DISCUSSION PAPER SERIES

Discussion paper No.299

Mechanism Design with Private Communication to Neutralize Fairness Constraints

Kohei Daido
(Kwansei Gakuin University)

Tomoya Tajika
(Nihon University)

September 2025

SCHOOL OF ECONOMICS

KWANSEI GAKUIN UNIVERSITY

1-155 Uegahara Ichiban-cho
Nishinomiya 662-8501, Japan

Mechanism Design with Private Communication to Neutralize Fairness Constraints

Kohei Daido* Tomoya Tajika†

September 10, 2025

We study mechanism design under auditable fairness mandates that constrain only the formal rule while allowing off-record private communication between the principal and agents. We model a two-layer environment: a formal rule that maps agents' reports to outcomes and must satisfy the mandate, and private advice in which the principal can provide type-contingent recommendations. We construct a format-preserving randomized encryption (FPRE): the principal randomizes over symmetry-constrained rules and pairs each realization with "password"-like advice. Under FPRE, any Bayesian incentive-compatible social choice function (SCF) is implementable by symmetric formal rules; if the SCF is dominant-strategy incentive-compatible (DSIC), the resulting mechanism achieves DSIC. In contrast, constraints that embed predictable structures—such as strict monotonicity and continuity—cannot be neutralized. We also present an approximate version: continuity is compatible with it. Our results highlight a regulatory-scope insight: if auditors can verify only the format of the rule, format-type fairness does not bind, whereas structure-revealing mandates (i.e., strict monotonicity and continuity) hinder the "encryption" that sustains obedience to private advice.

JEL Classification: C72; D82; D86

Keywords: mechanism design; symmetry; fairness; implementation; private communication; randomized encryption; dominant strategies; continuous rules

*Kwansei Gakuin University, email: daido@kwansei.ac.jp
†Nihon University, email: tajika.tomoya@nihon-u.ac.jp

1. Introduction

In many applications of mechanism design, the designer must comply with *auditable* fairness mandates—most prominently, symmetry, yet outcome-optimal social choice rules can be asymmetric even in symmetric environments (e.g., Kotowski, 2018). Such mandates are typically legal or procedural requirements that can be verified from the formal description of the rule, but they place no restrictions on off-record communication. This paper asks a first-order question: *when fairness constraints apply only to the auditable formal rule, which constraints truly bind implementability of a target social choice function (SCF), and which can be neutralized without violating the format?*

We model a two-layer environment. The *formal rule* is what can be audited; it maps a profile of reported types to an outcome and must satisfy a given constraint class (e.g., symmetry, surjectivity, continuity, monotonicity). In addition, the principal can privately communicate with each agent before the formal rule is executed. Our baseline assumes that the type space is standard Borel and has continuum cardinality, and the existence of symmetric worst outcomes.

We construct a *format-preserving randomized encryption* (FPRE): the principal draws a random seed, uses it to select a formal rule within the constrained class, and privately sends type-contingent *advice* ("passwords") that aligns the realized formal rule with the target outcome. Reports that match the advice implement the target SCF; otherwise, the formal rule returns a symmetric worst outcome. Since the correct passwords are drawn from a continuum, unrecommended reports hit a correct password with probability zero, which removes profitable deviations from advice. We formalize this as an *unimprovability* property.

This yields a sharp dichotomy. Fairness constraints that restrict only the formal rules are *neutralizable*: any Bayesian incentive compatible (BIC) SCF can be implemented by formal rules that themselves satisfy symmetry. Moreover, if the target SCF is (weakly) dominant-strategy incentive compatible (DSIC), the induced mechanism (which embeds the advice stage) attains weak dominant-strategy incentive compatibility: truthful type reports and advice-following are (weakly) dominant.

By contrast, constraints that embed *predictable structure* in the formal rule—strict monotonicity and continuity—are *non-neutralizable*. Such constraints enable agents to "guess" the

correct passwords. Predictability undermines unimprovability by making it possible to profitably deviate toward outcomes that the advised messages avoid. We also show an approximate result: continuity of the formal rule is compatible with *η-unimprovable* FPREs, in which the probability of a strictly improving deviation can be made arbitrarily small.

Our results isolate a regulatory-scope insight. If auditors can only verify the *format* of the formal rule, then format-type fairness (e.g., symmetry) does not bind: a designer with private advice can make a fair-looking rule behave as if it were discriminatory. In contrast, mandates that force the formal rule to reveal structure (monotonicity/continuity) hinder the "encryption" that sustains obedience to advice.

2. Related Literature

We build on four strands and depart in key ways.

Symmetric implementation. In auction theory, Deb and Pai (2017) show that discrimination can arise under symmetric auction rules. Azrieli and Jain (2018) and Korpela (2018) generalize beyond specific formats: Azrieli and Jain (2018) prove that any BIC SCF can be implemented by a *symmetric* mechanism in a Bayesian Nash equilibrium, but they also show that under a dominant-strategy equilibrium, only symmetric SCFs are implementable by symmetric mechanisms.[1] Our departure is to separate the *auditable formal rule* from *private communication*, and to impose symmetry only on the former. Thus, when the target SCF is asymmetric, private advice must (and in our construction does) break symmetry even though the auditable formal rule remains symmetric.

Virtual/robust implementation and randomization. Classic and robust implementation study what is achievable *without* audit-scope asymmetry (Abreu and Matsushima, 1992a,b, 1994; Tian, 1997; Bergemann and Morris, 2005, 2011; Bergemann, Morris, and Tercieux, 2011). Methodologically, our use of randomness is closest to Abreu and Matsushima (1992b):

[1] Azrieli and Jain (2018) use a strict version of weak dominance (strict for at least one profile of others' messages). Under a weaker notion (never requiring strict inequality), Chen and Knyazev (2023) exhibit asymmetric rules implementable by symmetric mechanisms; they do not give a general sufficient condition for dominant-strategy implementation.

both enlarge implementability via randomization in incomplete information. There are two differences: (i) *Goal/solution concept*: we obtain *exact*, format-preserving implementation under BIC or (weak) DSIC, whereas Abreu and Matsushima (1992b) obtain *virtual* implementation under iteratively undominated/rationalizable play. (ii) *Institutional scope*: our question is driven by *auditability*—only the format of the formal rule is regulated (symmetry), private advice is not. This asymmetry is absent in prior work. In dominant-strategy settings, (virtual) implementation requires the SCF to be DSIC (Tian, 1997); we identify which *format-type* constraints remain compatible with exact implementation when private communication is available.

Continuity. Our treatment differs in both object and consequence from *continuous implementation* (Oury and Tercieux, 2012): they impose continuity of the *equilibrium-induced outcome map* and derive monotonicity-type necessities for rationalizable implementation, whereas we impose continuity only on the *formal rule* and show exact non-neutralizability versus approximate neutralization within that class.

BCE and communication equilibrium. Our two-layer construction can be read as a Bayes correlated equilibrium (BCE) or communication equilibrium: the principal plays the role of a mediator who privately recommends reports, and unimprovability is the BCE obedience constraint with actions interpreted as reports to the formal rule (Forges, 1986, 1990; Bergemann and Morris, 2016a,b, 2019). The novelty here is institutional: auditors verify only the format (e.g., symmetry) of the formal rule, not the asymmetric private advice. We show that such format-type fairness is neutralizable (exactly, and even in DS when the SCF is DSIC), while structure-revealing mandates (strict monotonicity, exact continuity) are not—a dimension not captured by standard BCE analyses. Further, in the BCE framework, dominant-strategy obedience is not the usual focus. Our DS results strengthen this: advice-following is optimal regardless of beliefs about others' recommendations (belief-independent obedience).

Other Literature Our mechanism is institutionally distinct from perfect implementation (e.g., Izmalkov, Lepinski, and Micali (2010)) and zero-knowledge mechanism design (Canetti, Fiat,

and Gonczarowski, 2023). Perfect implementation augments the mechanism with cryptographic commitment/verification so that the designer can exactly implement target outcomes without relying on trusted mediators or violating the players' privacy. Zero-knowledge mechanisms design protocols that let a mediator or the mechanism prove incentive properties without disclosing agents' private information or the mechanism's sensitive details. Different from their focus, our interest is in auditable constraints on the format of the rule.

3. Model

The model features a principal and n agents. Let $N = \{1, 2, \dots, n\}$ denote the set of agents, and write $i = 0$ for the principal. Players derive utility from the profile of types and the implemented outcome. Let $\mathscr{T} = T^n$ be the set of type profiles and let $p \in \Delta(\mathscr{T})$ be a commonly known prior over $\mathscr{T}$. We assume that T is a standard Borel space with $|T| = |\mathbb{R}|$. Each $t_i \in T$ is privately observed by agent i. Let $\mathscr{O}$ be the set of outcomes; write each outcome as an n-tuple $o = (o_1, \dots, o_n)$. For each $i \in N^* := \{0\} \cup N$, let $u_i \colon \mathscr{T} \times \mathscr{O} \to \mathbb{R}$ denote player i's utility function.

Assumption 1 (Worst and symmetric punishments)**.** There exist outcomes $\{\underline{o}^i\}_{i \in N}$ and $o^* \in \mathscr{O}$ such that:

(i) *(Individual worst)* For each $i \in N$, $t \in \mathscr{T}$, and $o \in \mathscr{O}$, $u_i(t, o) \geqslant u_i(t, \underline{o}^i)$, and $\underline{o}^i_j = \underline{o}^i_k$ for any $j, k \neq i$.

(ii) *(Multi-deviation fallback)* o^* is symmetric, i.e., $o^*_i = o^*_j$ for all i, j, and for every $i \neq j$ and every $t \in \mathscr{T}$, $u_i(t, \underline{o}^j) \geqslant u_i(t, o^*)$.

Example 1. Consider an assignment problem without transfers with monotone preferences. Let $n = 3$ and $\mathscr{O} = \{(x_1, x_2, x_3) \in \mathbb{R}^3_+ \mid x_1 + x_2 + x_3 = 1\}$. Then, for instance, $\underline{o}^2 = (1/2, 0, 1/2)$ and $o^* = (1/3, 1/3, 1/3)$ satisfy Assumption 1.

We refer to a *rule* or *social choice function* (SCF) as a mapping $\mu \colon \mathscr{T} \to \mathscr{O}$. The set of all rules is $\mathscr{R} = \mathscr{O}^{\mathscr{T}}$. Typically, we use μ as a *target rule*: the principal's objective is to implement μ by a mechanism.

A subset $\mathscr{C} \subseteq \mathscr{R}$ is referred to as a *constraint class*. We use the following (format) constraints.

Definition 1 (Format constraints)**. Symmetry** Rule $\nu \in \mathscr{R}$ is *symmetric* if for any permutation $\pi\colon N \to N$, whenever $\nu(t_1, \ldots, t_n) = (o_1, \ldots, o_n)$ we also have $\nu(t_{\pi(1)}, \ldots, t_{\pi(n)}) = (o_{\pi(1)}, \ldots, o_{\pi(n)})$.

Surjectivity $\nu \in \mathscr{R}$ is *surjective* if for every $o \in \mathscr{O}$ there exists $t \in \mathscr{T}$ with $\nu(t) = o$.

Strict monotonicity Suppose T is linearly ordered by $>$ and $\mathscr{O}$ is partially ordered by $\succ$. Rule $\nu \in \mathscr{R}$ is *strictly monotone* if for any $t_i, t_i' \in T$ and $t_{-i} \in T^{n-1}$ with $t_i > t_i'$, it holds that $\nu(t_i, t_{-i}) \succ \nu(t_i', t_{-i})$.

Continuity Suppose (T, d_T) and $(\mathscr{O}, d_{\mathscr{O}})$ are metric spaces. Rule $\nu \in \mathscr{R}$ is *continuous* if it is continuous (with the product topology on T^n).

We say that a constraint class $\mathscr{C}$ *satisfies* properties A and B (e.g., symmetry and continuity) if $\mathscr{C} = \{\nu \in \mathscr{R} \mid \nu \text{ satisfies } A \text{ and } B\}$.

The principal uses a private communication and a *formal rule* $\nu^* \in \mathscr{C}$ to implement a target rule μ. In detail, the game proceeds in four stages:

1. The principal chooses $M \subseteq \mathscr{C}$ and draws $\nu^* \in M$ according to a countably additive probability measure P on (a Borel σ-algebra of) M. Drawn ν^* is hidden from the agents.

2. Each agent i observes $t_i \in T$ and reports $t_i' \in T$ to the principal.

3. Observing ν^* and t', the principal privately sends a recommendation $\tau_i^{\nu^*}(t') \in T$ to each agent i.

4. Each agent i reports $\tau_i' \in T$ to the formal mechanism, and the outcome $\nu^*(\tau')$ is implemented.

We say that a constraint class $\mathscr{C}$ is *neutralizable* if any target rule $\mu \in \mathscr{R}$ (possibly $\mu \notin \mathscr{C}$) can be implemented by selecting formal rules $\nu^* \in M \subseteq \mathscr{C}$ together with associated recommendation policies $\tau^{\nu^*}\colon \mathscr{T} \to \mathscr{T}$. The formal definition appears in section 5.

4. Format-Preserving Randomized Encryption

We now formalize the randomization-with-advice device used in our implementation results. Fix a constraint class $\mathscr{C}$. For each rule $\tilde{\nu} \in \mathscr{R}$, a map $\tilde{\tau}^{\tilde{\nu}} \colon \mathscr{O} \to \mathscr{T}$ is an *(outcome-based) recommendation function* for $\tilde{\nu}$ if $\tilde{\nu}(\tilde{\tau}^{\tilde{\nu}}(o)) = o$ for every $o \in \tilde{\nu}(\mathscr{T})$. Given the recommendation functions, the messages received, and a probability measure P over M, agents form beliefs about the chosen mechanism $\nu^* \in M$. We refer to a tuple consisting of a set of rules M, a probability measure P on M, and recommendation functions $\{\tilde{\tau}^{\tilde{\nu}}\}_{\tilde{\nu} \in M}$ as a *format-preserving randomized encryption (FPRE)* for $\mathscr{C}$ if $M \subseteq \mathscr{C}$. An FPRE is *unimprovable* if, conditional on the agent's received recommendation and the recommendation function, the probability that any unrecommended message yields a strictly better outcome is zero. The formal definition is as follows.

Definition 2 (Unimprovability)**.** A FPRE $(P, M, \{\tilde{\tau}^{\tilde{\nu}}\})$ is *unimprovable* if for any agent $i \in N$, $t \in \mathscr{T}$, $o \in \mathscr{O}$, $\tau_i \in T$, $\tau_i' \neq \tau_i$, and $\tau_{-i} \in T^{n-1}$, if

$$\Pr_{\tilde{\nu} \sim P}\left[\tau_i = \tilde{\tau}_i^{\tilde{\nu}}(o)\right] > 0,$$

we have

$$\Pr_{\tilde{\nu} \sim P}\Big[u_i(t, \tilde{\nu}(\tau_i', \tau_{-i})) > u_i(t, \tilde{\nu}(\tau_i, \tau_{-i})) \;\Big|\; \tau_i = \tilde{\tau}_i^{\tilde{\nu}}(o)\Big] = 0.$$

We show that the existence of an unimprovable FPRE for symmetric $\mathscr{C}$.

Lemma 1. *If $\mathscr{C} \neq \varnothing$ satisfies symmetry and surjectivity, there exists an unimprovable FPRE for $\mathscr{C}$.*

Let us briefly explain how to construct the FPRE. Fix a surjective and symmetric rule $\nu \in \mathscr{C}$. For a target outcome $o \in \mathscr{O}$, pick $\tau^* \in \mathscr{T}$ with $\nu(\tau^*) = o$. Since $|T| = |\mathbb{R}| = |\mathbb{R}^2|$, there exist Borel bijections $g \colon (0,1) \to T$ and $h \colon T \to (0,1)^2$. For each $s \in (0,1)$, draw an independent "password" ε_s following the uniform distribution on $(0,1)$, and let $\varepsilon = (\varepsilon_s)_{s \in (0,1)}$.

Given ε, define the formal rule ν^ε as follows. For any report profile $\tau = (\tau_i)_i$, write $h(\tau_i) = (s_i, w_i)$. If $w_i = \varepsilon_{s_i}$ for all i, implement $\nu(g(s_1), \ldots, g(s_n))$; otherwise implement the punishment

outcome specified in Assumption 1. For the target o, the principal recommends to agent i the message

$$\tilde{\tau}_i^{\nu^\varepsilon}(o) \;=\; h^{-1}(s_i^*, \varepsilon_{s_i^*}), \quad \text{where } h(\tau_i^*) = (s_i^*, w_i^*).$$

The password construction is independent of agents' identities, so each ν^ε is symmetric. Moreover, because passwords are drawn from a continuum, any unrecommended report matches a correct password with probability zero, which yields unimprovability. A formal proof appears in the Appendix.

5. Neutralization

5.1. Implementation via FPRE

Building on the FPRE defined above, we now present our implementation results. To state the conditions for implementation, we first recall standard incentive-compatibility notions.

Definition 3. A rule μ is *Bayesian incentive compatible (BIC)* if, for all i and $t_i, t_i' \in T$,

$$\mathbb{E}_{t_{-i}\sim p(\cdot|t_i)}\big[u_i((t_i, t_{-i}), \mu(t_i, t_{-i}))\big] \;\geqslant\; \mathbb{E}_{t_{-i}\sim p(\cdot|t_i)}\big[u_i((t_i, t_{-i}), \mu(t_i', t_{-i}))\big].$$

It is *weakly dominant-strategy incentive compatible (DSIC)* if the inequality holds pointwise in t_{-i} (and *DSIC* if it is strict for some t_{-i}).

We say that a constraint class $\mathscr{C}$ is *Bayesian (resp. weakly dominant-strategy, dominant-strategy) neutralizable* for μ if there exist $M \subseteq \mathscr{C}$, a probability measure P on M, and recommendation policies $\{\tau^\nu \colon \mathscr{T} \to \mathscr{T}\}_{\nu\in M}$ such that truthful reporting to the principal and obedience to τ^ν implement μ are optimal in the Bayesian (resp. weakly DS, DS) sense. More precisely:

Definition 4. Constraint class $\mathscr{C}$ is *Bayesian neutralizable* for $\mu \in \mathscr{R}$ if there exist (M, P) with $M \subseteq \mathscr{C}$ and $\tau^\nu \colon \mathscr{T} \to \mathscr{T}$ for each $\nu \in M$ such that

(a) For any $\nu \in M$, $i \in N$, and $t_i, t_i' \in T$,

$$\mathbb{E}_{t_{-i}\sim p(\cdot|t_i)}\big[u_i(t, \nu(\tau^\nu(t)))\big] \;\geqslant\; \mathbb{E}_{t_{-i}\sim p(\cdot|t_i)}\big[u_i(t, \nu(\tau^\nu(t_i', t_{-i})))\big].$$

(b) For any $i \in N$, $t \in \mathscr{T}$, $\tau_i' \in T$, and $\tau_{-i} \in T^{n-1}$,

$$\mathbb{E}_{\nu\sim P,\, t_{-i}\sim p(\cdot|t_i)}\left[u_i(t, \nu(\tau_i^\nu(t), \tau_{-i})) \,\middle|\, \tau_i = \tau_i^\nu(t)\right] \;\geqslant\; \mathbb{E}_{\nu\sim P,\, t_{-i}\sim p(\cdot|t_i)}\left[u_i(t, \nu(\tau_i', \tau_{-i})) \,\middle|\, \tau_i = \tau_i^\nu(t)\right].$$

Constraint class $\mathscr{C}$ is *weakly dominant-strategy neutralizable* for $\mu \in \mathscr{R}$ if the inequalities hold pointwise in t_{-i} (and *dominant-strategy neutralizable* if they are strict for some t_{-i} and τ_{-i}).

In our notion of neutralizability, agents are required to play a (weakly) dominant strategy in the communication stage, conditional on truthful reporting. This dominance requirement differs from Azrieli and Jain (2018), who impose dominance and symmetry on the *entire* mechanism, implying that only symmetric rules are implementable.[2]

With these notions in hand, we state a sufficient condition for format-preserving implementation. The next result follows immediately from unimprovability (Definition 2).

Proposition 1. *Suppose there exists an unimprovable FPRE for $\mathscr{C}$. Then $\mathscr{C}$ is Bayesian (resp. weakly dominant-strategy, dominant-strategy) neutralizable for any μ that is BIC (resp. weakly DSIC, DSIC).*

5.2. Neutralizable and Non-neutralizable Constraints

We now establish the neutralizability of the symmetry constraint. The next corollary follows immediately from Lemma 1 and Proposition 1.

Corollary 1. *If $\mathscr{C}$ satisfies symmetry and surjectivity, then $\mathscr{C}$ is Bayesian neutralizable (resp. dominant-strategy neutralizable) for any BIC (resp. DSIC) rule μ.*

By contrast, strict monotonicity and continuity are incompatible with unimprovability. The intuition is straightforward. If rules are required to be strictly monotone, then once an agent knows that the recommended report τ_i induces outcome $o \in \mathscr{O}$, any higher report $\tau_i' > \tau_i$ induces a different outcome $o' \succ o$ with probability one; if the agent prefers o' to o, unimprovability fails.

[2] As noted in section 2, Chen and Knyazev (2023) show that under a weaker notion of dominance, some asymmetric rules can be implemented by symmetric mechanisms. In our terminology, the definition of the dominant strategy follows Azrieli and Jain (2018), and that of the weak dominant strategy is the same as M-DSE defined in Chen and Knyazev (2023).

For continuity, if (τ_i', τ_{-i}) yields an outcome different from $\nu^*(\tau)$ and not equal to any worst outcome for some $\nu^* \in M$, then by continuity, a neighborhood of (τ_i', τ_{-i}) is also mapped to non-worst outcomes. Hence, under any probability measure over M, the conditional probability of a strict improvement after deviating is positive, contradicting unimprovability. Formally:

Proposition 2. *(1) Any FPRE over a strictly monotone constraint class $\mathscr{C}$ fails unimprovability for some utility profile.*

(2) Suppose T is separable and M consists of continuous *and nonconstant rules. Then any FPRE based on M fails unimprovability for some continuous utility profile.*

6. Extensions and Limitations

This section discusses the issues not fully captured in the main texts.

Approximate unimprovability We relax unimprovability to allow a (small) probability of a profitable deviation.

Definition 5. Fix $\eta > 0$. An FPRE $(P, M, \{\tilde{\tau}^\nu\}_{\nu \in M})$ is *η-unimprovable* if, conditional on receiving the recommended message, the probability that some unrecommended message yields a *strictly* better outcome (relative to Definition 2) is strictly less than η.

Even under η-unimprovability, some rules may fail to be implementable.

Example 2. Let $n = 2$ and $\mathscr{O} = \{(1,0), (1/2, 1/2), (0,1)\}$. Agent 2 prefers o to o' iff $o_2 > o_2'$, while agent 1 is indifferent across all outcomes. Take worst outcomes $\underline{o}^1 = (0,1)$, $\underline{o}^2 = (1,0)$ and symmetric fallback $o^* = (1/2, 1/2)$. Consider the (trivially DSIC) rule $\mu(t) \equiv (1,0)$ for all $t \in \mathscr{T}$. Suppose a symmetric formal rule ν^* implements μ. Then there is $\tau \in \mathscr{T}$ with $\nu^*(\tau) = (1,0)$, and by symmetry $\nu^*(\tau_1, \tau_1) = (1/2, 1/2)$. Since agent 2 never prefers $(1,0)$ to $(1/2, 1/2)$, any deviation that makes $\tau_2' = \tau_1$ deliver $(1/2, 1/2)$ with a positive probability. Even when this probability is small enough, μ is not implementable as deviation yields a strict improvement in expectation.

A sufficient way to preclude this problem is to ensure a strict gap between on-path outcomes and each agent's worst outcome. Let $O^* := \mu(\mathscr{T})$ denote the image of μ. We say that μ *induces α-strict better outcomes for the agents* if there exists $\alpha > 0$ such that, for every $i \in N$, every $t \in \mathscr{T}$, and every $o \in O^*$,

$$u_i(t, o) \;\geqslant\; u_i(t, \underline{o}^i) + \alpha.$$

If μ satisfies the property, then μ remains neutralizable even when M contains only approximately unimprovable rules (for sufficiently small η): the potential gain from any deviation that occasionally hits a non-worst outcome is dominated by the on-path gap α.

This concern is especially salient in assignment and matching problems. With indivisible objects (no transfers), "receiving nothing" is both an agent's worst outcome and an outcome that may arise under *any* rule when supply is scarce; similarly, in matching, "no match" is individually rational and may occur on path whenever market sides are unbalanced. In such environments, the strict gap condition typically fails, and approximate unimprovability alone will not secure neutralization.

Continuity As noted above, exact unimprovability is incompatible with continuity. Nevertheless, continuous rules can satisfy *approximate* unimprovability.

Proposition 3. *Fix $\eta > 0$. Suppose that $\mathscr{O}$ is convex, $T = \mathbb{R}$, and the symmetric worst outcomes are common (i.e., there exists $\underline{o} \in \mathscr{O}$ such that $\underline{o}^i = \underline{o}$ for each $i \in N$). Let $\mathscr{O}^* \subseteq \mathscr{O}$ be a countable set of outcomes. Then there exist a symmetric family M^* of continuous rules that are surjective onto $\mathscr{O}^*$ and a collection $\{\tilde{\tau}^\nu\}_{\nu \in M^*}$ such that $(M^*, P, \{\tilde{\tau}^\nu\})$ is η-unimprovable.*

Countable types When the type space is countable, exact unimprovability cannot be achieved. Indeed, suppose that for some profile $\tau \in \mathscr{T}$ and deviation $\tau_i' \in T$ we have, for every $\nu \in \mathscr{C}$, $\nu(\tau) \neq \nu(\tau_i', \tau_{-i}) \neq \underline{o}^i$. By countability of T and countable additivity of P over M, there exists τ_i' such that

$$\Pr_{\tilde{\nu} \sim P}\Big(u_i(t, \tilde{\nu}(\tau_i', \tau_{-i})) \;>\; u_i(t, \tilde{\nu}(\tau_i, \tau_{-i})) \;\Big|\; \tau_i = \tilde{\tau}_i^{\tilde{\nu}}(o)\Big) \;>\; 0.$$

In words, an agent can "guess" a password with strictly positive probability.

By contrast, for any $\eta > 0$, one can construct an FPRE that is η-unimprovable. The construction parallels Lemma 1 but draws passwords from a finite grid: instead of sampling from $(0,1)$, pick passwords from $(0,1) \cap \{1/m, 2/m, \ldots, m/m\}$ for some large $m \in \mathbb{N}$. Since $\mathscr{T}$ is countable, there is a bijection $\mathscr{T} \to ((0,1) \cap \mathbb{Q}) \times \{1/m, 2/m, \ldots, m/m\}$, which allows us to encode each type together with a password. Taking m sufficiently large makes the probability of a successful, unrecommended deviation arbitrarily small, yielding an η-unimprovable analogue of Lemma 1.

Incentive of the principal Absent commitment to the recommendation policy $\tau^\nu \colon \mathscr{T} \to \mathscr{T}$, the rule μ should satisfy ex post optimality to ensure truthful recommendations by the principal; that is, $u_0(t, \mu(t)) \geqslant u_0(t, o)$ for each $o \in \mathscr{O}$ and each $t \in \mathscr{T}$. Conversely, if μ is ex post optimal and incentive compatible, then some $\nu \in \mathscr{C}$ implements μ in a perfect Bayesian (Nash) equilibrium of the full game.

Coalitional deviations. Beyond individual deviations, coalitions can in principle manipulate the formal rule. Because the password test in our FPRE is applied *coordinate-wise* and the same password map $\varepsilon \colon (0,1) \to (0,1)$ is used for every coordinate, a coalition can permute the recommended pairs and mimic recommendations for others without triggering any punishment. Formally, fix $\nu^\varepsilon \in M$, let $o \in \mathscr{O}$ be the target outcome, and let $\tau = \tilde{\tau}^{\nu^\varepsilon}(o) \in \mathscr{T}$ denote the recommended message profile.

Consider $\tau' \in \{\tau_1, \ldots, \tau_n\}^n$. Since each coordinate (s_i, w_i) of $h(\tau')$ still satisfies $w_i = \varepsilon_{s_i}$, the password test passes at every coordinate and the punishments are not triggered, and the achievable outcomes is extended from $\nu^\varepsilon(\tau)$ to $\{\nu^\varepsilon(\tau') \mid \tau' \in \{\tau_1, \ldots, \tau_n\}^n\}$. By the symmetry, this set includes all permuted outcomes of o; $(o_{\pi(1)}, \ldots, o_{\pi(n)})$ for any permutation $\pi \colon N \to N$. Blocking such coalitional deviation would require additional structure that lies outside our format-preserving scope.

References

Abreu, D. and H. Matsushima (1992a): "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica*, 60, 993–1008.

——— (1992b): "Virtual Implementation in Iteratively Undominated Strategies: Incomplete Information," Unpublished manuscript, Princeton University.

——— (1994): "Exact Implementation," *Journal of Economic Theory*, 64, 1–19.

Azrieli, Y. and R. Jain (2018): "Symmetric Mechanism Design," *Journal of Mathematical Economics*, 74, 108–118.

Bergemann, D. and S. Morris (2005): "Robust Mechanism Design," *Econometrica*, 73, 1771–1813.

——— (2011): "Robust Implementation in General Mechanisms," *Games and Economic Behavior*, 71, 261–281.

——— (2016a): "Bayes Correlated Equilibrium and the Comparison of Information Structures," *Theoretical Economics*, 11, 487–522.

——— (2016b): "Information Design, Bayesian Persuasion, and Bayes Correlated Equilibrium," *American Economic Review: Papers & Proceedings*, 106, 586–591.

——— (2019): "Information Design: A Unified Perspective," *Journal of Economic Literature*, 57, 44–95.

Bergemann, D., S. Morris, and O. Tercieux (2011): "Rationalizable Implementation," *Journal of Economic Theory*, 146, 1253–1274.

Canetti, R., A. Fiat, and Y. A. Gonczarowski (2023): "Zero-Knowledge Mechanisms," ArXiv preprint.

Chen, B. and D. Knyazev (2023): "Symmetric Mechanism Design: Comment," *Journal of Mathematical Economics*, 109, 102910.

Deb, R. and M. M. Pai (2017): "Discrimination via Symmetric Auctions," *American Economic Journal: Microeconomics*, 9, 275–314.

Forges, F. (1986): "An Approach to Communication Equilibria," *Econometrica*, 54, 1375–1385.

——— (1990): "Universal Mechanisms," *Econometrica*, 58, 1341–1364.

Izmalkov, S., M. Lepinski, and S. Micali (2010): "Perfect Implementation," *Games and Economic Behavior*, 71, 121–140.

Korpela, V. (2018): "Procedurally Fair Implementation under Complete Information," *Journal of Mathematical Economics*, 77, 25–31.

Kotowski, M. H. (2018): "On asymmetric reserve prices," *Theoretical Economics*, 13, 205–237.

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

Tian, G. (1997): "Virtual Implementation in Incomplete Information Environments with Infinite Alternatives and Types," *Journal of Mathematical Economics*, 28, 313–339.

A. Proofs

Proof of Lemma 1. Step 1: Construction of a subfamily of symmetric formal rules.

Let $\varepsilon \in \Omega = (0,1)^{(0,1)}$. Note that as T has the same cardinality as $\mathbb{R}$, $|T| = |(0,1)^2|$. Therefore, there exist Borel bijections $g\colon (0,1) \to T$ and $h\colon T \to (0,1)^2$. Let $\nu \in \mathscr{C}$ be a symmetric and surjective rule.

We define ν^ε as follows:

1. For each $\tau_i \in T$, let $(\tau_i^*, w_i) = h(\tau_i)$. Then, $\nu^\varepsilon(\tau) = \nu(g(\tau_1^*), \ldots, g(\tau_n^*))$ if $w_i = \varepsilon_{\tau_i^*}$ for each $i \in N$.

2. If there is i such that $w_i \neq \varepsilon_{\tau_i^*}$ and $w_j = \varepsilon_{\tau_j^*}$ for each $j \in N \setminus \{i\}$, $\nu^\varepsilon(\tau) = \underline{o}^i$.

3. Otherwise, $\nu^\varepsilon(\tau) = o^*$.

We can show that each ν^ε is symmetric and surjective. If $\tau_j = \tilde{\tau}_j^{\nu^\varepsilon}(o)$ for any $j \in N$, the symmetry follows from the symmetry of ν^*. If $\left\{j \in N \mid \tau_j \neq \tilde{\tau}_j^{\nu^\varepsilon}(o)\right\} = \{i\}$, $\tau_i \neq \tau_j$ for any other $j \in N \setminus \{i\}$. As $\underline{o}_j^i = \underline{o}_k^i$ for each $j, k \in N \setminus \{i\}$, this specification preserves symmetry. If $\left|\left\{j \in N \setminus \{i\} \mid \tau_j \neq \tilde{\tau}_j^{\nu^\varepsilon}(o)\right\}\right| \geqslant 2$, the outcome o^* is symmetric and therefore, ν^ε is symmetric.

Surjectivity of ν^ε follows from that of ν. For each $o \in \mathscr{O}$, we can find $\nu(\tau) = o$ and let $\tau_i' = h^{-1}(g^{-1}(\tau_i), \varepsilon_{g^{-1}(\tau_i)})$ for each $i \in N$. Then, $\nu^\varepsilon(\tau') = o$. Therefore, each $\nu^\varepsilon \in \mathscr{C}$.

Step 2: Construction of the probability space.

Let $((0,1), \mathscr{B}, \lambda)$ denote the unit interval with its Borel σ-algebra and Lebesgue law. Define

$$\mathscr{F} := \sigma\left(\{\varepsilon \in \Omega \mid \varepsilon_{x_1} \in (0,1), \ldots, \varepsilon_{x_\ell} \in (0,1), \ell \in \mathbb{N}\}\right),$$

where $\sigma(X)$ denotes the σ-algebra generated by a set X. Let P be the (product) probability measure on $(\Omega, \mathscr{F})$ such that for every finite set $\{x_1, \ldots, x_\ell\} \subseteq (0,1)$, the random vector $(\varepsilon_{x_1}, \ldots, \varepsilon_{x_\ell})$ is i.i.d. with common law λ (uniform distribution on $(0,1)$).[3] We write $\varepsilon_x(\omega)$ for the coordinate map at index $x \in (0,1)$.

Let $f\colon \Omega \to M_\nu$ be the bijection $\varepsilon \mapsto \nu^\varepsilon$ and endow M_ν with the transported σ-algebra $\mathscr{M} := \{A \subseteq M_\nu \mid f^{-1}(A) \in \mathscr{F}\}$. Define $P_{M_\nu}(A) := P(f^{-1}(A))$ for each $A \in \mathscr{M}$.

Step 3: Construction of the recommendation functions.

For each $o \in \mathscr{O}$ such that $o = \nu^\varepsilon(t) = \nu(g(t_1^*), \ldots, g(t_n^*))$, let $\tilde{\tau}_i^{\nu^\varepsilon}(o) = h^{-1}(t_i^*, \varepsilon_{t_i^*})$ for each $i \in N$. As ν is surjective, this is well-defined.

Step 4: Proof of the unimprovability.

Now we show that $(P, M_\nu, \{\tilde{\tau}^{\nu'}\}_{\nu' \in M_\nu})$ is unimprovable.

For any $o \in \mathscr{O}$, by $\tau_i = \tilde{\tau}_i^{\nu^\varepsilon}(o)$, $h(\tau_i) = (x, \varepsilon_x)$ for some $x \in (0,1)$. For any deviation $\tau_i' \neq \tau_i$, let denote $(x', w') = h(\tau_i')$. Then, the probability that $w' = \varepsilon_{x'}$ is 0. We have three cases with respect to the others' reporting.

[3] By the Kolmogorov (Daniell-Kolmogorov) extension theorem, the consistent family of finite-dimensional distributions $\{\lambda^{\otimes \ell}\}_{\ell \in \mathbb{N}}$ determines a *unique* probability measure P on the product measurable space $(\Omega, \mathscr{F})$.

Case 1. Suppose that $\tau_j = \tilde{\tau}_j^{\nu^\varepsilon}(o)$ for any $j \in N \setminus \{i\}$. Then, $\nu^\varepsilon(\tau_i', \tau_{-i}) = \underline{o}^i$ with probability 1. Then, any deviation is not profitable.

Case 2. Suppose that $\left|\left\{j \in N \setminus \{i\} \mid \tau_j \neq \tilde{\tau}_j^{\nu^\varepsilon}(o)\right\}\right| = 1$. Note that if $\tau_j \neq \tilde{\tau}_j^{\nu^\varepsilon}(o)$, with probability 1, $w_j \neq \varepsilon_{\tau_j^*}$, where $(\tau_j^*, w_j) = h(\tau_j)$. Then, $\nu^\varepsilon(\tau_i', \tau_{-i}) = o^*$ with probability 1, while $\nu^\varepsilon(\tau_i, \tau_{-i}) = \underline{o}^j$. Then, any deviation is not profitable.

Case 3. Suppose that $\left|\left\{j \in N \setminus \{i\} \mid \tau_j \neq \tilde{\tau}_j^{\nu^\varepsilon}(o)\right\}\right| \geqslant 2$. Then, any deviation does not change the result. □

Proof of Proposition 1. Let $M = M^*$. For any $\nu \in M^*$, let $\tau^\nu \colon \mathscr{T} \to \mathscr{T}$ that satisfy $\tau^\nu(t) = \tilde{\tau}^\nu(\mu(t))$, and $\tau_i = \tau_i^\nu(t)$.

Consider stage 4. Suppose the message sent by the principal to agent i is τ_i^*. Then, the agent i believes that sending τ_i yields $\nu(\tau_i, \tau_{-i})$ when the others' report is τ_{-i}. By the unimprovability, agents have no incentive to misreport $\tau_i \neq \tau_i^*$. Furthermore, if $\tau_{-i} = \tau_{-i}^*$, the outcome is $\mu(t)$ when $\tau_i = \tau_i^*$.

Consider stage 2. Conditioned on the equilibrium strategy of the agents in stage 4, reporting t_i to the principal yields outcome $\mu(t_i, t_{-i})$. Then, by the incentive compatibility of μ, truth-telling is the optimal strategy. □

Proof of Proposition 2. **Proof for (1):** Consider a PRPE for strictly monotonic $\mathscr{C}$, $(M, P, \{\tilde{\tau}^\nu\})$.

Consider $\underline{o} \in \mathscr{O}$ and utility function for an agent $i \in \mathbb{N}$ such that $u_i(t, o) > u_i(t, o')$ if $o \succ o'$.

Let $\nu \in M$ be a strictly monotone rule. Suppose that $\nu(\tau)$ such that $\tau_i' > \tau_i$ for some $\underline{\tau}$ and $\overline{\tau}$. Then, any deviation $\tau_i' > \tau_i$ improves the utility with probability 1, and therefore, unimprovability fails.

Proof for (2): Consider a PRPE for continuous $\mathscr{C}$, $(M, P, \{\tilde{\tau}^\nu\})$.

Suppose that each $\nu \in M$ is continuous and nonconstant. Consider $\tau, \tau' \in \mathscr{T}$ such that $\nu(\tau) \neq \nu(\tau') = o'$. Then, there is $i \in N$ such that $\nu(\tau_i, \tau_{-i}'') = o \neq o' = \nu(\tau_i', \tau_{-i}'')$, where $\tau_j'' = \tau_j$ if $j < i$ and $\tau_j'' = \tau_j'$ if $i \geqslant j$. Consider the utility function such that $u_i(t, o) > u_i(t, o')$ for any $o \in \mathscr{O} \setminus \{o'\}$. For each $\nu \in M$, let $\tau_i^\nu = \tau_i$.

By the continuity, there is a $\delta > 0$ such that $u_i(t, \nu(\hat{\tau}_i, \tau_{-i}'')) > u_i(t, \nu(\tau''))$ for any $\hat{\tau}_i \in B_\delta(\tau_i^\mu)$, where B_δ is a δ-open ball in T.

As T is separable, let $D = \{d_1, d_2, \dots, \}$ be a countable dense set of T. For each $k \in \mathbb{N}$, let $M_k = \{\nu \in M \mid d_k \in B_\delta(\tau_i^\nu)\}$. Note that for each $\nu \in M$, as $B_\delta(\tau_i^\nu)$ is an open set of T, at least one $d_k \in B_\delta(\tau_i^\nu)$. Then, we have that $B_k \coloneqq \bigcap_{\nu \in M_k} B_\delta(\tau_i^\nu) \neq \varnothing$ and $\bigcup_{k \in \mathbb{N}} M_k = M$.

By the countable additivity of P, $P(M_k) > 0$ for some $k \in \mathbb{N}$. Now, let $\hat{\tau}_i \in B_k$ and this implies that $u_i(t, \nu(\hat{\tau}_i, \tau''_{-i})) > u_i(t, \nu(\tau''))$ with probability $P(M_k) > 0$. This contradicts the unimprovability. □

Proof of Proposition 3. Let $T^* = \mathbb{N}$, which is a countable subset of T. Then, there is a surjection $\nu^* \colon (T^*)^n \to \mathscr{O}^*$. Let $\varepsilon \in \Omega \coloneqq (\frac{1}{3}, \frac{2}{3})^{\mathbb{N}}$. For each $\tau_i \in \mathbb{R}$, let $k_i = \lfloor \tau_i \rfloor$ and $w_i = \tau_i - \lfloor \tau_i \rfloor$ for each $i \in N$.

Take real numbers $\beta = (\beta_1, \beta_0)$ with $1 > \beta_1 > \beta_0 > 0$, we define $q \colon (0,1) \times \mathbb{N} \to (0,1)$ as

$$q(w,k) = \begin{cases} 0 & \text{if} \quad |w - \varepsilon_k| \leqslant \beta_0 \\ \frac{|w-\varepsilon_k| - \beta_0}{\beta_1 - \beta_0} & \text{if} \quad \beta_0 < |w - \varepsilon_k| < \beta_1 \\ 1 & \text{if} \quad |w - \varepsilon_k| \geqslant \beta_1, \end{cases}$$

and

$$Q^{\varepsilon,\beta}(t) = 1 - \prod_{i=1}^{n} (1 - q(w_i, k_i)).$$

Then, we define $\nu^{\varepsilon,\beta}$ as follows:

$$\nu^{\varepsilon,\beta}(\tau) = Q^{\varepsilon,\beta}(\tau) \cdot \underline{o} + [1 - Q^{\varepsilon,\beta}(\tau)] \cdot \nu^*(k_1, \dots, k_n)$$

This implies that if $\left| w_i - \varepsilon_{\tau_i^*} \right| > \beta_1$, $\nu^\varepsilon(\tau) = \underline{o}$. Therefore, as $\beta_1 < 1/3$, $\nu^{\varepsilon,\beta}$ is continuous. Let $\tilde{\tau}^{\nu^{\varepsilon,\beta}}(o)$ be a recommendation function such that $\nu^*(k_1, \dots, k_n) = o$ and $\tilde{\tau}_i^{\nu^{\varepsilon,\beta}}(o) = k_i + \varepsilon_{k_i}$ for each $i \in N$.

We independently choose $\varepsilon_k \in (1/3, 2/3)$ from uniform distribution on $(1/3, 2/3)$ for each $k \in \mathbb{N}$. Let $z \in (0, 1/3)$ and $\ell \in \mathbb{N}$. We also choose β_1 from $\{z/\ell, 2z/\ell \dots, z\}$ with the same probability, and $\beta_0 = \beta_1 - 1/\ell$. Let $M = \left\{\nu^{\varepsilon,\beta} \mid \varepsilon \in (1/3, 2/3)^{\mathbb{N}}, \beta_1 \in \{z/\ell, 2z/\ell \dots, z\}\right\}$. Let (M, P) denote such a probability space.

Next, we check the deviation incentive. Let $\tau_i = \tilde{\tau}_i^{\nu^{\varepsilon,\beta}}(o)$, and $\tau_i' \neq \tau_i$. Let $k_i = \lfloor t_i \rfloor$, $w_i = t_i - k_i$, $k_i' = \lfloor t_i' \rfloor$, and $w_i' = t_i' - k_i'$. Consider $t_i' \in \mathbb{R}$ with $k_i' \neq k_i$. Then, the probability that $\left|w_i' - \varepsilon_{k_i'}\right| < \beta_1$ is at most $6z$.

Consider $t_i' \in \mathbb{R}$ with $k_i = k_i'$. Then, as the agent know ε_{k_i}, we consider the probability that $\nu^{\varepsilon,\beta}(\tau_i', \tau_{-i}) \in \{\underline{o}, \nu^{\varepsilon,\beta}(\tau)\}$, which is the probability that $\beta_0 < \left|w_i' - \varepsilon_{k_i'}\right| < \beta_1$. As β_1 is randomly chosen, the probability is at most $\frac{1}{\ell}$.

Based on the same discussion, the others' unrecommended reports change the result from $\{\nu^{\varepsilon,\beta}(\tau), \underline{o}\}$ is at most $\max\{6z, 1/\ell\}$. Therefore, the probability that $\nu^{\varepsilon,\beta}(\tau_i', \tau_{-i}) \in \{\underline{o}, \nu^{\varepsilon,\beta}(t)\}$ is at least $[1 - \max\{6z, 1/\ell\}]^n$. In conclusion, by taking z sufficiently small and ℓ sufficiently large, the probability of strict improvement can be arbitrarily small. and therefore, $(M, P, \{\tilde{\tau}^\nu\})$ is η-unimprovable.

Furthermore, if ν^* is symmetric, $\nu^{\varepsilon,\beta}$ is also symmetric as q is independent of agents' identities. □