Robust Coalitional Implementation

Huiyi Guo and Nicholas C. Yannelis*

October 1, 2020

Abstract

The paper introduces coalition structures to study belief-free full implementation. When the mechanism designer does not know which coalitions can be formed, 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 can be formed, we show that allowing for 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.

*Guo: Texas A&M University, Department of Economics, 4228 TAMU, College Station, TX 77843, huiyiguo@tamu.edu. Yannelis: The University of Iowa, Department of Economics, W380 John Pappajohn Business Building, Iowa City, IA 52242, nicholasyannelis@gmail.com. Part of the results reported in the current paper is based on the authors' earlier project entitled *Robust Strong Nash Implementation*.

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 coalitions that can be formed, is exogenously given by $\mathcal{S}$. When a coalition is formed, members of this coalition are allowed to jointly coordinate on a deviating strategy 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 generalizes the strong equilibrium of Aumann (1959) to 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 can be formed. 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 is neither robustly coalitionally implementable nor robustly $\bar{\mathcal{S}}$ implementable.

When the mechanism designer has information on the coalition pattern $\mathcal{S}$, she knows the equilibrium played by agents. Our robust $\mathcal{S}$ implementation question requires the social choice function to coincide with the interim $\mathcal{S}$ equilibrium regardless of agents' belief structures. Due to the additional information on the coalition pattern compared to robust coalitional implementation, we establish weaker 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 coalition patterns may help to implement social choice functions that are non-implementable under the non-cooperative framework.

The paper proceeds as follows. Section 1.1 discusses related literature. 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, are introduced in Sections 4 and 5. We then conclude in Section 6.

1.1 Literature Review

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 beliefs 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 implementation with coalitions. One is Hahn and Yannelis (2001). In exchange economies with general preferences, they generalize the strong equilibrium concept into 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 assumption that the belief structure is common knowledge between the mechanism designer and all agents.

Under complete information settings where 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 for non-trivial coalition patterns.

Recently, under the complete information setting, Koray and Yildiz (2018) and Korpela

[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.

et al. (2020) bring to the literature the idea of designing a rights structure or a code of rights, which specifies the collection of coalitions having the right to act cooperatively. We differ in our incomplete information setup and in our exogenous coalition structures. However, we find that there are social choice functions not robustly implementable under the non-cooperative framework but robustly implementable under certain coalition pattern. Our finding shares a similar implication with theirs in that non-trivial coalitions can bring value to institution design.

Instead of concerning coalitional manipulations, the recent full implementation literature has also explored how other non-standard assumptions affect the result of full implementation. For example, 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, Guo and Yannelis (2020) assume that agents hold maxmin expected utility with respect to each other's private information and show that the set of efficient and individually rational social choice functions becomes fully implementable.

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;
- $\Theta = \Theta_1 \times ... \times \Theta_n$ is a finite **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, $u_i(a,\theta) = \int_{x\in X} u_i(x,\theta)d\mu$.[2]

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.

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 notation $\sum_{k=1,2,...} \rho^k a^k$ 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 an onto mapping $\hat{\theta}_i : T_i \to \Theta_i$;
- 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 coalitions can be formed. 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 coalitions that can be formed. We assume that all singletons are included in $\mathcal{S}$, i.e., agents can always choose not to communicate with others. One example of a coalition pattern is the **minimal** (or trivial) coalition pattern, which we denote by $\underline{\mathcal{S}} = \{\{i\} : i \in I\}$. Another example is the **maximal** coalition pattern, denoted by $\bar{\mathcal{S}} = 2^I \backslash \{\emptyset\}$, which has the richest coalition possibility. In applications, one may be interested in other patterns formed

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

by partisanship, cultural differences, geographic isolation, etc.

When a coalition is formed, each member acquires new information on other members' private information, which may surprise him.[3] 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 $S \ni i$, let the notation $\pi_i(t_i^*)[t_{S\setminus\{i\}}^*]$ represent the marginal probability that coalition $S\setminus\{i\}$ has type profile $t_{S\setminus\{i\}}^* = (t_j^*)_{j\in S\setminus\{i\}}$. Whenever $\pi_i(t_i^*)[t_{S\setminus\{i\}}^*] = 0$, a **belief revising rule** specifies a posterior belief $(\pi_i(t_i^*)[t_{-i}|t_{S\setminus\{i\}}^*])_{t_{-i}\in T_{-i}}$ over T_{-i} whose marginal probability on the event that coalition $S\setminus\{i\}$ has type profile $t_{S\setminus\{i\}}^*$ is 1. The posterior belief is defined by the Bayes rule whenever $\pi_i(t_i^*)[t_{S\setminus\{i\}}^*] > 0$.

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 $(\sigma_i)_{i\in I}$.

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 and a type profile, such that under coalition members' pooled information, a deviating strategy profile 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 $i \in S$,*

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

We allow the coalition to pool members' information and adopt the most profitable deviating strategy. This is consistent with the coalition-proofness notions of Bennett and

[3]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.

Conn (1977), Green and Laffont (1979), Chen and Micali (2012), Safronov (2018), etc. An underlying assumption is that agents within a coalition act as a utility-maximizing pseudo agent without encountering within-coalition interactions. This assumption simplifies our analysis by helping us focus on the interaction between a coalition and all others out of the coalition. Although there are alternative models on coalition formation that can undermine the power of coalitional manipulations, when a mechanism designer does not know exactly the way coalitions are formed, our definition of interim $\mathcal{S}$ equilibrium imposes a strong stability requirement and serves as a benchmark to study robust coalitional implementation.

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 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.

When the mechanism designer does not know the coalition pattern and wishes to robustly $\mathcal{S}$ implement a social choice function under all coalition patterns, it is of interest to study the following implementation concept.

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

Notice that an interim strong equilibrium is an interim $\mathcal{S}$ equilibrium, and an interim $\mathcal{S}$ equilibrium is an interim equilibrium. Condition (i) of the definition above guarantees the existence of a "good" interim $\mathcal{S}$ equilibrium no matter what the coalition pattern $\mathcal{S}$ is. Condition (ii) implies that every "bad" interim $\mathcal{S}$ equilibrium can be dissolved by a singleton's deviation.

When the coalition pattern $\mathcal{S}$ is known to the mechanism designer, we present the definition of robust $\mathcal{S}$ implementation. It requires the set of interim $\mathcal{S}$ equilibrium outcomes to coincide with the social choice function under all type spaces and belief revising rules.

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 $g\big(\sigma(t)\big) = f(\hat{\theta}(t))$ for all $t \in T$;*

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

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 social planner can construct a public good, towards which agents have private valuation. To finance the costly good, the social planner can charge both agents. The utility of agent* i *is denoted by* $u_i(x, \theta) = \theta_i x_0 + x_i$, *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* $X = \{x \in \mathbb{R}^3 : x \leq f(\theta) \text{ for some } \theta \in \Theta\}$. *Notice that free disposal is allowed.*

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

In fact, f *is neither robustly* $\bar{\mathcal{S}}$ *implementable nor robustly coalitionally implementable because one cannot guarantee the existence of a good interim strong equilibrium under all belief structures.*[4] *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 by* $\sigma'_i(t_i) = \sigma_i(t'_i)$ *for all* $i \in I$ *and* $t_i \in T_i$, *each agent* $i \in I$ *in the grand coalition earns a payoff of* $u_i(g(\sigma'(t^*)), \theta^*) = u_i(g(\sigma(t')), \theta^*) = u_i(f(\theta'), \theta^*) = 0.5 > u_i(f(\theta^*), \theta^*) = 0.375$. *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.*

However, if we shrink the payoff type set 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* $\bar{\mathcal{S}}$

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

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 in Bergemann and Morris (2011) and proved to be necessary for robust implementation) is not necessary for robust $\mathcal{S}$ implementation in general.

Example 2: *Consider the same two-agent public good example as the one in Example 1 except that (i) Θ_i is $\{0,1\}$ for both agents, and (ii) agents have common valuation: the utility of agent i is denoted 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.*

Suppose the three deterministic feasible outcomes that do not involve free disposal are given by: $x^1 = (x_0^1, x_1^1, x_2^1) = (0,0,0)$, $x^2 = (x_0^2, x_1^2, x_2^2) = (2,-1,-1)$, and $x^3 = (x_0^3, x_1^3, x_2^3) = (4,-4,-4)$, which represent low, middle, and high levels of public good provision respectively. Free disposal is allowed and thus $X = \{x \in \mathbb{R}^3 : x \leq x^1, x^2, \text{ or } x^3\}$.

Define f by $f(0,0) = x^1$, $f(0,1) = f(1,0) = x^2$, and $f(1,1) = x^3$. Notice that f is the only ex-post efficient social choice function. Also, agents have common interest under f.

We claim that f is not robustly implementable in the sense of Bergemann and Morris (2011).[5] *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.*

[5] 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,* $g(\sigma'(t_1,t_2)) = g(\sigma(t_1^0,t_2^1)) = f(\hat{\theta}(t_1^0,t_2^1)) = f(0,1) = x^2$ *for all* $t \in T$, *but* $f(\hat{\theta}(t_1^1,t_2^1)) = f(1,1) = x^3$. *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*

$$\begin{aligned}
&\frac{\epsilon}{1+\epsilon}u_1(g(\sigma_1'(t_1^0),\sigma_2'(t_2^0)),(0,0)) + \frac{1}{1+\epsilon}u_1(g(\sigma_1'(t_1^0),\sigma_2'(t_2^1)),(0,1))\\
=&\frac{\epsilon}{1+\epsilon}u_1(g(\sigma_1(t_1^0),\sigma_2(t_2^1)),(0,0)) + \frac{1}{1+\epsilon}u_1(g(\sigma_1(t_1^0),\sigma_2(t_2^1)),(0,1))\\
=&\frac{\epsilon}{1+\epsilon}u_1(x^2,(0,0)) + \frac{1}{1+\epsilon}u_1(x^2,(0,1)),
\end{aligned}$$

which is close to $u_1(x^2,(0,1))$ *when* ϵ *is sufficiently small. In this case, since* x^2 *is the unique utility maximizing feasible outcome in the above expression, type-*t_1^0 *agent* 1 *is playing best response under* σ'. *Similarly, we can verify that each type of each agent plays best response under* σ', *which implies that* σ' *is an interim equilibrium. This contradicts the supposition that* (M,g) *robustly implements* f.

However, f *is robustly* $\bar{\mathcal{S}}$ *implementable under all type spaces and all belief revising rules. In fact, the direct mechanism that requires both agents to report their payoff types can robustly* $\bar{\mathcal{S}}$ *implement* f. *Truthfully reporting of both agents constitutes a "good" interim strong equilibrium. To see this, notice that under each* $\theta \in \Theta$, $f(\theta)$ *assigns the unique optimal 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 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.*

4 Robust Coalitional Implementation

We begin by assuming that the mechanism designer does not know the coalition pattern and thus does not know which equilibrium agents are playing. We will introduce three conditions that are jointly sufficient for robust coalitional implementation, and then construct a mechanism to establish the sufficiency of these conditions.

4.1 Sufficient 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 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 disincentives 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 $f(\Theta)$. As we focus on implementation of a social choice function, a global version of Pareto efficiency is unnecessary: for example, a constant inefficient social choice function is robustly coalitionally implementable.

The robust coalitional incentive compatibility condition is in general a strong condition. Allowing all coalitions to be formed 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 $S \ni i$, $\theta''_S \in \Theta_S$, and $\theta'_{S\backslash\{i\}} \in \Theta_{S\backslash\{i\}}$, there exists $j \in S$ such that

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

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 $(\theta_i, \theta'_{S\backslash\{i\}})$, there should exist $j \in S$ such that j is 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 and profitably deviate by proposing a successful coalitional reward function y regardless of his conjecture of other agents' true and reported payoff types.

Robust coalitional monotonicity is stronger than robust monotonicity of Bergemann and Morris (2011) because our coalitional reward set imposes a stronger requirement than their reward set. To see this, notice that a coalitional reward function proposed by the whistle-blower i cannot serve as a profitable deviation from truthfully consuming f for any coalition $S \ni i$ with payoff types θ_S. But in their paper, an element in the reward set merely should not serve as a profitable deviation for agent i with any payoff type θ_i. However, when agents have quasilinear utility functions (or under some other weak conditions), robust monotonicity implies robust coalitional monotonicity and the two conditions become equivalent. Intuitively, when i proposes a successful reward in the robust monotonicity condition, we can lower the transfers to all $j \neq i$ sufficiently but keep other parts of the reward unchanged. In this way, we can construct a successful coalitional reward function.

The proposition below shows that the robust coalitional monotonicity condition is necessary for robust coalitional implementation. Its proof is relegated to the Appendix.

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

We then introduce the interior coalitional reward property to complete the group of sufficient conditions. This condition is not necessary for robust coalitional implementation.

Definition 6: *A social choice function f satisfies the* ***interior coalitional reward property****, 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 $\psi_i \in \Delta(\{(\theta_{-i}, \theta'_{-i}) | \theta_{-i} \in \Theta_{-i}, \theta'_{-i} \in \beta_{-i}(\theta_{-i})\})$, 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 $(y^k \in \hat{Y}_i)_{k=1,2,...}$, and vector $(\rho^k)_{k=0,1,2,...} \in \Delta$, the function $\rho^0 y + \sum_{k=1,2,...} \rho^k y^k \in Y_i$.

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 every lottery over $\hat{Y}_i$ and some $y \in Y_i$ is still a coalitional reward function.

The interior coalitional reward property is a weak condition when agents have monotone preferences and free disposal. For example, for social choice function f and agent $i \in I$, we can consider two functions $\bar{y}, \underline{y} \in Y_i$, which always offer sufficiently low consumption to every agent. Let agents' consumption under $\bar{y}$ to be slightly higher than that under $\underline{y}$. 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 session. 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 condition 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 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 only provide a sketch of the proof. For convenience of notation, we decompose a strategy $\sigma_i : T_i \to M_i$ 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 $k = 1, 2, 3, 4$.

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 and to not blow 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 all type spaces and belief revising rules, agents do not blow a whistle. Suppose that there is 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 can show that t_j^* can be better off by proposing a better coalitional reward function $\hat{y}$ or decreasing the probability that the full-support lottery is realized.

Claim 3 further shows that in any interim equilibrium, agents follow an acceptable deception 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 that 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 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.

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 are playing an interim $\mathcal{S}$ equilibrium. We will provide a group of sufficient conditions for robust $\mathcal{S}$ implementation and construct a mechanism explicitly. The sufficient conditions are weaker than the ones for robust coalitional implementation. Furthermore, 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 for the mechanism designer to robustly $\mathcal{S}$ implement some social choice functions that are not robustly implementable in the non-cooperative framework.

5.1 Sufficient Conditions

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

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.

Then we define the $\mathcal{S}$ reward set and the robust $\mathcal{S}$ monotonicity condition. For each $S \in \mathcal{S}$, 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, we remark that in the special case $S = I$, the set Θ_{-S} degenerates and each $y : \Theta_{-S} \to A$ is viewed as a constant function with range in A. When f satisfies robust $\mathcal{S}$ incentive compatibility, the set $Y_S[\mathcal{S}]$ is non-empty for all coalition S. 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 $\theta_{-\bar{S}} \in \Theta_{-\bar{S}}$. Since $y \in Y_S[\mathcal{S}]$, $Y_S[\mathcal{S}]$ is non-empty again.

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.

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).

When agents have quasilinear utility functions, the robust monotonicity condition implies the robust $\mathcal{S}$ monotonicity condition for all $\mathcal{S}$. 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 the singleton whistle-blower $\{i\} \in \mathcal{S}$ can profitably propose the $\mathcal{S}$ reward function $\hat{y}$ to dissolve the unacceptable deception profile. Hence, robust $\mathcal{S}$ monotonicity condition is also satisfied.

The fact that robust monotonicity may be stronger than robust $\mathcal{S}$ monotonicity 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 a 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 a 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 but no singleton can play this role.

At last, we introduce a weak condition, the interior $\mathcal{S}$ reward property, to complete the group of sufficient conditions.

Definition 9: *A social choice function f satisfies the **interior $\mathcal{S}$ reward property**, if for any coalition S 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 $Y_S[\mathcal{S}]$.

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 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 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 $m_i^3 \in M_i^3 \equiv \prod_{S \ni i} Y_S[\mathcal{S}]$ 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 carnality $K < \infty$, define $y^k \equiv y^K$ for all $k > K$. Then let the outcome $g(m)$ be a lottery of realization $y(m_{-S^*}^1)$ with probability $\frac{m_{i^*}^4}{m_{i^*}^4+1}$ and of realization $y^k(m_{-S^*}^1)$ with probability $\frac{0.5^k}{m_{i^*}^4+1}$ 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. The robust $\mathcal{S}$ incentive compatibility condition prevents a coalition from profitably manipulating payoff types without leaving Rule 1. According to the definition of the $\mathcal{S}$ reward set, no coalition in $\mathcal{S}$ has the incentive to trigger Rule 2 either.

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^*, such that for all $t_{-j} \in T_{-j}$, the $\mathcal{S}$ reward functions proposed by t_j^* are used to determine the outcome allocations. In this case, we can identify a profitable unilateral deviation for type-t_j^* agent j.

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 profitable $\mathcal{S}$ 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 an additional 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.*

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 endogenous coalition design exercise for future study.

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 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 one-to-one mapping 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 exist 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 fully 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 each coalition $S \ni i$ 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 $\theta'_i\in\bar{\beta}_i(\theta_i)\backslash\beta_i(\theta_i)$, θ'_i has to be equal to θ_i. We can arbitrarily pick θ'_{-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:

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

With the information above, we 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:

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

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

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

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 $i\in I$, $t_i\in T_i$, $S\ni i$, and $t_{S\backslash\{i\}}$ happening with zero probability under distribution $\pi_i(t_i)$, we specify the following belief revising rule:

let the revised belief $\pi_i(t_i)[t_{S\setminus\{i\}}, t_{-S}|t_{S\setminus\{i\}}]$ be equal to the marginal belief $\pi_i(t_i)[t_{-S}]$ 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 $y(\theta_{-i}) = h\Big(\alpha_i(t_i^*), \big([\xi_j^1]^{-1}(\theta_j, \theta_j)\big)_{j\neq i}\Big)$ for all $\theta_{-i} \in \Theta_{-i}$. According to condition (ii), by having S go over every set $S \ni i$ 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, the two conditions in Definition 11 would contradict with each other when $S \equiv \{i\}$ given types in T_i^2 expect other agents to follow the profile of identity mappings under α_{-i}. Thus, 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. 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 all coalition $S \ni i$ 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 $t'_i \in T_i$, we further have

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

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

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, 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 an 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 $i \in S\setminus\underline{S}$. 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 $\sigma_j^2(t_j^*) = B$ 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 $y = \sigma_j^3(t_j^*)$. For each $t_{-j} \in T_{-j}$, $g(\sigma(t_j^*, t_{-j}))$ is a lottery of realization $y(\sigma_{-j}^1(t_{-j}))$ with probability $\frac{\sigma_j^4(t_j^*)}{1+\sigma_j^4(t_j^*)}$ and of realization $y^k(\sigma_{-j}^1(t_{-j}))$ with probability $\frac{0.5^k}{1+\sigma_j^4(t_j^*)} > 0$ 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

$$\psi_j(\theta_{-j}, \theta'_{-j}) \equiv \sum_{\hat{\theta}_{-j}(t_{-j})=\theta_{-j}, \sigma_{-j}^1(t_{-j})=\theta'_{-j}} \pi_j(t_j^*)[t_{-j}]$$

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

$$\frac{\sigma_j^4(t_j^*)}{1+\sigma_j^4(t_j^*)} \sum_{\theta_{-j} \in \Theta_{-j}, \theta'_{-j} \in \beta_{-j}(\theta_{-j})} u_j\big(y(\theta'_{-j}), (\theta_j^*, \theta_{-j})\big)\psi_j(\theta_{-j}, \theta'_{-j})$$
$$+ \frac{1}{1+\sigma_j^4(t_j^*)} \sum_{\theta_{-j} \in \Theta_{-j}, \theta'_{-j} \in \beta_{-j}(\theta_{-j})} \sum_{k=1,2,...} 0.5^k u_j\big(y^k(\theta'_{-j}), (\theta_j^*, \theta_{-j})\big)\psi_j(\theta_{-j}, \theta'_{-j}).$$

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 $\sigma_j'(t_j^*) = (\sigma_j^1(t_j^*), \sigma_j^2(t_j^*), y^k, \sigma_j^4(t_j^*))$ 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 $\sigma_j'(t_j^*) = (\sigma_j^1(t_j^*), \sigma_j^2(t_j^*), \sigma_j^3(t_j^*), \sigma_j^4(t_j^*) + 1)$ 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 $\beta_i(\theta_i) = \bigcup_{\{t_i \in T_i | \hat{\theta}_i(t_i) = \theta_i\}} \{\sigma_i^1(t_i)\}$ 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 $\sigma_i^1(t_i^*) = \theta_i'$. 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

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

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 $y(\sigma_{-i}^1(t_{-i}))$ 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. □

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. By robust $\mathcal{S}$ incentive compatibility, to show that σ_S is not a profitable deviation for t_S, 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 $y^k(\hat{\theta}_{S\setminus\underline{S}}(t_{S\setminus\underline{S}}), \hat{\theta}_{-S}(t_{-S}))$ 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 $S \ni j$, define $T_{-j}(S) \equiv \{t_{-j} \in T_{-j} : \exists \bar{S} \ni j \text{ s.t. } \sigma(t^*_j, t_{-j}) \in \hat{M}(\bar{S}), S^*[\bar{S}] = S\}$, 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 $S \ni j$ such that $\phi_j(S) > 0$, define a distribution $\psi_j[S] \in \Delta(\Theta_{-j} \times \Theta_{-S})$ 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} on conditional that $t_{-j} \in T_{-j}(S)$. Then, type-t^*_j agent j's expected utility is equal to

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

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

Case 1: suppose there exists $S \ni j$ 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 $S \ni j$ 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. This implies that σ is not an interim $\mathcal{S}$ equilibrium, a contradiction.

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 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

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

Fix any profile of types t^*_S such that $\hat{\theta}_S(t^*_S) = \theta_S$ and $\sigma^1_S(t^*_S) = \theta'_S$. For all $i \in S$, define a strategy σ'_i by $\sigma'_i(t^*_i) = (\sigma^1_i(t^*_i), B, m^3_i, K^*)$ 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 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 12: *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 $\bar{a}[\theta'_{-i}, \bar{\phi}_i[\theta'_{-i}]]$, $\underline{a}[\theta'_{-i}, \bar{\phi}_i[\theta'_{-i}]] \in R_i^*(\theta'_{-i})$. Multiply both sides of the above inequality by $\sum_{\theta''_{-i} \in \Theta_{-i}} \psi_i(\theta''_{-i}, \theta'_{-i})$ and sum up over different $\theta'_{-i} \in \Theta_{-i}$. Then we have

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

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. 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 a social choice function satisfies robust monotonicity and conditional no total indifference, sufficient conditions in our Theorem 2 hold under the minimal coalition pattern. By our Theorem 2, the social choice function 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.

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

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. and Yannelis, N. C. (2020). Full implementation under ambiguity. *American Economic Journal: Microeconomics*. Forthcoming.

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

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

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.

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

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. (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*.

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.

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.