theta_{-i},\theta'_{-i})|\theta_{-i} \in \Theta_{-i}, \theta'_{-i} \in \beta_{-i}(\theta_{-i})\})$, there exists $y \in Y_i$ such that

$$\sum_{\theta_{-i}\in\Theta_{-i},\theta'_{-i}\in\beta_{-i}(\theta_{-i})} u_i\big(y(\theta'_{-i}),(\theta^*_i,\theta_{-i})\big)\psi_i(\theta_{-i},\theta'_{-i}) > \sum_{\theta_{-i}\in\Theta_{-i},\theta'_{-i}\in\beta_{-i}(\theta_{-i})} u_i\big(f(\theta'_i,\theta'_{-i}),(\theta^*_i,\theta_{-i})\big)\psi_i(\theta_{-i},\theta'_{-i}). \quad (5)$$

Fix any type t^*_i such that $\hat{\theta}_i(t^*_i) = \theta^*_i$ and $\sigma^1_i(t^*_i) = \theta'_i$. Let the distribution $\psi_i \in \Delta(\{(\theta_{-i},\theta'_{-i})|\theta_{-i} \in$

$\Theta_{-i}, \theta'_{-i} \in \beta_{-i}(\theta_{-i})\}$) be defined by

$$\psi_i(\theta_{-i}, \theta'_{-i}) \equiv \sum_{\hat{\theta}_{-i}(t_{-i})=\theta_{-i}, \sigma^1_{-i}(t_{-i})=\theta'_{-i}} \pi_i(t_i^*)[t_{-i}].$$

Define a strategy σ'_i by $\sigma'_i(t_i^*) \equiv (\sigma_i^1(t_i^*), B, y, K^*)$ and $\sigma'_i(t_i) \equiv \sigma_i(t_i)$ for all $t_i \neq t_i^*$, where y satisfies expression (5) and $K^* > 0$ is sufficiently large. Thus, for each $t_{-i} \in T_{-i}$, $\big(\sigma'_i(t_i^*), \sigma_{-i}(t_{-i})\big) \in \hat{M}(\{i\})$ and the outcome is sufficiently close to $y(\sigma^1_{-i}(t_{-i}))$ when K^* is sufficiently large. According to expression (5), σ'_i is profitable for t_i^*, a contradiction.

In view of the three claims, (M, g) robustly coalitionally implements f. □

Example 6: *Suppose* $I = \{1, 2, 3\}$, $\mathcal{S} = \{\{1, 2\}, \{1\}, \{2\}, \{3\}\}$, $\Theta_1 = \Theta_2 = \{1, 2\}$, *and* $\Theta_3 = \{0, 1\}$. *There are four goods in the environment. The social planner can produce* $x^0 \in \{0, 1\}$ *unit of public good at the total cost of* $3x^0$. *The social planner can also allocate an indivisible private good, which is produced at no cost, to agents so that each agent* i *consumes* $x_i^1 \in \{0, 1\}$ *unit where* $\sum_{i \in I} x_i^1 \in \{0, 1\}$. *Each agent* i *receives a monetary transfer* x_i^2. *In addition, each agent* i *receives* $x_i^3 \in \{0, 1, 2\}$ *units of "pseudo" good, which is produced by the social planner at no cost. We remark that the pseudo good is introduced to capture an agent's intrinsic aversion to misreporting under some types. Also, the consumption of the pseudo good may cause externality. Define* $u_1(x, \theta) = \mathbb{1}_{\{\theta_3=0\}}\theta_1 x^0 + \mathbb{1}_{\{\theta_3=1\}}\theta_1 x_1^1 + x_1^2 - 0.1\mathbb{1}_{\{\theta_2=1 \, or \, \theta_3=1\}}|\theta_1 - x_1^3|$, $u_2(x, \theta) = \mathbb{1}_{\{\theta_3=0\}}\theta_2 x^0 + \mathbb{1}_{\{\theta_3=1\}}\theta_2 x_2^1 + x_2^2 - 0.1\mathbb{1}_{\{\theta_1=1 \, or \, \theta_3=1\}}|\theta_2 - x_2^3| - 0.2\mathbb{1}_{\theta=(2,2,1), x^3=(2,1,1)}$, *and* $u_3(x, \theta) = -|\theta_3 - x_3^3|$. *Agent* 3 *only cares about the pseudo good: he strictly prefers to consume* θ_3 *unit of pseudo good than any other number when his payoff type is* θ_3. *Thus, agent* 3 *has a strict incentive to truthfully reveal* θ_3. *When* $\theta_3 = 0$, *agents* 1 *and* 2 *essentially play a modified public good game where agent* 1 *(resp. agent* 2*) has a small intrinsic aversion to misreporting if* $\theta_2 = 1$ *(resp.* $\theta_1 = 1$*); when* $\theta_3 = 1$, *agents* 1 *and* 2 *essentially play a modified private good allocation game where they both have a small intrinsic aversion to misreporting. Moreover, when the true and reported payoff type profiles are* $(2, 2, 1)$ *and* $(2, 1, 1)$ *respectively, agent* 2 *has an additional aversion to misreporting, measured by the term* $-0.2\mathbb{1}_{\theta=(2,2,1), x^3=(2,1,1)}$. *The purpose of this construction is to show that the social choice function* f *defined below does not satisfy the robust* $\mathcal{S}$ *monotonicity condition.*

Now define a social choice function $f = (f^0, (f_i^1, f_i^2, f_i^3)_{i \in I})$. *First, define* $f_3^1(\theta) = f_3^2(\theta) =$

0 *and* $f_i^3(\theta) = \theta_i$ *for all* $i \in I$ *and* $\theta \in \Theta$. *Then we consider two cases. Case* 1*:* $\theta_3 = 0$. *Define* $f_1^1(\theta) = f_2^1(\theta) = 0$ *for all* $\theta_{1,2} \in \Theta_{1,2}$, *and other components follow a public good provision rule which taxes agents* 1 *and* 2. *Specifically, if* $\theta_1 = \theta_2 = 2$, $f^0(\theta) = 1$, $f_1^2(\theta) = f_2^2(\theta) = -1.5$*; otherwise,* $f^0(\theta) = f_1^2(\theta) = f_2^2(\theta) = 0$. *Case* 2*:* $\theta_3 = 1$. *Define* $f^0(\theta) = 0$ *for all* $\theta_{1,2} \in \Theta_{1,2}$, *and other components follow a second-price auction with the tie-breaking rule in favor of agent* 1. *In particular, if* $\theta_1 \geq \theta_2$, *then* $f_1^1(\theta) = 1$, $f_1^2(\theta) = -\theta_2$, $f_2^1(\theta) = f_2^2(\theta) = 0$*; if* $\theta_1 < \theta_2$, *then* $f_1^1(\theta) = f_1^2(\theta) = 0$, $f_2^1(\theta) = 1$, $f_2^2(\theta) = -\theta_1$. *The set of deterministic feasible outcomes is given by* $X \equiv f(\Theta)$.

We have two claims on the social choice function f*:* (i) f *does not satisfy the robust* $\mathcal{S}$ *monotonicity condition;* (ii) f *is robustly* $\mathcal{S}$ *implemented by the direct mechanism.*

To prove Claim (i)*, consider a deception profile* β *defined by* $\beta_1(1) = \beta_1(2) = \beta_2(1) = \beta_2(2) = \{1\}$ *and* $\beta_3(\theta_3) = \{\theta_3\}$ *for all* $\theta_3 \in \Theta_3$. *Agent* 1 *with payoff type* 2 *misreporting* 1 *cannot always be the (singleton) whistle-blower, since there is a conjecture* $\psi_1(\theta_2 = 2, \theta_3 = 0, \theta_2' = 1, \theta_3' = 0) = 1$ *under which there is no successful* $\mathcal{S}$ *reward function. Coalition* $\{1, 2\}$ *with payoff type profile* $(2, 2)$ *misreporting* $(1, 1)$ *cannot always be the whistle-blowers, since there are conjectures* $\psi_1(\theta_3 = 1, \theta_3' = 1) = \psi_2(\theta_3 = 1, \theta_3' = 1) = 1$ *under which no* $\mathcal{S}$ *reward function can be proposed to benefit both agents (recall that in the construction of* u_2*, there is an additional disutility term when the true and reported type profiles are* $(2, 2, 1)$ *and* $(2, 1, 1)$*, and we remark that this construction prevents the existence of a successful lottery-valued* $\mathcal{S}$ *reward function* $y : \Theta_3 \to A$ *to benefit both agents). Similarly, no other coalition can be whistle-blowers who can dissolve* β *regardless of members' conjectures.*

By following a similar argument to Examples 3 and 4, it is easy to show that f *satisfies the robust* $\mathcal{S}$ *incentive compatibility condition. To establish Claim* (ii)*, it thus suffices to show that under any type space, belief revising rule, and bad strategy profile* σ*, there always exists a coalition* $S \in \mathcal{S}$ *who can profitably deviate. The following four cases are exhaustive to support this point.*

Case 1*: suppose some type of agent* 3 *misreports. Due to the* $-|\theta_3 - x_3^3|$ *term, the misreporting type has a strict incentive to revert to truthfully report. We thus assume that agent* 3 *always truthfully reports in Cases* 2 *through* 4 *without explicitly stating it.*

Case 2*: suppose the following two conditions hold:*

(i) there exists $t_1 \in T_1$ *with* $\hat{\theta}_1(t_1) = 2$ *and* $\sigma_1(t_1) = 1$*, and for every such* t_1*, it holds that*

$$\sum_{t_{2,3} \in \{t_{2,3} \in T_{2,3} : \hat{\theta}_{2,3}(t_{2,3}) = (2,0), \sigma_{2,3}(t_{2,3}) = (1,0)\}} \pi_1(t_1)[t_{2,3}] = 1;$$

(ii) there exists $t_2 \in T_2$ *with* $\hat{\theta}_2(t_2) = 2$ *and* $\sigma_2(t_2) = 1$*, and for every such* t_2*, it holds that*

$$\sum_{t_{1,3} \in \{t_{1,3} \in T_{1,3} : \hat{\theta}_{1,3}(t_{1,3}) = (2,0), \sigma_{1,3}(t_{1,3}) = (1,0)\}} \pi_2(t_2)[t_{1,3}] = 1.$$

Then fix any $t_{1,2} \in T_{1,2}$ *for which* $\hat{\theta}_{1,2}(t_{1,2}) = (2,2)$*,* $\sigma_{1,2}(t_{1,2}) = (1,1)$*, and* $\pi_1(t_1)[t_2] > 0$*. Coalition* $\{1,2\}$ *with type profile* $t_{1,2}$ *can benefit from reverting to truthfully report. To see this, notice that* t_2 *is not a surprise for* t_1*, and thus the posterior belief of* t_1 *is that* t_3 *must have payoff type* 0 *with probability* 1*, which means that the game is a public good game from* t_1*'s view. Hence, jointly reverting to truthfully report is profitable for* t_1*. On* t_2*'s end,* t_1 *may or may not be a surprise to* t_2*, but under any belief revising rule, jointly reverting to truthfully report is also profitable for agent* 2*. To see this, it suffices to look at two extreme cases. In one extreme case where the posterior belief of type* t_2 *is that* t_3 *has payoff type* 0*, the public good game is played from type* t_2*'s view. In this case, jointly reverting to truthfully report with agent 1 is strictly profitable for agent* 2 *since the public good allocation and monetary transfer are improved. In the other extreme case where the posterior belief of type* t_2 *is that* t_3 *has payoff type* 1*, the modified private good game is played from type* t_2*'s view. In this case, jointly reverting to truthfully report with agent* 1 *is strictly profitable for agent* 2 *since the private good allocation and monetary transfer are unchanged for agent* 2 *but he benefits due to the strict incentive to truthfully report.*

Case 3*: suppose there exists a downward misreporting type of agent* 1 *or* 2*, but the two conditions required in Case* 2 *do not hold simultaneously. For instance, assume agent* 1 *has a downward misreporting type* t_1 *for which*

$$\sum_{t_{2,3} \in \{t_{2,3} \in T_{2,3} : \hat{\theta}_{2,3}(t_{2,3}) = (2,0), \sigma_{2,3}(t_{2,3}) = (1,0)\}} \pi_1(t_1)[t_{2,3}] < 1.$$

Then t_1 *has a strict incentive to revert to truthfully report, due to the* $-0.1\mathbb{1}_{\{\theta_2 = 1 \text{ or } \theta_3 = 1\}}|\theta_1 - x_1^3|$ *term or the fact that reverting improves the allocation of non-pseudo goods with positive*

probability. The case that agent 2 *has a downward misreporting type* t_2 *for which*

$$\sum_{t_{1,3}\in\{t_{1,3}\in T_{1,3}:\hat{\theta}_{1,3}(t_{1,3})=(2,0),\sigma_{1,3}(t_{1,3})=(1,0)\}} \pi_2(t_2)[t_{1,3}] < 1$$

can be analyzed in a similar way.

Case 4*: suppose neither agent* 1 *nor* 2 *has any downward misreport, but agent* 1 *or* 2 *has an upward misreport. Then the upward reporting type of agent* 1 *or* 2 *has a strict incentive to correct his misreport, due to a similar reason discussed in Case* 3*.*

Definition 12: *Given a type space and a belief revising rule, a social choice function* f *satisfies the* ***interim*** $\mathcal{S}$ ***monotonicity*** *condition if whenever a profile of mappings* $(\alpha_i : T_i \to T_i)_{i\in I}$ *is such that* $f\big(\hat{\theta}(\bar{t})\big) \neq f\big(\hat{\theta}(\alpha(\bar{t}))\big)$ *for some* $\bar{t} \in T$*, there exists a coalition* $S \in \mathcal{S}$*, a type profile* $t^*_S \in T_S$*, and a function* $h : T \to A$ *such that*

(*i*) *for all* $i \in S$*,*

$$\sum_{t_{-i}\in T_{-i}} u_i\Big(h(\alpha(t^*_S, t_{-S})), \hat{\theta}(t^*_S, t_{-S})\Big)\pi_i(t^*_i)[t_{-i}|t^*_{S\setminus\{i\}}]$$
$$> \sum_{t_{-i}\in T_{-i}} u_i\Big(f\big(\hat{\theta}\big(\alpha(t^*_S, t_{-S})\big)\big), \hat{\theta}(t^*_S, t_{-S})\Big)\pi_i(t^*_i)[t_{-i}|t^*_{S\setminus\{i\}}];$$

(*ii*) *for each coalition* $\bar{S}$ *such that* $S \subseteq \bar{S} \in \mathcal{S}$ *and type profiles* $t'_{\bar{S}}, t''_{\bar{S}} \in T_{\bar{S}}$*, there exists* $j \in \bar{S}$ *such that*

$$\sum_{t_{-j}\in T_{-j}} u_j\Big(f\big(\hat{\theta}(t''_{\bar{S}}, t_{-\bar{S}})\big), \hat{\theta}(t''_{\bar{S}}, t_{-\bar{S}})\Big)\pi_j(t''_j)[t_{-j}|t''_{\bar{S}\setminus\{j\}}]$$
$$\geq \sum_{t_{-j}\in T_{-j}} u_j\big(h(t'_{\bar{S}}, t_{-\bar{S}}), \hat{\theta}(t''_{\bar{S}}, t_{-\bar{S}})\big)\pi_j(t''_j)[t_{-j}|t''_{\bar{S}\setminus\{j\}}].$$

Proof of Theorem 2. We prove that (M, g) defined in the text robustly $\mathcal{S}$ implements f.

Claim 4: *Under any type space and any belief revising rule,* $\sigma^*_i(t_i) = (\hat{\theta}_i(t_i), NB, \cdot, \cdot)$ *for all* $i \in I$ *and* $t_i \in T_i$ *constitutes an interim* $\mathcal{S}$ *equilibrium of* (M, g)*.*

Proof: Fix any $S \in \mathcal{S}$, $t_S \in T_S$, and σ_S. To show that σ_S is not a profitable deviation for t_S, by robust $\mathcal{S}$ incentive compatibility, it suffices to focus on σ_S for which there exists an

non-empty set $\underline{S} \subseteq S$ such that $(\sigma_S(t_S), \sigma^*_{-S}(t_{-S})) \in \hat{M}(\underline{S})$ for all $t_{-S} \in T_{-S}$. For simplicity, denote the agent with the smallest index who blows a whistle, $i^*[\underline{S}]$, by i^* in the remainder of this claim. Denote the projection of $\sigma^3_{i^*}(t_{i^*})$ on $Y_{\underline{S}}[\mathcal{S}]$ by y and the elements in $\hat{Y}_{\underline{S}}[\mathcal{S}]$ by $y^1, y^2, ...$ For each $t_{-S} \in T_{-S}$, the outcome $g(\sigma_S(t_S), \sigma^*_{-S}(t_{-S}))$ is a lottery of realization $y(\sigma^1_{S\setminus\underline{S}}(t_{S\setminus\underline{S}}), \hat{\theta}_{-S}(t_{-S}))$ with probability $\frac{\sigma^4_{i^*}(t_{i^*})}{\sigma^4_{i^*}(t_{i^*})+1}$ and of realization $y^k(\sigma^1_{S\setminus\underline{S}}(t_{S\setminus\underline{S}}), \hat{\theta}_{-S}(t_{-S}))$ with probability $\frac{0.5^k}{\sigma^4_{i^*}(t_{i^*})+1}$ for $k = 1, 2, ...$ By condition (ii) of the interior $\mathcal{S}$ reward property, a lottery over $\{y\} \cup \hat{Y}_{\underline{S}}[\mathcal{S}]$ is in $Y_{\underline{S}}[\mathcal{S}]$ and thus σ_S is not a profitable deviation for t_S.

Claim 5: *Under any type space and any belief revising rule, if σ is an interim $\mathcal{S}$ equilibrium of the mechanism (M, g), then $\sigma(t) \in \bar{M}$ for all $t \in T$.*

Proof: Suppose by way of contradiction that we do not have $\sigma(t) \in \bar{M}$ for all $t \in T$. Let j be the agent with the smallest index for whom there exists $t^*_j \in T_j$ such that $\sigma^2_j(t^*_j) = B$. We fix one such type t^*_j and will show that t^*_j has a profitable deviation. Define $\theta^*_j \equiv \hat{\theta}_j(t^*_j)$.

For each coalition $S \subseteq I$ containing j, we define

$$T_{-j}(S) \equiv \{t_{-j} \in T_{-j} : \exists \bar{S} \subseteq I \text{ containing } j \text{ s.t. } \sigma(t^*_j, t_{-j}) \in \hat{M}(\bar{S}) \text{ and } S^*[\bar{S}] = S\},$$

which is the collection of all $t_{-j} \in T_{-j}$ such that the outcome is a lottery over $y(\sigma^1_{-S}(t_{-S}))$ and all $y^k(\sigma^1_{-S}(t_{-S}))$, where y is the projection of $\sigma_j(t^*_j)$ on $Y_S[\mathcal{S}]$ and each $y^k \in \hat{Y}_S[\mathcal{S}]$. Denote the measure of the set by $\phi_j(S) \equiv \sum_{t_{-j} \in T_{-j}(S)} \pi_j(t^*_j)[t_{-j}]$.

For any $S \subseteq I$ containing j such that $\phi_j(S) > 0$, define $\psi_j[S] \in \Delta(\Theta_{-j} \times \Theta_{-S})$ by

$$\psi_j[S](\theta_{-j}, \theta'_{-S}) \equiv \frac{\sum_{\hat{\theta}_{-j}(t_{-j}) = \theta_{-j}, \sigma^1_{-S}(t_{-S}) = \theta'_{-S}, t_{-j} \in T_{-j}(S)} \pi_j(t^*_j)[t_{-j}]}{\phi_j(S)},$$

which is the probability that t_{-j} has payoff type θ_{-j} and t_{-S} misreports θ'_{-S} conditional on $t_{-j} \in T_{-j}(S)$. If $\phi_j(S) = 0$, let $\psi_j[S] \in \Delta(\Theta_{-j} \times \Theta_{-S})$ be the uniform distribution. Then, type-t^*_j agent j's expected utility is equal to

$$\begin{aligned}
&\frac{\sigma^4_j(t^*_j)}{1+\sigma^4_j(t^*_j)} \sum_{S \ni j, S \subseteq I} \Big[\sum_{\theta_{-j} \in \Theta_{-j}, \theta'_{-S} \in \Theta_{-S}} u_j\big(y(\theta'_{-S}), (\theta^*_j, \theta_{-j})\big) \psi_j[S](\theta_{-j}, \theta'_{-S})\Big] \phi_j(S) \\
&+ \frac{1}{1+\sigma^4_j(t^*_j)} \sum_{S \ni j, S \subseteq I} \Big[\sum_{\theta_{-j} \in \Theta_{-j}, \theta'_{-S} \in \Theta_{-S}} \sum_{k=1,2,...} 0.5^k u_j(y^k(\theta'_{-S}), (\theta^*_j, \theta_{-j})) \psi_j[S](\theta_{-j}, \theta'_{-S})\Big] \phi_j(S).
\end{aligned} \tag{6}$$

We want to define a deviating strategy σ'_j by following a case-by-case discussion.

Case 1: suppose there exists $S \subseteq I$ containing j with $\phi_j(S) > 0$ such that

$$\sum_{\theta_{-j}\in\Theta_{-j},\theta'_{-S}\in\Theta_{-S}} u_j\big(y(\theta'_{-S}),(\theta^*_j,\theta_{-j})\big)\psi_j[S](\theta_{-j},\theta'_{-S}) \leq \sum_{\theta_{-j}\in\Theta_{-j},\theta'_{-S}\in\Theta_{-S}} \sum_{k=1,2,\ldots} 0.5^k u_j(y^k(\theta'_{-S}),(\theta^*_j,\theta_{-j}))\psi_j[S](\theta_{-j},\theta'_{-S}). \quad (7)$$

Fix one such S. Following a similar argument as in Claim 2, we can find some k such that

$$\sum_{\theta_{-j}\in\Theta_{-j},\theta'_{-S}\in\Theta_{-S}} u_j\big(y(\theta'_{-S}),(\theta^*_j,\theta_{-j})\big)\psi_j[S](\theta_{-j},\theta'_{-S}) < \sum_{\theta_{-j}\in\Theta_{-j},\theta'_{-S}\in\Theta_{-S}} u_j\big(y^k(\theta'_{-S}),(\theta^*_j,\theta_{-j})\big)\psi_j[S](\theta_{-j},\theta'_{-S})$$

by the interior $\mathcal{S}$ reward property. In this case, let σ'_j be identical to σ_j except that the component of ${\sigma'}^3_j(t^*_j)$ corresponding to $Y_S[\mathcal{S}]$ is y^k.

Case 2: if expression (7) does not hold for any $S \subseteq I$ containing j with $\phi_j(S) > 0$. Let σ'_j be identical to σ_j except that ${\sigma'}^4_j(t^*_j) = \sigma^4_j(t^*_j) + 1$.

It is easy to see that type t^*_j becomes better off under σ'_j, contradicting the supposition that σ is an interim $\mathcal{S}$ equilibrium.

Claim 6: *Under any type space and any belief revising rule, if σ is an interim $\mathcal{S}$ equilibrium of (M,g), then $g(\sigma(t)) = f(\hat{\theta}(t))$ for all $t \in T$.*

Proof: Suppose by way of contradiction that there exists $\bar{t} \in T$ such that $g(\sigma(\bar{t})) \neq f(\hat{\theta}(\bar{t}))$. For each $i \in I$, define a correspondence β_i in the same way as in the proof of Claim 3. Then the deception profile β is unacceptable. By the robust $\mathcal{S}$ monotonicity condition, there exists $S \in \mathcal{S}$, $\theta_S \in \Theta_S$, and $\theta'_S \in \beta_S(\theta_S)$ such that for any conjectures $(\psi_i \in \Delta(\{(\theta_{-S},\theta'_{-S})|\theta_{-S} \in \Theta_{-S}, \theta'_{-S} \in \beta_{-S}(\theta_{-S})\}))_{i\in S}$, there exists $y \in Y_S[\mathcal{S}]$ such that

$$\sum_{\theta_{-S}\in\Theta_{-S},\theta'_{-S}\in\beta_{-S}(\theta_{-S})} u_i\big(y(\theta'_{-S}),(\theta_S,\theta_{-S})\big)\psi_i(\theta_{-S},\theta'_{-S}) > \sum_{\theta_{-S}\in\Theta_{-S},\theta'_{-S}\in\beta_{-S}(\theta_{-S})} u_i\big(f(\theta'_S,\theta'_{-S}),(\theta_S,\theta_{-S})\big)\psi_i(\theta_{-S},\theta'_{-S}). \quad (8)$$

Fix any profile of types t_S^* such that $\hat{\theta}_S(t_S^*) = \theta_S$ and $\sigma_S^1(t_S^*) = \theta'_S$. For each $i \in S$, let the conjecture $\psi_i \in \Delta(\{(\theta_{-S}, \theta'_{-S}) | \theta_{-S} \in \Theta_{-S}, \theta'_{-S} \in \beta_{-S}(\theta_{-S})\})$ be defined by

$$\psi_i(\theta_{-S}, \theta'_{-S}) \equiv \sum_{\hat{\theta}_{-S}(t_{-S})=\theta_{-S}, \sigma^1_{-S}(t_{-S})=\theta'_{-S}} \pi_i(t_i^*)[(t^*_{S\setminus\{i\}}, t_{-S}) | t^*_{S\setminus\{i\}}].$$

Then there exists a function $y \in Y_S[\mathcal{S}]$ such that expression (8) holds. For all $i \in S$, define a strategy σ'_i by $\sigma'_i(t_i^*) = (\sigma_i^1(t_i^*), B, m_i^3, K^*)$ and $\sigma'_i(t_i) = \sigma_i(t_i)$ for all $t_i \neq t_i^*$, where the only restriction on $m_i^3 \in M_i^3$ is that its projection on $Y_S[\mathcal{S}]$ is the function y above. When K^* is sufficiently large, this deviation is profitable for S, a contradiction.

We thus have demonstrated that (M, g) robustly $\mathcal{S}$ implements f. □

Definition 13: *Given a social choice function f, for each $i \in I$ and $\theta'_{-i} \in \Theta_{-i}$, define*

$$R_i(\theta'_{-i}) \equiv \{a \in A : u_i\big(f(\theta''_i, \theta'_{-i}), (\theta''_i, \theta'_{-i})\big) \geq u_i\big(a, (\theta''_i, \theta'_{-i})\big) \forall \theta''_i \in \Theta_i\}.$$

The social choice function f is said to satisfy the ***conditional no total indifference property*** *if for all i, θ_i, θ'_{-i}, and $\phi_i \in \Delta(\Theta_{-i})$, there are outcomes $\bar{a}, \underline{a} \in R_i(\theta'_{-i})$ such that*

$$\sum_{\theta_{-i} \in \Theta_{-i}} u_i\big(\bar{a}, (\theta_i, \theta_{-i})\big)\phi_i(\theta_{-i}) > \sum_{\theta_{-i} \in \Theta_{-i}} u_i\big(\underline{a}, (\theta_i, \theta_{-i})\big)\phi_i(\theta_{-i}).$$

Proof of Corollary 1. **Step 1**. Suppose X is countable. We first prove that if f satisfies the conditional no total indifference property, then the interior $\underline{\mathcal{S}}$ reward property is satisfied.

For each i and θ'_{-i}, the set $R_i(\theta'_{-i})$ is convex. Since agents adopt expected utilities to evaluate lotteries, the set of extreme points of $R_i(\theta'_{-i})$, denoted by $R_i^*(\theta'_{-i})$, is a subset of X. Since X is countable, $R_i^*(\theta'_{-i})$ is countable, and so is the following set:

$$\hat{Y}_i[\underline{\mathcal{S}}] \equiv \{y : \Theta_{-i} \to X | y(\theta'_{-i}) \in R_i^*(\theta'_{-i}), \forall \theta'_{-i} \in \Theta_{-i}\}.$$

As $Y_i[\underline{\mathcal{S}}]$ is convex and $\hat{Y}_i[\underline{\mathcal{S}}] \subseteq Y_i[\underline{\mathcal{S}}]$, condition (ii) in the interior $\underline{\mathcal{S}}$ reward property holds.

To establish condition (i) in the interior $\underline{\mathcal{S}}$ reward property, we fix any i, θ_i, and $\psi_i \in \Delta(\Theta_{-i} \times \Theta_{-i})$ for the remainder of Step 1. For each $\theta'_{-i} \in \Theta_{-i}$, let a distribution $\bar{\phi}_i[\theta'_{-i}] \in \Delta(\Theta_{-i})$ be defined by $\bar{\phi}_i[\theta'_{-i}](\theta_{-i}) \equiv \frac{\psi_i(\theta_{-i}, \theta'_{-i})}{\sum_{\theta''_{-i} \in \Theta_{-i}} \psi_i(\theta''_{-i}, \theta'_{-i})}$ for all $\theta_{-i} \in \Theta_{-i}$ whenever $\sum_{\theta''_{-i} \in \Theta_{-i}} \psi_i(\theta''_{-i}, \theta'_{-i}) > 0$; let $\bar{\phi}_i[\theta'_{-i}] \in \Delta(\Theta_{-i})$ be the uniform distribution instead when

$\sum_{\theta''_{-i}\in\Theta_{-i}} \psi_i(\theta''_{-i},\theta'_{-i}) = 0$. Given i, θ_i, by the conditional no total indifference property, for each $\theta'_{-i} \in \Theta_{-i}$, there are outcomes $\bar{a}[\theta'_{-i}, \bar{\phi}_i[\theta'_{-i}]], \underline{a}[\theta'_{-i}, \bar{\phi}_i[\theta'_{-i}]] \in R_i(\theta'_{-i})$ such that

$$\sum_{\theta_{-i}\in\Theta_{-i}} u_i(\bar{a}[\theta'_{-i}, \bar{\phi}_i[\theta'_{-i}]], \theta)\bar{\phi}_i[\theta'_{-i}](\theta_{-i}) > \sum_{\theta_{-i}\in\Theta_{-i}} u_i(\underline{a}[\theta'_{-i}, \bar{\phi}_i[\theta'_{-i}]], \theta)\bar{\phi}_i[\theta'_{-i}](\theta_{-i}).$$

As agents adopt expected utilities, it is without loss of generality to assume that $\bar{a}[\theta'_{-i}, \bar{\phi}_i[\theta'_{-i}]]$, $\underline{a}[\theta'_{-i}, \bar{\phi}_i[\theta'_{-i}]] \in R_i^*(\theta'_{-i})$. A weighted sum of the strict inequalities gives

$$\sum_{\theta_{-i}\in\Theta_{-i},\theta'_{-i}\in\Theta_{-i}} u_i(\bar{a}[\theta'_{-i}, \bar{\phi}_i[\theta'_{-i}]], \theta)\psi_i(\theta_{-i},\theta'_{-i}) > \sum_{\theta_{-i}\in\Theta_{-i},\theta'_{-i}\in\Theta_{-i}} u_i(\underline{a}[\theta'_{-i}, \bar{\phi}_i[\theta'_{-i}]], \theta)\psi_i(\theta_{-i},\theta'_{-i}).$$

Define $\bar{y}(\theta'_{-i}) \equiv \bar{a}[\theta'_{-i}, \bar{\phi}_i[\theta'_{-i}]]$ and $\underline{y}(\theta'_{-i}) \equiv \underline{a}[\theta'_{-i}, \bar{\phi}_i[\theta'_{-i}]]$ for all $\theta'_{-i} \in \Theta_{-i}$. It is easy to see that $\bar{y}, \underline{y} \in \hat{Y}_i[\underline{\mathcal{S}}]$. Hence, we have established condition (i) of the interior $\underline{\mathcal{S}}$ reward property.

Step 2. We then prove Theorem 2 of Bergemann and Morris (2011).

Lemma 1 of Bergemann and Morris (2011) has proved that robust monotonicity implies ex-post incentive compatibility (equivalent to robust $\underline{\mathcal{S}}$ incentive compatibility). Hence, when the robust monotonicity condition is satisfied, both the robust $\underline{\mathcal{S}}$ incentive compatibility condition and the robust $\underline{\mathcal{S}}$ monotonicity condition hold. Taking into account our finding in Step 1, we know that whenever f satisfies robust monotonicity and conditional no total indifference, sufficient conditions in our Theorem 2 hold under the minimal coalition pattern. By our Theorem 2, f is robustly $\underline{\mathcal{S}}$ implementable, i.e., f is robustly implementable. □

References

Adachi, T. (2014). Robust and secure implementation: equivalence theorems. *Games and Economic Behavior*, 86:96–101.

Aumann, R. (1959). Acceptable points in general cooperative n-person games. *Contributions to the Theory of Games (AM-40)*, 4:287–324.

Bennett, E. and Conn, D. (1977). The group incentive properties of mechanisms for the provision of public goods. *Public Choice*, pages 95–102.

Bergemann, D. and Morris, S. (2008). Robust implementation in general mechanisms. Working Paper.

Bergemann, D. and Morris, S. (2009). Robust implementation in direct mechanisms. *The Review of Economic Studies*, 76(4):1175–1204.

Bergemann, D. and Morris, S. (2011). Robust implementation in general mechanisms. *Games and Economic Behavior*, 71(2):261–281.

Bierbrauer, F. J. and Hellwig, M. F. (2011). Mechanism design and voting for public-good provision. Working Paper.

Bierbrauer, F. J. and Hellwig, M. F. (2015). Public-good provision in a large economy. *Working Paper*.

Bierbrauer, F. J. and Hellwig, M. F. (2016). Robustly coalition-proof incentive mechanisms for public good provision are voting mechanisms and vice versa. *The Review of Economic Studies*, 83(4):1440–1464.

Che, Y.-K. and Kim, J. (2006). Robustly collusion-proof implementation. *Econometrica*, 74(4):1063–1107.

Chen, J. and Micali, S. (2012). Collusive dominant-strategy truthfulness. *Journal of Economic Theory*, 147(3):1300–1312.

de Castro, L. I., Liu, Z., and Yannelis, N. C. (2017a). Ambiguous implementation: the partition model. *Economic Theory*, 63(1):233–261.

de Castro, L. I., Liu, Z., and Yannelis, N. C. (2017b). Implementation under ambiguity. *Games and Economic Behavior*, 101:20–33.

Dutta, B. and Sen, A. (1991). Implementation under strong equilibrium: A complete characterization. *Journal of Mathematical Economics*, 20(1):49–67.

Dutta, B. and Sen, A. (2012). Nash implementation with partially honest individuals. *Games and Economic Behavior*, 74(1):154–169.

Green, J. and Laffont, J.-J. (1979). On coalition incentive compatibility. *The Review of Economic Studies*, 46(2):243–254.

Guo, H. (2019). Mechanism design with ambiguous transfers: An analysis in finite dimensional naive type spaces. *Journal of Economic Theory*, 183:76–105.

Guo, H. (2020). Coalition-proof ambiguous mechanism. Working Paper.

Guo, H. and Yannelis, N. C. (2020). Incentive compatibility under ambiguity. *Economic Theory*. Forthcoming, https://doi.org/10.1007/s00199-020-01304-x.

Guo, H. and Yannelis, N. C. (2021). Full implementation under ambiguity. *American Economic Journal: Microeconomics*, 13(1):148–78.

Hahn, G. and Yannelis, N. C. (2001). Coalitional Bayesian Nash implementation in differential information economies. *Economic Theory*, 18(2):485–509.

Hayashi, T., Jain, R., Lombardi, M., and Korpela, V. (2020). Behavioral strong implementation. Working Paper.

Jackson, M. O. (1991). Bayesian implementation. *Econometrica*, 59(2):461–477.

Jain, R. (2021). Rationalizable implementation of social choice correspondences. *Games and Economic Behavior*, 127:47–66.

Koray, S. and Yildiz, K. (2018). Implementation via rights structures. *Journal of Economic Theory*, 176:479–502.

Korpela, V. (2013). A simple sufficient condition for strong implementation. *Journal of Economic Theory*, 148(5):2183–2193.

Korpela, V., Lombardi, M., and Vartiainen, H. (2020). Do coalitions matter in designing institutions? *Journal of Economic Theory*, 185:104953.

Koutsougeras, L. C. (2020). Coalitions with limited coordination. *Economic Theory*, pages 1–18. Forthcoming, https://doi.org/10.1007/s00199-020-01302-z.

Li, S. (2017). Obviously strategy-proof mechanisms. *American Economic Review*, 107(11):3257–87.

Liu, Z. (2016). Implementation of maximin rational expectations equilibrium. *Economic Theory*, 62(4):813–837.

Liu, Z. and Yannelis, N. C. (2021). Persuasion in an asymmetric information economy: a justification of Wald's maxmin preferences. *Economic Theory*, 72(3):801–833.

Lombardi, M. and Yoshihara, N. (2018). Treading a fine line: (im)possibilities for Nash implementation with partially-honest individuals. *Games and Economic Behavior*, 111:203–216.

Lombardi, M. and Yoshihara, N. (2020). Partially-honest Nash implementation: a full characterization. *Economic Theory*, 70:871–904.

Maskin, E. (1977). Nash equilibrium and welfare optimality. Working paper.

Maskin, E. (1978). Implementation and strong Nash equilibrium. Working paper.

Maskin, E. (1979). Incentive schemes immune to group manipulation. Working paper.

Maskin, E., Hurwicz, L., Schmeidler, D., and Sonnenschein, H. (1985). The theory of implementation in Nash equilibrium: A survey. *Social Goals and Social Organization: Volume in Memory of Elisha Pazner*.

Moreno-García, E. and Torres-Martínez, J. P. (2020). Information within coalitions: risk and ambiguity. *Economic Theory*, 69(1):125–147.

Müller, C. (2016). Robust virtual implementation under common strong belief in rationality. *Journal of Economic Theory*, 162:407–450.

Ollár, M. and Penta, A. (2017). Full implementation and belief restrictions. *American Economic Review*, 107(8):2243–77.

Oury, M. and Tercieux, O. (2012). Continuous implementation. *Econometrica*, 80(4):1605–1637.

Pasin, P. (2009). *Essays on implementability and monotonicity*. PhD thesis, Bilkent University.

Penta, A. (2015). Robust dynamic implementation. *Journal of Economic Theory*, 160:280–316.

Pram, K. (2020). Weak implementation. *Economic Theory*, 69(3):569–594.

Safronov, M. (2018). Coalition-proof full efficient implementation. *Journal of Economic Theory*, 177:659–677.

Saijo, T., Sjostrom, T., and Yamato, T. (2007). Secure implementation. *Theoretical Economics*, 2(3):203–229.

Suh, S.-C. (1996). Implementation with coalition formation: A complete characterization. *Journal of Mathematical Economics*, 26(4):409–428.

Suh, S.-C. (1997). Double implementation in Nash and strong Nash equilibria. *Social Choice and Welfare*, 14(3):439–447.

Velez, R. A. and Brown, A. L. (2020). Empirical strategy-proofness. Working Paper.

Williams, S. R. (2001). Sufficient conditions for Nash implementation. *Review of Economic Design*, 6(3):325–342.