Robust Implementation with Costly Information

Harry Pei* Bruno Strulovici‡

January 21, 2022

Abstract: We study whether a planner can robustly implement a state-contingent social choice function when (i) agents must incur a cost to learn the state and (ii) the planner faces uncertainty regarding agents' preferences over outcomes, information costs, marginal utilities from transfers, and beliefs and higher-order beliefs about one another's payoffs. We propose mechanisms that can approximately implement the desired social choice function as long as the perturbations concerning agents' payoffs have small ex ante probability. We also design mechanisms that are robust when agents tremble with small probability and when agents' signals about the state are noisy.

Keywords: Robust Implementation, Partial Implementation, Critical Path Lemma.

JEL Codes: D82, D83.

1 Introduction

Implementation theory studies whether and how a state-contingent social choice function, such as convicting guilty defendants and acquitting innocent ones, can be achieved when (i) the information necessary to implement the objective is unknown to the planner and (ii) the social choice function is in conflict with the interests of the agents who do have access to the relevant information.

We study this question under two additional constraints: (iii) agents need to incur a cost to learn the relevant information,[1] and (iv) the planner knows agents' payoffs with probability close to one, but faces uncertainty regarding agents' payoffs with the remaining probability and regarding agents' beliefs and higher-order beliefs about one another's payoffs.

The motivation for our research question is twofold. First, in many situations of interest, agents do not possess the relevant information at the outset and must acquire it at some cost. For example,

*Department of Economics, Northwestern University. Email: harrydp@northwestern.edu

†Department of Economics, Northwestern University. Email: b-strulovici@northwestern.edu

‡We thank Gabriel Carroll, Yi-Chun Chen, Eddie Dekel, Dana Foarta, Yingni Guo, Matt Jackson, Takashi Kunimoto, Meg Meyer, Stephen Morris, David Rodina, Takuo Sugaya, Satoru Takahashi, and Olivier Tercieux for helpful comments. Pei thanks the NSF Grant SES-1947021 for financial support.

[1]As we explain by the end of Section 3, we are not aware of any existing work that examines our notion of robust implementation in environments where agents' costs of learning are zero. Although in some examples of interest, there exists a trivial solution to our robust implementation problem when agents' costs are zero, there is no straightforward solution to our robust implementation problem under generic payoffs even when agents' costs of learning are zero.

investigators need to exert effort in order to learn whether a defendant is guilty or innocent.

Second, the literature on robust implementation, pioneered by Bergemann and Morris (2005), has underscored the importance of implementing social choice functions when agents' preferences are not common knowledge. This literature has shown that only a restrictive subset of social choice functions can be implemented when robustness is required to hold *globally*: a social choice function can be implemented for arbitrary payoff functions and beliefs of the agents only if it is ex post incentive compatible (Bergemann and Morris 2005). When the notion of implementation is *local* but in an *interim* sense, i.e., for all profiles of agent-types close to a given profile, Oury and Tercieux (2012) show that implementable social choice functions must satisfy the monotonicity condition in Maskin (1999), a demanding property that is violated in a number of settings.

We consider a novel notion of robust implementation that builds on the concept of equilibrium robustness introduced by Kajii and Morris (1997). A Nash equilibrium of a complete information game is robust if it can be approximated by some equilibria in *all* nearby incomplete information games, which are all games where players' payoffs match those of the complete information game with probability close to one. Building on this approach, our concept of robust implementation is *local* and *ex ante*, rather than *interim*, in the sense that the perturbations considered have a small probability ex ante relative to the complete information game. As a result, our concept has the potential to avoid some of the stringent implications of global and interim concepts.

Our concept of robust implementation departs from Kajii and Morris (1997) by imposing a key restriction on the set perturbations considered by the planner. Since we study a mechanism design problem rather than a game, we focus on perturbations in which agents' payoffs do not depend per se on the *messages* that they send to the mechanism, i.e., it is common knowledge that messages are cheap talk. Instead, perturbations pertain to agents' preferences concerning the *outcomes* implemented as a result of their messages, their costs of learning the state, and their marginal utilities from transfers.

Our analysis focuses on the case of two agents.[2] A planner wishes to implement a social choice function that maps a finite set of states to a set of lotteries over outcomes.[3] The planner commits to a mechanism mapping agents' messages to outcomes and transfers *without* knowing the state as well as how the environment is perturbed. Agents observe the mechanism as well as their own

[2]Our main result is a possibility result. We construct mechanisms that can robustly implement the desired social choice function when two agents have the ability to learn the state. This mechanism can straightforwardly be extended to any arbitrary number agents, by applying it to two agents and ignoring the reports of all remaining agents.

[3]In Appendix B, we generalize our main result to a continuum of states when both the social choice function and agents' payoff functions in the unperturbed environment are continuous with respect to the state.

payoff functions under the perturbation. They independently decide whether to observe the state at some cost and then send messages to the planner.[4]

Our main result shows that every social choice function is robustly implementable under a generic assumption on the objective state distribution.[5] Our proof constructs a class of mechanisms called the *Augmented Status Quo Rules with Ascending Transfers*. These mechanisms treat states asymmetrically by (i) introducing the desired outcome in the ex ante most likely state as a *status quo outcome*, which is implemented whenever agents' reports are in conflict or when their reports match on the ex ante most likely state, and (ii) when agents' reports are the same, both of them receive a larger reward if they report an ex ante less likely state.

Another interesting feature of our mechanism is that each agent has $2n-1$ messages when there are n states, with one message corresponding to the ex ante most likely state and two messages corresponding to every other state. The two messages that represent the same state induce the same outcome regardless of the other agent's message but lead to different transfers. In Section 2, we use an example to explain why a larger message space makes the mechanism robust when some agent types have very high costs of learning or have very low marginal utilities from transfers.

Finally, we provide several results concerning stronger notions of robust implementation. First, Proposition 1 shows that it is impossible to approximately implement any non-constant social choice function when agents' payoff functions can be different from those in the original environment with probability bounded away from zero. This result motivates our notion of robust implementation which considers robustness to *local perturbations*, that is, agents' payoff functions coincide with those in the unperturbed environment with probability close to one.

Second, our notion of robustness concerns *partial* implementation: we only require that the desired social choice function be implemented by some (not necessarily all) equilibria of the game induced by our mechanism. In fact, when agents' payoff functions in the unperturbed environment put no weight on how outcomes relate to the state (e.g., when agents have *transparent motives* as in Lipnowski and Ravid 2020), or when agents' costs of learning in the unperturbed environment are above some cutoff, Proposition 2 shows that no mechanism can fully implement any non-constant social choice function. In these situations, there always exists an equilibrium in which no agent

[4]We extend our robust implementation result to settings where agents tremble with small probability when sending messages, and where agents' learn noisy signals about the state. See Section 4.3 for details.

[5]Our generic assumption requires that there exists a state that occurs with strictly higher probability compared to any other state. This generic assumption can be dropped when there is a known upper bound on the ratio between agents' costs of learning and their marginal utilities from transfers.

learns the state.[6] We also provide sufficient conditions for full implementation: When at least one agent's preference and the social choice function satisfy a strict version of Rochet (1987)'s cyclical monotonicity condition and this agent's cost of learning is small enough, the planner can robustly and fully implement that social choice function by ignoring the report of the other agent.

Section 2 uses an example to explain why mechanisms that (i) treat states symmetrically by rewarding agents a fixed amount when their reports match, and (ii) giving both agents no transfer and uniformly randomizing across outcomes when agents' reports mismatch, cannot robustly implement the desired social choice function. We then explain why mechanisms that (i) treat states asymmetrically, (ii) provide asymmetric rewards across different states, and (iii) include more messages than states, help address robustness issues. We introduce our general model in Section 3 and state our main results in Section 4. Section 5 presents impossibility results pertaining to stronger notions of implementation. Section 6 concludes and reviews the related literature.

2 Example

A planner faces a defendant who is either guilty or innocent, i.e., there are two states $\theta \in \{\text{innocent}, \text{guilty}\}$. The defendant's prior probability of guilt is $q \equiv \Pr(\theta = \text{guilty}) \in (0, 1)$.

The planner's objective is to convict guilty defendants and to acquit innocent ones. She commits to a mechanism $\mathcal{M} \equiv \{M_1, M_2, g, t_1, t_2\}$ in order to solicit information from two agents (e.g., investigators), where M_i is a finite set of messages for agent $i \in \{1, 2\}$, $g : M_1 \times M_2 \to [0, 1]$ is a mapping from agents' messages to the probability of conviction, and $t_i : M_1 \times M_2 \to \mathbb{R}$ is the transfer to agent i that depends only on the messages but not the realized state.

Each agent can, at some cost, conduct an investigation and learn whether the defendant is guilty or innocent. Each agent's decision to investigate and the observation resulting from his investigation are private in the sense that they are unbeknownst to the planner and the other agent.

Agent i's payoff function is $t_i - c\chi_i$, where $\chi_i \in \{0, 1\}$ denotes i's decision of whether to conduct investigation and $c > 0$ is the agent's cost of conducting investigation.

When agents' payoff functions are common knowledge, the planner can convict the guilty and acquit the innocent via the following mechanism: Each agent is asked to report whether the defendant is *guilty* or *innocent*. The outcome and the transfers are given by:

[6] This result echoes Strulovici (2021), who shows in a sequential model of learning that when agents' preferences are state independent, implementation is impossible even in a partial sense when signals about the state of the world are subject to an *information attrition* condition.

transfers	innocent	guilty
innocent	R, R	$0, 0$
guilty	$0, 0$	R, R

outcome	innocent	guilty
innocent	acquit	convict with prob 1/2
guilty	convict with prob 1/2	convict

where the reward $R > 0$ is large relative to agents' costs of learning c. One can verify that there exists an equilibrium where both agents conduct investigations and report their findings truthfully.

Robust Implementation with Biased Agents: The Maskin mechanism fails to implement the desired outcome when agents are subject to biases over outcomes and the planner does not know the direction and magnitude of agents' biases when designing the mechanism, nor does she know agents' beliefs and higher-order beliefs about each other's bias. This is the case even when agents are biased with arbitrarily small probability.

To fix ideas, suppose nature draws a random variable ω from a countable set $\Omega \equiv \{\omega_0, \omega_1, \omega_2, ...\}$ such that $\omega = \omega_t$ occurs with probability $\eta(1-\eta)^t$ for every $t \in \mathbb{N}$ where $\eta > 0$ is close to 0. The realization of ω is independent of whether the defendant is guilty or innocent. Agent 1 knows which element of the partition $\{\omega_0\}, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, ...$ the realized ω belongs to before deciding whether to conduct his investigation as well as what to report. Agent 2 knows which element of the partition $\{\omega_0, \omega_1\}, \{\omega_2, \omega_3\}, ...$ the realized ω belongs to before deciding whether to conduct his investigation as well as what to report. Each agent updates his belief about the other agent's knowledge of ω according to Bayes rule.[7]

Agent 2's payoff is $t_2 - \chi_2 c$ at every $\omega \in \Omega$. Agent 1's payoff is $t_1 - \chi_1 c$ at every $\omega \in \Omega \backslash \{\omega_0\}$. We consider two cases regarding agent 1's payoff at ω_0, leading to two classes of perturbations.

1. **Large-bias type:** When $\omega = \omega_0$, agent 1 receives a large benefit B from acquitting the defendant (e.g., agent 1 is the defendant's friend when $\omega = \omega_0$).

2. **Purely outcome-driven type:** When $\omega = \omega_0$, agent 1 receives a strictly positive (possibly small) benefit B from acquitting the defendant and does not care about transfers.

In both perturbations, agents' payoff functions coincide with those in the unperturbed environment when $\omega \neq \omega_0$, which happens with probability $1 - \eta$, i.e., these perturbations are *small*.

The Maskin mechanism fails to implement the desired objective even when η is close to 0. In the case with a large-bias type, for every $R \in \mathbb{R}_+$, there exists $B > R$ such that no matter how small η

[7]For example, the type of agent 2 who knows that $\omega \in \{\omega_0, \omega_1\}$ attaches probability $\frac{1}{2-\eta}$ to agent 1 being type $\{\omega_0\}$, the type of agent 1 who knows that $\omega \in \{\omega_1, \omega_2\}$ attaches probability $\frac{1}{2-\eta}$ to agent 2 being type $\{\omega_0, \omega_1\}$.

is, the perturbed environment has a unique equilibrium where agents always report *innocent* and the defendant is acquitted regardless of his guilt. In the case with a purely outcome-driven type, as long as $B > 0$, no matter how small η is, the perturbed environment has a unique equilibrium where agents always report *innocent* and the defendant is acquitted regardless of his guilt.

This is because when $\omega = \omega_0$, agent 1 wants to maximize the probability of acquittal if he is the large-bias type or the purely outcome-driven type, so he has a strict incentive to report innocent regardless of θ. When $\omega \in \{\omega_0, \omega_1\}$, agent 2 is unbiased, but he believes that agent 1 is biased with probability greater than $1/2$, so he believes that agent 1 will report innocent with probability greater than $1/2$ regardless of θ. Since agent 2 maximizes his expected transfer minus his cost of investigation, he has a strict incentive to report innocent regardless of θ. By induction, this contagion argument shows that all types of both agents will report *innocent* in the unique equilibrium of the perturbed environment.

In general, agents may be biased in either direction, e.g., some agent types may benefit from convicting the defendant, and with arbitrary magnitude. The planner faces uncertainty about the direction and magnitude of these biases as well as about agents' beliefs and higher-order beliefs about each other's biases. The planner aims to design a mechanism that approximately implements the desired objective when agents are unbiased with probability close to 1, but can have arbitrary biases with small but positive probability and may have arbitrary beliefs and higher-order beliefs as long as those beliefs can be derived from a common prior.

Status Quo Rule with Ascending Transfers: We propose a mechanism that can implement the desired outcome even when the planner does not know the direction and magnitude of agents' biases. For now, we maintain the assumption that agents' cost of learning the state is commonly known to be c, and we also *rule out* types who are purely outcome driven. We explain later how to modify our mechanism so that it can robustly implement the desired social choice function when there are types who have large learning costs and types who are purely outcome drive.

Our mechanism features two messages for each agent: *innocent* and *guilty*. The outcome and transfers are specified as follows:

outcome	innocent	guilty
innocent	acquit	acquit
guilty	acquit	convict

transfers	innocent	guilty
innocent	R^1, R^1	$0, 0$
guilty	$0, 0$	R^2, R^2

where $R^2 > R^1 > 0$ and $R^2 - R^1$ is bounded below by some function of c.

The above mechanism features a status quo outcome, *acquit*, which is implemented as long as one agent reports *innocent*. The defendant is convicted if and only if both agents report *guilty*. Agents receive strictly positive transfers only when their reports coincide and they receive a larger transfer when both of them report *guilty* compared to both of them report *innocent*.

To explain intuitively why this mechanism implements the desired objective, we restrict attention to the following class of perturbations and defer the general proof to Section 4. Nature draws a random variable ω from a countable set $\{\omega_0, \omega_1, \omega_2, ...\}$ according to distribution $\Pi \in \Delta\{\omega_0, \omega_1, \omega_2, ...\}$. Agent 1's information partition is $\{\omega_0\}, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, ...$ Agent 2's information partition is $\{\omega_0, \omega_1\}, \{\omega_2, \omega_3\}, ...$ Agent 2's payoff is $t_2 - c\chi_2$ at every ω. Agent 1's payoff is $t_1 - c\chi_1$ at every ω except for ω_0.

Each perturbation in the above class is characterized by Π and agent 1's payoff function at ω_0, under which we construct an equilibrium that approximately implements the desired outcome.

1. Suppose first that when $\omega = \omega_0$, agent 1 benefits from **acquitting** the defendant. This biased type can guarantee his desired outcome by reporting *innocent* no matter what. However, given that $R^1 < R^2$, $\Pi(\omega_1)$ needs to be strictly less than $\Pi(\omega_0)$ for type $\{\omega_0, \omega_1\}$ of agent 2 to have an incentive to report *innocent* no matter what, and $\Pi(\omega_2)$ needs to be strictly less than $\Pi(\omega_1)$ for type $\{\omega_1, \omega_2\}$ of agent 1 to have an incentive to report *innocent* no matter what, and so on. The upper bounds on these probabilities form a decaying geometric sequence, which means that the total probability of types that are *infected* by type ω_0 is at most $\sum_{t=0}^{+\infty} \left(\frac{R_1}{R_2}\right)^t \Pi(\omega_0) = \frac{R_2}{R_2 - R_1}\Pi(\omega_0)$. This expression vanishes to 0 as $\Pi(\omega_0) \to 0$.

2. Suppose now that when $\omega = \omega_0$, agent 1 benefits from **convicting** the defendant. If $\Pi(\omega_t) = \eta(1-\eta)^t$ for every $t \in \mathbb{N}$ *and* this biased type reports guilty no matter what, then all types of both agents will have a strict incentive to report guilty since $R^2 > R^1 > 0$.

   However, thanks to the outcome function of our mechanism, the defendant is convicted only if *both agents report guilty*, so the biased type cannot convict an innocent defendant when the other agent does not report guilty when the defendant is innocent.

   In fact, this biased type strictly prefers to be truthful when he believes that the other agent is truthful, since reporting innocent when the defendant is innocent leads to a strictly positive transfer and reporting guilty when the defendant is innocent leads to zero transfer. Hence, all types reporting truthfully is an equilibrium no matter how large the bias is.

Uncertainty about Biases and Costs: The planner may also face uncertainty about agents' costs of learning the state, agents' marginal utilities from transfers, and agents' beliefs and higher-order beliefs about each other's costs, biases, and marginal utilities.

We show that when the prior probability of guilt q is not $\frac{1}{2}$,[8] there is a mechanism that approximately implements the desired social choice function when agents' payoffs coincide with those in the unperturbed environment with probability close to 1 but can have arbitrary biases, costs, and marginal utilities from transfers with complementary probability.

For expositional simplicity, we focus on the case where the defendant's ex ante probability of guilty is strictly less than $\frac{1}{2}$. We also focus on the information structures considered before: Agent 1's information partition is $\{\omega_0\}, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, ...$ Agent 2's information partition is $\{\omega_0, \omega_1\}, \{\omega_2, \omega_3\}, ...$ Agent 2's payoff function is $t_2 - c\chi_2$ at every ω. Agent 1's payoff function is $t_1 - c\chi_1$ at every $\omega \neq \omega_0$. When $\omega = \omega_0$, agent 1's payoff is $B \cdot \mathbf{1}\{y = \text{convict}\} + \widetilde{b}t_1 - \widetilde{c}\chi_1$ for some $B > 0$, learning cost $\widetilde{c} \geq 0$, and marginal utility from transfer $\widetilde{b} \geq 0$.

We start by explaining why the status quo rule with ascending transfers *cannot* implement the desired outcome if agent 1's payoff at ω_0 includes a large benefit B from convicting the defendant and a large cost of learning $\widetilde{c}$ (we call him a *high-cost biased type*), or his payoff at ω_0 includes a strictly positive (possibly small) benefit from convicting the defendant but he does not care about transfers, i.e., $\widetilde{b} = 0$ (we call him a *purely outcome-driven type*).

Intuitively, when a high-cost biased type or a purely outcome-driven type believes that the other agent reports truthfully, he prefers to report *guilty* conditional on the defendant being guilty, since he benefits from convicting the defendant. If this type wants to report *innocent* when the defendant is innocent, then he needs to conduct an investigation, but when $\widetilde{c}$ is very large or when $\widetilde{b}$ is very small, his cost of doing so outweighs his benefit from the transfers. This explains why the high-cost biased type and the purely outcome-driven type prefer to *report guilty no matter what* even when he believes that the other agent reports truthfully. This causes contagion when the distribution of ω satisfies $\Pi(\omega_t) = \eta(1-\eta)^t$ for every $t \in \mathbb{N}$, no matter how small η is.

Augmented Status Quo Rule with Ascending Transfers: We propose another mechanism called the *Augmented Status Quo Rule with Ascending Transfers*, in which each agent has a third message which we denote $-$*guilty*. The outcome and transfers are given by:

[8] In general, our result requires that there exists a state whose prior probability of occurrence exceeds that of every other state. This property is generic among the set of all prior distributions.

outcome	$-$guilty	innocent	guilty
$-$guilty	convict	acquit	convict
innocent	acquit	acquit	acquit
guilty	convict	acquit	convict

transfers	$-$guilty	innocent	guilty
$-$guilty	R^0, R^0	R^0, R^0	$0, 0$
innocent	R^0, R^0	R^1, R^1	$0, 0$
guilty	$0, 0$	$0, 0$	R^2, R^2

where $\frac{R^0}{R^2}$ is strictly less than but is close to 1, $R^2 > R^1 > R^0 > 0$, and $R^2 - R^1$ and $R^1 - R^0$ are strictly positive and are bounded below by some affine function of agents' cost of learning c.

According to our new mechanism, (i) message $-$*guilty* implements the same outcome as message *guilty* regardless of the other agent's message, (ii) each agent can unilaterally implement the status quo outcome *acquit* by reporting *innocent*, and (iii) coordinating on message $-$*guilty* leads to a lower transfer compared to coordinating on any of the other two messages.

To understand why the message $-$*guilty* makes our mechanism robust to biased types that have high learning costs or types that have very low marginal utility from transfers, suppose that every non-biased type never reports *guilty* when the defendant is innocent (but may report $-$*guilty* and *innocent* with arbitrary probabilities). Then, type ω_0 of agent 1 receives an expected transfer of at least $(1-q)R^0$ if he reports $-$*guilty* and an expected transfer of at most qR^2 if he reports *guilty*. Given that $q < \frac{1}{2}$ and that $\frac{R^0}{R^2} \approx 1$, reporting $-$*guilty* leads to a greater expected transfer.

The assumption that $0 < R^0 < R^1 < R^2$ implies that coordinating on message $-$*guilty* leads to a lower expected transfer compared to coordinating on message *innocent* and coordinating on message *guilty*. As a result, for every type of agent whose payoff function coincides with that in the unperturbed environment, he prefers to conduct his investigation and to report the state truthfully (i.e., report *innocent* when the defendant is innocent and report *guilty* when the defendant is guilty) as long as he believes that (i) no type of the other agent reports *guilty* when the defendant is innocent, and (ii) with probability at least 1/2, the other agent reports truthfully.

We use this observation and the critical path lemma in Kajii and Morris (1997) to show that for every perturbation where agents' payoff functions coincide with those in the unperturbed environment with probability close to 1, there is an equilibrium where both agents report truthfully with probability close to 1 and the desired outcome is implemented with probability close to 1.

3 General Model

Primitives: A planner wants to implement a social choice function $f : \Theta \to \Delta(Y)$ where Θ is a finite set of states and Y is a finite set of outcomes.[9] The typical elements in these sets are $\theta \in \Theta$ and $y \in Y$. Let $n \equiv |\Theta|$ be the number of states. Let $q \in \Delta(\Theta)$ be the objective distribution of θ, with $q(\theta)$ the probability of state θ. We assume that $q(\theta) > 0$ for every $\theta \in \Theta$.

The planner does not know θ and elicits information about θ from two agents. She commits to a mechanism $\mathcal{M} \equiv \{M_1, M_2, t_1, t_2, g\}$, where M_i is a finite set of messages for agent $i \in \{1, 2\}$,[10] $t_i : M_1 \times M_2 \to \mathbb{R}$ is the transfer to agent i, and $g : M_1 \times M_2 \to \Delta(Y)$ is the implemented outcome.

After observing $\mathcal{M}$, agents simultaneously and independently decide whether to observe θ at some costs. Let $\chi_i \in \{0, 1\}$ be agent i's decision to obtain information, where $\chi_i = 1$ represents agent i obtaining information about θ and vice versa. Let $c_i \geq 0$ be his cost. We assume that information acquisition is *covert* in the sense that neither agent $-i$ nor the planner can observe χ_i.

Then the agents simultaneously send messages $(m_1, m_2) \in M_1 \times M_2$ to the planner, after which the planner makes transfers and implements an outcome according to $\mathcal{M}$. Agent i's payoff is:

$$t_i - c_i\chi_i + u_i(\theta, y). \tag{3.1}$$

A leading example is the case where agents have *transparent motives* (Lipnowski and Ravid 2020), that is, $u_i(\theta, y)$ does not depend on θ for $i \in \{1, 2\}$.

One assumption to highlight is that the agents' transfers *cannot* depend on the realized state. This stands in contrast to existing works on contracting with costly information acquisition, such as Zermeno (2011), Carroll (2019), and Clark and Reggiani (2021) where transfers can also depend on the realized state. Our model fits situations where either the planner cannot verify the state ex post, or additional information about the state takes a long time to arrive so that rewarding agents based on such late information is impractical.

Robustness Concerns: We examine whether the planner can *robustly* implement f when agents' preferences over outcomes, costs of obtaining information, and their beliefs and higher-order beliefs about each other's payoffs can be different from what the planner believed to be. Similar to Oury and Tercieux (2012), we focus on robust partial implementation: the planner only requires f to be

[9] We generalize our main result to a continuum of states in Appendix B, under the assumption that the social choice function f and agents' payoff functions in the unperturbed environment (u_1, u_2) are continuous in θ.

[10] See Jackson (1992) for a justification for focusing attention on finite mechanisms.

implemented in *one* equilibrium, not necessarily all equilibria.

Following Kajii and Morris (1997), a *perturbation*

$$\mathcal{G} \equiv \{\Omega, \Pi, Q_1, Q_2, \widetilde{u}_1, \widetilde{u}_2, \widetilde{c}_1, \widetilde{c}_2, \widetilde{b}_1, \widetilde{b}_2\} \tag{3.2}$$

consists of a countable set of *circumstances* Ω, a distribution $\Pi \in \Delta(\Omega)$ over the set of circumstances which we assume is independent of θ, a partition Q_i of Ω such that agent $i \in \{1,2\}$ knows which element of the partition Q_i the realized ω belongs to, as well as mappings $\widetilde{u}_i : \Omega \times \Theta \times Y \to \mathbb{R}$, $\widetilde{c}_i : \Omega \to [0, +\infty)$, and $\widetilde{b}_i : \Omega \to [0, +\infty)$ for $i \in \{1,2\}$. Agent i's payoff under perturbation $\mathcal{G}$ is

$$\widetilde{b}_i(\omega) t_i - \widetilde{c}_i(\omega)\chi_i + \widetilde{u}_i(\omega, \theta, y). \tag{3.3}$$

Intuitively, perturbations can affect agents' preferences over outcomes (θ, y), their costs of learning θ, and their marginal utilities from transfers. We assume that the latter two are non-negative. For every $\overline{c} > 0$, $\mathcal{G}$ is a $\overline{c}$-*bounded perturbation* if $\widetilde{c}_i(\omega)/\widetilde{b}_i(\omega) \leq \overline{c}$ for every i and $\omega \in \Omega$.

For every $\omega \in \Omega$, let $Q_i(\omega)$ be the partition element of Q_i that contains ω, which we call agent i's *type*. Type $Q_i(\omega)$ is a *normal type* if $\widetilde{u}_i(\omega', \theta, y) = u_i(\theta, y)$, $\widetilde{c}_i(\omega') = c_i$, and $\widetilde{b}_i(\omega') = 1$ for every $\omega' \in Q_i(\omega)$, i.e., type $Q_i(\omega)$ of agent i knows that his payoff in the perturbed environment coincides with his payoff in the unperturbed environment. We introduce our notion of *small perturbations*:

η-Perturbation. *For every $\eta \in (0,1)$, we say that $\mathcal{G}$ is an η-perturbation if*

$$\Pi\Big(\textit{both agents are normal types}\Big) \geq 1 - \eta. \tag{3.4}$$

We say that $\mathcal{G}$ is a $\overline{c}$-bounded η-perturbation if it is an η-perturbation and is $\overline{c}$-bounded.

Intuitively, a perturbation is *small* if agents' payoff functions coincide with those in the unperturbed environment with probability close to one, but their payoff functions can be very different with small but positive probability. Even though a normal type's payoff function coincides with that in the unperturbed environment, he may believe that the other agent is not normal, and may believe that the other agent thinks that he is not normal, and so on. The class of perturbations considered in Section 2 are η-perturbations since both agents are normal types when $\omega \in \Omega \backslash \{\omega_0\}$, and event $\Omega \backslash \{\omega_0\}$ occurs with probability $1 - \eta$ under Π.

The planner does not observe the perturbation $\mathcal{G}$ when she designs the mechanism. After observing the perturbation $\mathcal{G}$ and the mechanism $\mathcal{M}$, the two agents are playing an incomplete

information game, which we denote by $(\mathcal{M}, \mathcal{G})$. A typical strategy profile of this game is denoted by σ. Let $g_\sigma(\theta) \in \Delta(Y)$ be the implemented outcome conditional on the state being θ when the planner commits to outcome function g and agents behave according to σ.

Our main results in Section 4 focus on two notions of *local robust implementation*, that is, the planner designs a mechanism that approximately implements f for *all* small enough perturbations.

1. We say that $\mathcal{M}$ *robustly implements* f if for every $\varepsilon > 0$, there exists $\eta > 0$ such that for every η-perturbation $\mathcal{G}$, there exists an equilibrium $\sigma(\mathcal{G})$ of the incomplete information game induced by $(\mathcal{M}, \mathcal{G})$, such that

$$\max_{\theta \in \Theta} ||g_{\sigma(\mathcal{G})}(\theta) - f(\theta)||_{\text{TV}} < \varepsilon, \tag{3.5}$$

   where $||\cdot||_{TV}$ is the total variation distance between two distributions.

2. We say that $\mathcal{M}$ *robustly implements* f for all $\overline{c}$*-bounded perturbations* if for every $\varepsilon > 0$, there exists $\eta > 0$ such that for every $\overline{c}$-bounded η-perturbation $\mathcal{G}$, there exists an equilibrium $\sigma(\mathcal{G})$ of the incomplete information game induced by $(\mathcal{M}, \mathcal{G})$ such that inequality (3.5) holds.

By definition, our two notions of robust implementation differ only in terms of whether we allow for unbounded costs of learning or very low marginal utilities from transfers. If $\mathcal{M}$ robustly implements f, then it robustly implements f for all $\overline{c}$-bounded perturbations.

Section 5 considers two notions of *global robust implementation*, namely, the planner designs a mechanism that approximately implements f under all perturbations, including those where agents' payoff functions are different from those in the unperturbed environment with high probability.

1. Mechanism $\mathcal{M}$ *globally implements* f if for every $\varepsilon > 0$ and every perturbation $\mathcal{G}$, there exists an equilibrium $\sigma(\mathcal{G})$ of incomplete information game $(\mathcal{M}, \mathcal{G})$ such that inequality (3.5) holds.

2. Mechanism $\mathcal{M}$ *globally implements* f *for all* $\overline{c}$*-bounded perturbations* if for every $\varepsilon > 0$ and every $\overline{c}$-bounded perturbation $\mathcal{G}$, there exists an equilibrium $\sigma(\mathcal{G})$ of incomplete information game $(\mathcal{M}, \mathcal{G})$ such that inequality (3.5) holds.

By definition, if $\mathcal{M}$ globally implements f (for all $\overline{c}$-bounded perturbations), then it robustly implements f (for all $\overline{c}$-bounded perturbations). In Section 5.1, we show that for any $\overline{c}$ and any social choice function f that depends non-trivially on the state, there is no finite mechanism that can globally implement f for all $\overline{c}$-bounded perturbations.

Remark 1: Although our main focus is on settings where agents' costs of obtaining information are strictly positive, our mechanism also works when agents' costs of learning are zero. In addition, we are unaware of existing any work that examines our notion of robust implementation in environments where agents' costs of obtaining information are zero.

Nevertheless, when $c_1 = c_2 = 0$ and $u_1(\theta, y)$ and $u_2(\theta, y)$ do not depend on θ, there exists a trivial solution to our (partial) robust implementation problem: whenever the planner implements y, she promises agent 1 a transfer of $-u_1(y)$ and promises agent 2 a transfer of $-u_2(y)$. When the normal type of each agent is indifferent between all messages and their costs of learning θ are zero, there exists an equilibrium where all normal types learn the state and report truthfully.

This trivial solution does not work when $u_1(\theta, y)$ and $u_2(\theta, y)$ depend on θ. When $c_1 = c_2 = 0$, and (f, u_1, u_2) satisfies the Maskin monotonicity* condition,[11] Chen, Kunimoto, Sun, and Xiong (2021) show that there is a finite mechanism that fully implements f under the solution concept of correlated rationalizability, which implies that their mechanism also fully implements f under correlated equilibrium. According to Proposition 3.2 in Kajii and Morris (1997), their mechanism can also robustly implement f. Our main results in Section 4 construct a different class of finite mechanisms that can robustly implement f even when (f, u_1, u_2) violates Maskin monotonicity*.

Remark 2: Our formulation restricts attention to perturbations where (i) agents' utilities are linear and weakly increasing in their transfers, and (ii) their payoffs do not directly depend on their messages. Both assumptions are commonly made in the mechanism design literature, including Rochet (1987), Bergemann and Morris (2009), Chung and Ely (2007), and many others.

Our assumptions stand in contrast to the literature on robust prediction in games such as Kajii and Morris (1997) and Ui (2001), in which players' actions can directly affect their payoffs. Since we consider a mechanism design setting, agents' message spaces are endogenously chosen by the planner, so these messages have no meaning per se and can be viewed as cheap talk. In many applications, it is also reasonable to assume that all types of the agent weakly prefer more transfers, while agents' preferences over other aspects of the allocation (e.g., convict or acquit a defendant) are more subtle and may not be known to the planner.

A similar perspective is shared by Oury and Tercieux (2012), Chen, Mueller-Frank and Pai (2020), and Chen, Kunimoto and Sun (2020), all of whom use an *interim* approach to study robust

[11] Maskin monotonicity* is introduced by Bergemann, Morris and Tercieux (2011) when they study rationalizable implementation using infinite mechanisms. Jain (2021) shows that Maskin monotonicity* is strictly stronger than Maskin monotonicity. Theorem 1 in Chen, Kunimoto, Sun, and Xiong (2021) shows that Maskin monotonicity* is necessary and sufficient for full implementation under the solution concept of correlated rationalizability.

partial implementation where agents' messages are assumed to be cheap talk. These papers examine whether there exists a mechanism that partially implements a desired social choice function for *all* nearby interim types. By contrast, we take an *ex ante* approach and examine whether the planner can robustly implement a desired social choice function with probability close to one when she knows that agents' beliefs are derived from a common prior and that the agents' payoff functions coincide with those in her model with probability close to one.

4 Main Results

Theorem 1 shows that every f is robustly implementable when the ratio between agents' costs of learning and their marginal utilities from transfers is bounded from above. Theorem 2 shows that even when the aforementioned ratio can be arbitrarily large (e.g., some types of the agents have very high learning costs and/or do not care about transfers), every f is robustly implementable under a generic assumption on the objective state distribution. Theorem 3 extends our robust implementation results to situations where (i) agents tremble with small probability when sending messages, and (ii) agents observe noisy private signals about the state after paying their costs.

4.1 Robust Implementation under Bounded Cost

We show that every f is robustly implementable when the ratio between agents' costs of learning and their marginal utilities from transfers is uniformly bounded from above. Recall that $n \equiv |\Theta|$.

Theorem 1. *For every $\overline{c} > 0$ and $f : \Theta \to \Delta(Y)$, there exists a mechanism with n messages for each agent that robustly implements f for all $\overline{c}$-bounded perturbations.*

For illustration purposes, we prove this result as well as Theorems 2 and 3 in an example where $u_i(\theta, y) = 0$ for $i \in \{1, 2\}$ and $c_1 = c_2 = c$, i.e., each normal type's payoff equals his transfer minus his cost of learning and the normal types of both agents face the same cost c. Extending our proof to general (u_1, u_2) and heterogenous costs is straightforward, which we relegate to Appendix A.

Status Quo Rule with Ascending Transfers: Let us write the state space as $\Theta \equiv \{\theta^1, ..., \theta^n\}$. Each agent's message space is given by $M_1 = M_2 \equiv M \equiv \{1, 2, ..., n\}$. The outcome function is

$$g(m_1, m_2) = \begin{cases} f(\theta^{m_1}) & \text{if } m_1 = m_2 \\ f(\theta^1) & \text{otherwise.} \end{cases} \tag{4.1}$$

The transfer function for agent $i \in \{1,2\}$ is

$$t_i(m_i, m_{-i}) = \begin{cases} R^j & \text{if } m_1 = m_2 = j \\ 0 & \text{otherwise,} \end{cases} \tag{4.2}$$

where $R^n, ..., R^1 > 0$ and $R^j > R^1 + \frac{2\overline{c}}{q(\theta^j)}$ for every $j \geq 2$.

In the unperturbed game induced by our mechanism, an agent's pure strategy is an n-dimensional vector $(m^1, ..., m^n)$ where $m^j \in M$ is the message he sends when the state is θ^j. In order to capture agents' decisions to obtain information, each agent pays a penalty c when he chooses a non-constant vector. Let $\Sigma \equiv \{1, 2, ..., n\}^n$ be the set of pure strategies. An agent is *truthful* if he uses strategy $(1, 2, ..., n)$, that is, he reports the index of the realized state. Our proof consists of three steps.

Step 1: Restricted Game without Perturbation We start from examining a restricted game *without* any perturbation where agents are only allowed to use strategies in $\Delta(\Sigma^*)$, where $\Sigma^* \subset \Sigma$ is a subset of pure strategies defined as:

$$\Sigma^* \equiv \Big\{(m^1, ..., m^n) \in \Sigma \text{ such that } m^j \in \{1, j\} \text{ for every } j \geq 1\Big\}. \tag{4.3}$$

In this restricted game, each agent is only allowed to send the status quo message 1 or truthfully report the state. For example, when $n = 2$, $\Sigma^* = \{(1,1), (1,2)\}$ while $\Sigma = \{(1,1), (1,2), (2,1), (2,2)\}$.

Lemma 1. *In the restricted game without perturbation, there exists $\gamma < \frac{1}{2}$ such that both agents being truthful is a γ-dominant equilibrium.*

Proof. Conditional on $\theta = \theta^j$:

- If agent 1 sends message j, his expected transfer equals $\Pr(m_2 = j|\theta^j)R^j$.
- If agent 1 sends message 1, his expected transfer equals $\Pr(m_2 = 1|\theta^j)R^1$.

If agent 2's strategy belongs to $\Delta(\Sigma^*)$ and agent 2 is truthful with probability at least $\frac{1}{2}$, we have $\Pr(m_2 = j|\theta^j) \geq \frac{1}{2}$, so $\Pr(m_2 = j|\theta^j)R^j \geq \Pr(m_2 = 1|\theta^j)R^1$ given the condition that $R^j > R^1$. Since $R^j > R^1 + \frac{2\overline{c}}{q(\theta^j)}$, agent 1 strictly prefers to send message j to any $m_1 \leq 1$ in state θ^j as long as his cost of observing θ is no more than $\overline{c}$. Since agent 1 has a strict incentive when he believes that agent 2 is truthful with probability at least $\frac{1}{2}$, there exists $\gamma < \frac{1}{2}$ such that agent 1 strictly prefers to use strategy $(1, 2, ..., n)$ to any other pure strategy in Σ^* when agent 2's strategy belongs to $\Delta(\Sigma^*)$ and agent 2 is truthful with probability at least γ. □

Step 2: Restricted Game with Perturbation For any perturbation $\mathcal{G}$, consider a *restricted perturbed game* where type $Q_i(\omega)$ of agent i's payoff function is given by $\widetilde{u}_i(\omega,\theta,y)+\widetilde{b}_i(\omega)t_i-\widetilde{c}_i(\omega)\chi_i$ and all types of both agents are only allowed to use strategies in $\Delta(\Sigma^*)$. Since both agents being truthful is a γ-dominant equilibrium in the unperturbed restricted game for some $\gamma<\frac{1}{2}$, the critical path lemma in Kajii and Morris (1997) implies the following conclusion.

Lemma 2. *For every $\varepsilon>0$, there exists $\eta>0$, such that for every η-perturbation $\mathcal{G}$, there exists an equilibrium $\sigma(\mathcal{G})$ when the environment is perturbed by $\mathcal{G}$ and all types of both agents are only allowed to use strategies in $\Delta(\Sigma^*)$ such that in this equilibrium, the probability with which both agents are truthful is more than $1-\varepsilon$.*

Since $g(j,j)=f(\theta^j)$ for every $j\in\{1,2,...,n\}$, f is implemented with probability more than $1-\varepsilon$ if both agents behave according to $\sigma(\mathcal{G})$. What remains to be verified is that $\sigma(\mathcal{G})$ remains an equilibrium when agents can use any strategy in $\Delta(\Sigma)$.

Step 3: Unrestricted Game with Perturbation We show that for every $\mathcal{G}$, $\sigma(\mathcal{G})$ remains an equilibrium under perturbation $\mathcal{G}$ when agents can use any strategy in $\Delta(\Sigma)$.

Suppose by way of contradiction that there exists a type $Q_1(\omega)$ of agent 1 who strictly prefers strategy $(m^1,...,m^n)\notin\Sigma^*$ to any strategy in Σ^* when agent 2 behaves according to $\sigma(\mathcal{G})$. Let us define a new strategy $(m^1_*,...,m^n_*)$ where

$$m^j_*\equiv\begin{cases} m^j & \text{if } m^j\in\{1,j\} \\ 1 & \text{if } m^j\notin\{1,j\}\end{cases}\quad\text{for every } j\in\{1,2,...,n\}.$$

By construction, $(m^1_*,...,m^n_*)\in\Sigma^*$. We compare type $Q_1(\omega)$'s expected payoff from $(m^1,...,m^n)$ with his expected payoff from $(m^1_*,...,m^n_*)$. First, $(m^1,...,m^n)$ and $(m^1_*,...,m^n_*)$ lead to the same joint distribution of (θ,y) given that agent 2's strategy belongs to $\Delta(\Sigma^*)$, since conditional on $\theta=\theta^j$, agent 2 either sends either 1 or j, so the implemented outcome is $f(\theta^1)$ when agent 1's message is neither 1 nor j. Second, $(m^1_*,...,m^n_*)$ leads to weakly greater transfers conditional on each state, since the transfer is 0 when agent 1 sends message $m^j\notin\{1,j\}$ in state θ^j given that agent 2's message belongs to $\{1,j\}$. Third, $(m^1_*,...,m^n_*)$ leads to a strictly greater transfer compared to $(m^1,...,m^n)$ when $(m^1_*,...,m^n_*)$ requires strictly greater learning cost. To see this, note that $(m^1_*,...,m^n_*)$ requires a greater cost only if $m^1=...=m^n$ and, since $(m^1,...,m^n)\notin\Sigma^*$, it must be that $m^1=...=m^n\geq 2$. As a result, $(m^1_*,...,m^n_*)$ leads to strictly greater transfer

conditional on state θ^1, and the expected increase in transfer is at least $R^1 q(\theta^1)$, which is strictly greater than $\bar{c}$, the maximal ratio between the costs of learning and the marginal utilities from transfers. This suggests that every type of agent 1 prefers $(m_*^1, ..., m_*^n)$ to $(m^1, ..., m^n)$, which leads to a contradiction. Hence, $\sigma(\mathcal{G})$ remains an equilibrium when agents can use any strategy in $\Delta(\Sigma)$.

4.2 Robust Implementation under Generic State Distribution

We show that as long as the objective state distribution $q \in \Delta(\Theta)$ satisfies a generic condition, every f is robustly implementable even when we allow for perturbations where the ratio between agents' costs of learning and their marginal utilities from transfers is unbounded.

Definition 1. *The objective state distribution $q \in \Delta(\Theta)$ is generic if there exists $\theta^* \in \Theta$ such that $q(\theta^*) > q(\theta')$ for every $\theta' \neq \theta^*$.*

For example, when there are two states, our generic condition rules out q that assigns probability $\frac{1}{2}$ to each state but allows for any other full support distribution.

Theorem 2. *Suppose q is generic. For every social choice function $f : \Theta \to \Delta(Y)$, there exists a mechanism with $2n - 1$ messages for each agent that robustly implements f.*

Augmented Status Quo Rule with Ascending Transfers: When q is generic and has full support, we can write $\Theta \equiv \{\theta^1, ..., \theta^n\}$ where $q(\theta^1) > q(\theta^2) \geq q(\theta^3) \geq ... \geq q(\theta^n) > 0$.

Consider a mechanism where each agent's message space is given by $M_1 = M_2 = \{-n, ..., -2\} \cup \{1\} \cup \{2, ..., n\}$. The outcome function is

$$g(m_1, m_2) = \begin{cases} f(\theta^{|m_1|}) & \text{if } |m_1| = |m_2| \\ f(\theta^1) & \text{otherwise.} \end{cases} \tag{4.4}$$

The transfer function for agent $i \in \{1, 2\}$ is

$$t_i(m_i, m_{-i}) = \begin{cases} R^j & \text{if } m_1 = m_2 = j \geq 1 \\ R^0 & \text{if } m_1, m_2 \leq 1 \text{ but } (m_1, m_2) \neq (1, 1) \\ 0 & \text{otherwise,} \end{cases} \tag{4.5}$$

where $R^n > R^{n-1} > ... > R^2 > R^1 > R^0 > 0$ satisfy the following inequalities:

$$R^1 > R^0 + \frac{2c}{q(\theta^1)} \quad \text{and} \quad R^j > R^1 + \frac{2c}{q(\theta^j)} \text{ for every } j \geq 2, \tag{4.6}$$

and

$$\frac{R^0}{R^n} \geq \frac{q(\theta^2)}{q(\theta^1)}. \tag{4.7}$$

When q is generic, there exist $R^n, ..., R^1, R^0$ that satisfy both (4.6) and (4.7). When there are two states, we have displayed our mechanism in Section 2. When there are three states, our mechanism is:

g	-3	-2	1	2	3
-3	$f(\theta^3)$	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^3)$
-2	$f(\theta^1)$	$f(\theta^2)$	$f(\theta^1)$	$f(\theta^2)$	$f(\theta^1)$
1	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^1)$
2	$f(\theta^1)$	$f(\theta^2)$	$f(\theta^1)$	$f(\theta^2)$	$f(\theta^1)$
3	$f(\theta^3)$	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^3)$

t_1, t_2	-3	-2	1	2	3
-3	R^0, R^0	R^0, R^0	R^0, R^0	$0, 0$	$0, 0$
-2	R^0, R^0	R^0, R^0	R^0, R^0	$0, 0$	$0, 0$
1	R^0, R^0	R^0, R^0	R^1, R^1	$0, 0$	$0, 0$
2	$0, 0$	$0, 0$	$0, 0$	R^2, R^2	$0, 0$
3	$0, 0$	$0, 0$	$0, 0$	$0, 0$	R^3, R^3

An agent's pure strategy in the unperturbed environment is $(m^1, ..., m^n)$ where $m^j \in M$ represents the message he sends when the state is θ^j. Let $\Sigma \equiv \{-n, ..., -2, 1, 2, ..., n\}^n$ be the set of pure strategies. An agent is *truthful* if he uses strategy $(1, 2, ..., n)$, that is, he reports the index of the realized state. Our proof proceeds in three steps, which is similar to that of Theorem 1.

Step 1: Restricted Game without Perturbation We start from examining a game *without* any perturbation where agents are only allowed to use strategies in $\Delta(\Sigma^*)$, where

$$\Sigma^* \equiv \Big\{(m^1, ..., m^n) \in \Sigma \text{ such that } m^j \in \{-n, ..., -2, 1\} \cup \{j\} \text{ for every } j \geq 1\Big\}. \tag{4.8}$$

In words, each agent is only allowed to send negative messages, the status quo message 1, or a message that coincides with the index of the realized state. For example, when $n = 2$, $\Sigma^* = \{(-2, -2), (-2, 1), (-2, 2), (1, -2), (1, 1), (1, 2)\}$ while $\Sigma = \Sigma^* \bigcup \{(2, -2), (2, 1), (2, 2)\}$.

We show that there exists $\gamma < \frac{1}{2}$ such that both agents being truthful is a γ-dominant equilibrium in the restricted game without perturbation. Suppose agent 1 believes that agent 2 plays $(1, 2, ..., n)$ with probability at least $\frac{1}{2}$ and that agent 2's strategy belongs to $\Delta(\Sigma^*)$.

- For every $j \in \{2, 3, ..., n\}$, conditional on $\theta = \theta^j$. If agent 1 sends message j, his expected transfer equals $\Pr(m_2 = j|\theta^j)R^j$, which is at least $\frac{R^j}{2}$ given that agent 2 is truthful with probability at least $\frac{1}{2}$. If agent 1 sends any $m_1 \leq 1$, his expected transfer is no more than $\Pr(m_2 \neq j|\theta^j)R^1$, which is at most $\frac{R^1}{2}$. Since $R^j > R^1 + \frac{2c}{q(\theta^j)}$, agent 1 strictly prefers message j to any $m_1 \leq 1$ in state θ^j even taking into account the cost c of observing θ.

- Conditional on $\theta = \theta^1$. If agent 1 sends message 1, his expected transfer is $\Pr(m_2 = 1|\theta^1)R^1 + \Pr(m_2 \leq -2|\theta^1)R^0$, which is at least $\frac{R^1+R^0}{2}$ given that $\Pr(m_2 = 1|\theta^1) \geq \frac{1}{2}$. If agent 1 sends any negative message, he receives transfer R^0. Since $R^1 > R^0 + \frac{2c}{q(\theta^1)}$, agent 1 strictly prefers 1 to any negative message in state θ^1 even taking into account the cost of observing θ.

Since agent 1's incentives are strict when he believes that agent 2 is truthful with probability at least $\frac{1}{2}$, there exists $\gamma < \frac{1}{2}$ such that agent 1 strictly prefers $(1, 2, ..., n)$ to other strategies in Σ^* when agent 2's strategy belongs to $\Delta(\Sigma^*)$ and is truthful with probability at least γ.

Step 2: Restricted Game with Perturbation For any perturbation $\mathcal{G}$, consider a *restricted perturbed game* where type $Q_i(\omega)$ of agent i's payoff is $\widetilde{u}_i(\omega, \theta, y) + \widetilde{b}_i(\omega)t_i - \widetilde{c}_i(\omega)\chi_i$ and all types of both agents are only allowed to use strategies in $\Delta(\Sigma^*)$.

Since both agents being truthful is a γ-dominant equilibrium in the unperturbed restricted game for some γ less than $\frac{1}{2}$, the critical path lemma in Kajii and Morris (1997) implies that for every $\varepsilon > 0$, there exists $\eta > 0$, such that for every η-perturbation $\mathcal{G}$, there exists an equilibrium $\sigma(\mathcal{G})$ in the perturbed game where all types of both agents are only allowed to use strategies in $\Delta(\Sigma^*)$ such that the probability with which both agents are truthful under $\sigma(\mathcal{G})$ is at least $1 - \varepsilon$.

Step 3: Unrestricted Game with Perturbation We show that when q is generic and $\{R^n, ..., R^1, R^0\}$ satisfy (4.6) and (4.7), for every perturbation $\mathcal{G}$, every equilibrium $\sigma(\mathcal{G})$ in the restricted perturbed game remains an equilibrium when agents can use any strategy in $\Delta(\Sigma)$.

For this purpose, we only need to show that for every pure strategy that does not belong to Σ^*, there exists a pure strategy that belongs to Σ^* such that every type of agent 1 weakly prefers the latter to the former when he believes that agent 2 plays according to $\sigma(\mathcal{G})$. We consider two cases.

First, for every $(m^1, ..., m^n) \notin \Sigma^*$ that is non-constant, let $(m^1_*, ..., m^n_*)$ be such that

$$m^j_* \equiv \begin{cases} m^j & \text{if } m^j \in \{-n, ..., -2\} \cup \{1, j\} \\ -m^j & \text{if } m^j \notin \{-n, ..., -2\} \cup \{1, j\} \end{cases} \quad \text{for every } j \in \{1, 2, ..., n\}. \tag{4.9}$$

By construction, $(m^1_*, ..., m^n_*) \in \Sigma^*$, and moreover, $(m^1_*, ..., m^n_*)$ does not increase the cost of obtaining information compared to $(m^1, ..., m^n)$. Our construction of $g(m_1, m_2)$ ensures that $(m^1_*, ..., m^n_*)$ and $(m^1, ..., m^n)$ induce the same distribution of (θ, y) regardless of agent 2's message. Regardless of agent 1's type, as long as he believes that agent 2 behaves according to $\sigma(\mathcal{G})$, which implies that agent 2's strategy belongs to $\Delta(\Sigma^*)$, agent 1 receives weakly greater transfer from $(m^1_*, ..., m^n_*)$ compared to $(m^1, ..., m^n)$ since sending $m^j \notin \{-n, ..., -2\} \cup \{1, j\}$ leads to a transfer of 0 in state θ^j when agent 2's message belongs to $\{-n, ..., -2\} \cup \{1, j\}$.

Second, for every $(m^1, ..., m^n) \notin \Sigma^*$ that is constant, there exists $k \in \{2, 3, ..., n\}$ such that $(m^1, ..., m^n) = (k, ..., k)$. Compare any given type of agent 1's payoff from using strategy $(k, ..., k)$ and from using strategy $(-k, ..., -k)$. Our construction of $g(m_1, m_2)$ implies that $(k, ..., k)$ and $(-k, ..., -k)$ lead to the same distribution over (θ, y) and neither strategy requires agent 1 to learn θ. The expected transfer agent 1 receives is $\Pr(m_2 = k)R^k$ if he uses strategy $(k, ..., k)$ and is $\Pr(m_2 \leq 1)R^0$ if he uses strategy $(-k, ..., -k)$. When every type of agent 2's strategy belongs to $\Delta(\Sigma^*)$, $\Pr(m_2 \leq 1) \geq q(\theta^1)$ and $\Pr(m_2 = k) \leq q(\theta^k)$. Condition (4.7) implies that $\Pr(m_2 = k)R^k \leq \Pr(m_2 \leq 1)R^0$. Since θ and ω are independent, we know that conditional on every $\omega' \in Q_1(\omega)$, agent 1's expected transfer is weakly greater under $(-k, ..., -k)$ compared to that under $(k, ..., k)$.

4.3 Robustness to Trembles and Noisy Information

Motivated by the observation that our proofs of Theorems 1 and 2 construct mechanisms and equilibria in those mechanisms where no type of any agent uses strategies that do not belong to $\Delta(\Sigma^*)$, one may wonder whether our results are robust when agents tremble with small probability or when agents cannot perfectly learn θ even after paying the costs (so that the two agents' private signals about θ may not be perfectly correlated).

In this section, we extend our robust implementation result to the aforementioned situations, i.e., agents tremble with small probability when sending messages and agents' signals about the state are noisy. Our proof constructs a new mechanism that shares the same outcome function as

the *Augmented Status Quo Rule with Ascending Transfers*, but has a different transfer function.

Trembles: For any mechanism $\mathcal{M}$, suppose for every $i \in \{1,2\}$, when agent i intends to send message $m_i \in M_i$, the planner receives m_i with probability $1-\tau$ and receives a message that is drawn according to $F_i^{M_i} \in \Delta(M_i)$ with probability τ.

Throughout this section, we distinguish between an agent's *intended message* and his *realized message*. We write F_i instead of $F_i^{M_i}$ in order to simplify notation.

Noisy Information: Suppose $q \in \Delta(\Theta)$ is generic. Let $\Theta \equiv \{\theta^1, ..., \theta^n\}$ such that $q(\theta^1) > q(\theta^2) \geq ... \geq q(\theta^n) > 0$. For every $i \in \{1,2\}$, let $S_i \equiv \{s_i^1, ..., s_i^{|S_i|}\}$ be agent i's signal space, with $|S_i| \geq n$. Note that $|S_i|$ can be any finite number, i.e., there is no known upper bound on the number of signal realizations. Let $\pi \in \Delta(\Theta \times S_1 \times S_2)$ be the joint distribution of the state and agents' private signals. We say that π is of size $\tau > 0$ if

(a) The marginal distribution of π on Θ is $q \in \Delta(\Theta)$.

(b) For every $i \in \{1,2\}$, there exists a mapping $h_i : S_i \to \{1,2,...,n\}$ such that

$$\pi\Big(h_{-i}(s_{-i}) = h_i(s_i)\Big|s_i\Big) \geq 1-\tau \text{ for every } s_i \in S_i, \tag{4.10}$$

and

$$\sum_{j=1}^{n} \sum_{s_i \in \{h_i(s_i)=j\}} \pi(\theta^j, s_i) \geq 1-\tau. \tag{4.11}$$

Our first requirement says that the marginal distribution on θ equals q. Our second requirement is reminiscent of Chung and Ely (2003) and Sugaya and Takahashi (2013), in which every signal realization is linked to a particular state, given by the mapping h_i. One can think about h_i as endowing each signal realization with a *meaning* (each meaning corresponds to a state). The mappings from signal realizations to their meanings satisfy (i) every agent believes that the other agent receives a signal with the same meaning with probability close to 1, and (ii) the meaning of each agent's signal realization coincides with the state with ex ante probability close to 1.

The planner does not know the perturbation $\mathcal{G}$ as well as $\{\tau, F_1, F_2, \pi\}$. She would like to design a mechanism that can approximately implement the desired social choice function for all small enough perturbations, small enough trembles, and small enough noise in agents' signals about the state. Agents know the mechanism $\mathcal{M}$, the perturbation $\mathcal{G}$, their respective information about

ω under $\mathcal{G}$, as well as $\{\tau, F_1, F_2, \pi\}$ before deciding whether to learn θ and which messages they intend to send. Then the planner observes the *realized messages* not the *intended messages*, and implements an outcome and makes transfers according to the mechanism she committed to.

Theorem 3. *Suppose q is generic. For every $f : \Theta \to \Delta(Y)$, there exists a mechanism with $2n-1$ messages for each agent, such that for every $\varepsilon > 0$, there exist $\eta > 0$ and $\overline{\tau} > 0$ such that for every $\tau < \overline{\tau}$, every (F_1, F_2), every π that is of size $\overline{\tau}$, and every η-perturbation $\mathcal{G}$, there exists an equilibrium $\sigma(\mathcal{G})$ such that* $\max_{\theta\in\Theta} ||g_{\sigma(\mathcal{G})}(\theta) - f(\theta)||_{TV} < \varepsilon$.

Similar to our proof of Theorem 2, we rank the states $\theta^1, ..., \theta^n$ according to their ex ante probabilities, i.e., $q(\theta^1) > q(\theta^2) \geq ... \geq q(\theta^n)$, where the first strict inequality comes from our generic assumption. We consider a mechanism where each agent's message space is given by $M \equiv \{-n, ..., -2\} \cup \{1\} \cup \{2, 3, ..., n\}$. Agent i's pure strategy is an $|S_i|$-dimensional vector $(m^1, ..., m^{|S_i|})$ where $m^k \in M$ represents agent i's *intended message* when his private signal is $s_i = s_i^k$. That being said, conditional on $s_i = s_i^k$, agent i's *realized message* is m^k with probability $1-\tau$ and is randomly drawn according to $F_i \in \Delta(M_i)$ with probability τ. Let

$$\Sigma_i^* \equiv \Big\{(m^1, ..., m^{|S_i|}) \in \Sigma \text{ such that for every } k,\ m^k \in \{-n, ..., -2, 1\} \cup \{j\} \text{ when } h_i(m^k) = j\Big\}. \tag{4.12}$$

Intuitively, Σ_i^* is the subset of agent i's pure strategies such that conditional on every realization of his private signal, he either sends a negative message, or the status quo message 1, or the true meaning of his private signal. Agent i *intends to be truthful* if he sends the true meaning of his private signal for every $s_i \in S_i$.

When there are two states, our augmented status quo rule can robustly implement f when agents' tremble and their signals about the state are noisy. When there are three or more states, our augmented status quo rule cannot robustly implement f as can be illustrated by the following example. Suppose there are three states $\{\theta^1, \theta^2, \theta^3\}$ and three private signals $\{s_i^1, s_i^2, s_i^3\}$ for each agent $i \in \{1, 2\}$, where $h_i(s_i^j) = j$ for every $i \in \{1, 2\}$ and $j \in \{1, 2, 3\}$. For simplicity, let us also assume that each agent's private signal perfectly reveals the state. Suppose agent 1 observes $s_1 = s_1^3$ and he believes that every type of agent 2's strategy belongs to $\Delta(\Sigma_2^*)$, let us compare agent 1's expected transfer when he intends to send message -2 and when he intends to send message 2. When agent 1's realized message is 2, his expected transfer under our augmented status quo rule is $\Pr(m_2 = 2|\theta = \theta^3)R^2$. When agent 1's realized message is -2, his expected transfer under our augmented status quo rule is $\Pr(m_2 \leq 1|\theta = \theta^3)R^0$. If the trembling probability τ is 0, then

$\Pr(m_2 \leq 1|\theta = \theta^3)R^0 \geq \Pr(m_2 = 2|\theta = \theta^3)R^2$ when agent 2's strategy belongs to $\Delta(\Sigma^*)$. If $\tau > 0$ and agent 2 intends to send message 3 with probability 1 in state θ^3, then $\Pr(m_2 = 2|\theta = \theta^3)R^2 = \tau F_2(2)R^2$ can be strictly greater than $\Pr(m_2 \leq 1|\theta = \theta^3)R^0 = \tau F_2(m_2 \leq 1)R^0$.

We present a new mechanism called the *Modified Status Quo Rule with Ascending Transfers* which solves the above problem. Each agent has $2n-1$ messages with their message space given by $M \equiv \{-n, ..., -2\} \cup \{1\} \cup \{2, ..., n\}$. The outcome function is

$$g(m_1, m_2) = \begin{cases} f(\theta^{|m_1|}) & \text{if } |m_1| = |m_2| \\ f(\theta^1) & \text{otherwise} \end{cases}$$

The transfer functions are

$$t_1(m_1, m_2) = \begin{cases} R^j & \text{if } m_1 = m_2 = j \geq 1 \\ R^0 & \text{if } m_1 \leq 1 \text{ but } (m_1, m_2) \neq (1, 1) \\ R^0 - x & \text{if } m_1 \geq 2 \text{ and } m_2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$t_2(m_1, m_2) = \begin{cases} R^j & \text{if } m_1 = m_2 = j \geq 1 \\ R^0 & \text{if } m_2 \leq 1 \text{ but } (m_1, m_2) \neq (1, 1) \\ R^0 - x & \text{if } m_2 \geq 2 \text{ and } m_1 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

where $R^n, ..., R^0 > x > \frac{c}{q(\theta^n)}$ satisfy

$$R^1 - R^0 > \frac{4c}{q(\theta^1)}, \quad R^j - R^1 - x > \frac{2c}{q(\theta^j)} \text{ for every } j \in \{2, 3, ..., n\}, \tag{4.13}$$

and

$$\frac{x}{R^j - R^0} \geq \frac{q(\theta^2)}{q(\theta^1)}. \tag{4.14}$$

In an example with two states, our new mechanism is given by:

g	-2	1	2
-2	$f(\theta^2)$	$f(\theta^1)$	$f(\theta^2)$
1	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^1)$
2	$f(\theta^2)$	$f(\theta^1)$	$f(\theta^2)$

t_1, t_2	-2	1	2
-2	R^0, R^0	R^0, R^0	$R^0, R^0 - x$
1	R^0, R^0	R^1, R^1	$0, 0$
2	$R^0 - x, R^0$	$0, 0$	R^2, R^2

When there are three states, our new mechanism is given by:

g	-3	-2	1	2	3
-3	$f(\theta^3)$	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^3)$
-2	$f(\theta^1)$	$f(\theta^2)$	$f(\theta^1)$	$f(\theta^2)$	$f(\theta^1)$
1	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^1)$
2	$f(\theta^1)$	$f(\theta^2)$	$f(\theta^1)$	$f(\theta^2)$	$f(\theta^1)$
3	$f(\theta^3)$	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^1)$	$f(\theta^3)$

t_1, t_2	-3	-2	1	2	3
-3	R^0, R^0	R^0, R^0	R^0, R^0	$R^0, R^0 - x$	$R^0, R^0 - x$
-2	R^0, R^0	R^0, R^0	R^0, R^0	$R^0, R^0 - x$	$R^0, R^0 - x$
1	R^0, R^0	R^0, R^0	R^1, R^1	$R^0, R^0 - x$	$R^0, R^0 - x$
2	$R^0 - x, R^0$	$R^0 - x, R^0$	$R^0 - x, R^0$	R^2, R^2	$0, 0$
3	$R^0 - x, R^0$	$R^0 - x, R^0$	$R^0 - x, R^0$	$0, 0$	R^3, R^3

Intuitively, the outcomes under the augmented status quo rule and the modified status quo rule are the same. The only difference is in the transfer function: By sending the status quo message or any negative message, an agent is guaranteed to receive transfer R^0. When an agent sends a message at least 2, he faces a penalty x if the other agent sends the status quo message or a negative message, and receives zero transfer when the other agent sends a different message of at least 2.

The proof follows similar steps as before. First, there exists $\gamma < \frac{1}{2}$ such that both agents intending to be truthful is a γ-dominant equilibrium in the restricted unperturbed game where agents are only allowed to use strategies in $\Delta(\Sigma_1^*)$ and $\Delta(\Sigma_2^*)$, where the definition of Σ_i^* can be found in (4.12). To see this, suppose agent 2 intends to be truthful with probability at least $\frac{1}{2}$.

- For every $j \geq 2$, conditional on every $s_1 \in S_1$ with $h(s_1) = j$, if agent 1's realized message is j, then he receives an expected transfer of

$$\Pr(m_2 = j|s_1)R^j + \Pr(m_2 \leq 1|s_1)R^0,$$

and if agent 1's realized message is no more than 1, then he receives an expected transfer of

$$\Pr(m_2 = 1|s_1)R^1 + \Pr(m_2 \neq 1|s_1)R^0.$$

Since $\pi(h_2(s_2) = h_1(s_1)|s_1) \geq 1 - \overline{\tau}$ when π is of size $\overline{\tau}$, we have $\Pr(m_2 = j|s_1) \geq \frac{1-\tau}{2}(1 - \overline{\tau})$ and $\Pr(m_2 = 1|s_1) \leq 1 - \frac{1-\tau}{2}(1 - \overline{\tau})$. When $R^j - R^1 - x > \frac{2c}{q(\theta^j)}$, $\overline{\tau}$ is small enough, and

$\tau \leq \overline{\tau}$, we have

$$\Pr(m_2 = j|s_1)R^j + \Pr(m_2 \leq 1|s_1)R^0 > \Pr(m_2 = 1|s_1)R^1 + \Pr(m_2 \neq 1|s_1)R^0.$$

Hence, agent 1 strictly prefers to send message j when he receives signal s_1 such that $h(s_1) = j$.

- When $R^1 - R^0 > \frac{4c}{q(\theta^1)}$, conditional on agent 2's strategy belongs to $\Delta(\Sigma_2^*)$ and agent 2 intends to be truthful with probability at least $\frac{1}{2}$, agent 1 intending to send message 1 when $h(s_1) = 1$ leads to a strictly greater transfer compared to him intending to send any negative message.

The second step uses the critical path lemma. We can show that for every $\varepsilon > 0$, there exists $\eta > 0$ such that for every η-perturbation $\mathcal{G}$, there exists an equilibrium $\sigma(\mathcal{G})$ in the perturbed restricted game where both agents intend to be truthful with probability more than $1 - \frac{\varepsilon}{2}$. Under the outcome function g of our mechanism, if both agents behave according to $\sigma(\mathcal{G})$ and $\overline{\tau}$ is small compared to ε, then the implemented outcome is ε-close to $f(\theta)$ conditional on every θ.

For the third step, for every strategy of agent 1's $(m^1, ..., m^{|S_1|}) \notin \Sigma_1^*$ that is non-constant, let $(m_*^1, ..., m_*^{|S_1|}) \in \Sigma^*$ be defined as:

$$m_*^k \equiv \begin{cases} m^k & \text{if } m^k \in \{-n, ..., -2\} \cup \{1, j\} \text{ and } h_1(s_1^k) = j \\ -m^k & \text{if } m^k \notin \{-n, ..., -2\} \cup \{1, j\} \text{ and } h_1(s_1^k) = j \end{cases} \quad \text{for every } k \in \{1, 2, ..., |S_1|\}.$$

Intuitively, for every signal realization s_1^k, $m_*^k = m^k$ if m^k is no more than 1 or m^k coincides with the meaning of s_1^k; otherwise, $m_*^k = -m^k$. By construction, $(m^1, ..., m^{|S_1|})$ and $(m_*^1, ..., m_*^{|S_1|})$ induce the same joint distribution of (θ, y). Hence, types that are purely outcome driven are indifferent between $(m^1, ..., m^{|S_1|})$ and $(m_*^1, ..., m_*^{|S_1|})$.

For types that are not purely outcome driven, we compare agent 1's expected transfer from $(m^1, ..., m^{|S_1|})$ and from $(m_*^1, ..., m_*^{|S_1|})$. When agent 1's private signal s_1 is such that $h(s_1) = j$, agent 1's expected transfer when his realized message $m \notin \{-n, ..., -2\} \cup \{1, j\}$ is:

$$\Pr(m_2 = m|s_1)R^m + \Pr(m_2 \leq 1|s_1)(R^0 - x). \tag{4.15}$$

His expected transfer when his realized message is $-m$ is R^0. Since $\Pr(m_2 = m|s_1) \leq 2\overline{\tau}$ when agent 2's strategy belongs to $\Delta(\Sigma^*)$, the value of (4.15) is strictly less than R^0 when $\overline{\tau}$ is small. This implies that every type of agent 1 prefers $(m_*^1, ..., m_*^{|S_1|})$ to $(m^1, ..., m^{|S_1|})$.

For every constant vector $(m^1, ..., m^{|S_1|}) \notin \Sigma_1^*$, there exists $k \in \{2, 3, ..., n\}$ such that $(m^1, ..., m^{|S_1|}) = (k, ..., k)$. Compare agent 1's expected transfer (unconditioned on θ, s_1, and s_2) when his realized message is k and when his realized message is $-k$. When his realized message is k, he receives an expected transfer of $\Pr(m_2 = k)R^k + \Pr(m_2 \leq 1)(R^0 - x)$. When his realized message is $-k$, he receives an expected transfer of R^0. When agent 2's strategy belongs to $\Delta(\Sigma_2^*)$,

$$\Pr(m_2 = k) \leq \pi(h_2(s_2) = k) + \Big(1 - \pi(h_2(s_2) = k)\Big)\overline{\tau} \quad \text{and} \quad \Pr(m_2 \leq 1) \geq \pi(h_2(s_2) = 1)(1 - \overline{\tau}).$$

Hence, $\Pr(m_2 = k)R^k + \Pr(m_2 \leq 1)(R^0 - x) \leq R^0$ when $\overline{\tau}$ is small enough. Hence, conditional on agent 2 behaves according to $\sigma(\mathcal{G})$, for every $k \geq 2$, every type of agent 1 prefers $(-k, ..., -k)$ to $(k, k, ...k)$.

5 Stronger Notions of Implementation

First, we show that the planner *cannot* implement any state-contingent social choice function when we allow for perturbations where agents' payoffs do not coincide with those in the unperturbed environment with high probability. Second, we show that the planner cannot fully or virtually implement any state-contingent social choice function when agents' payoff functions do not depend on the state, or when agents' costs of learning are above some cutoff. We also provide a sufficient condition on agents' payoff functions under which full implementation is plausible when the costs of learning are sufficiently small. Throughout this section, we focus on non-constant f.

Definition 2. *Social choice function f is non-constant if there exist θ, θ' such that $f(\theta) \neq f(\theta')$.*

5.1 Impossibility of Global Robust Implementation

Suppose we allow for perturbations where $\widetilde{c}_i(\omega)$ is arbitrarily large and agents' payoffs can be different from those in the unperturbed environment with probability bounded away from zero, it is not surprising that there exists no finite mechanism that can robustly implement non-constant f. This is because for every finite mechanism $\mathcal{M}$, if both agents are types whose costs of obtaining information overwhelm the maximal transfer promised by the planner and their maximal benefits from the implemented outcome, then no agent has any incentive to learn the state, which implies that non-constant social choice function f is not implemented conditional on this event.

We show that even when we only allow for perturbations where the ratio between the costs of

obtaining information and the marginal utilities from transfers is bounded, no finite mechanism can approximately implement any non-constant social choice function if the probability of normal types can be arbitrary.

Proposition 1. *For every $\overline{c} > 0$ and every $f : \Theta \to \Delta(Y)$ that is non-constant, there exists no finite mechanism that can globally implement f for all $\overline{c}$-bounded perturbations.*

Proposition 1 implies the following corollary:

Corollary 1. *For every $f : \Theta \to \Delta(Y)$ that is non-constant, there exists $k(f) > 0$ such that for every finite mechanism $\mathcal{M}$ and every $\eta > 0$, there exists a $\overline{c}$-bounded η-perturbation $\mathcal{G}$, such that for every equilibrium $\sigma(\mathcal{G})$ of the game $(\mathcal{M}, \mathcal{G})$, we have $\max_{\theta\in\Theta} ||g_{\sigma(\mathcal{G})}(\theta) - f(\theta)||_{TV} \geq \eta k(f)$.*

The above corollary implies that when we allow for perturbations where agents' payoff functions do not coincide with those in the unperturbed environment with probability bounded away from zero, there exists such a perturbation under which every equilibrium of the game implements a social choice function that is bounded away from f. The distance between every implemented outcome and f is bounded below by a linear function of the size of perturbation η, with the coefficient depending only on f. For example, if $f(\theta)$ is a pure outcome for every $\theta \in \Theta$, then $k(f)$ equals 1. This result together with our previous results implies that the perturbed environment being *close* to the unperturbed environment is somewhat necessary for robust implementation.

Proof of Proposition 1: For any finite mechanism $\mathcal{M} \equiv \{M_1, M_2, g, t_1, t_2\}$, let

$$X(\mathcal{M}) \equiv \max_{(i,m_1,m_2)\in\{1,2\}\times M_1\times M_2} \Big|t_i(m_1, m_2)\Big|.$$

By definition, $X(\mathcal{M})$ exists and is finite. Since f is non-constant, there exists $\theta^* \in \Theta$ such that

$$f(\theta^*) \notin \underbrace{\text{co}\Big(\{f(\theta)\}_{\theta\in\Theta}\backslash\{f(\theta^*)\}\Big)}_{\equiv\mathcal{Y}}.$$

According to the separating hyperplane theorem, there exists $v : Y \to \mathbb{R}$ such that $v(f(\theta^*)) < \min_{y\in\mathcal{Y}} v(y)$. Hence, there exists $C > 0$ such that $\Big(\min_{y\in\mathcal{Y}} v(y) - v(f(\theta^*))\Big)C > 4X(\mathcal{M})$.

First, consider a perturbation $\mathcal{G}^+$ where $\widetilde{u}_1(\omega, \theta, y) = Cv(y)$ for all $\omega \in \Omega$. If $\mathcal{M}$ implements

$f(\theta^*)$ in state θ^*, then there exists $m_2^* \in \Delta(M_2)$ such that for every $m_1 \in M_1$,

$$\max_{m_1 \in \Delta(M_1)} \{Cv(g(m_1, m_2^*)) + t_1(m_1, m_2^*)\} \leq \underbrace{Cv(f(\theta^*)) + X(\mathcal{M})}_{\text{agent 1's highest payoff when the implemented outcome is } f(\theta^*)}. \tag{5.1}$$

This is because otherwise, agent 1 can secure himself a payoff strictly greater than the right-hand-side of (5.1), which means that he has no incentive to implement $f(\theta^*)$.

Next, consider another perturbation $\mathcal{G}^-$ where $\widetilde{u}_2(\omega, \theta, y) = -Cv(y)$ for all $\omega \in \Omega$. According to (5.1), agent 2's payoff by playing m_2^* is at least

$$\min_{m_1 \in \Delta(M_1)} \{-Cv(g(m_1, m_2^*)) + t_2(m_1, m_2^*)\}. \tag{5.2}$$

Since we have chosen C to satisfy $\Big(\min_{y \in \mathcal{Y}} v(y) - v(f(\theta^*))\Big)C > 4X(\mathcal{M})$ and $X(\mathcal{M}) \geq |t_i(m_1, m_2)|$ for every i and (m_1, m_2), inequalities (5.1) and (5.2) imply that

$$\min_{m_1 \in \Delta(M_1)} \{-Cv(g(m_1, m_2^*)) + t_2(m_1, m_2^*)\} \geq -Cv(f(\theta^*)) - 3X(\mathcal{M}) > \max_{y \in \mathcal{Y}} \{-Cv(y)\} + X(\mathcal{M}). \tag{5.3}$$

For any outcome in $\mathcal{Y}$ to be implemented in any state under perturbation $\mathcal{G}^-$, it must be the case that agent 2's payoff is no more than $\max_{y \in \mathcal{Y}} \{-Cv(y)\} + X(\mathcal{M})$. Since inequality (5.3) implies that agent 2 can secure himself a payoff strictly greater than $\max_{y \in \mathcal{Y}} \{-Cv(y)\} + X(\mathcal{M})$, no outcome in $\mathcal{Y}$ can be implemented in any state, which implies that every mechanism $\mathcal{M}$ that can implement f under perturbation $\mathcal{G}^+$ cannot implement f under perturbation $\mathcal{G}^-$. □

5.2 Full Implementation

Our theorems focus on robust *partial* implementation: The planner's objective is to design a mechanism so that for every small perturbation, there exists *one* equilibrium in the corresponding incomplete information game that induces an outcome close to f. This section examines whether the planner can fully or virtually implement f when agents need to pay costs to learn the state.

We say that f is *fully implementable* if there exists a finite mechanism $\mathcal{M} \equiv \{M_1, M_2, g, t_1, t_2\}$ such that $g_\sigma(\theta) = f(\theta)$ for every $\theta \in \Theta$ and every equilibrium σ under mechanism $\mathcal{M}$. We say that f is *virtually implementable* if for every $\varepsilon > 0$, there exists a finite mechanism $\mathcal{M}$ such that $||g_\sigma(\theta) - f(\theta)||_{TV} \leq \varepsilon$ for every $\theta \in \Theta$ and every equilibrium σ under mechanism $\mathcal{M}$. We say that f is *strongly-robust implementable* if for every $\varepsilon > 0$, there exists $\eta > 0$ such that for every

η-perturbation $\mathcal{G}$, $||g_{\sigma(\mathcal{G})}(\theta) - f(\theta)||_{TV} \leq \varepsilon$ for every $\theta \in \Theta$ and every equilibrium $\sigma(\mathcal{G})$ of $(\mathcal{M}, \mathcal{G})$. We introduce two conditions under which full and virtual implementation is impossible.

Proposition 2. *Suppose f is non-constant.*

1. *If (u_1, u_2) do not depend on θ and $c_1, c_2 > 0$, then f is not virtually implementable.*

2. *For every (u_1, u_2), there exists $\bar{c} > 0$ such that f is not virtually implementable when $c_1, c_2 > \bar{c}$.*

Proof of Proposition 2: When $c_1, c_2 > 0$, and (u_1, u_2) do not depend on θ, there always exists an equilibrium where neither agent pays the strictly positive cost to learn the state. In this equilibrium, the implemented outcome does not depend on the state, which implies that no mechanism can virtually implement any non-constant f.

Next, we show that no mechanism can virtually implement non-constant f when c_1 and c_2 are sufficiently large. For every u_1 and u_2, let

$$X(u_1, u_2) \equiv \max_{i \in \{1,2\}} \Big| \max_{\theta, y} u_i(\theta, y) - \min_{\theta, y} u_i(\theta, y) \Big|.$$

Fix any finite mechanism $\mathcal{M}$, for every $m_2 \in \Delta(M_2)$, let

$$T(m_2) \equiv \max_{m_1 \in M_1} t_1(m_1, m_2).$$

Suppose agent 1 believes that agent 2's message is m_2, the difference between his expected payoff when he learns θ and when he does not learn θ is

$$\mathbb{E}\Big[\max_{m_1 \in M_1} \{u_1(\theta, g(m_1, m_2)) + t_1(m_1, m_2)\} \Big] - \max_{m_1 \in M_1} \mathbb{E}\Big[u_1(\theta, g(m_1, m_2)) + t_1(m_1, m_2) \Big]. \tag{5.4}$$

By definition, if $m_1^* \in \arg\max_{m_1 \in M_1} \mathbb{E}\Big[u_1(\theta, g(m_1, m_2)) + t_1(m_1, m_2) \Big]$, then $t_1(m_1^*, m_2) \geq T(m_2) - X(u_1, u_2)$. This implies that the value of (5.4) is no more than $2X(u_1, u_2)$, and therefore, agent 1 has no incentive to learn θ when $c_1 > 2X(u_1, u_2)$. In addition, when agent 1 believes that agent 2's message is m_2, sending a message that belongs to $\arg\max_{m_1 \in M_1} \mathbb{E}\Big[u_1(\theta, g(m_1, m_2)) + t_1(m_1, m_2) \Big]$ regardless of the state is one of agent 1's best replies.

Similarly, suppose $c_2 > 2X(u_1, u_2)$. For every $m_1 \in \Delta(M_2)$, when agent 2 believes that agent 1's message is m_1, sending a message that belongs to $\arg\max_{m_2 \in M_2} \mathbb{E}\Big[u_2(\theta, g(m_1, m_2)) + t_2(m_1, m_2) \Big]$ regardless of the state is one of agent 2's best replies. For every $\mathcal{M}$, consider an auxiliary two-player

normal form game where agent $i \in \{1, 2\}$ has a finite set of pure strategies M_i and his payoff is

$$\mathbb{E}\Big[u_i(\theta, g(m_1, m_2)) + t_i(m_1, m_2)\Big]$$

when he uses strategy m_i and his opponent uses strategy m_{-i}. Since this is a finite game, a Nash equilibrium $(m_1, m_2) \in \Delta(M_1) \times \Delta(M_2)$ exists. By construction, agent 1 sending m_1 regardless of θ and agent 2 sending m_2 regardless of θ is an equilibrium under mechanism $\mathcal{M}$. In this equilibrium, the implemented outcome is the same for all θ, which means that it cannot fully implement f when f is non-constant. □

Proposition 2 and its proof imply that when agents' costs of learning θ are above some cutoff that depends only on u_1 and u_2, no matter which finite mechanism the planner commits to, there always exists an equilibrium where neither agent learns the state. This precludes the possibility of full implementation.

This conclusion is somewhat surprising since the planner can use transfers that are arbitrarily large relative to agents' costs of learning. Intuitively, suppose agent 1's message does not depend on θ and let us consider agent 2's incentive to learn θ. Since agents' transfers cannot depend on the realized state and depend only on the message profile, the only incentive for agent 2 to learn θ is to induce a more favorable state-contingent outcome. Hence, agent 2's benefit from learning depends only on u_2. When agent 2's cost of learning outweighs this benefit but is less relative to the transfers promised by the planner, he has no incentive to learn the state provided that agent 1's message does not depend on the state, which gives rise to equilibria where no agent learns the state and the implemented outcome is the same regardless of the state.

By contrast, our main results focus on robust partial implementation. In the equilibria we construct, both agents' messages depend on the state with probability close to 1. Each agent's benefit from learning not only comes from inducing a better state-contingent outcome (i.e., increasing the value of $u_i(\theta, y)$), but may also come from his incentive to receive a higher transfer. As a result, the planner can robustly implement any state-contingent f for any c_1 and c_2.

Nevertheless, f is fully implementable when one of the agent's payoff function satisfies a strict version of Rochet (1987)'s cyclical monotonicity condition and that agent's cost of learning is below some cutoff. Formally, for every $f : \Theta \to \Delta(Y)$ and $u_i : \Theta \times Y \to \mathbb{R}$, we say that u_i and f satisfy

strict cyclical monotonicity if for every permutation τ of Θ,

$$\sum_{\theta\in\Theta} u_i(\theta, f(\theta)) \geq \sum_{\theta\in\Theta} u_i\Big(\theta, f(\tau(\theta))\Big). \tag{5.5}$$

and for every permutation τ such that $f(\tau(\theta)) \neq f(\theta)$ for some $\theta \in \Theta$,

$$\sum_{\theta\in\Theta} u_i(\theta, f(\theta)) > \sum_{\theta\in\Theta} u_i\Big(\theta, f(\tau(\theta))\Big). \tag{5.6}$$

Condition (5.5) is the cyclical monotonicity condition in Rochet (1987). Condition (5.6) is novel. When f is constant, condition (5.6) has no bite. When f depends nontrivially on θ, condition (5.6) is violated when u_i does not depend on θ, such as the payoff functions in our leading example.

Proposition 3. *If there exists $i \in \{1, 2\}$ such that u_i and f satisfy strict cyclical monotonicity, then there exists $\overline{c} > 0$ such that when $c_i \leq \overline{c}$, there exists a finite mechanism $\mathcal{M}$ that can fully implement f and can strongly-robust implement f.*

Proof of Proposition 3: If f is constant, then fully implementing f is straightforward. The rest of the proof focuses on the case where f is non-constant. Consider a mechanism where $M_i = \Theta$, $M_{-i} = \{1\}$, $g(m_i, m_{-i}) = f(m_i)$, $t_i(m_1, m_2)$ depends only on m_1, and $t_{-i}(m_1, m_2) = 0$. Since f and u_i satisfy strict cyclical monotonicity, there exists $t_i : \Theta \to \mathbb{R}$ such that

1. $t_i(\theta) = t_i(\theta')$ for every $\theta, \theta' \in \Theta$ such that $f(\theta) = f(\theta')$,
2. $u_i(\theta, f(\theta)) + t_i(\theta) > u_i(\theta, f(\theta')) + t_i(\theta')$ for every $\theta, \theta' \in \Theta$ such that $f(\theta) \neq f(\theta')$.

Under such a mechanism, agent i chooses an outcome in $\{f(\theta)\}_{\theta\in\Theta}$ and receive an additional reward $t_i(\theta)$ for implementing $f(\theta)$. Agent i has a strict incentive to choose $f(\theta)$ in state θ, so he has a strict incentive to learn the state when c is small enough. Under every η-perturbation $\mathcal{G}$, every normal type of agent i has a strict incentive to learn the state and to induce outcome $f(\theta)$ in state θ. This completes the proof. □

6 Conclusion

We examine the problem faced by a social planner who wants to robustly implement a state-contingent social choice function when (i) agents need to incur costs to learn the state, (ii) the planner faces uncertainty about agents' costs of obtaining information, their biases over outcomes,

their marginal utilities from transfers, as well as their beliefs and higher-order-beliefs about each other's payoffs. We introduce mechanisms that robustly implement the desired social choice function when the state distribution satisfies a generic assumption, or when the planner knows an upper bound on the ratio between agents' costs of obtaining information and their marginal utilities from transfers. We conclude by discussing the related literature.

Our work contributes to the literature on robust implementation, pioneered by Bergemann and Morris (2005). Our paper is closely related to Oury and Tercieux (2012), who take an interim perspective and require the outcome induced by every nearby type to be close to that induced by the original type. They show that Maskin monotonicity is necessary for robust partial implementation.[12] By contrast, we take an ex ante perspective by requiring that the desired outcome to be implemented with probability close to one in all nearby type spaces. We show that all social choice functions are robustly implementable and our mechanisms are robust to small trembles and noisy information about the state. We propose mechanisms that can robustly elicit costly information when the planner faces uncertainty about agents' preferences over outcomes, costs of obtaining information, as well as beliefs and higher-order beliefs while assuming that agents' information acquisition technologies are known and their signals about the state are highly correlated. This stands in contrast to Carroll (2019) who focuses on robust contracting when the planner faces uncertainty about the agent's information acquisition technology and Aghion, Fudenberg, Holden, Kunimoto and Tercieux (2012) who perturb players' information about the state.

Our work is related to the literature on robust prediction in games, pioneered by Rubinstein (1989). Our notion of robustness builds on the notion of robust equilibrium in Kajii and Morris (1997).[13] Their notion of has been broadly applied to study the robustness of equilibria in potential games (Ui 2001, Morris and Ui 2005) and supermodular games (Oyama and Takahashi 2020). The key difference is that in our model, agents' payoffs in the perturbed game do not directly depend on their messages, which are their actions in our mechanism design setting. The assumption that agents' messages are cheap talk is common in the robust mechanism design literature, such as Bergemann and Morris (2005) and Aghion, Fudenberg, Holden, Kunimoto and Tercieux (2012).

Finally, our work is also related to the large literature on contracting for information acquisition

[12]The literature on full implementation (Maskin 1999) and virtual implementation (Abreu and Matsushima 1992) examine whether the planner can implement or approximately implement the desired social choice function in all equilibria. Unlike our model, they do not consider any perturbations and only require full or virtual implementation in the unperturbed environment. Hence, our result is neither stronger nor weaker than theirs.

[13]Our notion of robustness is also related to the notion of robust refinement in Fudenberg, Kreps and Levine (1988), in which an equilibrium should not be ruled out by a refinement if it is a strict equilibrium in some *nearby* game.

such as Zermeno (2011) and Clark and Reggiani (2021), as well as mechanism design with costly information acquisition such as Crémer and Khalil (1992), Persico (2000), Bergemann and Välimäki (2002), Li (2019), and Larionov, Pham and Yamashita (2021). Those papers study characterize the optimal mechanism under a fixed informational environment. In contrast, we examine whether it is possible to (approximately) implement a desired social choice function in *all* nearby environments.

A notable exception is Carroll (2019), who characterizes the optimal contract for information acquisition when the planner faces uncertainty about the set of information acquisition technologies available to the agent, but can condition the agent's transfer on the realized state. By contrast, the state cannot be verified ex post in our setting, so the planner cannot condition the transfers on the state. Unlike Carroll (2019), the planner in our model faces uncertainty not about agents' information acquisition technologies, but instead, she faces uncertainty about their costs of obtaining information, their biases over outcomes, their marginal utilities from transfers, and their beliefs and higher-order-beliefs about each other's payoffs.

A General Utility Functions

We explain how to generalize our proofs of Theorems 1, 2, and 3 to general $u_1(\theta, y)$ and $u_2(\theta, y)$. We will use Theorem 2 as an example and the proofs of other theorems are similar.

Consider the augmented status quo rule with ascending transfers constructed in Section 4, with the following modification of the conditions on $R^n, ..., R^1, R^0$. Suppose $R^n > R^{n-1} > ... > R^1 > R^0 > 0$ with

$$R^j > R^1 + \frac{2c}{q(\theta^j)} + 2 \max_{i\in\{1,2\}} \Big\{ \max_{y\in Y} u_i(\theta^j, y) - \min_{y\in Y} u_i(\theta^j, y) \Big\}, \tag{A.1}$$

for every $j \in \{2, 3, ..., n\}$,

$$R^1 > R^0 + \frac{2c}{q(\theta^1)} + 2 \max_{i\in\{1,2\}} \Big\{ \max_{y\in Y} u_i(\theta^1, y) - \min_{y\in Y} u_i(\theta^1, y) \Big\}, \tag{A.2}$$

and

$$\frac{R^0}{R^n} \geq \frac{q(\theta^2)}{q(\theta^1)}. \tag{A.3}$$

We modify the first step of our proof in which we show that truthfully revealing the state is a γ-dominant equilibrium for some $\gamma < \frac{1}{2}$.

Recall that each agent has $2n-1$ messages and that we are considering an *unperturbed restricted game* where both agents are only allowed to use strategies that belong to $\Delta(\Sigma^*)$ where Σ^* is defined

as

$$\Sigma^* \equiv \Big\{(m^1, ..., m^n) \in \Sigma \text{ such that } m^j \in \{-n, ..., -2, 1\} \cup \{j\} \text{ for every } j \geq 1\Big\}.$$

We show that in the unperturbed restricted game, both agents using strategy $(1, 2, ..., n)$ is a γ-dominant equilibrium for some $\gamma < \frac{1}{2}$.

- Conditional on $\theta = \theta^j$ for every $j \in \{2, 3, ..., n\}$. If agent 1 sends message j, his expected payoff is at least $\Pr(m_2 = j|\theta^j)R^j + \min_{y \in Y} u_1(\theta^j, y)$, If agent 1 sends any $m_1 \leq 1$, his expected payoff is at most $\Pr(m_2 \neq j|\theta^j)R^1 + \max_{y \in Y} u_1(\theta^j, y)$. When $\Pr(m_2 = j|\theta^j) \geq \frac{1}{2}$ and inequality (A.1) is satisfied, $\Pr(m_2 = j|\theta^j)R^j + \min_{y \in Y} u_1(\theta^j, y) > \Pr(m_2 \neq j|\theta^j)R^1 + \max_{y \in Y} u_1(\theta^j, y)$, and the difference is enough to cover the cost of learning the state.

- Conditional on $\theta = \theta^1$. If agent 1 sends message 1, his expected payoff is $\Pr(m_2 = 1|\theta^1)R^1 + \Pr(m_2 \leq -2|\theta^1)R^0 + \min_{y \in Y} u_1(\theta^1, y)$, which is at least $\frac{R^1 + R^0}{2} + \min_{y \in Y} u_1(\theta^1, y)$ given that $\Pr(m_2 = 1|\theta^1) \geq \frac{1}{2}$. If agent 1 sends any negative message, his expected payoff is at most $R^0 + \max_{y \in Y} u_1(\theta^1, y)$. Inequality (A.2) implies that $\frac{R^1 + R^0}{2} + \min_{y \in Y} u_1(\theta^1, y) > R^0 + \max_{y \in Y} u_1(\theta^1, y)$, and the difference is enough to cover the cost of learning the state.

The second and third steps are not affected by u_1 and u_2, which remain the same as in Section 4.

B Robust Implementation with a Continuum of States

We generalize our robust implementation results to environments with a continuum of states, as long as the state space Θ is compact and both the agents' payoff functions in the unperturbed environment and the desired social choice function f are continuous with respect to θ.

Formally, let Θ be a compact set in some normed vector space, with the endowed norm denoted by $||\cdot||$. Let $q \in \Delta(\Theta)$ be the objective distribution of θ which we assume has full support and has no atom. A social choice function $f : \Theta \to \Delta(Y)$ is *continuous* if for every $\varepsilon > 0$, there exists $\delta > 0$ such that $||f(\theta) - f(\theta')||_{TV} \leq \varepsilon$ for every $||\theta - \theta'|| \leq \delta$.[14]

Agent $i \in \{1, 2\}$ can observe the realization of θ at cost $c_i \in [0, +\infty)$. Agent i's payoff function in the unperturbed environment is $t_i - c_i\chi_i + u_i(\theta, y)$.

We say that $u_i(\theta, y)$ is continuous with respect to θ if for every $y \in Y$ and $\varepsilon > 0$, there exists $\delta > 0$ such that $|u_i(\theta, y) - u_i(\theta', y)| \leq \varepsilon$ for every $||\theta - \theta'|| \leq \delta$. The notion of η-perturbation

[14] Under our definition, f is uniformly continuous with respect to θ. This is without loss of generality since Θ is compact. The same comment applies when we introduce our notion of continuity for agents' payoff functions.

remains the same as in the baseline model, that is, agents' payoff functions coincide with those in the unperturbed environment with probability at least $1-\eta$, and with complementary probability, they can have arbitrary biases over outcomes, arbitrary costs of learning, arbitrary marginal utilities from transfers, and arbitrary beliefs and higher-order beliefs as long as these beliefs can be derived from a common prior belief.

One thing to note is that we do not require agent i's payoff in the perturbed game $\widetilde{u}_i(\omega, \theta, y)$ to be continuous with respect to θ when agent i is not normal at ω.

Corollary 2. *Suppose Θ is compact, q has full support and has no atom, and both f and (u_1, u_2) are continuous with respect to θ. For every $\varepsilon > 0$, there exist $\eta > 0$ and a finite mechanism $\mathcal{M}$ such that for every η-perturbation $\mathcal{G}$, there exists an equilibrium σ under $(\mathcal{M}, \mathcal{G})$ such that $\max_{\theta \in \Theta} ||g_\sigma(\theta) - f(\theta)|| \leq \varepsilon$.*

We explain how to modify the proof of Theorem 2 in order to show Corollary 2. For simplicity, we focus on the case where $u_1(\theta, y) = u_2(\theta, y) = 0$, and we comment on the proof for general $u_1(\theta, y)$ and $u_2(\theta, y)$ in Appendix A. Since the state space Θ is compact and the desired social choice function f is continuous, for every $\varepsilon > 0$, one can construct a finite partition of Θ using the finite cover theorem that satisfies the following three conditions:

1. Every partition element occurs with positive probability under q.
2. There exists a partition element that occurs with strictly higher probability compared to every other partition element.
3. $||f(\theta) - f(\theta')||_{TV} \leq \frac{\varepsilon}{2}$ for every pair θ, θ' that belong to the same partition element.[15]

Fix any partition that satisfies the above requirements. Denote the partition elements by $\{\Theta^1, ..., \Theta^n\}$. For every $j \in \{1, 2, ..., n\}$, let θ^j be an arbitrary element in Θ^j. We introduce a new social choice function $\widetilde{f} : \Theta \to \Delta(Y)$ such that $\widetilde{f}(\theta) = f(\theta^j)$ for every $\theta \in \Theta^j$ and $j \in \{1, 2, ..., n\}$.

Consider the mechanism we constructed in the proof of Theorem 2, in which every agent has $2n-1$ messages. In our new environment with a continuum of states, each agent is asked to report which element of the partition the realized θ belongs to. The proof of Theorem 2 implies that there exists a mechanism $\mathcal{M}$ such that for every η-perturbation $\mathcal{G}$, there exists an equilibrium $\sigma^*(\mathcal{G})$ such that $\max_{\theta \in \Theta} ||g_{\sigma^*(\mathcal{G})}(\theta) - \widetilde{f}(\theta)||_{\text{TV}} < \varepsilon/2$. Since $||\widetilde{f}(\theta) - f(\theta)||_{TV} = ||f(\theta^j) - f(\theta)||_{TV} \leq \varepsilon/2$, the

[15] For general $u_1(\theta, y)$ and $u_2(\theta, y)$ that are continuous with respect to θ, we can find a partition that satisfies the above requirements while also making sure that $|u_i(\theta, y) - u_i(\theta', y)| < \varepsilon/2$ for every $y \in Y$, $i \in \{1, 2\}$, and θ, θ' belonging to the same partition element.

triangular inequality implies that $\max_{\theta\in\Theta} ||g_{\sigma^*(\mathcal{G})}(\theta) - f(\theta)||_{\text{TV}} < \varepsilon$. Hence, the said mechanism robustly implements f.

References

[1] Abreu, D., Matsushima, H. (1992) "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica*, Vol. 60, pp. 993–1008.

[2] Aghion, P., Fudenberg, D., Holden, R., Kunimoto, T., Tercieux, O. (2012) "Subgame-Perfect Implementation Under Information Perturbations," *Quarterly Journal of Economics*, Vol. 127, pp. 1843–1881.

[3] Bergemann, D., Morris, S. (2005) "Robust Mechanism Design," *Econometrica*, Vol. 73, pp. 1771–1813.

[4] Bergemann, D., Morris, S. (2009) "Robust Implementation in Direct Mechanisms," *Review of Economic Studies*, Vol. 76, pp. 1175–1204.

[5] Bergemann, D., Morris, S., Tercieux, O. (2011) "Rationalizable Implementation," *Journal of Economic Theory*, Vol. 146, pp. 1253–1274.

[6] Bergemann, D., Välimäki, J. (2002) "Information Acquisition and Efficient Mechanism Design," *Econometrica*, Vol. 70, pp. 1007–1033.

[7] Chen, Y., Mueller-Frank, M. Pai, M. (2020) "Continuous Implementation with Direct Revelation Mechanisms," Working Paper.

[8] Chen, Y., Kunimoto, T. Sun, Y., Xiong, S. (2021) "Rationalizable Implementation in Finite Mechanisms," *Games and Economic Behavior*, Vol. 129, pp. 181–197.

[9] Chen, Y., Kunimoto, T. Sun, Y. (2020) "Continuous Implementation with Payoff Knowledge," Working Paper.

[10] Chung, K.S, Ely, J. (2003) "Implementation with Near-Complete Information," *Econometrica*, Vol. 71, pp. 857–871.

[11] Chung, K.S., Ely, J. (2007) "Foundations of Dominant-Strategy Mechanisms," *Review of Economic Studies*, Vol. 74, pp. 447–476.

[12] Carroll, G. (2019) "Robust Incentives for Information Acquisition," *Journal of Economic Theory*, Vol. 181, pp. 382–420.

[13] Clark, A., Reggiani, G. (2021) "Contracts for Acquiring Information," arXiv:2103.03911.

[14] Cremer, J,. Khalil, F. (1992) "Gathering Information Before Signing a Contract," *American Economic Review*, Vol. 82, pp. 566-578.

[15] Fudenberg, D,. Kreps, D., Levine, D. (1988) "On the Robustness of Equilibrium Refinements," *Journal of Economic Theory*, Vol. 44, pp. 354-380.

[16] Jackson, M. (1992) "Implementation in Undominated Strategies: A Look at Bounded Mechanisms," *Review of Economic Studies*, Vol. 59, pp. 757–775.

[17] Jain, R. (2021) "Rationalizable Implementation of Social Choice Correspondences," *Games and Economic Behavior*, Vol. 127, pp. 47–66.

[18] Kajii, A., Morris, S. (1997) "The Robustness of Equilibria to Incomplete Information," *Econometrica*, Vol. 65, pp. 1283–1309.

[19] Larionov, D., Pham, H., Yamashita, T. "First Best Implementation with Costly Information Acquisition," Working Paper.

[20] Li, Y. (2019) "Efficient Mechanisms with Information Acquisition," *Journal of Economic Theory*, Vol. 182, pp. 279–328.

[21] Lipnowski, E. and Ravid, D. (2020) "Cheap Talk with Transparent Motives," *Econometrica*, Vol. 88, pp. 1631–1660.

[22] Maskin, E. (1999) "Nash Equilibrium and Welfare Optimality," *Review of Economic Studies*, Vol. 66, pp. 23–38.

[23] Morris, S. and Ui, T. (2005) "Generalized Potentials and Robust Sets of Equilibria," *Journal of Economic Theory*, Vol. 124, pp. 45–78.

[24] Oury, M., Tercieux, O. (2005) "Continuous Implementation," *Econometrica*, Vol. 80, pp. 1605–1637.

[25] Oyama, D., Takahashi, S. (2020) "Generalized Belief Operator and Robustness in Binary-Action Supermodular Games," *Econometrica*, Vol. 88, pp. 693–726.

[26] Persico, N. (2000) "Information Acquisition in Auctions," *Econometrica*, Vol. 68, pp. 135–148.

[27] Rochet, J.C. (1987) "A Necessary and Sufficient Condition for Rationalizability in a Quasi-Linear Context," *Journal of Mathematical Economics*, Vol. 16, pp. 191-200.

[28] Rubinstein, A. (1989) "The Electronic Mail Game: Strategic Behavior Under Almost Common Knowledge," *American Economic Review*, Vol. 79, pp. 385–391.

[29] Strulovici, B. (2021) "Can Socitety Function without Ethical Agents? An Informational Perspective," *Working Paper*, Northwestern University.

[30] Sugaya, T. Takahashi, S. (2013) "Coordination Failure in Repeated Games with Private Monitoring," *Journal of Economic Theory,*, Vol. 148, pp. 1891–1928.

[31] Ui, T. (2001) "Robust Equilibria of Potential Games," *Econometrica*, Vol. 69, pp. 1373–1380.

[32] Zermeno, L. (2011) "A Principal-Expert Model and the Value of Menus," Working Paper, MIT.