Online Appendix
Robust Implementation with Costly Information

Harry Pei[*] Bruno Strulovici[†]

August 17, 2023

A Extension: Robust Implementation with a Continuum of States

This appendix extends Theorems 1 and 2 to environments in which (i) there is a continuum of states, (ii) the state space Θ is compact, and (iii) agents' payoff functions in the unperturbed environment and the social choice function f are all continuous with respect to θ.

Formally, let Θ be a compact set in some normed vector space with norm $||\cdot||$. Let $q \in \Delta(\Theta)$ denote the objective distribution of θ, which we assume to have full support and no atom. A social choice function $f : \Theta \to \Delta(Y)$ is *continuous* if for every $\varepsilon > 0$, there exists $\delta > 0$ such that $||f(\theta) - f(\theta')||_{TV} \leq \varepsilon$ for every $||\theta - \theta'|| \leq \delta$. This definition corresponds to uniform continuity, which is equivalent to continuity since Θ is compact. The same comment applies to the continuity of agents' payoff functions introduced below.

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

We say that $u_i(\theta, y)$ is continuous with respect to θ if for every $y \in Y$ and $\varepsilon > 0$, there exists $\delta > 0$ such that $|u_i(\theta, y) - u_i(\theta', y)| \leq \varepsilon$ for every $||\theta - \theta'|| \leq \delta$. The notion of η-perturbation remains the same as in the baseline model, that is, agents' payoff functions coincide with those in the unperturbed environment with probability at least $1 - \eta$, and with complementary probability, they can have arbitrary preferences over state-contingent outcomes $\widetilde{u}_i(\omega, \theta, y)$, arbitrary costs of learning $\widetilde{c}_i(\omega)$, and arbitrary beliefs and higher-order beliefs about each other's preferences over outcomes and costs of learning as long as these beliefs can be derived from a common prior.

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

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

We emphasize that we do not require agent i's payoff in the perturbed environment $\widetilde{u}_i(\omega, \theta, y)$ to be continuous with respect to θ when type $Q_i(\omega)$ of agent i is not a normal type.

Corollary 1. *Suppose Θ is compact, q has full support and has no atom, and both f and (u_1, u_2) are continuous with respect to θ. For every $\varepsilon > 0$, there exist $\eta > 0$ and a finite mechanism $\mathcal{M}$ such that for every η-perturbation $\mathcal{G}$, there exists an equilibrium σ under $(\mathcal{M}, \mathcal{G})$ such that $\max_{\theta\in\Theta} ||g_\sigma(\theta) - f(\theta)|| \leq \varepsilon$.*

When there is a continuum of states and the objective state distribution has no atom, we can dispense the generic assumption on q in the case where Θ is a finite set. Intuitively, this is because we can always partition Θ into several connected subsets such that the probability of one of these subsets is strictly greater than the probability of every other subset.

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

1. Every partition element occurs with positive probability under q.

2. There exists a partition element that occurs with strictly higher probability compared to every other partition element.

3. For every pair θ, θ' that belong to the same partition element, we have $||f(\theta) - f(\theta')||_{TV} \leq \frac{\varepsilon}{2}$.[1]

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

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

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

Remark: When there is a continuum of states, in order to implement a social choice function that is ε-close to f, our mechanism requires each agent to have more messages as ε goes to zero. This is because there are two sources of approximation errors: one of them is caused by the perturbation on agents' preferences and beliefs, and the other one is caused by approximating f via $\widetilde{f}$. The second source of approximation error vanishes to zero when the partition on Θ becomes finer. In another word, the number of messages in the mechanism depends on our tolerance of approximation errors ε. This stands in contrast to environments with a finite number of states, in which the designer can robustly implement the desired social choice function using a mechanism where each agent has $2|\Theta| - 1$ messages, regardless of the required approximation error.

B General Information Acquisition Technologies

We extend our main result to environments in which the agents can choose any partition of the state space Θ as their information structures, and different partitions of the state space may have different costs. We start from describing the general environment.

Let Θ be a finite set of states and $q \in \Delta(\Theta)$ denote the prior distribution of θ. Let Y denote the set of outcomes. The designer commits to a mechanism $\mathcal{M} \equiv \{M_1, M_2, t_1, t_2, g\}$, where M_i is a finite set of messages for agent $i \in \{1, 2\}$, $t_i : M_1 \times M_2 \to \mathbb{R}$ is the transfer to agent i, and $g : M_1 \times M_2 \to \Delta(Y)$ is the implemented outcome.

After observing $\mathcal{M}$, agents simultaneously and independently decide what information to acquire. Each agent can choose any partition of Θ as his information structure. Let $\mathcal{P}$ be the set of partitions of Θ. Let $P_i \in \mathcal{P}$ denote the partition chosen by agent i. Let P^* denote the finest partition. Agent $i \in \{1, 2\}$ observes the element of P_i the realized θ belongs to and sends a message $m_i \in M_i$. The designer makes transfers and implements an outcome according to $\mathcal{M}$. Agent i's payoff is:

$$u_i(\theta, y) + t_i - c_i(P_i), \tag{B.1}$$

where $c_i : \mathcal{P} \to [0, +\infty)$ is agent i's information acquisition cost function. A *perturbation* is characterized by

$$\mathcal{G} \equiv \Big\{\Omega, \Pi, (Q_i)_{i\in\{1,2\}}, (\widetilde{u}_i)_{i\in\{1,2\}}, (\widetilde{c}_i)_{i\in\{1,2\}}\Big\},$$

where Ω is a countable set of *circumstances*, whose typical element is denoted by $\omega \in \Omega$, $\Pi \in \Delta(\Omega)$ is the distribution of ω, and is assumed to be independent of θ, and Q_i is agent i's information partition on Ω. For every $\omega \in \Omega$, let $Q_i(\omega)$ denote the partition element of Q_i that contains ω. Agent i's payoff function is given by

$$\widetilde{u}_i(\omega, \theta, y) + t_i - \widetilde{c}_i(\omega, P_i). \tag{B.2}$$

Type $Q_i(\omega)$ is a *normal type* if $\widetilde{u}_i(\omega', \theta, y) = u_i(\theta, y)$ and $\widetilde{c}_i(\omega, P_i) = c_i(P_i)$ for every $\omega' \in Q_i(\omega)$. For every $\eta > 0$, we say that $\mathcal{G}$ is an η-perturbation if the probability of the event that both agents are normal is at least $1-\eta$. For every $\eta > 0$ and $\overline{c} > 0$, we say that $\mathcal{G}$ is a $\overline{c}$-bounded η-perturbation if it is an η-perturbation where $\widetilde{c}_i(\omega, P^*) \leq \overline{c}$ for every $i \in \{1,2\}$ and $\omega \in \Omega$.

For any $f : \Theta \to \Delta(Y)$, we describe a mechanism where each agent has n messages that can robustly implement f for all $\overline{c}$-bounded perturbations. We focus on the case where $u_1 = u_2 = 0$ since generalizing the proof to arbitrary u_1 and u_2 resembles the arguments in Appendix A.

Let $M_1 = M_2 = \{1, 2, ..., n\}$. The outcome function $g : M_1 \times M_2 \to \Delta(Y)$ is given by:

- $g(j,j) = f(\theta^j)$ for every $j \in \{1, 2, ..., n\}$.
- $g(i,j) = g(j,i)$ for every i, j.
- For every $j > i$, $g(j,i) = \sum_{k=1}^{j-1} \frac{1}{j-1} g(k,i)$, i.e., when agent 2 reports i, agent 1 reporting j and reporting 1 to $j-1$ uniformly at random lead to the same distribution over outcomes.

The last step of the construction also implies that for every $j > i$, when agent 2 reports i, agent 1 reporting j and reporting 1 to i uniformly at random lead to the same distribution over outcomes. The transfer function to agent $i \in \{1,2\}$ is given by

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

where $R_i^1, ..., R_i^n$ satisfy $R_i^j > R_i^{j-1} + \frac{2c_i(P^*)}{q(\theta^j)}$ for every $j \geq 2$, and $R_i^1 \geq \frac{(n-1)\overline{c}}{\min_{\theta \in \Theta} q(\theta)}$.

Step 1: Let $\Sigma \equiv M^n$ be the set of strategies, with $(1, 2, ..., n) \in \Sigma$ the *truthful strategy*. Let

$$\Sigma^* \equiv \Big\{ (m^1, ..., m^n) \in \Sigma \text{ such that } m^j \leq j \text{ for every } j \in \{1, 2, ..., n\} \Big\}. \tag{B.3}$$

Intuitively, Σ^* is the set of strategies where agent's report does not exceed the index of the state. We show that there exists $\gamma < 1/2$ such that in the auxiliary game where agents' payoffs are

$\{t_1 - c_1(P_1), t_2 - c_2(P_2)\}$ and both agents are only allowed to choose strategies supported in Σ^*, then both agents being truthful is a γ-dominant equilibrium. To see this, for every $i \in \{1,2\}$, suppose agent i believes that

1. agent $-i$'s strategy is supported in Σ^*,
2. agent $-i$ plays his truthful strategy $(1,2,...,n)$ with probability at least $1/2$.

Since $R_i^j > R_i^{j-1} > ... > R_i^1$ and $R_i^j > R_i^{j-1} + \frac{2c_i(P^*)}{q(\theta^j)}$, we know that conditional on the state being θ^j, agent i's expected transfer from reporting message j is strictly greater than his expected transfer from reporting any message strictly lower than j, and this difference in expected transfer is strictly greater than $c_i(P^*)$. When agent i is only allowed to report truthfully or to report a lower state, he strictly prefers his truthful strategy $(1,2,...,n)$ to any other strategy in Σ^*. Since Θ is finite, there exists $\gamma < 1/2$ such that both agents being truthful is a γ-dominant equilibrium in the auxiliary game.

Step 2: For any perturbation $\mathcal{G}$, consider a *perturbed auxiliary game* where agent i's payoff is $\widetilde{u}_i(\omega,\theta,y) + t_i - \widetilde{c}_i(\omega,P_i)$ and both agents are only allowed to use strategies supported in Σ^*. The critical path lemma in Kajii and Morris (1997) implies that for very $\varepsilon > 0$, there exists $\eta > 0$, such that for every η-perturbation $\mathcal{G}$, there exists an equilibrium $\sigma(\mathcal{G})$ in the perturbed auxiliary game where the probability with which both agents using the truthful strategy is at least $1-\varepsilon$. Since $g(j,j) = f(\theta^j)$ for every $j \in \{1,2,...,n\}$, social choice function f is implemented with probability more than $1-\varepsilon$ when agents behave according to $\sigma(\mathcal{G})$.

Step 3: We show that $\sigma(\mathcal{G})$ remains an equilibrium when both agents are allowed to choose any strategy supported in Σ, not only strategies that are supported in Σ^*. Suppose by way of contradiction that type $Q_1(\omega)$ strictly prefers some strategy $(m^1,...,m^n) \notin \Sigma^*$ to all strategies supported in Σ^*. Assuming that agent 2 behaves according to $\sigma(\mathcal{G})$, which means that his strategy is supported in Σ^*, we compare type $Q_1(\omega)$'s expected payoff from $(m^1,...,m^n)$ to his expected payoff from the following mixed strategy $(m_\dagger^1,...,m_\dagger^n)$, where

- if $m^j \leq j$, then $m_\dagger^j = m^j$;
- if $m^j > j$, then $m_\dagger$ is the mixed strategy of reporting $\{1,2,...,j\}$ each with probability $\frac{1}{j}$.

One can verify that $(m_\dagger^1,...,m_\dagger^n)$ is supported in Σ^*, and furthermore, as long as player 2's strategy is supported in Σ^*, the implemented outcome is the same no matter whether agent 1 uses strategy

$(m^1, ..., m^n)$ or strategy $(m^1_\dagger, ..., m^n_\dagger)$. In addition, for every j such that $m^j > j$, $m^j_\dagger$ attaches strictly positive probability to every element in the set $\{1, 2, ..., j\}$. Hence, by reporting $m^j_\dagger$ instead of m^j in state θ^j, agent 1's expected transfer increases by at least $\frac{q(\theta_j)R^1_1}{n-1}$. Type $Q_1(\omega)$ prefers $(m^1_\dagger, ..., m^n_\dagger)$ to $(m^1, ..., m^n)$ if $\frac{q(\theta_j)R^1_1}{n-1} > \overline{c}$. Hence, for every perturbation $\mathcal{G}$, the equilibrium in the auxiliary perturbed game $\sigma(\mathcal{G})$ remains an equilibrium when both agents are allowed to choose any strategy supported in Σ, which implies that our mechanism robustly implements f.

C Robustness to Trembles and Noisy Information

The proofs of Theorems 1 and 2 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 designer 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 $S_i \equiv \{s^1_i, ..., s^{|S_i|}_i\}$ 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

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

and

$$\sum_{j=1}^{n} \sum_{s_i \in \{h_i(s_i)=j\}} \pi(\theta^j, s_i) \geq 1 - \overline{\tau}. \tag{C.2}$$

Our first requirement is that the marginal distribution on θ be consistent with the objective state distribution q. Our second requirement is that 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 designer 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 designer 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* $\max_{\theta \in \Theta} ||g_{\sigma(\mathcal{G})}(\theta) - f(\theta)||_{TV} < \varepsilon$.

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.

C.1 Proof of Proposition 1

First, we prove Proposition 1 when $u_1 = u_2 = 0$ and $c_1 = c_2$. We explain later how to extend our proof to 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:

$$g(m_1, m_2) = \begin{cases} f(\theta^{|m_1|}) & \text{if } |m_1| = |m_2| \\ f(\theta^1) & \text{otherwise} \end{cases} \tag{C.3}$$

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

$$t_1(m_1, m_2) = \begin{cases} R^j & \text{if } m_1 = m_2 = j \geq 1 \\ R^0 & \text{if } m_1 \leq 1 \text{ but } (m_1, m_2) \neq (1, 1) \\ R^0 - x & \text{if } m_1 \geq 2 \text{ and } m_2 \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{C.4}$$

$$t_2(m_1, m_2) = \begin{cases} R^j & \text{if } m_1 = m_2 = j \geq 1 \\ R^0 & \text{if } m_2 \leq 1 \text{ but } (m_1, m_2) \neq (1, 1) \\ R^0 - x & \text{if } m_2 \geq 2 \text{ and } m_1 \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{C.5}$$

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

$$R^1 - R^0 > \frac{2c}{q(\theta^1)}, \quad R^j - R^1 - x > \frac{2c}{q(\theta^j)} \text{ for every } j \in \{2, 3, ..., n\}, \tag{C.6}$$

and

$$\frac{x}{R^j - R^0} > \frac{q(\theta^j)}{1 - q(\theta^j)} \text{ for every } j \in \{2, 3, ..., n\}. \tag{C.7}$$

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

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

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 (C.10), $(m^1, ..., m^{|S_1|})$ 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:

$$\Pr(m_2 = m|s_1)R^m + \Pr(m_2 \leq 1|s_1)(R^0 - x). \tag{C.8}$$

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 (C.8) 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,

$$\Pr(m_2 = k)R^k + \Pr(m_2 \leq 1)(R^0 - x)$$

$$\leq \Big(\pi(h_2(s_2)=k) + \big(1-\pi(h_2(s_2)=k)\big)\overline{\tau}\Big)R^k + \big(1-\pi(h_2(s_2)=k)\big)(1-\overline{\tau})(R^0-x) \tag{C.9}$$

When π is of size $\overline{\tau}$ and $\overline{\tau}$ converges to zero, the right-hand-side of (C.9) 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 (C.7). Therefore, the right-hand-side of (C.9) 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 $k \geq 2$.

Extension to Arbitrary (u_1, u_2, c_1, c_2) : We extend the proof of Proposition 1 to general utility functions u_1 and u_2 and general learning costs c_1 and c_2. Consider the mechanism whose outcome function is given by:

$$g(m_1, m_2) = \begin{cases} f(\theta^{|m_1|}) & \text{if } |m_1| = |m_2| \\ f(\theta^1) & \text{otherwise,} \end{cases} \tag{C.10}$$

and whose transfer functions are given by:

$$t_1(m_1, m_2) = \begin{cases} R^j & \text{if } m_1 = m_2 = j \geq 1 \\ R^0 & \text{if } m_1 \leq 1 \text{ but } (m_1, m_2) \neq (1,1) \\ R^0 - x & \text{if } m_1 \geq 2 \text{ and } m_2 \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{C.11}$$

$$t_2(m_1, m_2) = \begin{cases} R^j & \text{if } m_1 = m_2 = j \geq 1 \\ R^0 & \text{if } m_2 \leq 1 \text{ but } (m_1, m_2) \neq (1,1) \\ R^0 - x & \text{if } m_2 \geq 2 \text{ and } m_1 \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{C.12}$$

where the parameters $\{R^n, ..., R^1, R^0, x\}$ satisfy $R^n, ..., R^0 > x > \frac{\max\{c_1, c_2\}}{q(\theta^n)}$

$$R^1 - R^0 > \frac{2\max\{c_1, c_2\}}{q(\theta^1)} + 2\max_{i\in\{1,2\}}\Big\{\max_{y\in Y} u_i(\theta^1, y) - \min_{y\in Y} u_i(\theta^1, y)\Big\}, \tag{C.13}$$

$$R^j - R^1 - x > \frac{2\max\{c_1, c_2\}}{q(\theta^j)} + 2\max_{i\in\{1,2\}}\Big\{\max_{y\in Y} u_i(\theta^j, y) - \min_{y\in Y} u_i(\theta^j, y)\Big\} \text{ for every } j \in \{2, 3, ..., n\}, \tag{C.14}$$

and

$$\frac{x}{R^j - R^0} \geq \frac{q(\theta^j)}{1 - q(\theta^j)} \text{ for every } j \in \{2, 3, ..., n\}. \tag{C.15}$$

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

Recall that each agent has $2n-1$ messages and that we are considering a *restricted game without perturbation* where for every $i \in \{1,2\}$, agent i is only allowed to use strategies that belong to $\Delta(\Sigma_i^*)$ where

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

Suppose agent 1 believes that agent 2 intends to be truthful with probability at least $\frac{1}{2}$,

- For every $j \geq 2$, conditional on agent 1 receiving a signal $s_1 \in S_1$ that satisfies $h_1(s_1) = j$, 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 $\overline{\tau}$, we have $\Pr(m_2 = j|s_1) \geq \frac{1-\tau}{2}(1-\overline{\tau})$ and $\Pr(m_2 = 1|s_1) \leq 1 - \frac{1-\tau}{2}(1-\overline{\tau})$. When inequality (C.14) is satisfied, $\overline{\tau}$ is close to 0, and $\tau \leq \overline{\tau}$, we have $q(\theta^j)\Big(\Pr(m_2 = j|s_1)R^j + \Pr(m_2 \leq 1|s_1)(R^0 - x)\Big) - q(\theta^j)\Big(\Pr(m_2 = 1|s_1)R^1 + \Pr(m_2 \neq 1|s_1)R^0\Big) > \max\{c_1, c_2\} + q(\theta^j)\Big\{\max_{y,y' \in Y} u_1(\theta^j, y) - u_1(\theta^j, y')\Big\}$. Therefore, agent 1 strictly prefers to send message j when he receives any signal $s_1 \in S_1$ that satisfies $h_1(s_1) = j$ when he believes that agent 2 intends to be truthful with probability at least $\frac{1}{2}$.

- Conditional on agent 1 receiving a message s_1 that satisfies $h_1(s_1) = 1$, 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 inequality (C.13) is satisfied and $\overline{\tau}$ is close enough to 0, agent 1 prefers to message 1 as his intended message to any negative message as his intended message when he believes that 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 has a strict incentive to be truthful when he believes that agent 2 intends to be truthful with probability at least $\frac{1}{2}$, there exists $\gamma < \frac{1}{2}$ such that he also has a strict incentive to do so when he believes that agent 2 intends to be truthful with probability at least γ. Therefore, both agents intending to be truthful is a γ-dominant equilibrium.

D Uncertainty about the State Distribution

Our earlier proofs assume that the designer 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 designer may face uncertainty about the state distribution or about agents' beliefs about the state. This situation may arise, for instance, if the designer 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 designer 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.

The designer 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 designer 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 designer, even before they decide whether to pay the cost and to learn the state. The designer does not know the agents' information structures but knows that their interim beliefs belong to a certain range. The designer'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 designer 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*.*

Proposition 2 implies that even when (i) the designer does not know precisely what the objective state distribution is and (ii) agents may know more about the state than the designer 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 designer 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 designer 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 designer knows that the agents can rule out one state for free before paying the information acquisition cost but, (iii) the designer does not know which state the agents rule out.

D.1 Proof of Proposition 2

We show statement 1 focusing on the case where $u_1 = u_2 = 0$ and $c_1 = c_2 = c$. Extending the proof to general (u_1, u_2) and heterogeneous learning costs is analogous to the generalization in Appendix A of the main text, and modifying the proof of Theorem 2 to show Statement 2 follows a similar argument to the one given here. The details are available upon request.

Our proof uses Proposition 5.5 in Oyama and Tercieux (2010), that generalizes the critical path lemma in Kajii and Morris (1997) to environments with non-common priors. Let $\mathbf{q}_i$ denote the set of interim beliefs of player $i \in \{1, 2\}$. Player i's *pure strategy* is

$$\{m_i^1(q_i), ..., m_i^n(q_i)\}_{q_i \in \mathbf{q}_i}, \tag{D.1}$$

where $m_i^j(q_i)$ is the message he sends when his interim belief is q_i and the state is θ^j. Let Σ_i denote the set of pure strategies for agent i. Let $\Sigma_i^* \subset \Sigma_i$ be such that a strategy belongs to Σ_i^*

if and only if $m_i^j(q_i) \in \{1, j\}$ for every $j \in \{1, 2, ..., n\}$ and $q_i \in \mathbf{q}_i$. Agent i's strategy is *truthful* if $m_i^j(q_i) = j$ for every $j \in \{1, 2, ..., n\}$ and $q_i \in \mathbf{q}_i$. Consider the status quo rule with ascending transfers constructed in Section 4.1 of the main text where the parameters $R^n, ..., R^1$ satisfy

$$R^j > R^1 \text{ for every } j \geq 2, \tag{D.2}$$

$$\sum_{j=2}^{n} (R^j - R^1) q(\theta^j) > 2c \text{ and } R^1 q(\theta^1) \geq \bar{c} \text{ for every } q \in \mathbf{q}. \tag{D.3}$$

Such $R^n, ..., R^1$ exist when $\mathbf{q}$ is interior. The rest of the proof is similar to that of Theorem 1.

First, let us examine the restricted game without any perturbation where for every $i \in \{1, 2\}$, agent i is only allowed to choose strategies in $\Delta(\Sigma_i^*)$. If agent i believes that agent j is truthful with probability at least $\frac{1}{2}$, then conditional on each q_i, the expected transfer he receives is strictly greater when he uses his truthful strategy. Hence, there exists $\gamma < \frac{1}{2}$ such that both agents being truthful is a γ-dominant equilibrium. Next, let us consider the restricted game with perturbation $\mathcal{G}$. 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 $\sigma(\mathcal{G})$ induced by $(\mathcal{M}, \mathcal{G})$ in which both agents use their truthful strategies with probability more than $1 - \varepsilon$. In this equilibrium, f is implemented with probability more than $1 - \varepsilon$. In the last step, let us consider the unrestricted game with perturbation. Similar to the proof of Theorem 1, the second part of (D.3) implies that $\sigma(\mathcal{G})$ remains an equilibrium when agents can use any strategies in Σ_i, not just those in Σ_i^*. This verifies that our mechanism can robustly implement f for every $(q_1, q_2) \in \mathbf{q} \times \mathbf{q}$.

E Failure of Majority Rule

Consider an example with three agents. Let $\Theta = \{\theta^1, \theta^2\}$. Players' prior belief about θ is $q(\theta^1) = 2/3$ and $q(\theta^2) = 1/3$. Let $Y = \{y^1, y^2\}$ and $u_i(\theta, y) = 0$ for every $i \in \{1, 2, 3\}$ and $(\theta, y) \in \Theta \times Y$. Let $c > 0$ be agents' costs of learning. The designer would like to implement y^j in state θ^j.

A mechanism is a *majority rule* if there exists $T > 0$ such that (i) $M_1 = M_2 = M_3 = \{1, 2\}$, (ii) $g(m_1, m_2, m_3) = y^i$ if and only if there exist $1 \leq k < j \leq 3$ such that $m_k = m_j = i$, and (ii) $t_i(m_1, m_2, m_3) = T \cdot \mathbf{1}\{m_i = m_j \text{ for some } j \neq i\}$. Intuitively, each agent votes for the two alternatives. The alternative that receives at least two votes gets implemented and an agent is given a transfer T *if and only if* he voted for the alternative favored by the majority.

First, if there is no robustness concern, then for every $c > 0$, there exists $\underline{T} > 0$ such that a majority rule with $T > \underline{T}$ can partially implement the desired social choice function. This is because there exists an equilibrium where all agents learn the state and report their findings truthfully.

Next, we show that no majority rule can *robustly* implement the desired social choice function. For any majority rule parameterized by T, consider a perturbation where $\Omega \equiv \{\omega_0, \omega_1, ...\}$ with $\Pi(\omega_n) = (1-\delta)\delta^n$. Agents 1 and 2's information partitions are the same, given by $\{\omega_0\}, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}$,... and agent 3's information partition is $\{\omega_0, \omega_1\}, \{\omega_2, \omega_3\}, ...$ Agents are normal types except for type $\{\omega_0\}$ of agents 1 and 2, in which case their utility functions are $u_i(\omega_0, \theta, y) = B\mathbf{1}\{y = y^1\}$ and their cost of learning is C, where B is large enough relative to T such that

$$\left(\frac{B+T}{B+2T}\right)^2 > \delta \tag{E.1}$$

and $C >> B$. Since $C >> B$, types $\{\omega_0\}$ of agents 1 and 2 will not learn the state, so their strategies in any equilibrium must be supported in $\{(1,1),(2,2)\}$. Type $\{\omega_0\}$ of agent 1 strictly prefers $(1,1)$ to $(2,2)$ as long as he believes that type $\{\omega_0\}$ of agent 2 plays $(1,1)$ with probability more than $\frac{T}{B+2T}$. The same is true for type $\{\omega_0\}$ of agent 2. This implies that in any equilibrium, either types $\{\omega_0\}$ of agents 1 and 2 play $(1,1)$ for sure, or both types $\{\omega_0\}$ of agents 1 and 2 play $(2,2)$ with probability at least $\frac{B+T}{B+2T}$. We consider these two cases separately.

First, consider the case in which types $\{\omega_0\}$ of agents 1 and 2 play $(1,1)$ for sure. Type $\{\omega_0, \omega_1\}$ of agent 3 strictly prefers $(1,1)$ to all other strategies since the probability that his report matches the majority is strictly more than $1/2$ when he reports 1, conditional on every state. Type $\{\omega_1, \omega_2\}$ of agent $i \in \{1,2\}$ strictly prefers $(1,1)$ to all other strategies since the probability that his report matches the majority is strictly more than $1/2$ when he reports 1, conditional on every state... Iterate this process, we know that in equilibrium, all types of all agents play $(1,1)$ and outcome y^1 is implemented regardless of the state.

Second, consider the case in which types $\{\omega_0\}$ of agents 1 and 2 play $(2,2)$ with probability at least $\frac{B+T}{B+2T}$. Given inequality (E.1), type $\{\omega_0, \omega_1\}$ of agent 3 strictly prefers $(2,2)$ to all other strategies since the probability that his report matches the majority is strictly more than $1/2$ when he reports 1, conditional on every state. Type $\{\omega_1, \omega_2\}$ of agent $i \in \{1,2\}$ strictly prefers $(2,2)$ to all other strategies since the probability that his report matches the majority is strictly more than $1/2$ when he reports 1, conditional on every state... Iterate this process, we know that in equilibrium, all normal types of all agents play $(2,2)$ and outcome y^2 is implemented regardless of the state conditional on all agents being the normal type.

References

[1] Kajii, A., Morris, S. (1997) "The Robustness of Equilibria to Incomplete Information," *Econometrica*, 65(6), 1283-1309.

[2] Oyama, D. and Tercieux, O. (2010) "Robust Equilibria under Non-Common Priors," *Journal of Economic Theory*, 145(2), 752-784.