Mechanism Design with Informational Punishment[*]

Benjamin Balzer[†] Johannes Schneider[‡]

January 4, 2022

Abstract

We introduce *informational punishment* to the design of mechanisms that compete with an exogenous status quo: A signal designer can publicly communicate with all players even if some decide not to communicate with the designer. Optimal informational punishment ensures that full participation in the mechanism is optimal even if any single player can publicly enforce the status-quo mechanism. Informational punishment restores the revelation principle, is independent of the mechanism designer's objective, and operates exclusively off the equilibrium path. Informational punishment is robust to refinements and applies in informed-principal settings. We provide conditions that make it robust to opportunistic signal designers.

[*] Johannes Schneider gratefully acknowledges financial support from the German Research Foundation (DFG) through CRC TR 224 (Project B03), Agencia Estatal de Investigación (grant PID2019-111095RB-I00 and grant PID2020-118022GB-I00), Ministerio Economía y Competitividad (grant ECO2017-87769-P), and Comunidad de Madrid (grant MAD-ECON-POL-CM H2019/HUM-5891).

[†] University of Technology Sydney, benjamin.balzer@uts.edu.au

[‡] Universität Mannheim and uc3m, jschneid@eco.uc3m.es

1 Introduction

Economics aims at improving the efficiency of institutions to govern strategic parties. If those parties hold asymmetric information, the first welfare theorem fails—we may fail to attain efficiency through any institution. In these cases, mechanism design becomes a powerful tool. Mechanism design characterizes all outcomes that parties' strategic incentives permit. Via the revelation principle, a simple direct revelation mechanism is enough to determine the best results we can hope for under any institution. In reality, however, new institutions often replace a status-quo mechanism.

Changing to the new institution may require consent—if parties veto the proposal, the status quo prevails. Parties follow similar strategic incentives in their decision to veto the new institution as those at play when they decide on their behavior within any given institution. Vetoing a proposal can be strategic, too—a party may publicly veto the mechanism only to signal her private information.

This paper considers a setting where parties may veto a proposed mechanism and disclose their veto to others. If the proposal fails, parties revert to the play of a default game. The optimal mechanism in such settings may involve on-path rejections (Celik and Peters, 2011)—the revelation principle fails. As a result, complexity increases, and the mechanism design approach loses part of its power. We show, however, that if parties can store information and commit to releasing it at a later date, the revelation principle holds, and we can readily apply the tools of mechanism design. We refer to this technology as *informational punishment.* The information is stored and released in the event of a deviation. Its purpose is to punish the deviator by the release.

Informational punishment is a simple yet powerful tool. It requires only that each party has access to a signaling device that garbles the party's information and conceals the realization of that garbling for some time. The threat to release information later suffices to discipline others and ensures full participation. Informational punishment does not interfere with the design of the mechanism itself; a decentral implementation is straightforward. Moreover, informational punishment is robust both to equilibrium refinements and restrictions on the space of available mechanisms. In addition, it applies to informed-principal problems and is immune to a designer who suffers from informational opportunism.

Examples of our environment abound. Parties in a legal conflict can coordinate to settle through an arbitration mechanism. However, each party can unilaterally enforce the default game of a trial. Political parties can solve gridlocks through a bargaining procedure. However, they can refuse to cooperate and enter a stalemate until the next elections. Firms can work together to determine the standard of the industry. However, they can also rely on a non-cooperative standards war. Countries can negotiate free trade agreements. However, if one government refuses to engage, they fall back on the WTO trade regime.

In all of these cases, vetoing the mechanisms may signal a party's ability in the default game. The other parties interpret the signal and adjust their behavior. The change in behavior influences the default game's outcomes. Moreover, if vetoing signals information, participating signals information too. Thus, if we cannot rule out on-path rejections of the mechanism, we need to solve all combinations of vetoing and participating and the associated outcomes of the default game to compute the optimal mechanism—a cumbersome computational problem.

Informational punishment relaxes the computational burden. It restores full participation on the equilibrium path yet only affects the players' outside options. Incentives inside the mechanism remain unchanged. Thus, we relax participation constraints without affecting incentive constraints. Moreover, the off-path event of informational punishment solves a specific, well-defined information design problem: min-max the deviator's continuation payoff over Bayes' plausible information structures.

Informational punishment works because a signal realization about party i has two effects. The first effect is direct and distributional: The other parties update expected payoffs because i's private information is payoff relevant. The second effect is indirect and behavioral. A party's strategy is a function of the information set. Altering information alters the continuation strategy. Via equilibrium reasoning, the party's change in behavior alters the behavior of other parties too. Informational punishment exploits that channel.

Related Literature. Signaling through vetoes in mechanisms dates back to Cramton and Palfrey (1995). Like them, we consider a problem in which a proposed mechanism competes with a status-quo game. In line with their model, we assume that ratification is public—agents learn who vetoes the mechanism and who does not. Unlike them, we are not interested in refining the mechanism space to those that survive ratification without further ado.

Instead, and closer to Celik and Peters (2011), we assume perfect Bayesian equilibrium as our solution concept, take the space of mechanisms as given, and are interested in the set of outcomes that can arise in such a setting. Different to Celik and Peters (2011)'s setting where "equilibrium rejection" may be optimal, we allow parties to engage in informational punishment. We show that equilibrium rejection is of no concern with informational punishment available. The set of outcomes that we can implement with full participation contains those we can implement with equilibrium rejection.[1]

Gerardi and Myerson (2007) and Correia-da-Silva (2020) propose an alternative approach in settings with veto-constrained mechanisms. Instead of signaling information, they consider mechanisms that "tremble." Even if all parties decide to participate, the mechanism breaks down with a small probability and invokes the default game. The mechanism can get arbitrarily close to the full-participation optimum through such

[1] Dequiedt (2007) and Tan and Yilankaya (2007) are other examples of default games' threat to participation.

trembling. However, these mechanisms rely on the assumption that a deviating party cannot credibly signal that it was her veto that invoked the default game and not the mechanism's tremble. Instead, we allow parties to announce their veto publicly, making trembles insufficient to overcome the veto problem. Celik and Peters (2016) use reciprocal mechanisms to circumvent the equilibrium-rejection problem. The main difference to our approach is that parties make a public revelation on the equilibrium path through their mechanism proposal. Depending on the environment, these revelations can interfere with incentive compatibility. Because informational punishments affect only off-path events, these concerns are absent—the set of implementable outcomes nests those in Celik and Peters (2016). The reverse is not the case due to, for example, our weaker commitment assumption.

Our work on standard-setting organizations (Balzer and Schneider, 2021) applies informational punishment outside mechanism design. In that previous paper, we consider a setting where the available mechanisms contain only efficient take-it-or-leave-it offers with fixed shares. In Balzer and Schneider (2021) informational punishment enlarges the class of environments where full participation is feasible, yet informational punishment *cannot* guarantee full participation. The reason is that the space of available mechanisms is too small. In the current paper, we take a broader view and complement Balzer and Schneider (2021). On the one hand, we define the minimal set of available mechanisms in the designer's toolbox such that an optimal full-participation mechanism exists. On the other hand, we show how informational punishment simplifies the mechanism designer's task, particularly when the set of available mechanisms is large. Unlike those in Balzer and Schneider (2021), our results in this paper readily apply to various design problems. Moreover, we show that—given our minimal conditions—common restrictions to the designer's problem do not affect the power of informational punishment. Specifically, we consider equilibrium refinement concepts (Grossman and Perry, 1986; Cho and Kreps, 1987; Cramton and Palfrey, 1995), informed-principal problems (Myerson, 1983), and informational opportunism (Dequiedt and Martimort, 2015).

2 Setup

Players and Information Structure. There are N players, indexed by $i \in \mathcal{N} := \{1, ..., N\}$. Each player has a private type $\theta_i \in \Theta_i$ and $\Theta_i \subset \mathbb{R}$ is compact. The state $\theta := \theta_1 \times ... \times \theta_N \in \times_i \Theta_i =: \Theta$ is distributed according to a commonly-known distribution function $I^0 : \Theta \to \Delta(\Theta)$, the *prior information structure.* Let $\theta_{-i} := \theta \setminus \theta_i$, and define the marginal $I^0_i(\theta_i) := \int_{\Theta_{-i}} I(\theta_i, d\theta_{-i})$ with support $supp(I^0_i) \subseteq \Theta \setminus \Theta_i$.

An information structure $I : \Theta \to \Delta(\Theta)$ is a commonly-known joint distribution over the state θ. The only restriction we impose on I is that it is absolutely continuous w.r.t. I^0, that is, $supp(I) \subseteq supp(I^0)$. Given θ_i a player's *belief* about the other players' types is the conditional distribution $I(\theta_{-i}|\theta_i) := \frac{I(\theta_i, \theta_{-i})}{I_i(\theta_i)}$, where $I_i(\theta_i)$ is the marginal of I.

Let $\mathcal{I}^0$ be the set of all information structures for which I^0 is an expansion. That is, $I \in \mathcal{I}^0$ if and only if there exists a random variable $\widetilde{\Sigma} : \Theta \to \Delta(S)$ which maps types into distributions of signals such that the realization $\sigma \in S$ together with $\widetilde{\Sigma}$ and I^0 implies I via Bayes' rule.

Basic Outcomes, Decision Rules, and Payoffs. There is an exogenously given set of basic outcomes, $Z \subset \mathbb{R}^K$, with $K < \infty$. Player i values the outcome $z \in Z$ according to a Bernoulli utility function, u_i, defined over Z.

We represent the rules of a game by a decision rule,

$$\pi : \Theta \to \Delta(Z),$$

where $\Delta(Z)$ is the set of all distribution functions over the outcome space Z. Each rule π is a mapping from *type reports* to a distribution over outcomes represented by the distribution function $G_\pi(z|\theta)$.

Status quo. The status quo is an exogenous game of incomplete information. We assume an equilibrium in that game exists for any information structure $I \in \mathcal{I}^0$ and take the equilibrium selection as given. For any information structure I, the status quo induces a decision rule π_I^M. Under π_I^M the expected utility of a truthfully reporting player i with type θ_i is

$$\begin{aligned} v_i(\theta_i, I, \pi_I^M) &:= \int_{\Theta_{-i}} \int_Z u_i(z, \theta_i, \theta_{-i}) dG_{\pi_I^M}(z|\theta_i, \theta_{-i}) I(d\theta_{-i}|\theta_i) \\ &= \max_{m_i \in \Theta_i} \int_{\Theta_{-i}} \int_Z u_i(z, \theta_i, \theta_{-i}) dG_{\pi_I^M}(z|m_i, \theta_{-i}) I(d\theta_{-i}|\theta_i), \qquad (I\text{-}IC) \end{aligned}$$

almost everywhere conditional on I, that is, $\forall \theta_i \in supp(I_i)$. The second line follows because π_I^M is incentive compatible under information structure I (I-IC henceforth). Truthful reporting is optimal for all types of all players given π_I^M.

The existence of equilibrium under $\mathcal{I}^0$ implies that the collection of possible status-quo outcomes, $\Pi^M := \{\pi_I^M\}_{I \in \mathcal{I}^0}$ with π_I^M being I-IC, is well-defined.

Mechanism. The mechanism is an alternative to the status quo. Any mechanism is a game of incomplete information represented by a decision rule. The collection of decision rules is Π. Given π and I, we define each player's optimal reporting strategy $m_{i,I}(\theta_i)$. We collect players' reports in $m_I(\theta)$. An equilibrium of π implements the decision rule $\pi_I := \pi \circ m_I : \Theta \to \Delta(Z)$ which is I-IC.[2]

[2] Although any decision rule in Π represents a direct revelation mechanism, a truthful implementation may not be guaranteed. Indeed, Π is shorthand for all game forms with m_i a player's action. The equilibrium play of each $\pi \in \Pi$ under I then induces some I-IC decision rule.

The set of available mechanisms, Π, may be restricted by legal or institutional constraints, or particular outcomes may simply be infeasible. We assume two minimal requirements on the set of available mechanisms:

(i) $\Pi^M \subseteq \Pi$, and

(ii) Π is closed under convex combinations, that is, if $\pi, \pi' \in \Pi$, then for any $\lambda : \Theta \to [0,1]$ it holds that $\lambda\pi + (1-\lambda)\pi' =: \pi^\lambda \in \Pi$.

The first property implies that the mechanism can replicate the status quo. The second property implies that if two games (1 and 2) are part of the available mechanism, so is the game in which game 1 is played for specific type reports and game 2 for the remaining type reports.

Apart from these requirements, we do not restrict Π. Instead, we allow for both a classical mechanism design setting and the possibility that the designer's set of mechanisms is exogenously limited. The latter includes pure 'mediation' within the status quo.

Informational Punishment. We assume all players have access to a signaling device Σ. The N-dimensional random variable $\Sigma : \Theta \to S$ maps type reports into realizations in signal space $S \equiv S_1 \times S_2 \times .. \times S_N$ with $|S_i| \geq |\Theta_i|$. We denote the realization of Σ by $\sigma \in S$, that of element Σ_i by $\sigma_i \in S_i$.

Timing. First, players learn their types and observe (π, Σ). Second, they simultaneously send a message m_i^Σ to Σ. Third, players simultaneously decide whether to veto the mechanism. If at least one player vetoes the mechanism, the set V of vetoing players becomes common knowledge, and the signal realizations σ become public. Players use that information to update to an information structure $I^{V,\sigma} \in \mathcal{I}^0$ and the status quo implements $\pi^M_{I^{V,\sigma}}$. If players unanimously ratify the mechanism, they report m_i to the mechanism which implements π.

Solution Concept and Veto Beliefs. We consider all mechanism-signaling device combinations (π, Σ) implementable as the grand game's perfect Bayesian equilibrium (PBE) using the definition from Fudenberg and Tirole (1988).

We use *veto information structures*, I^V: the information structure that arises *after* an observed veto, but *before* the realization σ. PBE implies that $I^V(\theta_{-i}|\theta_i) = I^{V\setminus i}(\theta_{-i}|\theta_i)$ for any $i \in V$ and $I^V(\theta_{-i}|\theta_i) = I^{V\cup i}(\theta_{-i}|\theta_i)$ for any $i \notin V$. In addition, all but first-node off-path beliefs on deviators follow Bayes' rule. The remaining off-path beliefs are arbitrary.[3]

[3] In our setting, a player is at most observed to deviate once. Off-path belief cascades (see Sugaya and Wolitzky (2020)) are thus not possible in our model.

3 Analysis

3.1 Main Result

An optimal full-participation mechanism may not exist, absent informational punishment, even if Π is large. Consider the case in which all players participate on the equilibrium path. A deviator i who vetoes guarantees herself the prior belief about the other players through that deviation. At the same time, the deviator is most punished by the "worst" off-path belief assigned to her. If i's outside option exhibits concavities in the information structure, full participation *without* informational punishment is not always optimal.

With informational punishment, the designer can relax i's participation constraint without relying on on-path rejections and (possibly) beyond what is implementable with rejections.

Proposition 1. *It is without loss of generality to focus on mechanisms that ensure full participation if informational punishment is available.*

Proof. The proof is constructive. Take any π, and a *veto equilibrium* in which the mechanism is vetoed with positive probability on the equilibrium path. We first characterize the decision rule of the veto equilibrium. Then, we show it can be implemented with full participation using *informational punishment.*

Let $\xi(\theta)$ be the probability that π is vetoed given type profile θ. Moreover, $\xi_i(\theta_i)$ is the likelihood that type θ_i vetoes π on the equilibrium path. If players mix regarding their veto decision, the set of players that vetoed, V, might be random. After a veto, players observe the set of players V_k that vetoed and update to information structure I^{V_k}. Outcomes realize according to $\pi^M_{I^{V_k}}$. Taking expectations over all realizations of V, V_k, the ex-ante expected continuation game conditional on a veto is a lottery $(P(V_k), \pi^M_{I^{V_k}})$ defined over all V_k. $P(V_k)$ is the on-path likelihood that a veto is caused by the set V_k and not by any other set. Because $\Pi^M \in \Pi$ and Π is closed under convex combinations, the lottery implies $\pi^M_{I^{\mathbb{E}V_k}} = \sum P(V_k)\pi^M_{I^{V_k}} \in \Pi$.

Conditional on no veto, the information structure is I^a, and π_{I^a} is the decision rule.

The grand game implements an I^0-IC decision rule $\pi'_{I^0} := \xi\pi^M_{I^{\mathbb{E}V_k}} + (1-\xi)\pi_{I^a}$. Again, $\pi' \in \Pi$ because Π is closed under convex combinations.

We now construct a signal Σ such that the mechanism π'_{I^0} is implementable under full participation. By construction, π' is feasible and I^0-IC. What remains is to show that no player has an incentive to veto π'.

We construct the following signaling device $\Sigma_i : \Theta_i \to \Delta(\{0,1\})$ where $\sigma_i(\theta_i) = 1$ with probability $\xi_i(\theta_i)$ and 0 otherwise. When observing off-path behavior (i.e., a veto) by i, j believes that i has randomized uniformly over the entire type-space when reporting to Σ_i. Thus, she disregards σ_i. We choose the off-path belief on i identical to the belief attached to i after observing her unilateral veto in the veto equilibrium.

No player i has an incentive to veto the mechanism. If a player vetoes the mechanism, Σ provides her with the same lottery over information structures that she expects from a veto in the veto equilibrium. Participation, in turn, gives the same outcome as the veto equilibrium. No player can improve the outcome of the veto equilibrium by vetoing π'.

Truthful reporting to Σ_i is a best response. Σ_i is on-path payoff-irrelevant. Thus, under (π', Σ) an equilibrium with full participation in π' exists that implements the same outcome as the veto equilibrium. □

Optimal Informational Punishment

Suppose the task is to design the optimal mechanism under some objective. Proposition 1 shows that a mechanism with informational punishment and full participation can replicate the outcome of any mechanism with on-path rejection. However, it may not be immediate from Proposition 1 how that property simplifies the mechanism design problem. Here, we outline the implied simplification. Informational punishment is most effective if the signaling device conditions on the deviator's identity. That is, $\Sigma^V : \times_{j \in N \setminus V} \theta_j \to \Delta(S)$ is the signaling device used if the set of vetoing players is V. Specifically, Σ^i is the mapping from reports to realizations used if (only) i vetoes the mechanism.

Informational punishment separates the full-participation problem from the mechanism design problem. The reason is straightforward in light of the proof of Proposition 1. Informational punishment affects the participation constraints only. Moreover, by Proposition 1 it is without loss to restrict attention to optimal mechanisms in which vetoes are off-path events. Therefore, we can limit our attention to unilateral vetoes when constructing the optimal mechanism.

What remains is to find Σ^i that punishes deviator i the most. If player i vetoes, all other players hold an off-path belief about player i's type and use information structure I^i to update any signal sent by Σ^i. Let $P(I) \in [0,1]$ be the probability that information structure $I \in \mathcal{I}^0$ prevails if player i vetoes the mechanism. To solve for the optimal Σ^i, we can solve for the lottery over information structures $(P(I), I)$, such that $\sum_I P(I) = 1$. Formally, we solve the following simple information design problem.[4]

$$\min_{P(I)} \sum_I P(I) \sum_{\theta_i} \alpha(\theta_i) v_i(\theta_i, I, \pi_I^M)$$
$$s.t. \sum_I P(I) I = I^i, \text{and}$$
$$I(\theta_i|\theta_{-i}) = I^i(\theta_i|\theta_{-i}),$$

where $\alpha(\cdot) \in [0,1]$ with $\sum_{\theta_i} \alpha(\theta_i) = 1$ are weights corresponding to the Lagrangian

[4] See, e.g., Bergemann and Morris (2016) and references therein for more information on information design problems.

multipliers of the binding participation constraints in the mechanism design problem (see Jullien (2000)).[5] The first constraint implies that every I results from some feasible signal Σ^i: the signal is *Bayes' plausible.* The second constraint means that Σ^i cannot reveal information about the deviator who vetoed the mechanism: the belief on the deviator, $I^i(\theta_i|\theta_{-i})$, is constant. Players do not observe whether the deviator has deviated only at the ratification stage or already before that. Therefore, they attach a single belief to the deviator's type distribution.

The solution to the information design problem relaxes player i's participation constraint the most. Thus, we can determine each player's least binding participation constraint, player-by-player. Once participation constraints under informational punishment are determined, we can design the mechanism taking each player's participation constraint as exogenously given. Using the same arguments as those in the concavification literature (Aumann and Maschler, 1995), optimal informational punishment *convexifies* the deviation payoffs and thus minimizes the gains from a veto.

3.2 Other Environments

This section shows that our results extend straightforwardly to more complex settings. We begin by looking at refined equilibrium concepts. Then we consider informed-principal problems. Finally, we reduce the designer's commitment.

Refinements

Proposition 1 assumes that the designer can freely pick off-path beliefs under the PBE restriction. She can select any first-node off-path belief of the continuation game. Depending on the context and the application, such an equilibrium selection may not be reasonable. Using a refinement could instead make on-path vetoes unavoidable because it limits the designer's equilibrium choice set in the first place.[6]

Our second finding is that Proposition 1 is robust to most common refinements. Specifically, whenever we refine the equilibrium concept according to

$$(\star) \in \{\text{Perfect Sequential Equilibrium, Intuitive Criterion, Ratifiability}\},$$

full participation remains optimal.

[5] In many standard environments, it holds that $\alpha(\theta_i) = 1$ for some θ_i, i.e. only the participation constraint of one type binds. Moreover, for many default games, it is the case that the solution of the information design problem is independent of α. Such a situation occurs in settings where there is a 'worst' information structure for all types of the deviator. This property holds, in particular, in games that feature strategic complements or strategic substitutes.

[6] Correia-da-Silva (2020) provides additional discussions on this issue and how it may interfere with the design of a mechanism.

Proposition 2. *Suppose the solution concept is perfect Bayesian equilibrium with refinement concept $(\star)$, and informational punishment is available. Then, focusing on mechanisms that imply full participation is without loss of generality.*

Proof. Ratifiability requires full participation in the mechanism and therefore holds trivially, as the designer can always choose a degenerate signaling device. It thus is without loss of generality to show full participation under refinement $(\star)' \in \{$Perfect Sequential Equilibrium, Intuitive Criterion$\}$.

Consider the veto equilibrium used in the proof of Proposition 1. Suppose this equilibrium satisfies refinement $(\star)'$.

We show that the full-participation equilibrium constructed in the proof of Proposition 1, (π', Σ), satisfies the same refinement criterion. Two aspects are crucial. First, compare the equilibrium with vetoing and that with full participation. On-path (expected) outcomes and those that are off-path but can be reached by a unilateral deviation are identical between these two equilibria for every state θ. First, take any state θ in which the mechanism is unanimously accepted in both equilibria. Then both outcomes coincide, and so does the credibility of the beliefs. Second, consider a state θ in which the mechanism is rejected in the veto equilibrium with positive probability. For the same state, suppose that π' is rejected in the full-participation equilibrium—an off-path event. The resulting *off-path belief* on the deviator θ_i coincides with the *on-path belief* on the same θ_i in the veto equilibrium.

Thus, the constructed off-path beliefs put positive mass only on those types that *weakly prefer to deviate*, while no such type *strictly prefers to deviate*. Thus, any off-path belief for type θ_i is credible in the sense of Grossman and Perry (1986), *and* off-path beliefs do not violate the intuitive criterion. □

Informed-Principal Problems

The informed-principal environment assumes that a privately informed player proposes the mechanism which should replace the status quo.

Formally, instead of a non-strategic third party, one of the players, say $i = 0$, proposes a mechanism as an alternative to the default game. The setting becomes an informed-principal problem. Players $i = 1, ..., N$ are the agents.

A key concept to solve informed-principal problems is the concept of inscrutability (see Myerson, 1983). It states that it is without loss to assume that the informed principal, player 0, selects a mechanism that does not allow the other players $1, ..., N$ to learn about the principal's type from the proposed mechanism. That is, inscrutability means it is without loss to restrict attention to pooling solutions in which each principal type offers the same mechanism.

The default game can depend non-linearly on beliefs about player 0's type. Consequently, the principle of inscrutability might fail. Player 0 may have strict incentives to signal private information via the mechanism proposal. Thereby player 0 relaxes the other players' participation constraints. The following result states that these concerns are irrelevant if informational punishment is available.

Proposition 3. *The principle of inscrutability holds if informational punishment is available.*

Proof. Consider an equilibrium of the grand game such that different types of player 0 propose different mechanisms πs. Let $\mathcal{M}$ be the set of πs that are proposed with strictly positive probability. Let $\xi_0^\pi(\theta_0)$ denote the probability that player 0 type θ_0 proposes mechanism $\pi \in \mathcal{M}$.

Consider the case in which at least one type of one player vetoes some $\pi \in \mathcal{M}$ on the equilibrium path. We refer to this equilibrium as the separate-and-veto equilibrium. Recall that if π is vetoed, some rule in Π^M results. Let the probability that π is vetoed be ξ^π. Moreover, $\xi_i^\pi(\theta_i)$ is the probability that θ_i vetoes π. The separate-and-veto equilibrium implements a I^0-IC decision rule, $\pi_{I^0}^E$. $\pi_{I^0}^E \in \Pi$ because Π is closed under convex combinations.

We prove the existence of the following equilibrium. All types of player 0 propose $\pi_{I^0}^E$ and every player accepts it. This equilibrium leads to the I^0-IC decision rule $\pi_{I^0}^E$. We construct a signaling device Σ to support acceptance of $\pi_{I^0}^E$. Let $o : \mathcal{M} \to \mathbb{R}$ be some invertible function. For $i = 0$, we construct the signal $\Sigma_0(\theta_0)$ with support $\{o(\Pi')\}_{\Pi' \in \mathcal{M}}$ and associated probabilities $Pr(\Pi'|\theta_0) = \xi_0^{\Pi'}(\theta_0)$. For any $i > 0$, let the signal be $\Sigma_i(\theta_i) = 1$ with probability $\xi_i(\theta_i)$ and $\Sigma_i(\theta_i) = 0$ with remaining probability. Whenever player $i > 0$ vetoes, a signal realizes according to Σ. Thus, the reason why no player rejects Π^E is the same as in the proof of Proposition 1. The only difference is that the signaling function also replicates the potential signal-by-mechanism-choice behavior of the principal, captured by Σ_0. □

Informational Opportunism

Our design approach on the side of the signal assumes that the signaling device Σ is fixed. In other words, the signal designer has commitment power. However, if the signal is created through passing information to a third player, e.g., a journalist, then that assumption may not always hold. In such a setting, the person that keeps the information may be a strategic player. In addition, she could be unable to commit to both the signaling device and report its outcome. Dequiedt and Martimort (2015) refer to such a setting as *informational opportunism.*

To adequately address the situation with informational opportunism, we need to take a stance on two aspects. First, how large is the signal designer's commitment power? Second, does the designer's objective change once she sees a deviation? Is it credible for her to use the information to punish the deviator?

An extreme form of informational opportunism occurs if the signal designer chooses the *signal realization*, σ, after receiving a report rather than the *signaling device*, Σ. In that case, the signal cannot commit to a mapping from reports to realizations. Instead, the signal becomes a cheap-talk announcement. The results of Propositions 1 to 3 trivially do not hold without any commitment power, even if the signal designer wants to *punish the deviator*.

Instead, we focus on a signal designer whose action space is still the set of signaling devices. We make the following assumption.

Assumption 1 (No Fabricated Data)**.** The choice of signaling device, Σ, becomes public together with its realization, σ.

Under Assumption 1 the signal designer can back up her claims by providing evidence. Indeed, all interested players can see the designer's method to reach her conclusion, σ. Within this setting, informational opportunism is best seen when considering the timing of the grand game. In our baseline model, the signal designer commits to Σ *before* players decide about vetoing. She is thus an impartial third-player and not an interested player in the grand game. In contrast, we assume now that the designer is an interested player in the game.

First, we assume that the signal designer can commit to an objective: to punish a potential deviator. However, she *cannot* commit to her signaling device at the beginning of the game. Instead, she picks Σ after players have made their acceptance decision and the designer has elicited the information. That is, the signal designer faces *ex-post incentive constraints.*

Definition 1 (Informational Opportunism)**.** The designer of the signaling device suffers from *informational opportunism* if she cannot commit to a signaling device Σ before players' participation decision.

We are interested in situations in which our previous results are *immune* to informational opportunism.

Definition 2 (Immunity)**.** A result is *immune to informational opportunism* if it is implementable by a signal designer that suffers from *informational opportunism.*

Allowing for informational opportunism comes at a cost. We cannot make a general statement on immunity. However, we state a set of definitions that restricts the environment. These restrictions allow us to state Proposition 4, which determines a condition for immunity.

We identify a signal designer by her *type.* The designer's type is the information she has elicited from the (participating) players. Observe that any signal designer type *could* fully reveal her type by choice of the appropriate signaling device.

Yet, if the signal designer chooses not to reveal her type, the interpretation of realization σ depends not only on Σ. It also depends on the belief that players form about the signal designer. To form the belief, players observe Σ. Each Σ triggers a belief which, together with Σ, leads to a lottery over information structures. Different designer types potentially have different preference rankings over lotteries for a given signal designer's objective. We assume, however, that preferences are aligned, and all types share the same order. Thus, there is a common understanding of which information structures better achieve the desired goal.

To achieve immunity to informational punishment, we impose two properties on the environment. First, the players' types are distributed independently. Second, the signal designer has *aligned preferences* across her types.

Definition 3 (Aligned Preferences)**.** Fix arbitrary distributions over a collection of $(N-1)$ players' types. Let F and F' be two (possible) distributions of the remaining player i's type. The signal designer has aligned preferences if every designer type prefers F to F' whenever F first-order stochastically dominates F'.

Finally, we define an extreme notion of the desire to separate.

Definition 4 (Unraveling Pressure)**.** A signal designer faces unraveling pressure under signaling device Σ if she strictly prefers to verify her type to the lottery induced by Σ.

Suppose types are independently distributed, and preferences are aligned. In that case, the absence of unraveling pressure is necessary and sufficient to guarantee that a signaling device is implementable, as the following proposition shows.

Proposition 4. *Suppose players' types are independently distributed, and the signal designer's preferences are aligned. Then, a signaling device Σ is implementable under informational opportunism if and only if no signal designer type faces unraveling pressure.*

Proof. The "only if" direction follows from Definition 4. If a signal designer type faces unraveling pressure, she prefers to reveal her type over the signaling device Σ.

For the "if" direction, consider a signaling device Σ. Assume player i has vetoed the mechanism. The signal designer elicited the information θ_{-i} from the non-deviating players. We want to show that no designer type, θ_{-i}, has an incentive to announce a different device than Σ.

Suppose signal designer type θ_{-i} deviates by announcing Σ' which does not verify θ_{-i}. Players observe the deviation Σ' and its realization σ'. Using these objects, they form off-path beliefs about the types of all $N-1$ players. The symmetry of PBE and the independence of players' types imply the following. Any subset of players has identical beliefs about those not in that subset.

The off-path beliefs on the signal designer's type are only restricted by the signaling function Σ'. If a realization σ' occurs with probability 0 given a type θ_{-i}, then players exclude that type from the set of possible signal designer types. Denote the set of not excluded types by $\Theta^{\sigma'}$. The distribution $F^{\sigma'}: \Theta^{\sigma'} \to [0,1]$ is arbitrary. That is, for every

Σ', there always exists an off-path belief about the deviating designer that rationalizes $F^{\sigma'}$.

By assumption Σ' does not verify θ_{-i}. Thus, $|\Theta^{\sigma'}| > 1$. Types have aligned preferences. Thus, we can find a signal designer type $\tilde{\theta}$ such that a degenerate belief on $\tilde{\theta}$ makes every designer type other than $\tilde{\theta}$ worse off compared to that signal designer revealing her type. No unraveling pressure implies that no type benefits from the deviation. Σ is implementable under informational opportunism. □

The immunity of our results to informational opportunism is a straightforward corollary to Proposition 4. Moreover, it provides a simple way to test for immunity given a candidate Σ.

Corollary 1. *Suppose players' types are independently distributed, and the designer's preferences are aligned. Propositions 1 to 3 are immune to informational opportunism if no signal designer type faces unraveling pressure under* Σ.

We conclude our discussion by addressing a weaker notion of informational opportunism.

The signal designer may commit to a signaling device Σ but may choose to conceal the realization. We refer to this as weak informational opportunism.

Definition 5 (Weak Informational Opportunism)**.** The designer of the signaling function Σ suffers from *weak informational opportunism* if she can commit to a signaling function Σ at the beginning of the game, but not to the disclosure of the realization σ.

It is straightforward to see that immunity to informational opportunism implies immunity to weak informational opportunism. In addition, we can drop the no-unraveling pressure condition. The reason is that—with aligned preferences—there is a common worst realization $\underline{\sigma}$ across signal designer types. If a signal designer hides information off the equilibrium path, an off-path belief assuming a (hidden) realization of $\underline{\sigma}$ punishes every designer type the most. Consequently, using the standard unraveling arguments from the persuasion literature, (see, e.g., Grossman, 1981; Milgrom, 1981), no designer has an incentive to hide her information. Moreover, the result applies even if players are unaware of the signal designer's objective.

Corollary 2. *Suppose players' types are independently distributed, and the designer's preferences are aligned. Propositions 1 to 3 are immune to weak informational opportunism even when players have ambiguity over the signal designer's objective function.*

The reason for Corollary 2 is that if the signal designer's types preferences are aligned, then given any objective, Σ has a common worst signal realization. Not revealing the signal realization leads to an arbitrary off-path belief. Thus, even if the designer's objective is unclear, parties can coordinate on an off-path belief in some PBE that puts all probability mass on the worst signal.

4 Final Remarks

Mechanism design can facilitate policy advice. It provides a simple benchmark that informs what is possible theoretically. The power of mechanism design derives from its simplicity in calculations. Invoking the revelation principle, we can derive strong results even in a complex environment.

However, suppose the environment is such that the designer cannot control the entire strategic setting, and parties can block the implementation of the mechanism. In that case, the revelation principle for the part the designer controls can fail. The reason is that parties can use their veto power to signal private information strategically. To obtain the desired benchmarks, we would have to either restrict the environment to settings in which signals through vetoes are non-profitable or delve into complex case distinctions.

In this paper, we argue that restricting to the setting in which strategic vetoes are of no concern is without loss provided that parties have access to a tool we call *informational punishment.* Informational punishment allows parties to store information for some time and release a garbled version of it in case of a deviation. Furthermore, we show that through informational punishment, we can transform every environment in which strategic vetoes are relevant into an equivalent setting in which they are not. Thus, we can—without loss—restrict ourselves to full-participation mechanisms under an (appropriate) outside option.

Our results go beyond classical applications of mechanism design. We derive a minimum condition on the available mechanism space that determines whether informational punishment guarantees full participation at an optimum. Informational punishment works off the equilibrium path, does not affect incentive compatibility directly, and allows for publicly verifiable rejections. We can implement informational punishment through a centralized signal, through the designer of the mechanism who could also be an informed principal or decentralized through the parties individually. Furthermore, informational punishment is robust to various additional constraints on the setting.

References

Aumann, R. J. and M. Maschler (1995). *Repeated games with incomplete information.* Cambridge, MA: MIT press.

Balzer, B. and J. Schneider (2021). "Persuading to Participate: Coordinating on a Standard". *International Journal of Industrial Organization* 78, p. 102764.

Bergemann, D. and S. Morris (2016). “Bayes correlated equilibrium and the comparison of information structures”. *Theoretical Economics*, pp. 487–522.

Celik, G. and M. Peters (2011). “Equilibrium rejection of a mechanism”. *Games and Economic Behavior* 73 (2), pp. 375–387.

— (2016). “Reciprocal relationships and mechanism design”. *Canadian Journal of Economics* 49 (1), pp. 374–411.

Cho, I.-K. and D. M. Kreps (1987). “Signaling Games and Stable Equilibria”. *Quarterly Journal of Economics* 102 (2), pp. 179–221.

Correia-da-Silva, J. (2020). “Self-rejecting mechanisms”. *Games and Economic Behaviour* 104, pp. 434–457.

Cramton, P. C. and T. R. Palfrey (1995). “Ratifiable mechanisms: learning from disagreement”. *Games and Economic Behavior* 10 (2), pp. 255–283.

Dequiedt, V. (2007). “Efficient collusion in optimal auctions”. *Journal of Economic Theory* 136 (1), pp. 302–323.

Dequiedt, V. and D. Martimort (2015). “Vertical Contracting with Informational Opportunism”. *American Economic Review* 105 (7), pp. 2141–2182.

Fudenberg, D. and J. Tirole (1988). “Perfect Bayesian and Sequential Equilibria - A clarifying Note”. *mimeo.*

Gerardi, D. and R. B. Myerson (2007). “Sequential equilibria in Bayesian games with communication”. *Games and Economic Behavior* 60 (1), pp. 104–134.

Grossman, S. J. and M. Perry (1986). “Perfect sequential equilibrium”. *Journal of Economic Theory* 39 (1), pp. 97–119.

Grossman, S. J. (1981). “The informational role of warranties and private disclosure about product quality”. *The Journal of Law and Economics* 24 (3), pp. 461–483.

Jullien, B. (2000). “Participation constraints in adverse selection models”. *Journal of Economic Theory* 93 (1), pp. 1–47.

Milgrom, P. R. (1981). “Good News and Bad News: Representation Theorems and Applications”. *Bell Journal of Economics* 12 (2), pp. 380–391.

Myerson, R. B. (1983). “Mechanism design by an informed principal”. *Econometrica*, pp. 1767–1797.

Sugaya, T. and A. Wolitzky (2020). “Revelation Principles in Multistage Games”. *The Review of Economic Studies* 88 (3), pp. 1503–1540.

Tan, G. and O. Yilankaya (2007). “Ratifiability of efficient collusive mechanisms in second-price auctions with participation costs”. *Games and Economic Behavior* 59 (2), pp. 383–396.