Robust Implementation with Costly Information

Harry Pei[*] Bruno Strulovici[‡]

October 23, 2022

Abstract: We design mechanisms that robustly implement any desired social choice function when (i) agents must incur a cost to learn the state of the world, (ii) with small probability, agents' preferences can be arbitrarily different from some baseline known to the social planner, and (iii) the planner does not know agents' beliefs and higher-order beliefs about one another's preferences. The mechanisms we propose have a natural interpretation, and are robust to trembles in agents' reporting strategies, to the introduction of a small amount of noise affecting agents' signals about the state, and to uncertainty concerning the state distribution and agents' prior beliefs about the state. We also establish impossibility results for stronger notions of robust implementation.

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

JEL Codes: D82, D83.

1 Introduction

Theories of robust implementation study whether a state-contingent social choice function, such as one aiming to convict guilty defendants and acquit innocent ones, can be implemented when information about the state must be elicited from agents whose objective may be misaligned with the social choice function, and the planner who designs the mechanism faces uncertainty about agents' preferences and their beliefs and higher-order beliefs about one another's preferences.

Whether a social choice function can be *robustly* implemented depends on the notion of robustness considered. When robustness is required to hold *globally*, in the sense that agents' preferences and beliefs may be arbitrary, Bergemann and Morris (2005) show that a social choice function is robustly implementable only if it is ex post incentive compatible. Oury and Tercieux (2012) consider a less demanding notion of robust implementation, which concerns *local* perturbations of

[*] 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, Nina Fluegel, Dana Foarta, Yingni Guo, Matt Jackson, Navin Kartik, Takashi Kunimoto, Elliot Lipnowski, Meg Meyer, Stephen Morris, Kota Murayama, Kyohei Okumura, David Rodina, Larry Samuelson, Takuo Sugaya, Satoru Takahashi, Olivier Tercieux, Takashi Ui, Siyang Xiong, and Boli Xu for helpful comments. Pei thanks the NSF Grant SES-1947021 for financial support.

agents' preferences and beliefs in an *interim* sense. They require that the desired social choice function be approximately implemented for all profiles of agent types *close* to a given type profile. They show that robustly implementable social choice functions must satisfy Maskin monotonicity (Maskin 1999)—a demanding property that is violated in a number of settings.

1.1 An Ex Ante Notion of Robust Implementation

This paper proposes a new, ex ante notion of robust implementation, and examines whether a given social choice function can be robustly implemented when agents need to incur a cost to learn the state.[1]

Our notion of robust implementation builds on the concept of equilibrium robustness introduced by Kajii and Morris (1997). According to this concept, a Nash equilibrium of a complete information game is robust if it can be approximated by some equilibria in *every* incomplete information game close to the complete information game in the sense that players' payoffs match those of the complete information game with probability close to one. In these incomplete information games, agents can have arbitrary payoffs with small probability and arbitrary beliefs and higher-order beliefs about one another's payoffs as long as these beliefs are consistent with a common prior.

Building on this definition, we propose a *local* and *ex ante* notion of robust implementation. First, we focus on perturbations in which agents' payoffs differ from those of the unperturbed environment with small probability. Second, we require that the desired social choice function be implemented only with probability close to one rather than approximately implemented for all nearby types. Our notion of robust implementation relaxes some restrictive requirements of the global and interim notions and thus can potentially avoid some of the most stringent implications of these notions, as we will show in this paper.

Our notion of robust implementation departs from Kajii and Morris (1997) by imposing a key restriction on the set of perturbations considered by the planner: We focus on perturbations in which agents' payoffs do not directly depend on the messages that agents send to the mechanism. Precisely, we assume that it is common knowledge that messages are cheap talk. This restriction is motivated by our focus on mechanism design problems, in which agents' actions amount to messages designed by the planner, which do not have intrinsic value, and matter only through the outcomes and transfers in which they result. We do allow perturbations to affect agents' preferences

[1] Our mechanisms also work when agents' costs of learning are zero. Even in this simpler case, we are unaware of any existing result that solves the robust implementation problem for our notion, except when agents' payoffs and the social choice function satisfy a condition that is stronger than Maskin monotonicity (see Chen et al. (2021)).

concerning the *outcomes* implemented as a result of their messages as well as their costs to learn the state.

Our analysis focuses for simplicity on the case of two agents.[2] There are n states of the world. A planner knows the objective distribution of the state and wishes to implement a social choice function that maps each state to a lottery over a set of outcomes. She commits to a mechanism mapping agents' messages to lotteries over outcomes and transfers without knowing the state as well as how the environment is perturbed. Each agent observes the mechanism offered by the planner, the realized perturbation, and his type under that perturbation. Agents then independently decide whether to observe the state at some cost and then send messages to the planner.

1.2 Main Results

Our main contribution is to construct mechanisms that robustly implement the desired social choice function when there exists a state whose ex ante probability of occurrence is strictly higher than that of any other state,[3] *or* when the planner is concerned only about perturbations in which agents' costs of learning are uniformly bounded above by some commonly known bound. The mechanisms we construct for generic state distributions have three noteworthy features.

First, each agent is given a message space with $2n - 1$ messages. One of these messages corresponds to the ex ante most likely state, and is called the *status quo message.* The remaining $2n - 2$ messages are divided into pairs that are associated with each of the $n - 1$ remaining states. The two messages corresponding to state θ have the following interpretations: One of them is a *confident message* which means "*I am confident that the state is θ*". The other one is a *confession message* which means "*I am uninformed but I would like to implement the desired outcome in state θ.*"

Second, the desired outcome in the ex ante most likely state is treated as a *status quo outcome.* This outcome is implemented whenever the messages sent by the two agents correspond to different states. When agents' messages correspond to the same state, the planner implements the desired outcome in that state. Therefore, the *confession message* and the *confident message* corresponding to the same state induce the same outcome for any message of the other agent.

[2] We construct mechanisms that robustly implement any given social choice function when there are at least two agents who may have the ability to learn the state. These mechanisms can be easily modified to account for the presence of more agents, e.g., by applying them to two of the agents and ignoring the reports of remaining agents. Even when there are three or more agents, it is unclear whether using the majority rule can easily resolve the robust implementation problem when agents' learning costs are strictly positive. See footnote 13 on page 16 for details.

[3] This condition is satisfied for generic state distributions. In Online Appendix A, we extend our robust implementation results to environments with a continuum of states without requiring any equivalent of the generic condition.

Third, if an agent sends any *confession message* or the *status quo message*, he receives a strictly positive transfer as long as the other agent does not send any *confident message*. If an agent sends any *confident message*, then he faces more risk in the sense that he receives a strictly positive transfer if and only if the other agent sends the same *confident message*, but the transfer he receives conditional on matching the other agent's message is strictly greater than the maximal transfer he can receive from sending any *confession message* or the *status quo message*.

We then address other robustness concerns. First, our mechanisms continue to implement the desired social choice function when the planner faces uncertainty about the objective state distribution or about agents' prior beliefs about the state (Proposition 2 in Section 5.2). Such uncertainty arises, for instance, when agents observe noisy private signals about the state before deciding whether to fully learn it at some cost, and the planner does not know agents' information structures. Second, robust implementation survives the introduction of trembles when agents send messages and of a small amount of noise in agents' information about the state. This may capture situations in which it is infinitely costly to learn the state perfectly.

1.3 Impossibility Results for Stronger Notions of Robust Implementation

We provide several results pertaining to stronger notions of robust implementation. First, we show that even if we require only *approximate* implementation of a non-constant social choice function, it is impossible to achieve such implementation when agents' payoffs can differ from those of the unperturbed environment with non-negligible probability.

Second, we examine the possibility of full implementation and virtual implementation. We show that when agents' costs of learning the state in the unperturbed environment are above some cutoff, or when agents' payoff functions in the unperturbed environment are independent of the state, under every finite mechanism, there exists an equilibrium in which no agent learns the state. This result implies that under each of these two conditions, no finite mechanism can virtually implement any non-constant social choice function.[4] We also provide a sufficient condition 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.

[4] 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.

Third, we examine the possibility of robust (partial) implementation in an interim sense. We adapt the notion of robust interim implementation in Oury and Tercieux (2012) to our setting.[5] We show that no finite mechanism can robustly implement any non-constant social choice function when agents' costs of learning the state in the unperturbed environment are above some cutoff, even when the planner can use unbounded transfers.

1.4 The Cost of Implementation

We construct mechanisms that robustly implement desired social choice functions, but we do not characterize the lowest transfer needed to achieve such robust implementation. Although it would be valuable to determine the lowest cost of implementation, computing this cost seems challenging because it would require precise knowledge of the set of robust equilibria under every mechanism. However, to the best of our knowledge, there is no full characterization of the set of Kajii-Morris robust equilibria in the existing literature.

We view our results showing that it is *possible* to robustly implement the desired outcome via mechanisms with relatively few messages as an important first step for the study of robust implementation in the ex ante sense.[6] These results offer a new perspective on locally robust implementation. They stand in contrast to impossibility results under interim notions of robust implementation, such as Theorem 6 in the present paper and the results contained in Oury and Tercieux (2012). While we do not compute the minimum cost required to robustly implement a given social choice function in the ex ante sense, the mechanisms that we construct provide an upper bound on this cost.

1.5 Outline

Section 2 presents an example to illustrate our results. First, mechanisms that (i) reward agents a fixed amount when their reports match, and (ii) give agents no transfer and randomize across outcomes when their reports mismatch, *cannot* robustly implement the desired outcome. Second, we introduce new mechanisms and explain why they are robust against types that are biased in favor of certain outcomes and types that have high costs of learning. The general model is introduced in

[5] Oury and Tercieux (2012) consider environments without costly learning and with bounded utilities, which stands in contrast to our setting where agents need to learn the state at some cost and agents' utilities are unbounded.

[6] This first step is, in spirit, similar to the results of Vickrey, Clarke, and Groves, who show that the socially efficient outcome is dominant-strategy implementable but leave open the question of finding the lowest cost to implement the socially efficient outcome. Similarly, in the dynamic mechanism design literature, one of Pavan, Segal and Toikka (2014)'s main contributions is to provide a necessary condition for an allocation to be implementable.

Section 3. Our main results appear in Section 4. Section 5 extends our results in several directions, allowing for the possibilities that the planner faces uncertainty about the objective state distribution and about agents' beliefs about the state, that agents tremble with small probability, and that agents can only observe noisy signals about the state after paying their learning costs. Section 6 presents impossibility results for stronger notions of robust implementation. Section 7 reviews the related literature. Extensions to a continuum of states and general information acquisition technologies are given in an online appendix.

2 Example

Consider a planner facing a defendant who is either guilty or innocent of a crime. The state of the world θ is binary: $\theta \in \Theta = \{\text{innocent}, \text{guilty}\}$. Let $q = \Pr(\theta = \text{guilty}) \in (0, 1)$ denote the prior probability of guilt.

The planner knows q but not θ.[7] Her objective is to convict guilty defendants and to acquit innocent ones. She commits to a mechanism $\mathcal{M} = \{M_1, M_2, g, t_1, t_2\}$ in order to elicit information from two agents in charge of investigating the crime. Here, 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. We assume that transfers are non-negative. Importantly, t_1 and t_2 depend only on agents' messages, not on the realized state.

Each agent can conduct an investigation at cost c and learn whether the defendant is guilty or innocent. The decision made by an agent and the information that may result from his investigation are private: they are observed neither by the planner nor by the other agent.

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

Partial Implementation without Robustness: When agents' payoffs are common knowledge, the planner can implement the desired social choice function (convict the guilty and acquit the innocent) via a *Maskin mechanism*: Each agent is asked to report whether the defendant is *guilty* or *innocent* (we use *italics* for messages). The outcome and the transfers are given by:

[7] Section 5.2 extends our main results when the planner does not know q precisely or agents prior beliefs about θ.

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

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

When the reward $R > 0$ is large relative to agents' cost c, there is an equilibrium in which both agents conduct investigations and report their findings truthfully.

Failure of Maskin Mechanisms with Biased Agents: Maskin mechanisms fail to implement the desired social choice function—*even approximately*—when agents can have biases over outcomes with small but positive probability.

We illustrate such failures with a class of perturbations inspired by Rubinstein (1989)'s email game. Suppose that nature draws a random variable ω from a countable set $\Omega \equiv \{\omega_0, \omega_1, \omega_2, ...\}$ according to the geometric distribution $\Pr(\omega = \omega_t) = \eta(1-\eta)^t$ for every $t \in \mathbb{N}$, where $\eta > 0$ is a parameter close to 0. We assume that ω is independent of the state θ. Agent 1 observes 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 and what message to send. Likewise, agent 2 observes which element of the partition $\{\omega_0, \omega_1\}, \{\omega_2, \omega_3\}, \{\omega_4, \omega_5\}, ...$ the realized ω belongs to before deciding whether to conduct his investigation and what message to send. An agent's type is defined by the partition cell that he observes. After observing his own type, each agent updates his belief about the other agent's type according to Bayes rule.[8] The distribution of ω has the property that, whenever an agent observes a cell $\{\omega_k, \omega_{k+1}\}$ of his partition, this agent assigns strictly higher probability to $\omega = \omega_k$ than to $\omega = \omega_{k+1}$.

Agent 2's payoff is $t_2 - cd_2$ at every $\omega \in \Omega$. Agent 1's payoff is $t_1 - cd_1$ at every $\omega \neq \omega_0$. When $\omega = \omega_0$, agent 1's payoff is $t_1 - cd_1 + B \cdot \mathbf{1}\{\text{defendant is acquitted}\}$, i.e., he receives a benefit $B > 0$ if the defendant is acquitted. This perturbation is *small* when η is close to 0 in the sense that agents' payoffs coincide with those in the unperturbed environment when $\omega \neq \omega_0$, and $\Pr(\omega \neq \omega_0) = 1 - \eta$.

For large enough biases, Maskin mechanisms fail to implement the desired social choice function even when η is arbitrarily close to 0: For any reward $R \in \mathbb{R}_+$, there exists a bias $B > R$ such that no matter how close η is to 0, the perturbed game has a unique equilibrium in which no agent conducts any investigation and both agents report *innocent* regardless of the state.[9] As a result,

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

[9] For Maskin mechanisms to fail, we do not need type ω_0's bias B to be arbitrarily large. Our contagion argument applies when agent 1's payoff when $\omega = \omega_0$ is $-cd_1 + b \cdot \mathbf{1}\{\text{the defendant is acquitted}\}$, i.e., type ω_0 of agent 1 is

the defendant is acquitted regardless of his guilt in the unique equilibrium.

This conclusion comes from the following contagion argument. When $\omega = \omega_0$, agent 1 is biased in favor of acquitting the defendant, so he has an incentive to report *innocent* regardless of θ if B is large enough. When $\omega \in \{\omega_0, \omega_1\}$, agent 2 is unbiased, but he believes that agent 1 is biased with probability greater than $\frac{1}{2}$, so he believes that agent 1 will report *innocent* with probability greater than $\frac{1}{2}$ for every 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, all types of both agents will report *innocent* regardless of θ in the unique equilibrium of the perturbed game.

In general, agents may be biased in either direction: some agent types may benefit from convicting the defendant while others may benefit from acquitting the defendant, and these biases may have arbitrary magnitudes. 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 can approximately implement the desired social choice function under every perturbation where agents are unbiased with probability close to 1, but may have arbitrary biases with small probability and may entertain arbitrary beliefs and higher-order beliefs about these biases, as long as those beliefs can be derived from a common prior.

Status Quo Rule with Ascending Transfers. We propose a mechanism that implements the desired social choice function when the planner does not know the direction and magnitude of agents' biases. From now on, we assume that agents' costs of learning are commonly known and equal to some constant c. We later introduce mechanisms to address the case in which the planner also faces uncertainty about the cost of learning.

Our mechanism asks each agent to report the state, i.e., whether the defendant is *innocent* or *guilty*. The outcome and the transfers are given by:

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 the magnitude of transfers R^2 and R^1 satisfy $R^2 - R^1 > \frac{2c}{q}$ and $R^1 > \frac{c}{1-q}$.

This mechanism features a status quo outcome, *acquit*, which is implemented as long as one

purely outcome-driven in the sense that he does not care about the transfers, and receives a strictly positive benefit $b > 0$ from acquitting the defendant. Maskin mechanisms fail even when b is arbitrarily small. Our *Augmented Status Quo Rule with Ascending Transfers* can robustly implement the desired social choice function when perturbations can also affect agents' marginal utilities from transfers. The details are available upon request.

agent reports *innocent.* The defendant is convicted if and only if both agents report *guilty.* Agents receive strictly positive transfers only if their reports coincide. Moreover, they receive a larger transfer when they both report *guilty* than when they both report *innocent.*

To see why this mechanism is robust to the existence of biased types, let us revisit the email game perturbations introduced above: Nature draws a random variable ω from $\Omega = \{\omega_0, \omega_1, \omega_2, ...\}$ according to distribution $\Pi \in \Delta(\Omega)$ independently of θ. 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 $t_2 - cd_2$ at every ω. Agent 1's payoff is $t_1 - cd_1$ at every $\omega \neq \omega_0$. Therefore, every email game perturbation is characterized by the distribution Π and by agent 1's payoff at ω_0.

1. Suppose first that when $\omega = \omega_0$, agent 1 receives a large benefit if the defendant is **acquitted**. This type can guarantee outcome *acquit* by reporting *innocent* regardless of θ. Since $R^2 - R^1 > \frac{2c}{q}$, however, there exists $\lambda \in (0, 1)$ such that $\Pi(\omega_1)$ needs to be less than $\lambda\Pi(\omega_0)$ in order for type $\{\omega_0, \omega_1\}$ of agent 2 to have an incentive to report *innocent* regardless of θ. Likewise, $\Pi(\omega_2)$ needs to be less than $\lambda\Pi(\omega_1)$ in order for type $\{\omega_1, \omega_2\}$ of agent 1 to have an incentive to report *innocent* regardless of θ, and so on. The upper bounds on these probabilities form a decaying geometric sequence, so the total probability of types that are *infected* by type ω_0 is at most $\sum_{t=0}^{+\infty} \lambda^t \Pi(\omega_0) = \frac{1}{1-\lambda}\Pi(\omega_0)$. This expression vanishes to 0 as $\Pi(\omega_0) \to 0$.

2. Suppose now that when $\omega = \omega_0$, agent 1 receives a large benefit if the defendant is **convicted**. If $\Pi(\omega_t) = \eta(1-\eta)^t$ for every $t \in \mathbb{N}$ and type ω_0 reports *guilty* regardless of the state, then all types of both agents have a strict incentive to report *guilty* regardless of the state, because $R^2 > R^1 > 0$.

   However, according to the outcome function specified by the mechanism, the defendant is convicted only if both agents report *guilty.* Therefore, an agent who is biased in favor of convicting the defendant *cannot impose a conviction* when the defendant is innocent and the other agent is truthful. In this case, paying the cost c of learning the defendant's guilt, and reporting *innocent* when the defendant is innocent, leads to a strictly positive transfer. Moreover, the expected value of this transfer exceeds the cost of learning c when $R^1 > \frac{c}{1-q}$.

Of course, the previous argument only shows why the mechanism we propose may be able to avoid some type of contagion for some specific perturbations. To address the general case, we show in the proof of Theorem 1 that under our mechanism, for every perturbation in which both agents are unbiased with probability close to 1, which includes but is not limited to email game perturbations, there always exists an equilibrium in which (i) agents never report *guilty* when the defendant is innocent, and (ii) both agents report the state truthfully with probability close to 1. This equilibrium approximately implements the desired social choice function.

Uncertainty about Agents' Costs of Learning: The planner may also face uncertainty about agents' costs of learning the state. For example, one of the agents may be "inept" in the sense of being unable to conduct investigations.

We show that, as long as the prior probability of guilt q is not exactly equal $\frac{1}{2}$, there exists a mechanism that approximately implements the desired social choice function when, with probability close to 1, agents are unbiased and have cost of learning c, but with some small probability can have arbitrary biases and costs of learning.

We start by explaining why the *Status Quo Rule with Ascending Transfers*, which was introduced earlier to address agents' biases, is unable to deal with inept types. For any $0 < R^1 < R^2$, consider an email game perturbation where agent 1's payoff at ω_0 is $t_1 - \widetilde{c}d_1 + B \cdot \mathbf{1}\{\text{defendant is convicted}\}$. We consider perturbations where his benefit from convicting the defendant $B > 0$ and his cost of learning $\widetilde{c} > 0$ are large relative to the transfers promised by the mechanism.

When this *high-cost biased type* of agent 1 believes that agent 2 reports *guilty* when the defendant is guilty, he prefers to report *guilty* when the defendant is guilty, since he receives a large benefit B from convicting the defendant. If this type wants to report *innocent* when the defendant is innocent, then he needs to conduct an investigation, but his cost of doing so $\widetilde{c}$ outweighs the highest transfer promised by the mechanism. Hence, this type prefers to report *guilty* regardless of θ even when he believes that agent 2 reports truthfully. Since $R^2 > R^1$, 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 close η is to 0.

Augmented Status Quo Rule with Ascending Transfers: We propose another mechanism called the *Augmented Status Quo Rule with Ascending Transfers* that solves the problem caused by high-cost biased types. Without loss of generality, we focus on the case in which $q < \frac{1}{2}$. Under this new mechanism, each agent has a third message, which we denote $-$*guilty*, and which we interpret as the agent *confessing* that (i) he does not know the state and (ii) he prefers to convict the defendant. Under this new mechanism, the outcome and transfers are given by:

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 $\frac{R^0}{R^2} \approx 1$, and $R^2 - R^1, R^1 - R^0$, and R^0 are bounded below by some affine function of c.

Notice that (i) the confession 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 the confession message leads to a lower transfer R^0 than coordinating on any other message, but reporting the confession message leads to a positive transfer as long as the other agent does not report *guilty*. By contrast, reporting *guilty* leads to a positive transfer if and only if the other agent also reports *guilty*.

We now explain why including the confession message makes the mechanism robust to high-cost biased types. First, we note that if agent 1 believes that agent 2 will never send message *guilty* when the defendant is innocent (but agent 2 may send $-$*guilty* and *innocent*), then regardless of agent 1's preference over outcomes and his cost of learning, agent 1 prefers sending $-$*guilty* in both states to sending *guilty* in both states. The reason is that (i) both strategies induce the same outcome regardless of agent 2's message, (ii) none of the two strategies requires any cost of learning, and (iii) agent 1's expected transfer for sending $-$*guilty* in both states equals $R^0 \Pr(m_2 \neq \textit{guilty})$ and agent 1's expected transfer for sending *guilty* in both states equals $R^2 \Pr(m_2 = \textit{guilty})$. As long as agent 2 does not send message *guilty* when the defendant is innocent, we have $\Pr(m_2 \neq \textit{guilty}) \geq \Pr(\theta = \text{innocent}) = 1 - q$ and $\Pr(m_2 = \textit{guilty}) \leq \Pr(\theta = \text{guilty}) = q$. Since $q < \frac{1}{2}$ and $\frac{R^0}{R^2} \approx 1$, reporting $-$*guilty* in both states leads to a higher expected transfer than reporting *guilty* in both states. Hence, the high-cost biased type prefers sending $-$*guilty* in both states over sending *guilty* in both states.

The second key observation is that when a type sends $-$*guilty* in both states, the total probability of types that it can infect is bounded above by a linear function of the probability of this type. This is because sending message $-$*guilty* leads to a transfer of at most R^0, while coordinating on message *innocent* or coordinating on message *guilty* results in strictly greater transfers R^1 and R^2. Every type of agent $i \in \{1, 2\}$ whose payoff is $t_i - cd_i$ prefers to conduct his investigation and to report *innocent* when the defendant is innocent and to 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 $\frac{1}{2}$, the other agent reports *innocent* when the defendant is innocent and reports *guilty* when the defendant is guilty.

The proof of Theorem 2, which covers perturbations in which agents may have arbitrarily high learning costs, generalizes the argument of the previous paragraph and shows that under every perturbation in which, with probability close to 1, agents are unbiased and have costs of learning equal to c, there is an equilibrium in which (i) agents never send *guilty* when the defendant is innocent, and (ii) with probability close to 1, agents send *guilty* when the defendant is guilty and send *innocent* when the defendant is innocent. Such an equilibrium implements the desired social choice function with probability close to 1.

High-probability cost perturbations: Although we focus for simplicity on perturbations in which agents' costs of learning coincide with their costs in the unperturbed environment with probability close to 1, our mechanisms are in fact also robust as long as agents' costs of learning are *no more than* c with probability close to 1. Even if the planner does not know agents' exact learning cost, knowing some upper bound that applies with probability close to 1 suffices to achieve robust implementation. See Footnote 11 of Section 3 for details.

Cheap-talk messages: We assume throughout that it is common knowledge that *messages are cheap talk*. In the example, this means that we rule out types who directly benefit, say, from sending message *guilty*. If we allowed such types, and the benefit for sending *guilty* were large enough, then sending *guilty* regardless of θ is these types' dominant strategy, which would cause contagion since $R^2 > R^1 > R^0 > 0$.

3 Model

Unperturbed Environment: 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 set of outcomes.[10] 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 $\theta \in \Theta$.

[10] In general, agents' ability to learn the state of the world may be limited, creating a discrepancy between what agents can learn and what the planner cares about. In this case, we interpret θ as what agents *can* learn, since it is the only information that can be elicited from any mechanism. Moreover, our results extend when agents observe noisy private signals about the state after paying their costs of learning, as shown in Proposition 1. Our main result also holds when there is a continuum of states, as shown in Online Appendix A, under the assumption that the social choice function f and agents' payoff functions in the unperturbed environment (u_1, u_2) are continuous with respect to θ.

The planner knows q but does not know θ. She commits to a mechanism $\mathcal{M} \equiv \{M_1, M_2, t_1, t_2, g\}$ in order to elicit θ from two agents, where M_i is a *finite* set of messages for agent i, $t_i : M_1 \times M_2 \to \mathbb{R}_+$ is the transfer to agent i, and $g : M_1 \times M_2 \to \Delta(Y)$ is the implemented outcome. Our restriction to finite mechanisms makes our robust implementation results stronger. It is also motivated by the fact that mechanisms with infinitely many messages have undesirable properties.

After observing $\mathcal{M}$, agents simultaneously and independently decide whether to observe θ at some cost. Let $d_i \in \{0, 1\}$ be agent i's decision to obtain information, where $d_i = 1$ represents agent i obtaining information about θ and vice versa. Let $c_i \geq 0$ be agent i's cost of learning.[11] We assume that learning is *covert* in the sense that neither agent $-i$ nor the planner can observe d_i.

Agents then 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:

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

Robust Implementation: We examine whether the planner can *robustly* implement f when agents' preferences over outcomes, their costs of learning the state, and their beliefs and higher-order beliefs about each other's payoffs can differ from those of the baseline setting.

Following Kajii and Morris (1997), a *perturbation* $\mathcal{G} \equiv \{\Omega, \Pi, Q_1, Q_2, \widetilde{u}_1, \widetilde{u}_2, \widetilde{c}_1, \widetilde{c}_2\}$ 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 $\widetilde{u}_i : \Omega \times \Theta \times Y \to \mathbb{R}$, and $\widetilde{c}_i : \Omega \to [0, +\infty]$ for $i \in \{1, 2\}$, where $\widetilde{c}_i(\omega) = +\infty$ means that agent i does not have the ability to learn θ at ω. Agent i's payoff under perturbation $\mathcal{G}$ is

$$\widetilde{u}_i(\omega, \theta, y) - \widetilde{c}_i(\omega) d_i + t_i. \tag{3.2}$$

For given $\overline{c} > 0$, we say that $\mathcal{G}$ is a *$\overline{c}$-bounded perturbation* if $\widetilde{c}_i(\omega) \leq \overline{c}$ for every i and ω.

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 $\widetilde{u}_i(\omega', \theta, y) = u_i(\theta, y)$ and $\widetilde{c}_i(\omega') = c_i$ 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

[11] In our baseline model, each agent either fully learns the state or learns nothing. Proposition 1 in Section 5 generalizes the main result to situations where agents can only observe *noisy signals* about the state after paying their learning costs. In Online Appendix B, we generalize our result by allowing agents to choose any partition of the state space as their information structures, and different partitions may incur different costs.

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*

$$\Pi\Big(\textit{both agents are normal types}\Big) \geq 1 - \eta. \tag{3.3}$$

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

Intuitively, a perturbation is *small* if agents' payoffs coincide with those in the unperturbed environment with probability close to one, but their payoffs can be very different from the unperturbed environment with small but positive probability. Even though every normal-type agent's payoff coincides with his payoff 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 email game perturbations considered in Section 2 are η-perturbations since both agents are normal types when $\omega \in \Omega \backslash \{\omega_0\}$, and the event $\Omega \backslash \{\omega_0\}$ occurs with probability $1 - \eta$ under Π.

The planner faces uncertainty about 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 lottery over outcomes conditional on the state being θ when the planner commits to outcome function g and agents behave according to σ.

Like Oury and Tercieux (2012), we focus on *partial* implementation: the planner requires only that f be implemented in at least *one* equilibrium, not necessarily all equilibria. Our main results in Section 4 examine whether the planner can design a mechanism that approximately implements f for *all* small enough perturbations.[12]

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 game induced by $(\mathcal{M}, \mathcal{G})$, such that

$$\max_{\theta \in \Theta} ||g_{\sigma(\mathcal{G})}(\theta) - f(\theta)||_{\text{TV}} < \varepsilon, \tag{3.4}$$

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

[12] Theorems 1 and 2 extend to a larger class of perturbations in which $\widetilde{u}_i(\omega', \theta, y) = u_i(\theta, y)$ and $\widetilde{c}_i(\omega') \leq c_i$ with probability close to 1, i.e., to settings in which the planner knows that agents' costs of learning are (with probability close to 1) no more than c_1 and c_2, but may not know agents' exact learning costs. The proof uses similar ideas, but the analysis is more cumbersome as one needs to redefine agents' strategies as mappings from states *and learning costs* to messages.

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 (3.4) holds.

We do not characterize the lowest expected transfer needed to robustly implement f. Doing so would likely require knowing the set of games for which there exist robust equilibria that implement f. However, to the best of our knowledge, there is no characterization of the set of games that have Kajii-Morris robust equilibria. We do compute the expected aggregate transfer that is needed to robustly implement f under the mechanisms we propose. This transfer may be viewed as an *upper bound* on the cost needed to robustly implement f. Formally, say that mechanism $\mathcal{M}$ robustly implements f with cost no more than $T \in \mathbb{R}_+$ if for every $\varepsilon > 0$ and $\xi > 0$, there exists $\eta > 0$ such that for every η-perturbation $\mathcal{G}$, there exists an equilibrium $\sigma(\mathcal{G})$ of the game induced by $(\mathcal{M}, \mathcal{G})$ such that inequality (3.4) is satisfied and, moreover, $\mathbb{E}\Big[t_1(m_1, m_2) + t_2(m_1, m_2)\Big| \mathcal{M}, \mathcal{G}, \sigma(\mathcal{G})\Big] \leq T + \xi$.

Relation to the Existing Literature: In our analysis, the planner cannot condition transfers on the realized state. This assumption 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 depend on the realized state. Our model fits situations (and, in this respect, is consistent with the large literature on implementation) 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 information is impractical.

We also restrict attention to perturbations in which agents' payoffs are quasi-linear and do not directly depend on their messages. These assumptions are commonly made in the mechanism design literature, including Rochet (1987), Chung and Ely (2007), and Bergemann and Morris (2009). Our assumption that messages are cheap talk stands 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, in our mechanism design setting, agents' message spaces are endogenously chosen by the planner, these messages have no meaning per se and we thus believe that it is reasonable to view them as cheap talk.

Oury and Tercieux (2012), Chen, Kunimoto and Sun (2020), and Chen, Mueller-Frank and Pai (2022) adopt an *interim* approach to study robust partial implementation. Similar to our paper, these works all assume that agents' messages are 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 some commonly known baseline with probability close to one.

Knowledge about the Realized Perturbation: We assume that the realized perturbation is common knowledge among the agents but is unknown to the planner. This assumption is standard in the robust mechanism design literature (e.g., Chung and Ely 2007). It fits applications in which (i) the planner sets rules (the mechanism) in advance without knowing the specific circumstances that the society will be facing (e.g., when designing laws, constitutions, or corporate rules) but (ii) agents do know the particular circumstances they are facing when they decide on how to react to the mechanism.

Since both agents can observe the perturbation $\mathcal{G}$, one may wonder whether the planner could ask both agents to report $\mathcal{G}$, and punish both agents if their reports do not coincide (e.g., by implementing a particular outcome or by giving them negative transfers).

Although this possibility would be worth exploring, we make two observations. First, our focus on *finite mechanisms* precludes such a possibility, since there are infinitely many perturbations. Soliciting information about the realized perturbation requires the use of infinite mechanisms. Moreover, asking agents to report complicated objects such as the realized perturbation may be difficult and intractable in practice. Second, when only two agents can learn the state, it is unclear whether there exists a mechanism that can induce all agent types to report the realized perturbation truthfully. The reason is that agents' preferences in the perturbed environment can be arbitrary, so it is impossible to design a punishment that deters all types from lying. For example, the outcome that punishes some types may constitute an arbitrarily large reward for other types who are strongly biased in favor of this outcome, and may encourage these latter types to lie about the perturbation they observed just for the sake of getting this outcome implemented. Moreover, agents' coordination motives would then imply that a type's incentive to lie about the perturbation may encourage other types to lie as well.

The Case of Zero Learning Cost: While our mechanisms achieve robust implementation with positive learning costs, we are unaware of similar results even when agents have zero learning cost.[13]

[13] When there are three or more agents and the planner is concerned about perturbations for which the learning costs of *all* types are equal to zero, there is a trivial solution to the robust implementation problem. The solution

If $c_1 = c_2 = 0$ *and* $u_1(\theta, y)$ and $u_2(\theta, y)$ are independent of θ, there is a trivial solution to the robust implementation problem: The planner promises agent 1 a transfer of $-u_1(y)$ and agent 2 a transfer of $-u_2(y)$ whenever she implements outcome y. The normal type of each agent is indifferent between all messages, so there exists an equilibrium where all normal types learn the state and report it truthfully. However, this solution does not work when u_1 or u_2 depends on θ, or when agents have positive costs of learning.

When $c_1 = c_2 = 0$, and (f, u_1, u_2) satisfies a Maskin monotonicity* condition, which is strictly stronger than the Maskin monotonicity condition in Maskin (1999), Chen, Kunimoto, Sun, and Xiong (2021) show that there exists a finite mechanism that fully implements f under the solution concept of correlated rationalizability, which implies that their mechanism can fully implement f under the solution concept of correlated equilibrium. According to Proposition 3.2 in Kajii and Morris (1997), the mechanism in Chen et al (2021) can robustly implement f when (f, u_1, u_2) satisfies Maskin monotonicity*. By contrast, our results in Section 4 construct a different class of finite mechanisms that can robustly implement f without any restriction on (f, u_1, u_2).

4 Main Results

This section presents two main results. Theorem 1 shows that every f is robustly implementable when agents' costs of learning are uniformly bounded from above across all the perturbations considered by the planner. Theorem 2 shows that even when arbitrarily large (or infinite) learning costs are allowed, every f is robustly implementable, as long as the state's prior distribution satisfies a generic assumption.

4.1 Robust Implementation with Bounded Perturbations

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

For simplicity, in the main text, we prove all our results under the assumptions that $u_1(\theta, y) = u_2(\theta, y) = 0$ and $c_1 = c_2 = c$, i.e., that each *normal type*'s payoff is equal to his transfer minus

consists in using the majority rule and zero transfers. In this case, no agent is pivotal when other agents are truthful and truthtelling is an equilibrium. However, such mechanisms can fail when some types have positive costs of learning in the perturbed environment. To see this, note that When a type has a positive learning cost, he has no incentive to learn the state if he believes that he is never pivotal. For example, suppose that there are three agents and that type $\{\omega_0\}$ of agent 1 and type $\{\omega_0, \omega_1\}$ of agent 2 both have positive learning costs. These types' messages do not depend on the state when these types believe that they are not pivotal. This, in turn, affects agent 3's incentive to make a state-contingent report. The contagion process that ensues may lead to an undesirable implementation outcome.

his cost of learning and the normal types of both agents face the same cost c. Types that are not normal can have arbitrary $\widetilde{u}_i(\omega,\theta,y)$ and $\widetilde{c}_i(\omega)$. The proofs for general utility functions (u_1,u_2) and heterogeneous costs of learning are in Appendix A and do not present additional conceptual challenges.

Proof. We propose a mechanism called the *Status Quo Rule with Ascending Transfers.* Let $\Theta \equiv \{\theta^1,...,\theta^n\}$. 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 $R^1 > \frac{\bar{c}}{q(\theta^1)}$, $R^j > R^1$ for every $j \geq 2$, and $\sum_{j=2}^n (R^j - R^1)q(\theta^j) > 2c$.[14]

In the *unperturbed game* induced by our mechanism, an agent's pure strategy can be described as an n-dimensional vector $(m^1,...,m^n)$, where $m^j \in M$ is the message that the agent sends when the state is θ^j. If agent 1 uses strategy $(m_1^1,...,m_1^n)$ and agent 2 uses strategy $(m_2^1,...,m_2^n)$, then agent i's expected payoff equals

$$\sum_{j=1}^n q(\theta^j)\Big\{t_i(m_1^j,m_2^j) + u_i(\theta^j, g(m_1^j,m_2^j))\Big\} - \Big(1 - \mathbf{1}\{m_i^1 = ... = m_i^n\}\Big)c_i. \tag{4.3}$$

When $u_1 = u_2 = 0$ and $c_1 = c_2 = c$—the case we focus on in the main text—agent i's expected payoff is equal to

$$\sum_{j=1}^n q(\theta^j)t_i(m_1^j,m_2^j) - \Big(1 - \mathbf{1}\{m_i^1 = ... = m_i^n\}\Big)c. \tag{4.4}$$

In the perturbed game induced by our mechanism and $\mathcal{G} = \{\Omega,\Pi,Q_1,Q_2,\widetilde{u}_1,\widetilde{u}_2,\widetilde{c}_1,\widetilde{c}_2\}$, a *type*'s pure strategy is also given by $(m^1,...,m^n)$, where $m^j \in M$ is the message that this type sends when the state is θ^j. A *pure strategy profile* $\{(m_i^1(\omega),...,m_i^n(\omega))\}_{i\in\{1,2\},\omega\in\Omega}$ describes each agent i's message $m_i^j(\omega)$ for each state θ^j and circumstance ω, and must satisfy the restriction that $m_i^j(\omega)$ be measurable with respect to Q_i for every $i \in \{1,2\}$ and $j \in \{1,2,...,n\}$. For every $i \in \{1,2\}$ and

[14] The mechanism for general (u_1,u_2,c_1,c_2) has the same outcome function. The transfers satisfy (A.1) and (A.2).

$\omega^* \in \Omega$, the expected payoff for type $Q_i(\omega^*)$ of agent i's is given by

$$\sum_{j=1}^{n} q(\theta^j)\mathbb{E}_\omega\Big[t_i(m_1^j(\omega), m_2^j(\omega)) + \widetilde{u}_i\Big(\omega, \theta^j, g(m_1^j(\omega), m_2^j(\omega))\Big)\Big|Q_i(\omega^*)\Big]$$

$$- \Big(1 - \mathbf{1}\{m_i^1(\omega^*) = ... = m_i^n(\omega^*)\}\Big)\mathbb{E}_\omega\Big[\widetilde{c}_i(\omega)\Big|Q_i(\omega^*)\Big]. \tag{4.5}$$

Let $\Sigma \equiv \{1, 2, ..., n\}^n$ denote the set of pure strategies. Agent $i \in \{1, 2\}$ is *truthful* if he uses strategy $(1, 2, ..., n)$, i.e., if he truthfully reports the index of the realized state. We define $\Sigma^* \subset \Sigma$ as:

$$\Sigma^* \equiv \Big\{(m^1, ..., m^n) \in \Sigma \text{ such that } m^j \in \{1, j\} \text{ for every } j \geq 1\Big\}. \tag{4.6}$$

If an agent's strategy belongs to Σ^*, then in every state θ^j, this agent either sends the status quo message 1 or reports the state truthfully by sending message j. For example, when $n = 2$, $\Sigma^* = \{(1, 1), (1, 2)\}$ while $\Sigma = \{(1, 1), (1, 2), (2, 1), (2, 2)\}$. The rest of the proof consists of three steps.

Step 1: The first step examines a *restricted game without perturbation* where both agents are only allowed to use (mixed) strategies in $\Delta(\Sigma^*)$ and it is common knowledge that agents' payoffs are $t_1 - cd_1$ and $t_2 - cd_2$. For any given $\gamma \in [0, 1]$, a *γ-dominant equilibrium* is a Nash equilibrium where every agent finds it strictly optimal to play his equilibrium strategy when he believes that the other agent will play their equilibrium with probability at least γ.

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. In the restricted game without perturbation, agents can only send message 1 conditional on $\theta = \theta^1$ and, for every $j \geq 2$, agents can only send message 1 or message j conditional on $\theta = \theta^j$.

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

Suppose agent 2 is truthful with probability at least $\frac{1}{2}$, $\Pr(m_2 = j|\theta^j) \geq \frac{1}{2}$ and $\Pr(m_2 = 1|\theta^j) \leq \frac{1}{2}$. Since $R^j > R^1$ for every $j \geq 2$, agent 1 strictly prefers strategy $(1, 2, ..., n)$ to any other strategy $(m^1, ..., m^n)$ that belongs to Σ^* but is neither $(1, 2, ..., n)$ nor $(1, 1, ..., 1)$. Since $\sum_{j=2}^{n}(R^j - R^1)q(\theta^j) > 2c$, agent 1's expected payoff under $(1, 2, ..., n)$ minus that under $(1, 1, ..., 1)$ is at least

$\sum_{j=2}^{n} \frac{1}{2}(R^j - R^1)q(\theta^j) - c$, which is strictly positive. Since each agent *strictly* prefers $(1, 2, ..., n)$ to any other strategy in Σ^* when he believes that the other agent is truthful with probability at least $\frac{1}{2}$, there exists $\gamma < \frac{1}{2}$ such that both agents being truthful is a γ-dominant equilibrium. □

Step 2: For any $\mathcal{G}$, consider a *restricted game with perturbation* $\mathcal{G}$ where agent $i \in \{1, 2\}$'s payoff is $\widetilde{u}_i(\omega, \theta, y) - \widetilde{c}_i(\omega)d_i + t_i$, and agents are only allowed to use strategies in $\Delta(\Sigma^*)$. Since there exists $\gamma < \frac{1}{2}$ such that both agents being truthful is a γ-dominant equilibrium in the restricted game without perturbation, the Critical Path Lemma in Kajii and Morris (1997) implies that:

Lemma 2. *For every $\varepsilon > 0$, there exists $\eta > 0$, such that for every η-perturbation $\mathcal{G}$, there exists an equilibrium $\sigma(\mathcal{G})$ in the restricted game with perturbation $\mathcal{G}$, under which the probability with which both agents being truthful is greater than $1 - \varepsilon$.*

Since $g(j, j) = f(\theta^j)$ for every $j \in \{1, 2, ..., n\}$, f is implemented when both agents are truthful, which occurs with probability at least $1 - \varepsilon$ when agents behave according to $\sigma(\mathcal{G})$.

Step 3: We show that for every $\mathcal{G}$, the equilibrium $\sigma(\mathcal{G})$ constructed in the previous step remains an equilibrium under perturbation $\mathcal{G}$ when agents can use any strategy in the set $\Delta(\Sigma)$.

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

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

By construction, $(m_*^1, ..., m_*^n) \in \Sigma^*$. We compare type $Q_1(\omega)$'s expected payoff from $(m^1, ..., m^n)$ to his expected payoff from $(m_*^1, ..., m_*^n)$.

1. First, $(m^1, ..., m^n)$ and $(m_*^1, ..., m_*^n)$ lead to the same joint distribution of (θ, y) when agent 2's strategy belongs to $\Delta(\Sigma^*)$. This is because $m_*^j = m^j$ when $m^j \in \{1, j\}$; and when $m^j \notin \{1, j\}$, agent 2 sends either 1 or j when the realized state is θ^j. Given the outcome function (4.1), the implemented outcome is $f(\theta^1)$ whenever agent 1 sends a message other than j.

2. Second, conditional on each state, $(m_*^1, ..., m_*^n)$ gives a weakly greater transfer to agent 1 than does $(m^1, ..., m^n)$. This is because when the state is θ^j and agent 2's message belongs to $\{1, j\}$, agent 1 receives zero transfer when he sends any message that is neither 1 nor j.

3. Third, if $(m_*^1, ..., m_*^n)$ requires a strictly greater learning cost compared to $(m^1, ..., m^n)$, then $m^1 = ... = m^n \geq 2$. Conditional on $\theta = \theta^j$, the transfer under m_*^1 is R^1 and the transfer under m^1 is 0 when the other agent's strategy belongs to $\Delta(\Sigma^*)$. Since $q(\theta^1)R^1 \geq \bar{c}$, the expected transfer from $(m_*^1, ..., m_*^n)$ is greater than $\bar{c}$ plus the expected transfer from $(m^1, ..., m^n)$.

Since each agent's learning cost is no more than $\bar{c}$ when $\mathcal{G}$ is a $\bar{c}$-bounded perturbation, every type prefers $(m_*^1, ..., m_*^n)$ to $(m^1, ..., m^n)$. This contradicts the hypothesis that type $Q_1(\omega)$ strictly prefers $(m^1, ..., m^n)$ to all strategies in Σ^*. Since $\sigma(\mathcal{G})$ is an equilibrium in the restricted game with perturbation when agents are only allowed to use strategies in $\Delta(\Sigma^*)$, $\sigma(\mathcal{G})$ remains an equilibrium in the unrestricted game with perturbation $\mathcal{G}$ in which agents can use any strategy in $\Delta(\Sigma)$. □

Implementation Cost: We bound the expected cost to implement f focusing on the case where $u_1 = u_2 = 0$ and $c_1 = c_2 = c$. The cost for the general case is in Appendix A. The expected cost $\mathbb{E}[t_1 + t_2]$ under our mechanism equals $2\sum_{j=1}^n q(\theta^j)R^j$, which can be as low as $\frac{2\bar{c}}{\max_{\theta\in\Theta} q(\theta)} + 4c$. This is because when θ^1 maximizes $q(\theta)$, R^1 can be as low as $\frac{\bar{c}}{\max_{\theta\in\Theta} q(\theta)}$, and the requirement that $\sum_{j=2}^n (R^j - R^1)q(\theta^j) > 2c$ implies that $\sum_{j=2}^n q(\theta^j)R^j$ can be as low as $2c + R^1 \sum_{j=2}^n q(\theta^j)$.

4.2 Robust Implementation with an Ex Ante Most Likely State

We show that as long as the objective state distribution $q \in \Delta(\Theta)$ satisfies a generic condition, stated below, every f is robustly implementable even when some types have arbitrarily large biases or learning costs, or when some types are inept in the sense that they do not have the ability to learn the state.

Definition 1. *$q \in \Delta(\Theta)$ is* generic *if there exists $\theta^* \in \Theta$ such that $q(\theta^*) > q(\theta')$, $\forall \theta' \neq \theta^*$.*

When there are two states, for instance, this condition rules out the prior q that assigns probability $\frac{1}{2}$ to each state, but allows any other full support distribution. In Online Appendix A, we generalize our result to environments in which (i) there is a continuum of states, (ii) the objective distribution q has no atom, and (iii) (f, u_1, u_2) are continuous with respect to the state. In that environment, the generic condition is no longer required and our result holds for all full support distributions.

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

Proof. We propose a mechanism called the *Augmented Status Quo Rule with Ascending Transfers.* When q is generic, we can write $\Theta \equiv \{\theta^1, ..., \theta^n\}$ such that $q(\theta^1) > q(\theta^2) \geq ... \geq q(\theta^n) > 0$.

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

$$g(m_1, m_2) = \begin{cases} f(\theta^{|m_1|}) & \text{if } |m_1| = |m_2| \\ f(\theta^1) & \text{otherwise.} \end{cases} \tag{4.7}$$

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

$$t_i(m_i, m_{-i}) = \begin{cases} R^j & \text{if } m_1 = m_2 = j \geq 1 \\ R^0 & \text{if } m_1, m_2 \leq 1 \text{ but } (m_1, m_2) \neq (1, 1) \\ 0 & \text{otherwise,} \end{cases} \tag{4.8}$$

where $R^n, ..., R^0$ satisfy $\min\{R^n, ..., R^2\} > R^1 > R^0 > 0$,

$$\sum_{j=2}^{n} q(\theta^j)(R^j - R^1) > 2c, \tag{4.9}$$

and

$$\frac{R^0}{R^j} > \frac{q(\theta^j)}{q(\theta^1)} \text{ for every } j \geq 2. \tag{4.10}$$

When q is generic, there exist $R^n, ..., R^0$ that satisfy these inequalities. Our mechanism when there are two states is presented in Section 2. When there are three states, our 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	$0, 0$	$0, 0$
-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 (or an agent type's) *pure strategy* is $(m^1, ..., m^n)$, where $m^j \in M$ represents the message he sends when the state is θ^j. He pays the cost of learning unless $m^1 = ... = m^n$. The truthful strategy is $(1, 2, ..., n)$, according to which the agent reports the index of the realized state.

Let $\Sigma \equiv \{-n, ..., -2, 1, 2, ..., n\}^n$ be the set of pure strategies. Let

$$\Sigma^* \equiv \Big\{(m^1, ..., m^n) \in \Sigma \text{ such that } m^j \in \{-n, ..., -2, 1\} \cup \{j\} \text{ for every } j \geq 1\Big\}. \tag{4.11}$$

By definition, if an agent's strategy belongs to Σ^*, then conditional on each state θ^j, he either sends a negative message, or sends the status quo message 1, or sends message j. For example, when $n = 2$, $\Sigma^* = \{(-2, -2), (-2, 1), (-2, 2), (1, -2), (1, 1), (1, 2)\}$ while $\Sigma = \Sigma^* \bigcup \{(2, -2), (2, 1), (2, 2)\}$.

Step 1: We examine a restricted game *without* perturbation in which it is common knowledge that payoffs are $t_1 - cd_1$ and $t_2 - cd_2$,[15] and both agents are only allowed to use strategies in $\Delta(\Sigma^*)$.

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 $j \geq 2$, conditional on $\theta = \theta^j$, agent 1's expected transfer equals $\Pr(m_2 = j|\theta^j)R^j$ if he sends message j, and is at most $\Pr(m_2 \leq 1|\theta^j)R^1$ if he sends message 1 or any negative message. If he believes that agent 2 is truthful with probability at least $\frac{1}{2}$, then $\Pr(m_2 = j|\theta^j)R^j > \Pr(m_2 \leq 1|\theta^j)R^1$ given that $R^j > R^1$.
- Conditional on $\theta = \theta^1$, agent 1's expected transfer equals $\Pr(m_2 = 1|\theta^1)R^1 + \Pr(m_2 \leq -2|\theta^1)R^0$ if he sends message 1 and equals R^0 if he sends any negative message. If he believes that agent 2 is truthful with probability at least $\frac{1}{2}$, then $\Pr(m_2 = 1|\theta^1)R^1 + \Pr(m_2 \leq -2|\theta^1)R^0 > R^0$ given that $R^1 > R^0$.

The discussion above implies that agent 1 strictly prefers the truthful strategy to any other non-constant strategy that belongs to Σ^*. Agent 1's expected payoff from using a constant strategy in Σ^* is at most $\sum_{j=1}^n q(\theta^j)R^1 \Pr(m_2 \leq 1|\theta^j)$, while his expected payoff from being truthful is at least $\sum_{j=1}^n q(\theta^j)R^n \Pr(m_2 = j|\theta^j)$. Inequality (4.9) implies that $\sum_{j=1}^n q(\theta^j)R^n \Pr(m_2 = j|\theta^j) > \sum_{j=1}^n q(\theta^j)R^1 \Pr(m_2 \leq 1|\theta^j)$ when agent 2 is truthful with probability at least $\frac{1}{2}$. Since agent 1 strictly prefers to be truthful 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 any other strategy in Σ^* when agent 2's strategy belongs to $\Delta(\Sigma^*)$ and is truthful with probability at least γ.

[15] Recall that in the main text, our proof focuses on the case where $u_1 = u_2 = 0$ and $c_1 = c_2 = c$. We explain how to generalize our proof to arbitrary $u_1(\theta, y)$ and $u_2(\theta, y)$, and to heterogeneous costs of learning in Appendix A.

Step 2: For any perturbation $\mathcal{G}$, consider a *restricted game with perturbation* $\mathcal{G}$ where agent i's payoff is $\widetilde{u}_i(\omega, \theta, y) - \widetilde{c}_i(\omega)d_i + t_i$, and agents are only allowed to use strategies in $\Delta(\Sigma^*)$.

Since there exists $\gamma < \frac{1}{2}$ such that both agents being truthful is a γ-dominant equilibrium in the restricted game without perturbation, the Critical Path Lemma 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 restricted game perturbed by $\mathcal{G}$ in which both agents are truthful with probability more than $1 - \varepsilon$.

Since $g(j, j) = f(\theta^j)$ for every $j \in \{1, 2, ..., n\}$, f is implemented when both agents are truthful, which occurs with probability more than $1 - \varepsilon$ when agents behave according to $\sigma(\mathcal{G})$.

Step 3: We show that when q is generic and $\{R^n, ..., R^1, R^0\}$ satisfy (4.9) and (4.10), the equilibrium $\sigma(\mathcal{G})$ in the restricted game with perturbation $\mathcal{G}$ remains an equilibrium in the *unrestricted game with perturbation* $\mathcal{G}$ in which agents can use any strategy in $\Delta(\Sigma)$, not just those 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 defined as

$$m^j_* \equiv \begin{cases} m^j & \text{if } m^j \in \{-n, ..., -2\} \cup \{1, j\} \\ -m^j & \text{if } m^j \notin \{-n, ..., -2\} \cup \{1, j\} \end{cases} \quad \text{for every } j \in \{1, 2, ..., n\}. \tag{4.12}$$

By construction, $(m^1_*, ..., m^n_*) \in \Sigma^*$. Since $(m^1, ..., m^n)$ is non-constant, $(m^1_*, ..., m^n_*)$ does not increase the cost of learning compared to $(m^1, ..., m^n)$. The outcome function (4.7) ensures that, regardless of whether agent 1 uses strategy $(m^1_*, ..., m^n_*)$ or strategy $(m^1, ..., m^n)$, he will induce the same joint distribution of (θ, y) regardless of agent 2's strategy. When agent 1 believes that agent 2's strategy belongs to $\Delta(\Sigma^*)$, which is the case when agent 2 plays according to $\sigma(\mathcal{G})$, agent 1 receives a weakly greater transfer from $(m^1_*, ..., m^n_*)$ compared to $(m^1, ..., m^n)$. This is because sending any message that does not belong to $\{-n, ..., -2\} \cup \{1, j\}$ leads to a transfer of 0 in state θ^j when agent 2's message in state θ^j belongs to $\{-n, ..., -2\} \cup \{1, j\}$.

Second, for every $(m^1, ..., m^n) \notin \Sigma^*$ that satisfies $m^1 = ... = m^n$, there exists $k \in \{2, 3, ..., n\}$ such that $(m^1, ..., m^n) = (k, ..., k)$. Let us compare the expected payoff that any given type of agent 1 receives with strategies $(k, ..., k)$ and $(-k, ..., -k)$. The outcome function in (4.7) implies that $(k, ..., k)$ and $(-k, ..., -k)$ lead to the same joint distribution over (θ, y). None of these strategies requires agent 1 to learn θ. The expected transfer is $\Pr(m_2 = k)R^k$ if agent 1 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 $\Delta(\Sigma^*)$, we have $\Pr(m_2 \leq 1) \geq q(\theta^1)$ and $\Pr(m_2 = k) \leq q(\theta^k)$. Condition (4.10) then implies that $\Pr(m_2 = k)R^k \leq q(\theta^k)R^k < q(\theta^1)R^0 \leq \Pr(m_2 \leq 1)R^0$. Hence, type $Q_1(\omega)$'s expected transfer is weakly greater under $(-k, ..., -k)$ compared to that under $(k, ..., k)$. □

Implementation Cost: In the case where $u_1 = u_2 = 0$ and $c_1 = c_2 = c$, the expected cost $\mathbb{E}[t_1 + t_2]$ under our mechanism equals $2\sum_{j=1}^{n} q(\theta^j)R^j$. A tight lower bound for this is

$$\frac{4c}{\sum_{j=2}^{n}(q(\theta^1) - q(\theta^j))} + 4c. \tag{4.13}$$

The calculations are in Appendix A, together with the implementation cost in the general case.

5 Trembles, Noisy Signals, & Unknown State Distributions

Section 5.1 modifies our mechanism so that it can robustly implement f when (i) agents tremble with small probability, and (ii) agents observe noisy private signals about the state after paying their costs of learning. This extension captures situations in which learning the state perfectly is prohibitively costly and agents can only learn an imperfect signal about the state. Section 5.2 examines the robustness of our results when the planner does not know the state distribution q or faces uncertainty about agents' beliefs about the state (e.g., when agents receive noisy private signals about the state before observing the mechanism and the planner does not know the agents' information structures).

5.1 Robustness to Trembles and Noisy Information

In the proofs of Theorems 1 and 2, we construct equilibria in which no type uses any strategy that does not belong to $\Delta(\Sigma^*)$. One may wonder whether our results are robust when agents tremble with small probability or when agents cannot perfectly observe θ even after paying their costs of learning, in which case agents may not know each others' private beliefs. This section shows that our results are robust when the trembling probabilities and the noise in agents' private signals are *small.*

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 $F_i \in \Delta(M_i)$ with probability τ, where $\tau \in (0,1)$ is the probability with which agents tremble. Throughout this section, we distinguish between an agent's *intended message* and his *realized message*. We suppress the dependence of F_i on $\mathcal{M}$ in order to simplify notation.

Imperfect Signals about the State: 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. Note that $|S_i|$ can be any finite number, i.e., we do not impose any 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. For every $\overline{\tau} > 0$, we say that π is of size $\overline{\tau}$ if

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

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

$$\pi\Big(h_{-i}(s_{-i}) = h_i(s_i)\Big|s_i\Big) \geq 1 - \overline{\tau} \text{ for every } s_i \in S_i, \tag{5.1}$$

and

$$\sum_{j=1}^{n} \sum_{s_i \in \{h_i(s_i)=j\}} \pi(\theta^j, s_i) \geq 1 - \overline{\tau}. \tag{5.2}$$

Our first requirement is that the marginal distribution on θ be consistent with the objective state distribution q. Our second requirement is reminiscent of Chung and Ely (2003), Aghion, Fudenberg, Holden, Kunimoto and Tercieux (2012), and Sugaya and Takahashi (2013), in which every signal that can be observed by agent $i \in \{1,2\}$ is linked to a particular state, given by the mapping h_i. One can think about h_i as endowing each of agent i's realized signal with a *meaning*, where each meaning corresponds to a state. According to requirement (b), the mappings from realized signal to their meanings satisfy (i) no matter which signal an agent observes, he 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 coincides with the state with probability close to 1.

The planner knows neither $\mathcal{G}$ nor $\{\tau, F_1, F_2, \pi\}$. She would like to design a mechanism $\mathcal{M}$ that can approximately implement f for all small enough perturbations, small enough trembles, and small enough noise in agents' private signals. Agent i knows the mechanism $\mathcal{M}$, the perturbation $\mathcal{G}$, his information about ω under $\mathcal{G}$, as well as $\{\tau, F_1, F_2, \pi\}$. He decides whether to pay a cost c_i in order to learn s_i and, after this decision and possibly the observation of s_i, which message in M_i he intends to send. The planner observes the *realized messages* but not the *intended messages*.

Proposition 1. *Suppose q is generic. For every $f : \Theta \to \Delta(Y)$, there exists a mechanism with $2|\Theta| - 1$ messages for each agent, such that for every $\varepsilon > 0$, there exist $\eta > 0$ and $\overline{\tau} > 0$ such that for every trembling probability $\tau < \overline{\tau}$, 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 $\max_{\theta \in \Theta} ||g_{\sigma(\mathcal{G})}(\theta) - f(\theta)||_{TV} < \varepsilon$.*

The proof is in Appendix F. When there are two states, our *Augmented Status Quo Rule with Ascending Transfers* can robustly implement f when agents tremble with small probability and there is a small amount of noise in their private signals about the state, and the proof is similar to that of Theorem 2. When there are three or more states, we propose a new mechanism that has the same outcome function as the mechanism in the proof of Theorem 2 but has a different transfer function.

5.2 Uncertainty about the State Distribution

Our earlier proofs assume that the planner knows the objective state distribution and that this prior distribution is equal to both agents' prior belief before they take their actions—including their decision of whether to learn the state. In some applications, the planner may face uncertainty about the state distribution or about agents' beliefs about the state. This situation may arise, for instance, if the planner faces Knightian uncertainty about the state, or if each agent privately and freely observes a noisy signal about the state before deciding whether to pay an additional cost to learn θ and the planner does not know agents' information structures.

To model this situation, suppose that agent i's belief is $q_i \in \Delta(\Theta)$ when he decides whether to pay cost c_i in order to fully learn θ. We assume these beliefs are obtained as follows: agents have a prior belief q about θ, and form their respective interim beliefs q_1 and q_2 after receiving some informative signals. Intuitively, agent $i \in \{1, 2\}$ privately observes a signal s_i for free and his *interim belief* q_i is derived according to Bayes rule. We allow q_1 and q_2 to have arbitrary correlations as long as they satisfy the martingale condition $\mathbb{E}[q_1] = \mathbb{E}[q_2] = q$.

The planner knows neither q nor the realizations of q_1 and q_2. She only knows that q, q_1, and q_2 belong to a subset $\mathbf{q} \subset \Delta(\Theta)$. Our baseline model from earlier sections corresponds to the special case in which $\mathbf{q}$ is a singleton. In the more general formulation, the planner need not know the exact state distribution. Rather, she knows that this distribution belongs to some subset. This formulation also allows agents to have more information about the state relative to the planner, even before they decide whether to pay the cost and to learn the state. The planner does not know the agents' information structures but knows that their interim beliefs belong to a certain range. The planner's objective is to design a mechanism $\mathcal{M}$ that can robustly implement f for *all* $(q_1, q_2) \in \mathbf{q} \times \mathbf{q}$ and for *all* small enough ($\bar{c}$-bounded) perturbations.

Whether the planner can achieve her objective depends on $\mathbf{q}$, i.e., on the extent to which she knows the agents' interim beliefs. When $\mathbf{q}$ is larger, the robust implementation problem becomes harder. We say that $\mathbf{q}$ is *interior* if there exists $\tau > 0$ such that $q(\theta) > \tau$ for every $\theta \in \Theta$ and $q \in \mathbf{q}$. Let $B(q, \tau) \equiv \left\{ q' \in \Delta(\Theta) \middle| ||q' - q||_{TV} \leq \tau \right\}$ denote the τ-neighbourhood of q.

Proposition 2. *For any given social choice function $f : \Theta \to \Delta(Y)$:*

1. *Suppose $\mathbf{q}$ is interior. For every $\bar{c} > 0$, there exists a mechanism with n messages for each agent that robustly implements f for all $\bar{c}$-bounded perturbations.*

2. *For every generic $q \in \Delta(\Theta)$, there exists $\tau > 0$ such that if $\mathbf{q} \subset B(q, \tau)$, then there exists a mechanism with $2n - 1$ messages for each agent that robustly implements f.*

The proof is in Online Appendix D. Proposition 2 implies that even when (i) the planner does not know precisely what the objective state distribution is and (ii) agents may know more about the state than the planner does even before they pay the cost of learning, the desired social choice function is still robustly implementable as long as one of the two conditions is satisfied:

1. The planner is confident that agents' interim beliefs are not arbitrarily precise (i.e., assign probability close to 0 to some states) and agents' costs of learning are bounded from above.

2. The planner knows what the ex ante most likely state is and is confident that the signals freely received by the agents are sufficiently noisy.

Proposition 2 can be extended to the case in which $\mathbf{q}$ includes degenerate beliefs that assign probability 1 to some particular state. Nevertheless, we do need to rule out situations such as the following one: (i) $\Theta = \{\theta^1, \theta^2, \theta^3\}$, (ii) the planner knows that the agents can rule out one state for free before paying the information acquisition cost but, (iii) the planner does not know which state the agents rule out.

6 Stronger Notions of Robust Implementation

This section considers the robust implementation of *non-constant* social choice functions, defined as follows.

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

Section 6.1 shows that the planner *cannot* robustly implement any non-constant social choice function when we allow for perturbations where agents' payoffs do not coincide with those in the unperturbed environment with high probability. Sections 6.2 and 6.3 show that when agents' costs of learning in the unperturbed environment are above some cutoff, the planner *cannot* approximately implement any non-constant social choice function in all equilibria, and she *cannot* robust-partially implement any non-constant social choice function in an interim sense.

6.1 Impossibility of Global Implementation

First, suppose that perturbations for which $\widetilde{c}_i(\omega)$ is arbitrarily large are allowed, and that agents' payoffs may differ from those in the unperturbed environment with probability bounded away from zero. In this case, it is easy to see that no finite mechanism can approximately implement any non-constant social choice function. To this end, fix any finite mechanism $\mathcal{M}$. Clearly, no agent has any incentive to learn the state when agents' learning costs exceed the maximal transfer promised by mechanism $\mathcal{M}$ plus their maximal benefit from implementing specific outcomes. This implies that f *cannot* be implemented conditional on this event, which can occur with probability bounded above 0.

Next, we show that even when we only consider $\overline{c}$-bounded perturbations, or even when we only consider perturbations where it is common knowledge that agents' costs are c_1 and c_2, no finite mechanism can approximately implement any non-constant social choice function if the probability of normal types is not close to 1. To state the result formally, we will say that 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 $\max_{\theta\in\Theta} ||g_{\sigma(\mathcal{G})}(\theta) - f(\theta)||_{\mathrm{TV}} < \varepsilon$.

Theorem 3. *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.*

The proof is in Appendix B. Here we provide some general intuition. For every f that is non-constant, one can find $\theta \in \Theta$ such that $f(\theta)$ does not belong to the convex hull of $\{f(\theta')\}_{\theta'\neq\theta}$. For a mechanism $\mathcal{M}$ to implement f in a perturbation where all types of agent 1 dislike $f(\theta)$ and like outcomes in $\{f(\theta')\}_{\theta'\neq\theta}$, there must exist a distribution of agent 2's messages under which agent 1's payoff cannot exceed his payoff from $f(\theta)$ no matter which message he sends. This implies that under another perturbation where all types of agent 2 like $f(\theta)$, agent 2 can guarantee his payoff from $f(\theta)$ regardless of agent 1's message, which means that mechanism $\mathcal{M}$ cannot implement any outcome in $\{f(\theta')\}_{\theta' \neq \theta}$.

In fact, the proof of Theorem 3 implies the following corollary, which shows that even if one focuses on *virtual* implementation, no mechanism can virtually implement f when payoff perturbations have a probability that is bounded away from zero.

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 $\bar{c}$-bounded η-perturbation $\mathcal{G}$, such that for every equilibrium $\sigma(\mathcal{G})$ of the game $(\mathcal{M}, \mathcal{G})$, we have $\max_{\theta \in \Theta} ||g_{\sigma(\mathcal{G})}(\theta) - f(\theta)||_{TV} \geq \eta k(f)$.*

Corollary 1 shows that for every finite mechanism $\mathcal{M}$, there exists a perturbation $\mathcal{G}$ under which every equilibrium of the incomplete information game induced by $(\mathcal{M}, \mathcal{G})$ implements a social choice function that is bounded away from f. This corollary shows that, even if one focuses on partial and virtual implementation, robust implementation is possible only if the perturbed environment is *close* to the unperturbed environment.

6.2 Full Implementation and Virtual Implementation

We now examine whether the planner can approximately implement f in *all* equilibria under all small enough perturbations. Say that f is *virtually implementable* if for every $\varepsilon > 0$, there exists a mechanism $\mathcal{M}_\varepsilon$, such that $||g_\sigma(\theta) - f(\theta)||_{TV} \leq \varepsilon$ for every $\theta \in \Theta$ and every equilibrium σ under $\mathcal{M}_\varepsilon$.[16] Our first result provides two sufficient conditions under which every non-constant social choice function is not virtually implementable, even with no robustness concern.

Theorem 4. *Suppose f is non-constant.*

1. *If (u_1, u_2) do not depend on θ, then f is not virtually implementable.*

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

The proof, in Appendix C, shows that as long as c_1 and c_2 are above some cutoff $\bar{c} > 0$, even when the planner can use arbitrarily large transfers, there always exists an equilibrium where no

[16] Our definition of virtual implementation is similar to that of Abreu and Matsushima (1992) except that we require the desired outcome to be implemented in every Nash equilibrium while they require the desired outcome to be implemented in every rationalizable strategy. Since our goal is to show a negative result—every non-constant social choice function is *not* virtually implementable—using a stronger solution concept makes our result stronger.

agent learns the state. Intuitively, suppose agent 1's message does not depend on θ. Since agents' transfers depend only on the messages, the only incentive for agent 2 to learn θ is to induce a more favorable joint distribution of (θ, y) in order to increase $u_2(\theta, y)$. Therefore, agent 2's benefit from learning the state depends only on u_2. When agent 2's cost of learning outweighs this benefit from increasing $u_2(\theta, y)$, he has no incentive to learn provided that agent 1's message does not depend on θ, no matter how large the promised transfers are. This logic gives rise to equilibria where no agent learns the state and the implemented outcome is the same regardless of the state.

We also provide sufficient conditions under which the desired social choice function f can be robustly implemented in *all* equilibria. Say that f is *robust-fully implementable* if there exists a finite mechanism $\mathcal{M}$ such that (i) every equilibrium of $\mathcal{M}$ in the unperturbed environment implements f, and (ii) 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 $(\mathcal{M}, \mathcal{G})$.

First, when $c_1 = c_2 = 0$ and (f, u_1, u_2) satisfies Maskin monotonicity$*$, a condition that is strictly stronger than Maskin monotonicity, Chen, Kunimoto, Sun, and Xiong (2021) construct a finite mechanism that fully implements f under the solution concept of correlated rationalizability. Since the unique correlated equilibrium is robust in the sense of Kajii and Morris (1997), the mechanism in Chen, Kunimoto, Sun, and Xiong (2021) robust-fully implements f.

When $c_1, c_2 > 0$, f is robust-fully implementable when one of the agent's payoff function satisfies a strict version of Rochet (1987)'s cyclical monotonicity condition and that c_1 and c_2 are below some cutoff. Formally, (u_i, f) satisfies *strict cyclical monotonicity* if for every permutation $\xi : \Theta \to \Theta$, we have

$$\sum_{\theta \in \Theta} u_i(\theta, f(\theta)) \geq \sum_{\theta \in \Theta} u_i\Big(\theta, f(\xi(\theta))\Big), \tag{6.1}$$

with strict inequality for every ξ that satisfies $f(\xi(\theta)) \neq f(\theta)$ for some $\theta \in \Theta$. Condition (6.1) is the cyclical monotonicity condition. The strict inequality condition has no bite when f is constant, but can be violated when f is non-constant (e.g., when u_i does not depend on θ).

Theorem 5. *If (u_i, f) satisfies strict cyclical monotonicity for some $i \in \{1, 2\}$, then there exists $\overline{c} > 0$ such that when $c_i \leq \overline{c}$, there is a finite mechanism that robust-fully implements f.*

The proof is in Appendix D.

6.3 Interim Notion of Robust Implementation

Finally, we show that robust implementation in the interim sense is impossible when the costs of learning c_1 and c_2 lie above some cutoff that depends only on the unperturbed preferences u_1 and u_2, and which holds regardless of whether the planner can use arbitrarily large transfers.

We adapt the notion of interim robust implementation in Oury and Tercieux (2012) to our setting. Agents need to pay a cost to learn the state $\theta \in \Theta$. The planner knows the objective state distribution $q \in \Delta(\Theta)$ but faces uncertainty about agents' payoffs and costs of learning, and can use transfers to motivate the agents. Let Y be the set of outcomes, with $y \in Y$. Let Ω be a countable set of circumstances. Agent i's payoff is $\widetilde{u}_i(\omega, \theta, y) - \widetilde{c}_i(\omega)d_i + t_i$. Let $\omega^* \in \Omega$ be such that $\widetilde{u}_i(\omega^*, \theta, y) = u_i(\theta, y)$ and $\widetilde{c}_i(\omega^*) = c_i$ for every (θ, y) and $i \in \{1, 2\}$.

A *model* is denoted by $\mathcal{Z} \equiv (Z, \kappa)$ where $Z \equiv Z_1 \times Z_2$ is a countable type space and $\kappa_i(z_i) \in \Delta(\Omega \times Z_{-i})$ is the belief associated with type $z_i \in Z_i$. For two models $\mathcal{Z} \equiv (Z, \kappa)$ and $\mathcal{Z}' \equiv (Z', \kappa')$, $\mathcal{Z}' \subset \mathcal{Z}$ if $Z' \subset Z$ and $\kappa'_i(z'_i)[(\Omega \times Z'_{-i}) \cap E] = \kappa_i(z'_i)(E)$ for every $z'_i \in Z'_i$ and measurable event $E \subset \Omega \times Z_{-i}$. For each type z_i, one can compute his first-order belief (i.e., his belief about ω), his second-order belief (i.e., his belief about ω and the first-order belief of agent $-i$), and so on.[17] Let $h_i^k(z_i)$ denote the kth-order belief of type z_i. A sequence of types $\{z_i[n]\}_{n=0}^{+\infty}$ converges to type z_i (under the product topology) if for every $k \in \mathbb{N}$, $h_i^k(z_i[n])$ converges to $h_i^k(z_i)$ as $n \to +\infty$.

The model that corresponds to our unperturbed environment is denoted by $\mathcal{Z}^* = (Z^*, \kappa^*)$ where $Z^* = \{(z_1^*, z_2^*)\}$ and $\kappa_i^*(z_i^*)$ assigns probability 1 to $\omega = \omega^*$ and $z_{-i} = z_{-i}^*$. That is, $\omega = \omega^*$ is common knowledge in this model. A mechanism $\mathcal{M}$ *robustly implements* $f : \Theta \to \Delta(Y)$ *in the interim sense* if for every model $\mathcal{Z}$ with $\mathcal{Z}^* \subset \mathcal{Z}$, there is an equilibrium in the game induced by $(\mathcal{M}, \mathcal{Z})$ such that (i) f is implemented when agents' types are (z_1^*, z_2^*), and (ii) for every sequence of types in $\mathcal{Z}$ that converge to (z_1^*, z_2^*), the implemented social choice function converges to f.

Theorem 6. *For every (u_1, u_2) and non-constant f, there exists $\overline{c} > 0$ that depends only on (u_1, u_2) such that when $c_1, c_2 > \overline{c}$, no finite mechanism robustly implements f in the interim sense.*

The proof is in Appendix E.

[17] We omit the mathematical details of computing belief hierarchies. We refer readers to Weinstein and Yildiz (2007) and Oury and Tercieux (2012) for rigorous treatments.

7 Related Literature

Our paper contributes to the literature on robust implementation. We take an *ex ante* perspective and show that *all* social choice functions are robustly implementable under generic state distributions or under bounded costs of learning.[18] We also show that no non-constant social choice function is robustly implementable in the interim sense of Oury and Tercieux (2012) when agents' costs of learning are above some threshold, even if the planner can use unbounded transfers.

We require the desired outcome to be implemented with probability close to 1. This is related to the literature on virtual implementation such as Abreu and Matsushima (1992). They construct, for *each* ε, a mechanism that fully implements the desired outcome with probability more than $1-\varepsilon$. The number of messages in their mechanisms goes to infinity as $\varepsilon \to 0$. By contrast, we construct for *all* ε, a mechanism that partially implements the desired outcome with probability more than $1-\varepsilon$ when the perturbation on agents' preferences and costs of learning is small enough. The number of messages in our mechanism either equals the number of states n, or equals $2n-1$.

Kim (2021) proposes a monotonicity condition and shows that it is necessary for partial implementation in p-dominant strategies when the environment is quasi-linear. By contrast, we show that every social choice function is robustly implementable. This is because $\frac{1}{2}$-dominance is only sufficient but not necessary for equilibrium robustness in the sense of Kajii and Morris (1997). Indeed, although the truthful equilibrium in our mechanism is robust to small perturbations, it is not a $\frac{1}{2}$-dominant equilibrium in the game induced by our mechanism.

Our mechanisms can robustly elicit costly information when the planner (almost) knows agents' learning technologies but faces uncertainty about their payoffs as well as their beliefs and higher-order beliefs.[19] Our research question stands in contrast to Carroll (2019)'s which examines robust contracting when the planner faces uncertainty about the agent's information acquisition technology. The mechanisms we propose can robustly implement the desired social choice function when agents can either perfectly observe the state or observe signals that are highly correlated with the state. As explained in Propositions 1 and 2, our results do not require the planner to know the

[18] This echoes the findings in the literature on robust predictions in games. Weinstein and Yildiz (2007) show that an equilibrium is robust in the interim sense if and only if it is strictly dominant. Kajii and Morris (1997) provide sufficient conditions for an equilibrium to be robust in the ex ante sense, which are more permissible than the ones in Weinstein and Yildiz (2007). Oyama and Tercieux (2010) drop the common prior assumption and show that the two approaches become essentially equivalent in terms of the characterization of robust equilibrium outcomes.

[19] Our work is related to the literature on the optimal contracts for information acquisition. Zermeno (2011), Clark and Reggiani (2021), and Larionov, Pham and Yamashita (2021) examine the optimal contracts for information acquisition in fixed informational environments. By contrast, we examine whether it is possible to implement a desired social choice function in *all* nearby informational environments.

agents' interim beliefs and are robust to small trembles in agents' reporting strategies.[20]

Finally, our work is related to the literature on robust prediction in games (e.g., Rubinstein 1989, Kajii and Morris 1997, Weinstein and Yildiz 2007) and the literature on the robustness of equilibrium refinements (e.g., Fudenberg, Kreps and Levine 1988). Our notion of robust implementation builds on the notion of robust equilibrium in Kajii and Morris (1997), which is 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 do not directly depend on their messages, which are their actions in our mechanism design setting.

A Proofs of Theorems 1 and 2: General Utility Functions

We generalize the proofs of Theorems 1 and 2 to arbitrary $u_1(\theta, y)$, $u_2(\theta, y)$, c_1, and c_2.

Proof of Theorem 1: The outcome function is the same as the *Status Quo Rule with Ascending Transfers* in Section 4.1. Agents receive 0 transfer if their messages do not coincide. If both of them report message j, then agent i receives R_i^j which satisfies $R_i^1 \geq \frac{\overline{c}}{q(\theta_1)}$,

$$R_i^j + u_i(\theta^j, f(\theta^j)) - R_i^1 - u_i(\theta^j, f(\theta^1)) > 0 \text{ for every } j \geq 2 \tag{A.1}$$

$$\sum_{j=2}^{n} q(\theta^j)\Big\{R_i^j + u_i(\theta^j, f(\theta^j)) - R_i^1 - u_i(\theta^j, f(\theta^1))\Big\} > 2c_i. \tag{A.2}$$

We modify the first step of our proof in which we show that both agents using their truthful strategies is a γ-dominant equilibrium for some $\gamma < \frac{1}{2}$. The second and third steps remain the same. Let $\Sigma \equiv \{1, 2, ..., n\}^n$ and let

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

In the restricted game without perturbation where agents can only use strategies in $\Delta(\Sigma^*)$, they can only send message 1 conditional on $\theta = \theta^1$, and for every $j \in \{2, 3, ..., n\}$, agents send either message 1 or message j conditional on $\theta = \theta^j$

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

If agent 2 is truthful with probability at least $\frac{1}{2}$, then $\Pr(m_2 = j|\theta^j) \geq \frac{1}{2}$ and $\Pr(m_2 = 1|\theta^j) \leq \frac{1}{2}$. Hence, conditional on knowing that $\theta = \theta^j$, agent 1's expected payoff from sending message j is:

$$\Pr(m_2 = j|\theta^j)\Big(u_1(\theta^j, f(\theta^j)) + R^j\Big) + \Pr(m_2 = 1|\theta^j)u_1(\theta^j, f(\theta^1)),$$

[20] Our results are also related to the results in Chung and Ely (2003) and Aghion, Fudenberg, Holden, Kunimoto and Tercieux (2012), which examine the robustness of undominated strategy and subgame perfect implementation.

and his expected payoff from sending message 1 is:

$$\Pr(m_2 = j|\theta^j)u_1(\theta^j, f(\theta^1)) + \Pr(m_2 = 1|\theta^j)\Big(u_1(\theta^j, f(\theta^1)) + R^1\Big).$$

The former is greater than the latter if (A.1) is satisfied. Inequality (A.2) implies that agent i strictly prefers $(1, 2, ..., n)$ to $(1, 1, ..., 1)$ when he believes that agent $-i$ uses the truthful strategy with probability at least $\frac{1}{2}$. Hence, there exists $\gamma < \frac{1}{2}$ such that agent 1 strictly prefers $(1, 2, ..., n)$ to any other strategy that belongs to Σ^* when he believes that agent 2's strategy belongs to $\Delta(\Sigma^*)$ and agent 2 is truthful with probability at least γ. The second and third steps are not affected by u_1 and u_2, which remain the same as in Section 4.1. The expected cost of implementation $\sum_{i=1}^{2}\sum_{j=1}^{n} q(\theta^j)R_i^j$ can be as low as

$$\min_{\theta^*\in\Theta}\sum_{i=1}^{2}\Big\{\frac{\bar{c}}{q(\theta^*)} + 2c_i + \sum_{\theta\in\Theta} q(\theta)(u_i(\theta, f(\theta^*)) - u_i(\theta, f(\theta)))\Big\} \tag{A.3}$$

Hence, in order to lower the implementation cost, one needs to choose a status quo state θ^* that occurs with high ex ante probability but agents receive low utilities from outcome $f(\theta^*)$ when $\theta \neq \theta^*$.

Proof of Theorem 2: Without loss of generality, let $\Theta = \{\theta^1, ..., \theta^n\}$ with $q(\theta^1) > q(\theta^2) \geq ... \geq q(\theta^n) > 0$. The outcome function remains the same as before. The transfers are similar although we replace R^j with R_1^j and R_2^j for every $j \in \{0, 1, 2, ..., n\}$ such that for every $i \in \{1, 2\}$

$$D_i(j) \equiv R_i^j + u_i(\theta^j, f(\theta^j)) + \min_\tau u_i(\theta^j, f(\theta^\tau)) - R_i^1 - 2\max_\tau u_i(\theta^j, f(\theta^\tau)) > 0 \text{ for every } j \geq 2, \tag{A.4}$$

$$D_i(1) \equiv R_i^1 + u_i(\theta^1, f(\theta^1)) - R_i^0 - \max_\tau u_i(\theta^1, f(\theta^\tau)) > 0, \tag{A.5}$$

$$\sum_{j=2}^{n} q(\theta^j)D_i(j) > 2c_i, \quad \sum_{j=2}^{n} q(\theta^j)\Big(D_i(j) + R_i^1 - R_i^0\Big) + q(\theta^1)D_i(1) > 2c_i, \tag{A.6}$$

and

$$\frac{R_i^0}{R_i^j} > \frac{q(\theta^j)}{q(\theta^1)} \text{ for every } j \geq 2. \tag{A.7}$$

We modify the first step of our proof in which we show that both agents being truthful is a γ-dominant equilibrium for some $\gamma < \frac{1}{2}$. Consider a *restricted game without perturbation* 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 restricted game without perturbation, both agents using $(1, 2, ..., n)$ is a γ-dominant equilibrium for some $\gamma < \frac{1}{2}$. Suppose agent 2 is truthful with probability at least $\frac{1}{2}$,

- Conditional on $\theta = \theta^j$ for every $j \in \{2, 3, ..., n\}$. Agent 1's payoff when he sends j is at least $\frac{1}{2}(R_1^j + u_1(\theta^j, f(\theta^j))) + \frac{1}{2}\min_\tau u_1(\theta^j, f(\theta^\tau))$. His payoff when he sends 1 is at most $u_1(\theta^j, f(\theta^1)) + \frac{1}{2}R_1^1$, and his payoff when he sends any negative message is at most $\frac{1}{2}R_1^0 + \max_\tau u_1(\theta^j, f(\theta^\tau))$. Inequality (A.4) implies that his expected payoff is strictly greater when he sends message j.

- Conditional on $\theta = \theta^1$. Agent 1's payoff when he sends 1 is at least $u_1(\theta^1, f(\theta^1)) + \frac{1}{2}(R_1^1 + R_1^0)$ and his payoff when he sends any negative message is at most $R_1^0 + \frac{1}{2}u_1(\theta^1, f(\theta^1)) + \frac{1}{2}\max_\tau u_1(\theta^1, f(\theta^\tau))$. Inequality (A.5) implies that his expected payoff is strictly greater when he sends message 1.

The above discussion implies that agent 1 prefers to be truthful compared to any other non-constant strategy that belongs to Σ^*. Inequality (A.6) implies that he prefers to be truthful to any constant strategy that belongs to Σ^*. Since agent 1 has a strict incentive to be truthful when he believes that agent 2 is truthful with probability at least $\frac{1}{2}$, there exists $\gamma < \frac{1}{2}$ such that both agents being truthful is a γ-dominant equilibrium in the restricted game without perturbation. The second and third steps are not affected by u_1 and u_2, which remain the same as in Section 4.2.

We provide a tight lower bound on expected cost of implementation $\sum_{i=1}^2 \sum_{j=1}^n q(\theta^j) R_i^j$ given constraints (A.4), (A.5), (A.6), and (A.7). First, we bound R_i^1 from below. Inequality (A.6) implies that

$$2c_i < \sum_{j=2}^n q(\theta^j)\Big\{R_i^j - R_i^1 + \min_\tau u_i(\theta^j, f(\theta^\tau)) - 2\max_\tau u_i(\theta^j, f(\theta^\tau)) + u_i(\theta^j, f(\theta^j))\Big\}.$$

Inequality (A.7) implies that $R_i^0 q(\theta^1) > R_i^j q(\theta^j)$ for every $j \geq 2$, and using plugging in inequality (A.5) to substitute R_i^0 with R_i^1, we obtain:

$$2c_i < \sum_{j=2}^n (q(\theta^1) - q(\theta^j))R_i^1 + \sum_{j=2}^n q(\theta^j)\Big\{\min_\tau u_i(\theta^j, f(\theta^\tau)) - 2\max_\tau u_i(\theta^j, f(\theta^\tau)) + u_i(\theta^j, f(\theta^j))\Big\}$$
$$- \sum_{j=2}^n q(\theta^1)\Big\{\max_\tau u_i(\theta^1, f(\theta^\tau)) - u_i(\theta^1, f(\theta^1))\Big\}. \tag{A.8}$$

Inequality (A.8) leads to a tight bound on R_i^1. Next, we compute a tight lower bound on $\sum_{j=1}^n q(\theta^j) R_i^j$.

$$\sum_{j=1}^n q(\theta^j) R_i^j = R_i^1 + \sum_{j=2}^n q(\theta^j)(R_i^j - R_i^1)$$

$$= R_i^1 + \sum_{j=2}^n q(\theta^j) D_i(j) + \sum_{j=2}^n q(\theta^j)\Big\{2\max_\tau u_i(\theta^j, f(\theta^j)) - \min_\tau u_i(\theta^j, f(\theta^j)) - u_i(\theta^j, f(\theta^j))\Big\}$$

$$> 2c_i + R_i^1 + \sum_{j=2}^n q(\theta^j)\Big\{2\max_\tau u_i(\theta^j, f(\theta^j)) - \min_\tau u_i(\theta^j, f(\theta^j)) - u_i(\theta^j, f(\theta^j))\Big\}.$$

Plugging in the tight lower bound on R_i^1, we obtain a tight lower bound on the implementation cost. In the special case where $u_1 = u_2 = 0$ and $c_1 = c_2 = c$, the lower bound is given by (4.13).

B Proof of Theorem 3

For any finite mechanism $\mathcal{M} \equiv \{M_1, M_2, g, t_1, t_2\}$, let

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

be the highest transfer promised to any agent by $\mathcal{M}$. By definition, $X(\mathcal{M})$ exists. Recall that Y is the set of outcomes, $\Delta(Y)$ is the set of lotteries over outcomes, and $f(\theta) \in \Delta(Y)$. We use $\operatorname{co}(\cdot)$ to denote the convex hull of a set. Since f is non-constant, there exists $\theta^* \in \Theta$ such that

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

According to the separating hyperplane theorem, there exists $v : Y \to \mathbb{R}$ such that $v(f(\theta^*)) < \min_{y\in\mathcal{Y}} v(y)$.[21] 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 $\mathcal{G}^+$ in which $\widetilde{u}_1(\omega, \theta, y) = Cv(y)$ for all $(\omega, \theta) \in \Omega \times \Theta$. If $\mathcal{M}$ implements $f(\theta^*)$ in state θ^* under perturbation $\mathcal{G}^+$, there must exist $m_2^* \in \Delta(M_2)$ such that

$$\max_{m_1\in\Delta(M_1)} \Big\{Cv(g(m_1, m_2^*)) + t_1(m_1, m_2^*)\Big\} \le \underbrace{Cv(f(\theta^*)) + X(\mathcal{M})}_{\text{agent 1's highest possible payoff if the planner implements } f(\theta^*)}. \tag{B.1}$$

This is because otherwise, agent 1 can secure himself a payoff strictly greater than the right-hand-side of (B.1), in which case $f(\theta^*)$ cannot be implemented in any state under $\mathcal{G}^+$.

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

$$\min_{m_1\in\Delta(M_1)} \Big\{ - Cv(g(m_1, m_2^*)) + t_2(m_1, m_2^*)\Big\}. \tag{B.2}$$

Since we have chosen $C > 0$ in order to satisfy $\Big(\min_{y\in\mathcal{Y}} v(y) - v(f(\theta^*))\Big)C > 4X(\mathcal{M})$ and moreover, $X(\mathcal{M}) \ge |t_i(m_1, m_2)|$ for every i and (m_1, m_2), inequality (B.1) implies that

$$\min_{m_1\in\Delta(M_1)} \Big\{ - Cv(g(m_1, m_2^*)) + t_2(m_1, m_2^*)\Big\} \ge \underbrace{-Cv(f(\theta^*)) - 3X(\mathcal{M}) > -C\min_{y\in\mathcal{Y}}\{v(y)\} + X(\mathcal{M})}_{\text{since } \left(\min_{y\in\mathcal{Y}} v(y) - v(f(\theta^*))\right)C > 4X(\mathcal{M})}.$$

Therefore, agent 2 can secure a payoff strictly greater than $-C\min_{y\in\mathcal{Y}}\{v(y)\} + X(\mathcal{M})$, which implies that no outcome in $\mathcal{Y}$ can be implemented under perturbation $\mathcal{G}^-$. Hence, every finite mechanism $\mathcal{M}$ that can implement non-constant f under $\mathcal{G}^+$ cannot implement f under $\mathcal{G}^-$.

C Proof of Theorem 4

Suppose that u_1 and u_2 are independent of θ. Under any finite mechanism $\mathcal{M}$, there always exists an equilibrium in which both agents use state-independent strategies, since agents' preferences over messages are independent of the state regardless of the mechanism. In this equilibrium, the implemented outcome does not depend on the state, which means that for every non-constant f, there exists a state $\theta \in \Theta$ such that the implemented outcome is bounded away from $f(\theta)$.

Next, we show that f is not virtually implementable when c_1 and c_2 are above some cutoff $\overline{c}$ that depends only on (u_1, u_2). For every (u_1, u_2), let

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

[21] For every distribution over outcomes $\widetilde{y} \in \Delta(Y)$, we let $v(\widetilde{y})$ denote the expected value of $v(y)$ when y is distributed according to $\widetilde{y}$.

Fix any finite mechanism $\mathcal{M}$ and, for every $m_2 \in \Delta(M_2)$, let $T(m_2) \equiv \max_{m_1 \in M_1} t_1(m_1, m_2)$ be the maximal transfer received by agent 1 when agent 2's message is m_2. Suppose that agent 1 believes that agent 2's message is m_2 regardless of θ. Then, the difference between agent 1's expected payoff when he learns θ and when he does not learn θ is

$$\mathbb{E}\Big[\max_{m_1 \in M_1} \{u_1(\theta, g(m_1, m_2)) + t_1(m_1, m_2)\}\Big] - \max_{m_1 \in M_1} \mathbb{E}\Big[u_1(\theta, g(m_1, m_2)) + t_1(m_1, m_2)\Big]. \tag{C.1}$$

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 (C.1) 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.

Fix any finite mechanism $\mathcal{M}$ and 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 $\mathbb{E}_\theta\Big[u_i(\theta, g(m_1, m_2)) + t_i(m_1, m_2)\Big]$ when he uses strategy m_i and his opponent uses strategy m_{-i}. Since this auxiliary game is finite, 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 $\mathcal{M}$. This equilibrium implements a constant social choice function. For every non-constant social choice function f, there exists $\beta > 0$ such that for every constant social choice function g, there exists $\theta \in \Theta$ such that $||f(\theta) - g(\theta)||_{TV} > \beta$. This implies that f is not virtually implementable.

D Proof of Theorem 5

If f is constant, then robust-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 receives a transfer $t_i(\theta)$ for implementing $f(\theta)$. Under every η-perturbation $\mathcal{G}$, every normal type of agent i has a strict incentive to learn θ and to choose $f(\theta)$ in state θ for every $\theta \in \Theta$, provided that his cost of learning c_i is small enough. This implies that the above mechanism can robust-fully implement f.

E Proof of Theorem 6

As shown in the proof of Theorem 4, for every (u_1, u_2), there exists $\overline{c} > 0$ such that for *all* finite mechanisms (even when transfers can be arbitrarily large), when $c_i > \overline{c}$, agent i finds it strictly suboptimal to learn θ when he believes that agent $-i$'s message does not depend on θ. Suppose $c_1, c_2 > \overline{c}$. For every finite mechanism $\mathcal{M}$, let $\overline{U} \equiv 2\max_{\{i,\theta,y,m_1,m_2\}} \Big|u_i(\theta, y) + t_i(m_1, m_2)\Big|$. We construct a sequence of types that converge to (z_1^*, z_2^*) under the product topology but for which the implemented outcome is bounded away from f. Let $z_2[0]$ be a type that assigns probability 1 to ω' where $\widetilde{u}_1(\omega', \cdot, \cdot) = u_1(\cdot, \cdot)$, $\widetilde{u}_2(\omega', \cdot, \cdot) = u_2(\cdot, \cdot)$, $\widetilde{c}_1(\omega') = c_1$, and $\widetilde{c}_2(\omega') > \overline{U}$. Let $z_1[0]$ be a type that assigns probability β to player 2 being type $z_2[0]$ and probability $1-\beta$ to $\omega = \omega^*$, where β is close enough to 1 such that it is strictly suboptimal for type $z_1[0]$ to learn the state, regardless of his belief about type $z_2[0]$'s message. For every $j \geq 1$, let $z_2[j]$ be a type who knows that $\omega = \omega^*$ but assigns probability β to player 1 being type $z_1[j-1]$ and probability β to player 1 being type $z_1[j]$, and let $z_1[j]$ be a type who knows that $\omega = \omega^*$ but assigns probability β to player 2 being type $z_2[j]$ and probability $1-\beta$ to player 2 being type $z_2[j+1]$. These sequence of types converge to (z_1^*, z_2^*) under the product topology. By construction, type $z_2[0]$ finds it strictly suboptimal to learn θ, so his message is independent of θ. Type $z_1[j]$ finds it strictly suboptimal to learn θ since type $z_2[j]$'s message is independent of θ with probability at least β. Type $z_2[j]$ finds it strictly suboptimal to learn θ since type $z_1[j-1]$'s message is independent of θ with probability at least β. Therefore, the implemented outcome under this sequence of types is independent of θ, which is bounded away from f since f is non-constant.

F Proof of Proposition 1

First, we prove Proposition 1 when $u_1 = u_2 = 0$ and $c_1 = c_2$. In Online Appendix C, we extend our proof by allowing for arbitrary u_1, u_2, c_1, c_2. We rank the states according to their ex ante probabilities, i.e., $q(\theta^1) > q(\theta^2) \geq ... \geq q(\theta^n) > 0$, where the first strict inequality comes from our generic assumption. 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 given by:

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

The transfer functions when $u_1 = u_2 = 0$ and $c_1 = c_2 = c$ are given by:

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

$$t_2(m_1, m_2) = \begin{cases} R^j & \text{if } m_1 = m_2 = j \geq 1 \\ R^0 & \text{if } m_2 \leq 1 \text{ but } (m_1, m_2) \neq (1,1) \\ R^0 - x & \text{if } m_2 \geq 2 \text{ and } m_1 \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{F.3}$$

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

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

and

$$\frac{x}{R^j - R^0} > \frac{q(\theta^j)}{1 - q(\theta^j)} \text{ for every } j \in \{2, 3, ..., n\}. \tag{F.5}$$

When there are two states, our *Augmented Status Quo Rule with Modified Transfers* 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	R^0, R^0-x
2	R^0-x, R^0	R^0-x, R^0	R^2, R^2

When there are three states, our *Augmented Status Quo Rule with Modified Transfers* 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

Since $M \equiv \{-n, ..., -2\} \cup \{1\} \cup \{2, 3, ..., n\}$, 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 about the state is s_i^k. Hence, 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. This implies that agent i prefers m to m' as his intended message *if and only if* he receives a higher expected payoff when m is his realized message compared to when m' is his realized message. Let

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

In words, Σ_i^* is the set of pure strategies of agent i such that, conditional on each of agent i's private signal s_i^k, agent i intends to send either a negative message, or the status quo message 1, or message $h_i(s_i^k)$ that matches the meaning of his private signal. Agent i *intends to be truthful* if his strategy $(m^1, ..., m^{|S_i|})$ satisfies $m^k = h_i(s_i^k)$ for every $k \in \{1, ..., |S_i|\}$, i.e., agent i intends to send the message that matches the meaning of his private signal for each of his private signals.

First, we show that 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 $\Delta(\Sigma_2^*)$. 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 $h_1(s_1) = j$, if agent 1's realized message is j, then he receives an expected transfer of

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

  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 $\overline{\tau}$, and agent 2 intends to be truthful with probability at least $\frac{1}{2}$, we know that $\Pr(m_2 = j|s_1) \geq \frac{1-\tau}{2}(1-\overline{\tau})$ and $\Pr(m_2 \leq 1|s_1) \leq 1 - \frac{1-\tau}{2}(1-\overline{\tau})$. When $R^j - R^1 - x > \frac{2c}{q(\theta^j)}$, $\overline{\tau}$ is close to 0, and $\tau \leq \overline{\tau}$, we have

  $$q(\theta^j)\Big\{\Pr(m_2 = j|s_1)R^j + \Pr(m_2 \leq 1|s_1)(R^0 - x)\Big\} > q(\theta^j)\Big\{\Pr(m_2 = 1|s_1)R^1 + \Pr(m_2 \neq 1|s_1)R^0\Big\} + c.$$

Therefore, if agent 1 believes that agent 2's strategy belongs to $\Delta(\Sigma_2^*)$ and that agent 2 intends to be truthful with probability at least $\frac{1}{2}$, then agent 1 strictly prefers sending message j over sending the status quo message or any negative message whenever he receives a signal s_1 that satisfies $h_1(s_1) = j$. Moreover, this statement holds even after taking into account agent 1's cost of learning the state.

- Conditional on agent 1 receiving a signal s_1 such that $h_1(s_1) = 1$, his expected transfer when his realized message is 1 is $\Pr(m_2 = 1|s_1)R^1 + \Pr(m_2 < 0|s_1)R^0$ and his expected transfer when his realized message is negative is R^0. When $R^1 - R^0 > \frac{2c}{q(\theta^1)}$ and $\overline{\tau}$ is close enough to 0, $\Pr(m_2 = 1|s_1)R^1 + \Pr(m_2 < 0|s_1)R^0$ is at least $\frac{R^1+R^0}{2}$ given that agent 2 is truthful with probability at least $\frac{1}{2}$. Since $q(\theta^1)\Big(\frac{R^1+R^0}{2} - R^0\Big) > c$, agent 1 strictly prefers to send message 1 to any negative message when agent 2's strategy belongs to $\Delta(\Sigma_2^*)$ and agent 2 intends to be truthful with probability at least $\frac{1}{2}$, even taking into account his cost of learning c.

Since agent 1 strictly prefers to be truthful when agent 2's strategy belongs to $\Delta(\Sigma_2^*)$ and agent 2 intends to be truthful with probability at least $\frac{1}{2}$, there exists $\gamma < \frac{1}{2}$, such that both agents intending to be truthful is a γ-dominant equilibrium in the restricted game without perturbation.

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 restricted game with perturbation $\mathcal{G}$ 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 for every θ, outcome $f(\theta)$ is implemented with probability at least $1 - \varepsilon$.

In the third step, we show that $\sigma(\mathcal{G})$ remains an equilibrium in the game induced by our mechanism and perturbation $\mathcal{G}$ where agents can use *any strategy*, not restricted to strategies in $\Delta(\Sigma_1^*)$ and $\Delta(\Sigma_2^*)$. We consider two cases.

First, for any of agent 1's strategy $(m^1, ..., m^{|S_1|}) \notin \Sigma_1^*$ that is non-constant, let us define a new strategy $(m_*^1, ..., m_*^{|S_1|})$ that belongs to Σ_1^*:

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

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, $m_*^k = -m^k$. According to the mechanism's outcome function (F.1), $(m^1, ..., m^{|S_1|})$ and $(m_*^1, ..., m_*^{|S_1|})$ induce the same joint distribution of (θ, y). 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 $h_1(s_1) = j$, his expected transfer when his realized message $m \notin \{-n, ..., -2\} \cup \{1, j\}$ is:

$$\Pr(m_2 = m|s_1)R^m + \Pr(m_2 \leq 1|s_1)(R^0 - x). \tag{F.6}$$

Agent 1's expected transfer when his realized message is $-m$ is R^0. When agent 2's strategy belongs to $\Delta(\Sigma_2^*)$, he intends to send message m only if the meaning of his signal is m. When π is of size $\overline{\tau}$, we have $\Pr(m_2 = m|s_1) \leq 2\overline{\tau}$. If this is the case, the value of (F.6) is strictly less than R^0 when $\overline{\tau}$ is close to 0. This implies that every type of agent 1 prefers $(m_*^1, ..., m_*^{|S_1|})$ to $(m^1, ..., m^{|S_1|})$.

Second, for any strategy $(m^1, ..., m^{|S_1|}) \notin \Sigma_1^*$ that is a constant vector, there exists $k \in \{2, 3, ..., n\}$ such that $(m^1, ..., m^{|S_1|}) = (k, ..., k)$. Compare any given type of agent 1's expected payoff from strategies $(k, ..., k)$ and $(-k, ..., -k)$. These strategies induce the same joint distribution over (θ, y) and neither of them requires any cost of learning. In terms of the transfers, when agent 1's 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 $\Delta(\Sigma_2^*)$, agent 2 intends to send message k only when his signal has meaning k. Therefore,

$$\begin{aligned} & \Pr(m_2 = k)R^k + \Pr(m_2 \leq 1)(R^0 - x) \\ \leq & \Big(\pi(h_2(s_2) = k) + \big(1 - \pi(h_2(s_2) = k)\big)\overline{\tau}\Big)R^k + \big(1 - \pi(h_2(s_2) = k)\big)(1 - \overline{\tau})(R^0 - x) \end{aligned} \tag{F.7}$$

When π is of size $\overline{\tau}$ and $\overline{\tau}$ converges to zero, the right-hand-side of (F.7) converges to $q(\theta^k)R^k + (1 - q(\theta^k))(R^0 - x)$, which is strictly smaller than R^0 given our condition on the transfers (F.5). Therefore, the right-hand-side of (F.7) is strictly smaller than R^0 for all $\overline{\tau}$ close enough to 0. This implies that when agent 2 behaves according to $\sigma(\mathcal{G})$, every type of agent 1 receives a strictly greater transfer from strategy $(-k, ..., -k)$ to strategy $(k, k, ...k)$ for every $k \geq 2$.

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] Chen, Y., Mueller-Frank, M. Pai, M. (2022) "Continuous Implementation with Direct Revelation Mechanisms," *Journal of Economic Theory*, Vol. 201, 105422.

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

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

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

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

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

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

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

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

[14] Kim, D. (2021) "p-Dominant Implementation," Working Paper.

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

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

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

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

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

[20] Oyama, D., Tercieux, O. (2010) "Robust Equilibria under Non-Common Priors," *Journal of Economic Theory*, Vol. 145, pp. 752-784.

[21] Pavan, A., Segal, I. Toikka, J. (2014) "Dynamic Mechanism Design: A Myersonian Approach," *Econometrica*, Vol. 82, pp. 601-653.

[22] 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.

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

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

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

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

[27] Weinstein, J., Yildiz, M. (2007) "A Structure Theorem for Rationalizability with Application to Robust Predictions of Refinements," *Econometrica*, Vol. 75, pp. 365-400.

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