Persuasion in Evidentiary Mechanisms

Sam Kapon*

September, 2024

Abstract

How should a regulator reveal private evidence of a crime committed by multiple agents (e.g., a cartel) to spur whistleblowing by members of that group? I formalize this question using a model of information design in games, in which a regulator sends private signals to agents, who can then communicate amongst themselves before simultaneously deciding whether to reveal evidence to the regulator. I first show that, with two firms, the regulator can do as well as if she could shut down communication between the firms. Interpreting the state as the probability of conviction *without* a whistleblower, I characterize optimal outcomes and show that the likelihood of whistleblowing is increasing in this probability, and that the principal can facilitate more whistleblowing in groups with more asymmetrically distributed gains from crime. Finally, I demonstrate a class of simple information structures that improve over public communication.

1 Introduction

Regulatory agencies often use whistleblower rewards to destabilize groups of misbehaving agents. For instance, the Department of Justice (DOJ) operates a leniency policy, whereby cartel members may provide evidence against their partners and secure lenient treatment during prosecution of the cartel. Similar policies are operated by the European Commission, and many other antitrust authorities. These policies are key to the success of antitrust

*Haas School of Business, UC Berkeley: skapon@berkeley.edu.

efforts; indeed, the DOJ calls its leniency policy "its most important investigative tool for detecting cartel activity."[1,2] The standard of evidence required to prove antitrust violations is high; as a result, cooperation from a member of the conspiracy is valuable.

Prior to a whistleblower coming forward, the regulator often possesses evidence of wrongdoing; in the case of cartels, this could be information from third parties, such as aggrieved buyers, documents discovered in unannounced inspections (dawn raids), or suspicious market activity. This evidence can be used to encourage whistleblowers. An agent who learns that the regulator has strong rather than weak evidence may be more incentivized to approach the regulator with information to avoid harsh punishment. The central question of this paper is then, how should a regulator reveal its private evidence to a group of misbehaving agents, to encourage whistleblowing? I focus on three sets of sub-questions. How harmful is communication amongst agents to the principal? How do optimal outcomes vary with the underlying features of the misbehaving group? Finally, can simple information structures improve over public communication without exact knowledge of primitives?

To address these questions, I study a problem of information design in games, with a principal (sender) and two agents (receivers). The principal commits to an information structure over a *state*, sending a private signal to each agent. In the motivating settings, the state represents the principal's private evidence, and so I refer to it as the *evidence* state. Agents share a prior over possible evidence states and, after communication from the principal, each agent chooses one of two actions: inform the principal about the group's misbehavior, or not. Payoffs are state-contingent, and in every evidence state, the principal prefers more agents inform. I make two assumptions on agents' preferences: (i) each agent prefers their partner *not* inform the principal, and (ii) each agent prefers to inform the principal if their partner does. Given the assumptions on agents' preferences, equilibrium multiplicity may arise. While there always exists a favorable equilibrium for the principal in which both agents inform, there may also exist an unfavorable equilibrium in which neither does. I evaluate information structures robustly, by the worst possible equilibrium for the principal that they generate.[3,4]

[1] https://www.justice.gov/atr/leniency-program

[2] In the literature studying antitrust leniency policies, a cartel member who brings forward evidence to the regulator with the objective of receiving leniency is typically called a *leniency applicant*, while the term *whistleblower* is often instead reserved for those outside the cartel who provide the regulator with evidence. In this paper, I instead use the term *whistleblower* to refer to the former.

[3] More precisely, information structures are evaluated by taking the infimum over equilibria, but I abuse terminology until formally describing the model.

[4] Other equilibrium selection devices work as well. For instance, the principal's worst equilibrium among the set of equilibria that are Pareto efficient for agents leads to identical results.

While Bayesian Nash Equilibrium (BNE) is a standard solution concept in related settings, a natural feature of this environment is that agents can communicate. At least in the case of cartels, they are already communicating about various aspects of the crime, and may even have access to an explicit mediator of communication.[5] As a result, the set of outcomes over which the principal's worst case is evaluated is the set of *communication equilibria*, allowing for the possibility that players communicate private information supplied to them by the principal before acting. Communication is formulated generally, as in Myerson (1982)—a communication equilibrium is a mapping from type reports by agents into a distribution over private recommendations to agents (which can be interpreted as coming from a mediator) of whether to inform or not, such that reporting one's type and obeying the recommendation is incentive compatible.

Turning first to the question of the effects of communication on the principal's payoff, I begin by showing that a class of well-known information structures (see Halac, Lipnowski, and Rappoport (2022) and Morris et al. (2024)) is immune to communication amongst firms—they generate the same worst-case outcome for the regulator whether firms can communicate or not.[6] These information structures have the following structure. There is one public signal accompanied by no private signals, and agents both choose not to whistleblow in the principal's worst BNE. There is another public signal accompanied by private signals that ensure whistleblowing is iteratively dominant (as in the email game or global games). As long as each private type in the latter case believes that there is at most one type of his partner who could be lower in the iterative dominance order, the information structure is immune to communication. I then show that, in supermodular environments—whistleblowing by one agent increases the incentive for the other to whistleblow—it is without loss of value to restrict to information structures implementing *perfectly coordinated* outcomes, in which either both agents inform or neither does in the principal's worst equilibrium. Finally, I combine these results with results in Morris et al. (2024) to conclude that, in supermodular environments, the principal incurs no loss from communication between firms.

I then ask, how do optimal outcomes look and how do they vary with underlying features of the environment? To provide sharp results, I specialize to *linear* environments, in which evidence states are identified as real numbers and agents' preferences are affine in the evidence state. I show that, as long as whistleblowing is dominant for both agents at the

[5]See for instance, the activities of AC-Treuhand as a facilitator of cartels (Vallery and Schell, 2016).

[6]A closely related argument, showing that communication is impossible in an electronic mail game with negative spillovers (in the language of this paper, each agent prefers his partner not inform) first appears in Baliga and Morris (1998).

highest evidence state, there exists an optimal outcome in which the likelihood that agents whistleblow takes a threshold form: both agents whistleblow when the state is above the threshold, and neither does when the state is below the threshold. I then show that the principal's optimal value increases as the agents' payoffs when neither informs become *more asymmetric*. I discuss interpretations in the context of antitrust, in particular how shocks to a market can create asymmetries in cartels, making them more susceptible to breakdown via information design.

Finally, I show that, if the regulator can identify one agent as having weaker whistleblowing incentives—dominance of whistleblowing for that agent implies dominance for his partner—then a simple adjustment to any public information structure weakly improves the principal's value. In this adjustment, the principal reveals the state to the agent with weaker whistleblowing incentives if and only if whistleblowing in that state is dominant, and reveals nothing to his partner.

All results are presented for 2 firms, and it is natural to consider the case of more than 2 firms. In Section 8, I show that the information structures that would naturally extend perfectly coordinated information structures described for 2 firms can perform poorly with 3 firms. In particular, as the slack in incentive constraints in these information structures disappears—as would typically happen as the regulator's choice approaches optimality—worst communication equilibrium outcomes converge to the regulator's worst possible outcome.

The paper proceeds as follows: after describing the literature, I describe the model in Section 2. I define the relevant information structures and their key property without communication in Section 3, and establish that they are communication-proof if they satisfy an additional restriction on the type distribution in Section 4. I establish that the class of such information structures is rich enough to solve the principal's problem in supermodular environments in Section 5. I study linear environments in Section 6, describe simple information structures in Section 7, and provide results for 3 firms in Section 8.

Literature This paper is related to the literatures on the optimal design of self-reporting policies, especially in the context of collusion, and joins a growing literature concerned with information design in games under adversarial equilibrium selection and the closely related literatures on contracting with externalities and unique implementation.

The optimal design of self-reporting, amnesty, whistleblowing and leniency policies in group settings—a primary example being cartels—has spawned a large literature, including

Spagnolo (2000), Motta and Polo (2003), Harrington Jr (2008), Miller (2009), Harrington Jr (2013), Gamba, Immordino, and Piccolo (2018), and Landeo and Spier (2020).[7] Many of these papers study design questions, but most focus on the design of the *payoff* environment, whereas I focus on the design of the *information* environment for a fixed payoff environment. Harrington Jr (2013) studies an environment in which firms can choose whether to apply for leniency and have private information about the likelihood of conviction without a leniency applicant. In discussing potential future work in its conclusion, Harrington Jr (2013) poses the question: when the regulator has its own private information, how should it reveal it to the firms to encourage them to come forward? This is the overarching question taken up in this paper. A closely related paper, in motivation, in this literature is Sauvagnat (2015), which studies the problem of an antitrust regulator who privately observes a binary signal about the strength of its case, and can commit to a policy of opening a costly investigation as a function of the signal. The regulator can also design a leniency policy, that rewards cartel members for reporting information on the cartel after an investigation has begun. If the regulator opens an investigation always when evidence is strong and sometimes when it is weak, this can entice leniency applicants, and hence create cartel breakdown, even when the principal's evidence is weak. This paper provides a complementary analysis, pursuing further the idea of the regulator signaling the strength of her evidence to cartel members, allowing for general information policies and private communication.[8]

The literature on unique implementation, contracting with externalities and divide-and-conquer schemes includes Abreu and Matsushima (1992), Winter (2004), Segal (2003), Bernstein and Winter (2012), Halac, Kremer, and Winter (2019), Halac, Lipnowski, and Rappoport (2020), Moriya and Yamashita (2020), Chan (2023), Camboni and Porcellacchia (2024), Chassang, Del Carpio, and Kapon (2022), Halac, Lipnowski, and Rappoport (2022). This literature studies how a principal can use incentives, or both information design and incentives, to uniquely implement a desirable outcome. Closely related (and often overlapping) is the literature on information design with adversarial equilibrium selection, for instance recently in Bergemann and Morris (2019) (Section 7.1), Mathevet, Perego, and Taneva (2020) (Section V), Ziegler (2020), Sandmann (2021), Li, Song, and Zhao (2022), Hoshino (2022),

[7] For a comprehensive survey, see Marvão and Spagnolo (2018).

[8] Another relevant paper is Chassang and Ortner (2022), which details the process of regulating collusion, and identifies a number of avenues for future research, one of which is to better understand how a regulator can leverage privately held evidence to facilitate cartel breakdown. Chassang and Ortner (2022) also provide a discussion of cases relating to the standards of evidence required by a court, in particular comments by Judge Richard Posner of the U.S. Court of Appeals of the Seventh Circuit in re Text Messaging Litigation (2010), as well as the Supreme Court case Bell Atlantic v. Twombly (2007).

Morris, Oyama, and Takahashi (2024), and Inostroza and Pavan (2023).[9] Key in both is the idea that to implement a desirable action profile, the design tool is deployed to make it dominant for some agents to take their assigned action, (iteratively) dominant for another group of agents to take their assigned actions given the behavior of the first group, and so on. This logic features centrally in the analysis of this paper. A novel aspect of this paper is the communication allowed between agents, as well as results provided regarding linear environments and simple information structures. In the aforementioned literature, the principal seeks unique or worst-case implementation under BNE (or rationalizability). In this paper instead, the principal designs under a worst-case communication equilibrium criterion, a concept that allows agents to communicate private information supplied to them by the principal. I use a result in Morris, Oyama, and Takahashi (2024), which studies unique and smallest BNE implementation in two action supermodular games, to prove that the principal's optimal value is independent of whether agents can communicate or not in supermodular environments (Proposition 2).

The paper is also related to the literature on global games and robustness of equilibria to incomplete information, as well as the investigation of cheap talk in that context. Early papers include Rubinstein (1989), Carlsson and Van Damme (1993), Kajii and Morris (1997), and a large literature has followed. The argument that, because of incentives to deceive other agents, communication is impossible in an electronic mail game with negative spillovers appears in Baliga and Morris (1998), and a closely related argument underlies the failure of communication in this paper.[10]

The paper is also more broadly related to recent theoretical work on reporting in crime such as Chassang and Padró i Miquel (2019), Dannay (2019), Lee and Suen (2020), Pei and Strulovici (2024) and Angelucci and Russo (2022). For instance, Pei and Strulovici (2024) study the informativeness of accusations of wrong-doing against a potential criminal, when accusers may have an incentive to lie and face retaliation if their accusations do not lead to conviction. Chassang and Padró i Miquel (2019) study how a principal can incentivize a monitor to blow the whistle on a misbehaving agent when the agent can retaliate against the whistleblower.

[9]For a survey of information design, with a comprehensive literature review of information design with adversarial equilibrium selection, as well as adversarial mechanism selection, see Bergemann and Morris (2019).

[10]Similar arguments also appear in Acharya and Ramsay (2013), which also analyzes cheap talk in other types of information structures.

2 Model

States. There is a finite set of states Θ, with arbitrary element denoted θ. The principal and agents share a full support common prior $\mu \in \Delta(\Theta)$. In the motivating environments, the state is interpreted as the principal's private evidence, and so I refer to it as the *evidence state*.

Agents. Agents $i \in I = \{1, 2\}$ play a simultaneous-move game. Each agent i takes action $a_i \in A \equiv \{w, n\}$.[11,12] Agent i's payoff in state θ from action profile (a_i, a_{-i}) is denoted $u_i(a_i, a_{-i}, \theta)$, and an arbitrary action profile is denoted $\mathbf{a} = (a_i)_{i \in I}$.

Assumption 1 (Negative Spillovers)**.** *For each $i \in I, a_i \in A, \theta \in \Theta$,*

$$u_i(a_i, n, \theta) - u_i(a_i, w, \theta) > 0.$$

This assumption guarantees that i prefers that $-i$ not inform, independent of i's choice.

Assumption 2 (Jointly Informing)**.** *For each $i \in I$, $\theta \in \Theta$,*

$$u_i(w, w, \theta) - u_i(n, w, \theta) > 0.$$

This assumption guarantees that i prefers to inform if $-i$ informs. Let $\textsc{Dom}_i \subset \Theta$, be the set of states θ such that w is strictly dominant for i in state θ.

Information. An *information structure* is a pair (T, π) such that $T = T_1 \times T_2$ for some pair (T_1, T_2) with T_i countable, and $\pi \in \Delta(T \times \Theta)$ such that for each $\theta \in \Theta$,

$$\sum_{\boldsymbol{t} \in T} \pi(\boldsymbol{t}, \theta) = \mu(\theta).$$

Prior to choosing an action, each agent privately observes $t_i \in T_i$ (henceforth called agent i's *type*), with $\boldsymbol{t} = (t_1, t_2) \in T$ drawn according to π. Denote an arbitrary information structure by $\mathcal{I}$. I write (t_i, t_{-i}) to denote the element of T in which i observes t_i and $-i$ observes t_{-i}. Unless otherwise noted, I will assume wlog that $T_i = \mathbb{N}_\infty = \mathbb{N} \cup \{\infty\}$, and when it risks no confusion I will denote an information structure simply by π.

[11] w is for informing—or whistleblowing—and n is for not informing.

[12] Much of the literature deals with binary-action games, see for instance Morris et al. (2024), Halac et al. (2020), and Halac et al. (2022).

Communication Between Agents. A *communication mechanism* is a function

$$\sigma : T \to \Delta\left(A^I\right)$$

with the interpretation that each agent reports type $m_i \in T_i$ to a mediator, which then sends *recommendation* $a_i \in A$ to agent i according to distribution $\sigma(m_1, m_2)$. Given an information structure $\mathcal{I}$ and communication mechanism σ, if truthully reporting one's type and obeying the recommendation is incentive compatible assuming that others do, σ is called a *communication equilibrium* given $\mathcal{I}$.[13] Let $C(\mathcal{I})$ denote the set of communication equilibria given an information structure $\mathcal{I}$.

Principal. The principal chooses the information structure, π, that determines agents' private types. Let $v(\mathbf{a}, \theta)$ denote the principal's value in state θ for action profile $\mathbf{a}$.

Assumption 3. *For each* $\theta \in \Theta$, $\boldsymbol{a} \in A^I$,

$$v((w, w), \theta) \geq v(\boldsymbol{a}, \theta) \geq v((n, n), \theta).$$

This assumption ensures that in each state, the principal prefers more agents inform. Letting $v^*(\pi) \equiv \inf_{\sigma \in C(\pi)} \mathbb{E}_{\sigma,\pi}\left(v(\mathbf{a}, \theta)\right)$, the principal's problem is:

$$V^* \equiv \sup_{\pi} v^*(\pi)$$

I will also call this the principal's problem *with group communication*.[14]

It is useful to define another problem in which agents are *not* allowed to communicate. Let $\text{BNE}(\pi)$ be the set of BNE in the game induced by information structure π. Then, the principal's problem *without group communication* is

$$V^0 \equiv \sup_{\pi} \inf_{\sigma \in \text{BNE}(\mathcal{I})} \mathbb{E}_{\sigma,\pi}\left(v(\mathbf{a}, \theta)\right).$$

An information structure π is called *communication-proof* if:

$$\inf_{\sigma \in \text{BNE}(\pi)} \mathbb{E}_{\sigma,\pi}\left(v(\mathbf{a}, \theta)\right) = \inf_{\sigma \in C(\pi)} \mathbb{E}_{\sigma,\pi}\left(v(\mathbf{a}, \theta)\right)$$

[13] For a revelation principle justifying the restriction to such mechanisms, see Myerson (1982).

[14] All results remain true if rather than taking infimum over equilibria, I take infimum over equilibria that are Pareto efficient for agents.

Remark 1. *Observe that by Assumption 2, both agents choosing w (i.e., informing) is always a BNE. Hence, if the principal could choose her preferred equilibrium, revealing no information would be optimal.*

3 Information Structures: No Communication

In this section, I define a class of information structures, unraveling information structures, and describe their main attractive property *without* communication: the unique equilibrium outcome under an unraveling information structure is (w, w). I then define *partially* unraveling information structures, which have similar properties but allow for some probability of (n, n). As I show in Section 5, under the additional assumption of supermodular payoffs, the principal loses no value restricting to partially unraveling information structures satisfying an additional property guaranteeing communication-proofness (described in Section 4).

For any information structure π, let π_i denote the marginal distribution of π along dimension i, and π_I the marginal distribution along I. Fix an information structure π, an agent $i \in I$ and a type $t \in \text{supp}(\pi_i) \subseteq T_i$. Action $a \in A$ is an *interim strict best-response* (BR) for agent i given belief $\beta \in \Delta(T_{-i} \times A)$ over $-i$'s action if for each $a' \in A$,

$$\mathbb{E}_{\pi,\beta}\left(u_i(a, a_{-i}, \theta)|t_i = t\right) > \mathbb{E}_{\pi,\beta}\left(u_i(a', a_{-i}, \theta)|t_i = t\right)$$

where $\mathbb{E}_{\pi,\beta}(.|t_i = t)$ is the conditional expectation given $t_i = t$ and $(t_{-i}, a_{-i}) \sim \beta$.

For any $i \in I$ and type $t \in T_i$, let

$$\Phi_i(t) \equiv \left\{ \beta \in \Delta(T_{-i} \times A) \middle| \beta(t', n) = 0 \quad \forall t' \in T_{-i} \text{ s.t. } (t, t') \in \text{supp}(\pi_I) \text{ and } t' < t \right\}.$$

In words, $\Phi_i(t)$ is the set of beliefs i of type t can hold about $-i$'s action such that if $-i$'s type is smaller than i's, $-i$ chooses w.

Definition 1. *An information structure (T, π) is an **unraveling information structure** if $\infty \notin$ supp(π_i) for each i, and for any $i \in I$ and $t_i \in$ supp(π_i), w is an interim strict-BR for any belief $\beta \in \Phi_i(t_i)$ about $-i$'s action s.t. $\sum_{\boldsymbol{a} \in A^I} \beta(t_{-i}, \boldsymbol{a}) = \pi_i(t_{-i})$.*

The key property of unraveling information structures is they uniquely implement (w, w).

Lemma 1. *Any unraveling information structure implements (w, w) as the unique BNE.*

Unsurprisingly, unraveling information structures may not exist. For instance, if w is dominant with only small probability then for some payoff structures the principal will be unable to implement (w, w) as the unique BNE. Therefore, to solve the principal's problem, it is necessary to study a larger class of information structures, *partially* unraveling information structures. Partially unraveling information structures can be described by a two-step procedure: the principal sends a binary public signal, after one of the public signals agents face an unraveling information structure (hence (w, w) is the unique BNE), and after the other public signal (n, n) is a BNE (and hence is the principal's worst BNE).

Given an information structure (T, π) and $S \subset T$, let $\pi_S(.)$ denote the distribution of types t conditional on $t \in S$, and let μ_S denote the distribution of θ conditional on $t \in S$.

Definition 2. *An information structure (T, π) is a **partially unraveling information structure** if*

- *If $\pi(\boldsymbol{t}) > 0$, then $\boldsymbol{t} \in \mathbb{N}^2$ or $\boldsymbol{t} = (\infty, \infty)$*
- *$\tilde{\pi}$ is an unraveling information structure given prior $\tilde{\mu}$ over Θ, where $\tilde{\mu} = \mu(\theta | \boldsymbol{t} \in \mathbb{N}^2)$ and $\tilde{\pi}(\boldsymbol{t}, \theta) = \tilde{\pi}((\boldsymbol{t}, \theta) | \boldsymbol{t} \in \mathbb{N}^2)$*
- *If $\pi(\infty, \infty) > 0$, n is an interim BR for $t_i = \infty$ given belief that $t_{-i} = \infty$ chooses n.*

In words, the first requirement states that there is 0 probability that one agent's type is ∞ while another agent's type is *not* ∞. The second requirement states that on $\mathbb{N}^2$, agents face an unraveling information structure. The third requirement states that if there is a positive probability that both types are ∞, then n is a best-response for an agent with type ∞ given the belief that $-i$ with $t_{-i} = \infty$ chooses n, or equivalently there exists a BNE in which $t_i = \infty$ chooses n for each $i \in I$.[15]

The difference from an unraveling information structure is that in a partially unraveling information structure, there may be positive probability types who choose n (in the principal's worst BNE), while in an unraveling information structure all types choose w in the unique BNE.

Lemma 2. *Fix a partially unraveling information structure, (T, π). For each $i \in I$, in the principal's worst BNE, type $t_i \in T_i$ chooses w if $t_i \neq \infty$ and n if $t_i = \infty$.*

[15]The equivalence follows from the second requirement: if $\pi(\infty, \infty) > 0$, then if $t_i = \infty$, i believes $t_{-i} = \infty$ w.p. 1.

4 Information Structures: With Communication

It is not difficult to find examples of unraveling information structures that fail to be communication-proof. In this section, I show that unraveling information structures *are* communication-proof if they satisfy an additional restriction.

Recall that for any information structure π, π_i denotes the marginal distribution of agent i's type. For any information structure π, let $\pi^i_{t_i}$ denote the distribution of $-i$'s type, t_{-i}, conditional on i having type $t_i \in \text{supp}\,(\pi_i) \subset T_i$.

Proposition 1. *Fix a partially unraveling information structure, π, and suppose that*

$$\left|\{t_i | \text{supp}\,(\pi^i_{t_i}) \text{ and } t_{-i} > t_i\}\right| \leq 1.$$

Then π is communication-proof, and the principal's worst BNE outcome is identical to the principal's worst communication equilibrium outcome.

Note that if π is an unraveling information structure, the result implies that (w, w) is the unique and hence the principal's worst communication equilibrium.

Remark 2. *Partially unraveling information structures have the* ***perfect coordination*** *property in the principal's worst communication equilibrium: either both agents choose w or both agents choose n.*

5 Optimality under Supermodularity

In this section, I show that under the additional assumption that the game played by the agents is supermodular in each state, the principal's optimal value can be approximated arbitrarily well by partially unraveling information structures satisfying the premise of Proposition 1.

Agents' payoffs are *supermodular* if for all $\theta \in \Theta$ and $i \in I$,

$$u_i(w, w, \theta) - u_i(n, w, \theta) \geq u_i(w, n, \theta) - u_i(n, n, \theta)$$

In words, the difference in payoffs between choosing w and choosing n is *larger* when one's partner chooses w than when one's partner chooses n.[16]

[16]Whenever (n, n) is an equilibrium, supermodularity is a consequence of Assumption 2: $u_i(w, w, \theta) - u_i(n, w, \theta) > 0 \geq u_i(w, n, \theta) - u_i(n, n, \theta)$. But, if w is a dominant strategy for a given θ, then supermodularity may fail, and so must be imposed as an additional assumption.

Definition 3. *A set of information structures, B, **implements** V^* if*

$$V^* = \sup_{\pi \in B} \inf_{\sigma \in C(\pi)} \mathbb{E}_{\sigma,\pi}(v(\boldsymbol{a}, \theta))$$

*and **implements** V^0 if*

$$V^0 = \sup_{\pi \in B} \inf_{\sigma \in BNE(\pi)} \mathbb{E}_{\sigma,\pi}(v(\boldsymbol{a}, \theta)).$$

Proposition 2. *Suppose that agents' payoffs are supermodular. The principal's optimal value with group communication is the same as without group communication, i.e.,*

$$V^* = V^0.$$

The set of partially unraveling information structures π satisfying $\left|\{t_i | \text{supp}\,(\pi^i_{t_i}) \text{ and } t_{-i} > t_i\}\right| \leq 1$ implements V^* and V^0.

The proof is given in the appendix and follows by combining Morris et al. (2024)'s Theorem 1 with Proposition 1, and arguing that the principal can restrict without loss of value to information structures that have the perfect coordination property in the principal's worst equilibrium.

6 Linear Environments

In this section, I specialize to environments in which evidence states are real numbers and preferences are affine in the evidence state, and show that if w is dominant for both agents in the highest evidence state, there exists a solution to the principal's problem in which the likelihood of informing is monotonically increasing in the evidence state. Affine preferences are natural when the state, θ, is the probability that the principal's evidence is strong enough to prove misbehavior when neither agent informs.

An environment is called *linear* if $\Theta \subset \mathbb{R}$ and v, u_i are *affine* in θ. The starting point for the analysis is the linear programming formulation of the principal's problem in Morris et al. (2024). An implication of Proposition 2 is that the linear programming formulation remains valid, and can be further specialized using the second part of Proposition 2, stating that partially unraveling information structures—which satisfy the perfect coordination property—implement the principal's value. With this in hand, the monotonic characteriza-

tion of an optimal policy below is proved by examining the dual of the linear program. The proof is given in Appendix B.

A *consistent outcome* is a distribution $\nu \in \Delta(A^I \times \Theta)$ such that the marginal of θ equals the prior μ. An outcome ν is called *optimal* if (i) $V^* = \mathbb{E}_\nu(v(\mathbf{a}, \theta))$ and (ii) there exists a sequence $(\nu_m)_{m\in\mathbb{N}}$ such that $\nu_m \to \nu$, and for each m, ν^m is the outcome induced by the principal's worst communication equilibrium under some information structure. Define $\overline{\theta} \equiv \max\{\Theta\}$ and $\underline{\theta} \equiv \min\{\Theta\}$.

Proposition 3. *Fix any linear environment in which agents' payoffs are supermodular and w is dominant for each $i \in I$ at $\overline{\theta}$. Then, there exists $\theta^* \in \Theta$ and and an optimal outcome $\nu \in \Delta(A^I \times \Theta)$ satisfying*

$$\begin{aligned}
\nu((w,w),\theta) &= \begin{cases} \mu(\theta) & \theta \in \left(\theta^*, \overline{\theta}\right] \\ 0 & \theta \in [\underline{\theta}, \theta^*) \end{cases} \\
\nu((n,n),\theta) &= \mu(\theta) - \nu((w,w),\theta) \\
\nu((n,w),\theta) &= \nu((w,n),\theta) = 0.
\end{aligned}$$

In the settings motivating this paper, where θ is the likelihood that the principal can prove misbehavior without either agent informing, a natural case is when each agent's preference for w is *increasing* in θ, and the principal's preference for w is *decreasing* in θ. That is, when agents face a *greater* likelihood that the principal will be able to prove their misbheavior, they have the strongest incentives to inform, and when the principal faces a greater likelihood of being able to prove misbehavior without an informant, her value for an informant is lower. In this case, the result implies that, unless the principal can achieve her first best, it is optimal not to induce agents to inform when θ is small, where the principal's value for an informant is largest but the cost of providing incentives to inform is also largest.

6.1 Comparative Statics

It is convenient for comparative statics to define $g_i(\mathbf{a}), \ell_i(\mathbf{a}) : A^I \to \mathbb{R}$ such that

$$u_i(\mathbf{a}, \theta) = g_i(\mathbf{a})(1-\theta) + \ell_i(\mathbf{a})\theta \tag{1}$$

for each $\mathbf{a} \in A^I$, $\theta \in \Theta$, $i \in I$.

Let $\mathcal{G}$ denote any linear environment, and let $V^*(\mathcal{G})$ be the principal's value in this environment. Denote by $u_i(\mathbf{a}, \theta; \mathcal{G})$ agent i's payoff in this environment, and $g_i(\mathbf{a}; \mathcal{G})$ and

$\ell_i(\mathbf{a}; \mathcal{G})$ the coefficients on agent i's payoff, as defined in equation (1).

It is clear that increasing the payoffs when neither agent informs in every state, holding all else fixed, has an unambiguously *negative* effect on the principal's optimal value. The following result formalizes this observation.

Proposition 4. *Fix any linear environments $\mathcal{G}$ and $\mathcal{G}'$. Suppose that $u_i(\boldsymbol{a}, \theta; \mathcal{G}) = u_i(\boldsymbol{a}, \theta; \mathcal{G}')$ for each $i \in I$, $\theta \in \Theta$, and $\boldsymbol{a} \in A^I$ with $\boldsymbol{a} \neq (n, n)$. Then,*

$$u_i(n, n, \theta; \mathcal{G}) \geq u_i(n, n, \theta; \mathcal{G}') \text{ for each } i \in I, \theta \in \Theta \implies V^*(\mathcal{G}') \geq V^*(\mathcal{G}).$$

In the remainder of the section, I study another comparative static; increasing the asymmetry between firms.

Let $\mathcal{G}$ be a linear symmetric environment—a linear environment with symmetric payoffs for the agents. Let $\mathcal{G}^{\epsilon,\delta}$ denote the perturbed environment that is identical to $\mathcal{G}$, except that

$$g_1(n, n; \mathcal{G}^{\epsilon,\delta}) = g_1(n, n; \mathcal{G}) - \epsilon \tag{2}$$

$$g_2(n, n; \mathcal{G}^{\epsilon,\delta}) = g_2(n, n; \mathcal{G}) + \delta \tag{3}$$

Observe that under perturbation $\mathcal{G}^{\epsilon,\epsilon}$ for any $\epsilon \geq 0$, the total payoff of the group is unchanged i.e., $\sum_{i \in I} u_i(\mathbf{a}, \theta; \mathcal{G}) = \sum_{i \in I} u_i(\mathbf{a}, \theta; \mathcal{G}^{\epsilon,\epsilon})$, and only the distribution of payoffs when both choose n is affected. Say that a perturbation is *admissible* if it has supermodular payoffs for the agents, and w is dominant for both agents at $\theta = \overline{\theta}$. After stating the comparative static in the following proposition, I interpret these perturbations in the context of antitrust.

Proposition 5. *Fix a symmetric linear environment, $\mathcal{G}$, in which payoffs are supermodular for the agents, and w is dominant for both agents at $\overline{\theta}$. Then, for any triple of admissible perturbations $\mathcal{G}^{\epsilon,\epsilon}$, $\mathcal{G}^{\delta,\delta}$ and $\mathcal{G}^{\epsilon,\delta}$ with $\epsilon, \delta \in \mathbb{R}_+$,*

$$\epsilon \geq \delta \implies V^*(\mathcal{G}^{\epsilon,\epsilon}) \geq V^*(\mathcal{G}^{\epsilon,\delta}) \geq V^*(\mathcal{G}^{\delta,\delta}) \geq V^*(\mathcal{G}).$$

This result shows that *more asymmetric* environments are more susceptible to disruption through information design, leading to a higher value for the principal, and, given the monotonic characterization of an optimal policy in Proposition 3, a greater likelihood of informing.

Antitrust. In this section, I discuss the interpretation of comparative statics in the context of antitrust. Suppose that $\Theta \subset [0, 1]$, which I interpret as the probability that the

principal can prove misbehavior *without* an informant, $u_i((w,n),\theta) = 0$, $u_i((w,w),\theta) = -\frac{\ell_i}{2}$, $u_i((n,w),\theta) = -\ell_i$, and $u_i((n,n),\theta) = g_i \times (1-\theta) - \ell_i \times (\theta)$. The value g_i is interpreted as agent i's *profit* from misbehavior when neither agent informs and the principal cannot prove misbehavior, while ℓ_i is interpreted as agent i's *punishment* if the principal is able to prove misbehavior.

Most immediately in the context of cartels in antitrust, a *decrease* in cartel profits, g_i, for both agents, leads to an increase in the principal's value, a consequence of Proposition 4. If, for instance, a new entrant reduces cartel profits or demand shrinks, the cartel is more susceptible to breakdown through strategic information revelation by the principal. Such market events can be thoughts of as "markers" and can be used by the regulator to direct resources to the most susceptible cartels.

Next, consider the perturbation $\mathcal{G}^{\epsilon,\epsilon}$; such perturbations have the following interpretation: in the event that both agents choose n and the principal fails to prove misbehavior, agent 1's payoff increases by ϵ, and agent 2's payoff decreases by ϵ. Proposition 5 implies that if $\epsilon \geq \delta \geq 0$, the principal is better off in $\mathcal{G}^{\epsilon,\epsilon}$ than in $\mathcal{G}^{\delta,\delta}$, a more symmetric environment. In the context of cartels in antitrust, market events can generate such asymmetries. To see how, observe first that cartel punishments are a multiple of past illicit gains; as a result, if a cartel is initially symmetric but an event occurs that affects the future profits of cartelization, this change is *not* reflected in the payoffs when at least one firm applies to the regulator for leniency (effectively ending the cartel). Recent market events therefore only affect payoffs if neither firm informs, as required by the perturbation.

A number of events may trigger transfers of this form. For instance, cartels often operate by splitting markets geographically, agreeing not to invade each others' markets.[17] If demand in one market grows while demand in another shrinks, firms may face a situation similar to that represented by the perturbation in Proposition 5—the firm with a growing market faces a greater value from cartelization, while the firm with a shrinking market faces a smaller value from cartelization. Alternatively, even though firms try not to poach each other's customers, a (possibly large) customer may switch from one firm to another; this transfer can lead to an increase in the value of cartelization for one firm in the cartel and a decrease for the other firm, of the form described in Proposition 5. Each of these examples is a special case of a more general "marker" the regulator can use for allocating resources to strategic information provision; any market-event that shifts potential future profits of cartelization.

In the examples described, firms could re-allocate buyers or geographies in such a way

[17] See for instance, the copper plumbing tubes cartel, and others described in Sugaya and Wolitzky (2018).

as to re-balance the cartel profits. However, bargaining problems have been identified as a key obstacle for cartel success. For instance, in the cartel sample of Levenstein and Suslow (2006), approximately one quarter of all cartels in the paper's cartel sample ended because of bargaining problems. As the paper states, "successful cartels have developed organizational designs that allow the agreement to accommodate fluctuations in the external environment without requiring costly renegotiations." Cartels that are successful in this regard are therefore also more immune to the regulator's attempt to destroy the cartel with information. In contrast, cartels that struggle to re-bargain after market-shifting events are more susceptible to destruction through information provision by the regulator.

7 Simple Information Structures and Implementation

The information structures required to achieve the principal's optimal value may involve complex private communication with the agents. In this section, I consider a simple information structure that only requires private communication by fully disclosing the state or disclosing nothing.

Definition 4. *An information structure (T, π) is called **simple** if there exists some $i^* \in I$ s.t. $T_{i^*} = \{\emptyset\} \cup \mathrm{DOM}_{i^*}$, $T_{-i^*} = \{\emptyset\}$, and*

$$\pi(t_{i^*}, t_{-i^*}, \theta) = \mu(\theta)\mathbf{1}_{\theta \notin \mathrm{DOM}_{i^*}, t_{i^*} = \emptyset} + \mu(\theta)\mathbf{1}_{\theta \in \mathrm{DOM}_{i^*}, t_{i^*} = \theta}.$$

In words, a simple information structure picks some i^* and reveals the state to them when it is dominant for them to whistleblow, and reveals nothing otherwise.

Definition 5. *Agent i is least tempted if*

$$\mathbb{E}_\mu \left(u_i(w, n, \theta) - u_i(n, n, \theta) \right) \geq 0 \implies \mathbb{E}_\mu \left(u_{-i}(w, n, \theta) - u_{-i}(n, n, \theta) \right) \geq 0$$

Proposition 6. *Suppose agents' payoffs are supermodular and i^{least} is least tempted. Then the simple information structure with $i^* = i^{least}$ (weakly) improves the principal's value relative to no information μ.*

Observe that if $\Theta \in [0, 1]$, $u_i((w, n), \theta) = 0$, and $u_i((n, n), \theta) = g_i(1 - \theta) - l_i\theta$, then i is least tempted if $\frac{g_i}{l_i} \geq \frac{g_{-i}}{l_{-i}}$. Such an agent is least tempted for any prior μ over states. The regulator therefore need not know the agents' common prior to implement an improving simple information structure.

7.1 Antitrust and Cartels

In this section, I discuss some of the issues involved in implementing these information structures, in the context of antitrust leniency.

Commitment. To implement a simple information structure, the regulator needs to commit to releasing private information to only one agent.

Revealing the state to one agent (say, agent 1) may fail to create the contagion inherent in unraveling information structures; as a result, only the agent to whom the state was revealed, and who subsequently believes that w is dominant, chooses w. The problem then is that if the regulator is meant to leave player 2 uninformed, she may be tempted to reveal to player 2 that $\theta \in \mathrm{DOM}_1 \cup \mathrm{DOM}_2$, which may spur player 2 to report when he otherwise may not have. If player 2 anticipates this, interpreting no signal from the regulator as indication that $\theta \notin \mathrm{DOM}_1 \cup \mathrm{DOM}_2$, then he and player 1 may be able to avoid informing when $\theta \notin \mathrm{DOM}_1 \cup \mathrm{DOM}_2$.

One way to create some commitment is to implement a "first-in" rule that only the first agent to inform is granted leniency, and thus commit the regulator to only extract evidence voluntarily from the first reporting agent.[18] In that case, if $\theta \in \mathrm{DOM}_1$, the regulator is indifferent about revealing this information to player 2, since the second agent who informs is not granted any leniency and therefore provides no useful evidence to the regulator, so commitment becomes unnecessary (though if $\theta \in \mathrm{DOM}_2 \cap \mathrm{DOM}_1^C$, the temptation will remain). Of course, a first-in rule may lead to losses if the additional evidence provided by the second-in applicant would have turned a defeat into a victory in the case against the cartel, and so whether such a first-in rule is ideal depends on specifics of the environment.[19]

For reasons outside the model another issue is that the regulator may be tempted, after revealing information to agent 1 and observing that agent 1 does not apply for leniency, to reveal the same evidence to agent 2, in the hopes that it spurs him to apply for leniency. This is somewhat less problematic; as long there is a lag between the time agent 2 knows that evidence has been revealed to agent 1 and the time the regulator reveals evidence to agent 2, then to observe evidence, agent 2 must wait and potentially be preempted by agent 1, losing the benefits of being first to the authority. Thus, as long as the regulatory authority can order its communication sequentially, it can avoid this issue.

[18] The largest benefit to betrayal is always conferred on the first one to inform, but in some instances of antitrust leniency the second-in can also receive lenient treatment in exchange for evidence.

[19] Whether to restrict leniency to only the first-in applicant is a question that has been studied in the leniency literature, with benefits and costs beyond those considered here.

Choosing i^*. To generate improving simple information structures, one needs to identify an agent who is least tempted, i^*, and provide them with private signals. One way to identify i^* is to track changes to the composition of an industry. A firm that shrinks is likely to have relatively more to lose from being detected than a firm that grows: it has larger past illicit gains but expects little in the future. As a result, if one firm shrinks while another grows, then setting $i^* =$ FIRM THAT GROWS appears to be a good choice.

7.2 The Mechanics of Implementation

In this section, I discuss two ways that an antitrust regulator could possibly implement a private information policy such as the simple information structures discussed above.

Initial Investigations. To obtain authorization from a court to initiate an investigative action against a possible cartel—e.g. an unannounced inspection—an antitrust authority only need to present evidence of suspicious market behavior (OECD, 2013). Alternatively, an antitrust authority may initiate an action after receiving information from a third-party whistleblower. Even if the evidence the regulator has at this stage is not enough to successfully prove the existence of a cartel, it may lead to an inspection and the collection of potentially more serious hard evidence. In the context of the model, the initiation and continuance of an investigation are *public signals* of the strength of the regulator's evidence and suspicion. Since investigating takes resources—inspections as well as continued investigations are costly—they are credible signals of the antitrust authority's belief that it can obtain a successful conviction. To the extent that no cartel member knows what the regulator knows, the strength of the evidence is private information of the regulator.

A simple information structure can be implemented in addition to a public signal. To do this, the regulator must *commit*, at some stage during the investigation, to *privately releasing* evidence it discovers to only one of the cartel members. If, as discussed in the previous section, such commitment is possible, the only thing left for the regulator to determine is which of the cartel members to target with information. As already described, the model provides a rationale for the informed member to be the one who is *least tempted* to inform absent any private communication from the principal.

Affirmative Amnesty. One environment in which an antitrust regulator can potentially implement a simple information structure is in the context of *affirmative amnesty*, a practice of the Department of Justice. When a cartel is discovered, investigators may find evidence

of a second cartel.[20] Affirmative amnesty refers to the practice of revealing this evidence to one of the cartel members and offering them amnesty, in the hopes of inducing one of them to inform. Since in these instances the regulator is already approaching cartel members privately and providing them with evidence, implementing a simple information structure only requires the additional feature that the regulator make an ex-ante commitment to reveal the evidence to only one cartel member (and commit to which cartel member it will be).

8 Three Firms

A natural question is, to what extent does Proposition 1 extend to more than 2 firms? Here, I consider the case of 3 firms, i.e. $I \equiv \{1, 2, 3\}$. For each $i \in I$, and $\boldsymbol{a} \in \{n, w\}^3$, denote by $\boldsymbol{a}^j$ the action profile that replaces j's action with n.

It is necessary to adjust assumptions 1, 2, and 3.

Assumption 4 (Negative Spillovers 3 Firms)**.** *For any i, $\boldsymbol{a} \in \{n, w\}^I$, and $\theta \in \Theta$, $u_i(a_i, a^j_{-i}) \geq u_i(\boldsymbol{a})$for each $j \neq i$ with strict inequality if $a_j = w$.*

Assumption 5 (Jointly Whistleblowing 3 Firms)**.** *For each $i \in I$, then $u_i(w, a_{-i}, \theta) > u_i(n, a_{-i}, \theta)$ if $a_{-i} \in \{(w, w), (w, n), (n, w)\}$.*

Assumption 6 (Regulator's Value 3 Firms)**.** *For any i, $\boldsymbol{a} \in \{n, w\}^3$ and $\theta \in \Theta$, $v(\boldsymbol{a}^j) \geq v(\boldsymbol{a})$.*

The construction in Morris et al. (2024) is not limited to 2 firms, and so if some modification of the construction for more than 2 firms is communication-proof, then Proposition 1 could be extended to more than 2 firms. As I show below, a natural extenstion of the information structures that prove sufficiently rich to implement the principal's optimal value in the 2 firm case can perform poorly in the case of more than 2 firms; in particular, under some conditions, importantly that the slack in incentive constraints induced by the information structure shrinks to 0, the worst communication equilibrium converges to the regulator's worst outcome.

To state the result, it is necessary to recall some (slightly modified) definitions from Morris et al. (2024). Let $\Gamma = \{123, 132, 231, 213, 312, 321\}$, the set of all permutations of I.

[20]The likelihood is high; at least as of statistics published in 2006, around 50% (see https://www.justice.gov/atr/speech/measuring-value-second-cooperation-corporate-plea-negotiations).

Let

$$a_{-i}(\gamma) \equiv \begin{cases} w & \text{if } -i \text{ is ranked higher than } i \text{ in } \gamma \\ n & \text{otherwise} \end{cases}$$

An *ordered outcome* is a $\nu_\Gamma \in \Delta(\Gamma \cup \{\emptyset\} \times \Theta)$, satisfies *sequential obedience* if

$$s_i(\nu_\gamma) \equiv \sum_{\gamma \in \Gamma, \theta \in \Theta} \nu_\Gamma(\gamma, \theta)(u_i(w, a_{-i}(\gamma), \theta) - u_i(n, a_{-i}(\gamma), \theta)) > 0,$$

is n-obedient if

$$\sum_{\theta \in \Theta, a_{-i} \in A} (u_i(n, a_{-i}, \theta) - u(w, a_{-i}, \theta))\nu_\Gamma(\emptyset, \theta) \geq 0,$$

and is *consistent* if

$$\sum_{\mathbf{a} \in A^I, \gamma \in \Gamma \cup \{\emptyset\}} \nu_\Gamma(\mathbf{a}, \theta) = \mu(\theta)$$

for each $\theta \in \Theta$.

Given type profile $\boldsymbol{t}$, define

$$f^\gamma(\mathbf{t}) = \begin{cases} ijk \text{ if } t_k = t_j + 1 = t_i + 2 \\ 0 \text{ if no such } ijk \text{ exists} \end{cases}$$

Finally, consider the following information structures, natural extensions (slightly modified) of the 2 firm information structures used for implementation:

$$\pi^{p,\nu_\Gamma,\eta}(\boldsymbol{t}, \theta) = \begin{cases} (1-p) \times \eta(1-\eta)^{m-1}\nu_\Gamma(f^\gamma(\boldsymbol{t}), \theta) & \text{if } \min_i(\boldsymbol{t}_i) = m \text{ and } m \geq 1 \textbf{ and } f^\gamma(\mathbf{t}) \in \Gamma \\ (1-p) \times \nu_\Gamma(\emptyset, \theta) + p \times \mathbf{1}_{\theta \notin \bigcup_i \mathrm{DOM}_i} & \text{if } \boldsymbol{t} = (\infty, \infty, \infty) \\ \frac{\mathbf{1}_{\theta \in \mathrm{DOM}_i}}{\sum_{j \in I} \mathbf{1}_{\theta \in \mathrm{DOM}_j}} \frac{p}{2} & \text{if } t_i = 0 \text{ and } f^\gamma(\mathbf{t}) \in \Gamma \\ 0 & \text{otherwise} \end{cases}$$

Denote $\boldsymbol{n} = (n, n, n)$.

Proposition 7. *There exists $\epsilon > 0$ s.t. for all ordered outcomes ν_Γ with $\max_{i \in I} |s_i(\nu_\Gamma)| < \epsilon$, if ν_Γ is n-obedient, sequentially obedient, and consistent then*

$$v^*(\pi^{p,\nu_\Gamma,\eta}) \xrightarrow[\eta \to 1]{} \mathbb{E}_\mu\big[(1-p) \times v(\boldsymbol{n}, \theta) + p \times g(\theta)\big]$$

where $g(\theta) \leq \max_{\boldsymbol{a} \in A^I} \{v(\boldsymbol{a}, \theta)\}$.

The proposition implies that, if the premise is true, then as η approaches 1, the regulator is held down to her worst possible outcome, plus a term converging to 0 as p does.

References

Abreu, D. and H. Matsushima (1992): "Virtual implementation in iteratively undominated strategies: complete information," *Econometrica: Journal of the Econometric Society*, 993–1008.

Acharya, A. and K. W. Ramsay (2013): "The calculus of the security dilemma," *Quarterly Journal of Political Science*, 8, 183–203.

Angelucci, C. and A. Russo (2022): "Petty corruption and citizen reports," *International Economic Review*.

Baliga, S. and S. Morris (1998): "Cheap Talk and Co-ordination with Payoff Uncertainty," .

Bergemann, D. and S. Morris (2019): "Information design: A unified perspective," *Journal of Economic Literature*, 57, 44–95.

Bernstein, S. and E. Winter (2012): "Contracting with heterogeneous externalities," *American Economic Journal: Microeconomics*, 4, 50–76.

Camboni, M. and M. Porcellacchia (2024): "Monitoring Team Members: Information Waste and the Self-Promotion Trap," .

Carlsson, H. and E. Van Damme (1993): "Global games and equilibrium selection," *Econometrica: Journal of the Econometric Society*, 989–1018.

Chan, L. T. (2023): "Weight-Ranked Divide-and-Conquer Contracts," *Available at SSRN 3780434*.

Chassang, S., L. Del Carpio, and S. Kapon (2022): "Using Divide and Conquer to Improve Tax Collection," Tech. rep., NBER Working Paper.

Chassang, S. and J. Ortner (2022): "Regulating Collusion," *Annual Review of Economics*, Forthcoming.

Chassang, S. and G. Padró i Miquel (2019): "Crime, intimidation, and whistleblowing: A theory of inference from unverifiable reports," *The Review of Economic Studies*, 86, 2530–2553.

Dannay, G. (2019): "Information Design Against Petty Corruption," Ph.D. thesis.

Gamba, A., G. Immordino, and S. Piccolo (2018): "Corruption, organized crime and the bright side of subversion of law," *Journal of Public Economics*, 159, 79–88.

Halac, M., I. Kremer, and E. Winter (2019): "Raising capital from heterogeneous investors," *American Economic Review*.

Halac, M., E. Lipnowski, and D. Rappoport (2020): "Rank Uncertainty in Organizations," *Available at SSRN 3553935*.

——— (2022): "Addressing Strategic Uncertainty with Incentives and Information," in *AEA Papers and Proceedings*, vol. 112, 431–37.

Harrington Jr, J. E. (2008): "Optimal corporate leniency programs," *The Journal of Industrial Economics*, 56, 215–246.

——— (2013): "Corporate leniency programs when firms have private information: the push of prosecution and the pull of pre-emption," *The Journal of Industrial Economics*, 61, 1–27.

Hoshino, T. (2022): "Multi-Agent Persuasion: Leveraging Strategic Uncertainty," *International Economic Review*, 63, 755–776.

Inostroza, N. and A. Pavan (2023): "Adversarial coordination and public information design," *Available at SSRN 4531654*.

Kajii, A. and S. Morris (1997): "The robustness of equilibria to incomplete information," *Econometrica: Journal of the Econometric Society*, 1283–1309.

Landeo, C. M. and K. E. Spier (2020): "Optimal law enforcement with ordered leniency," *The Journal of Law and Economics*, 63, 71–111.

Lee, F. X. and W. Suen (2020): "Credibility of crime allegations," *American Economic Journal: Microeconomics*, 12, 220–59.

Levenstein, M. C. and V. Y. Suslow (2006): "What determines cartel success?" *Journal of economic literature*, 44, 43–95.

Li, F., Y. Song, and M. Zhao (2022): "Global manipulation by local obfuscation," *Journal of Economic Theory*, 105575.

Marvão, C. and G. Spagnolo (2018): "Cartels and leniency: Taking stock of what we learnt," in *Handbook of Game Theory and Industrial Organization, Volume II*, Edward Elgar Publishing.

Mathevet, L., J. Perego, and I. Taneva (2020): "On information design in games," *Journal of Political Economy*, 128, 1370–1404.

Miller, N. H. (2009): "Strategic leniency and cartel enforcement," *American Economic Review*, 99, 750–68.

Moriya, F. and T. Yamashita (2020): "Asymmetric-information allocation to avoid coordination failure," *Journal of Economics & Management Strategy*, 29, 173–186.

Morris, S., D. Oyama, and S. Takahashi (2024): "Implementation via Information Design in Binary-Action Supermodular Games," *Econometrica*, 92, 775–813.

Motta, M. and M. Polo (2003): "Leniency programs and cartel prosecution," *International journal of industrial organization*, 21, 347–379.

Myerson, R. B. (1982): "Optimal coordination mechanisms in generalized principal–agent problems," *Journal of mathematical economics*, 10, 67–81.

OECD (2013): "Ex officio cartel investigations and the use of screens to detect cartels," .

Pei, H. and B. Strulovici (2024): "When to Convict Defendants Facing Multiple Accusations? A Strategic Analysis," .

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

Sandmann, C. (2021): "Recursive information design," Tech. rep., Mimeo.

Sauvagnat, J. (2015): "Prosecution and leniency programs: the role of bluffing in opening investigations," *The Journal of Industrial Economics*, 63, 313–338.

Segal, I. (2003): "Coordination and discrimination in contracting with externalities: Divide and conquer?" *Journal of Economic Theory*, 113, 147–181.

Spagnolo, G. (2000): "Optimal leniency programs," *FEEM Working Paper*.

Sugaya, T. and A. Wolitzky (2018): "Maintaining privacy in cartels," *Journal of Political Economy*, 126, 2569–2607.

Vallery, A. and C. Schell (2016): "AC-Treuhand: Substantial Fines for Facilitators of Cartels," *Journal of European Competition Law & Practice*, 7, 254–257.

Winter, E. (2004): "Incentives and discrimination," *American Economic Review*, 94, 764–773.

Ziegler, G. (2020): "Adversarial bilateral information design," Tech. rep., Working paper.

A Proofs of Section 4

Denote by

$$Upper_i(t) \equiv \{s \in T | s < t \text{ and } \mathbb{P}_\pi(t_i = t, t_{-i} = s) > 0\}$$

and

$$Lower_i(t) \equiv \{s \in T | s > t \text{ and } \mathbb{P}_\pi(t_i = t, t_{-i} = s) > 0\}.$$

Observe that Proposition 1 assumes that $|Lower_i(t)| \leq 1$ for all $t \in \text{supp}\,(\pi_i)$.

Proof of Proposition 1: It is sufficient to prove the result for unraveling information structures. To see why, observe that if $t_i = t_{-i} = \infty$, then (n, n) is a BNE and, hence, the principal's worst communication equilibrium. Further, the information structure conditional on $t_i \neq \infty$ for some i is an unraveling information structure, and so the analysis for unraveling information structures would apply.

Then, fix an unraveling information structure, (T, π) satisfying the premise of the proposition. Let π_i denote the marginal distribution of t_i. Let $\pi^\theta_{\boldsymbol{t}}$ denote the distribution of θ, conditional on $t_i \in T_i$ for any $\boldsymbol{t}$ such that $\pi_i(\boldsymbol{t}) > 0$.

Recall that a communication equilibrium is defined by a map $\sigma : T \to \Delta(A^I)$; agents report types $m_i \in T_i$ to a mediator, who then generates a recommendation $a \in A^I$ according to distribution $\sigma(m)$, privately shows recommendation a_i to agent i, and each agent i finds

it optimal to *truthfully report* his type and *obey* the recommendation. Let $\mathbf{r}^{\sigma(\mathbf{m})} = (r_i^{\sigma(\mathbf{m})})_{i \in I}$ denote a random variable distributed according to $\sigma_i(\mathbf{m})$ i.e., recommendation to agent i, and let r_i denote the realization of the recommendation revealing to agent i.

To prove the result, I will show that in any communication equilibrium σ, $\sigma(\mathbf{m}) = \delta_{(n,n)}$ for each $\mathbf{m}$ with positive probability under π or, equivalently, $\mathbb{P}(r_i^{\sigma(\mathbf{m})} = n) = 0$ for any positive probability $\mathbf{m} \in T$. The proof proceeds by induction on $t_i \in \{0, 1, ...\}$. In an unraveling information structure, it is wlog to assume that $\mathbb{P}_\pi(t_i = 0) > 0$ for some i, and so I proceed under that assumption.

Base Case: Fix any $i \in I$ and $t_i \in T_i$ such that $t_i = 0$ and $\mathbb{P}_\pi(t_i) > 0$. The definition of unraveling information structure implies that t_i has a strict-BR to choose w, independent of $-i$'s action. By assumption, $|Lower_i(t_i)| = 1$.[21] But then, for any r s.t. $\sigma(\mathbf{m})(\{r\}) > 0$, $\pi^\theta_{t_i}(.|r^{\sigma(\mathbf{m})}) = \pi^\theta_{t_i}(.)$ for any $\mathbf{m}$ with $m_i = t_i$. Since i had a strict-BR to choose w before observing the recommendation *independent of $-i$'s action* and the recommendation does not change i's belief about θ, then to satisfy obedience, it must be that $\mathbb{P}(r_i^{\sigma(\mathbf{m})} = n) = 0$ for any $\mathbf{m}$ such that $m_i = t_i$.

Inductive Step: Suppose that $\mathbb{P}(r_i^\sigma(\mathbf{m})) = n) = 0$ for any $\mathbf{m}$ such that $m_i < k$. I will prove the statement for any $\mathbf{m}$ such that $m_i = k$. To this end, fix any type profile $\mathbf{t}$ such that $t_i = n$ such that $\pi(\mathbf{t}) > 0$. If no such profile exists, we are done. Otherwise, let $t \equiv t_i$. By assumption, $|Lower_i(t)| \in \{0, 1\}$.

Case 1: $|Lower_i(t)| = 0$**.** In this case, player i with $t_i = t$ believes that player $-i$ chooses w with probability 1, a result of our inductive hypothesis and the definition of an unraveling information structure that implies $\mathbb{P}_\pi(t_i = t_{-i}) = 0$.

Case 2: $|Lower_i(t)| = 1$**.** Abusing notation, denote by $Lower_i(t)$ the unique element in $Lower_i(t)$.

Claim 1. *For each pair $n, n' \in Upper_i(t)$,*

$$\mathbb{P}(r_i^{\sigma(\boldsymbol{m})} = n) = \mathbb{P}(r_i^{\sigma(\boldsymbol{m}')} = n)$$

[21] Note that if $\lambda(t_i) = 0$ and t_i has positive probability, then $|Lower_i(t_i)| = 0$ is ruled out.

for $m_i = m_i' = t$, $m_{-i} = s$, *and* $m_{-i}' = s'$. *Further, for each* $s \in Upper_i(t)$,

$$\mathbb{P}(r_i^{\sigma(\boldsymbol{m})} = n) \geq \mathbb{P}(r_i^{\sigma(\boldsymbol{m}')} = n)$$

for $m_i = m_i' = t$, $m_{-i} = s$, *and* $m_{-i}' = Lower_i(t)$.

Proof of Claim: If $Upper_i(t) = \emptyset$, there is nothing to show. Otherwise, to prove the claim, observe first that by the inductive hypothesis, truth-telling and obedience requires that for all $s \in Upper_i(t)$, each agent's payoff is maximized by reporting type truthfully and choosing w. Consider now the payoff to agent $-i$ with any type $s \in Upper_i(t)$ from reporting type $\hat{m} \in Upper_i(t) \cup Lower_i(t)$ and choosing w:

$$\begin{aligned}
&\mathbb{P}(t_i \in Upper_{-i}(s)|t_{-i} = s) \times \mathbb{E}\Big(u_{-i}(w, w, \theta)|t_{-i} = s, t_i \in Upper_{-i}(s)\Big) \\
&+ \\
&\mathbb{P}(t_i \in Lower_{-i}(s)|t_{-i} = s) \times \Big(\\
&\quad \mathbb{P}(r_i = s|m_i \in Lower_{-i}(s), m_{-i} = \hat{m})\mathbb{E}\left(u_{-i}(w, n, \theta)|t_i \in Lower_{-i}(s), t_{-i} = s\right) \\
&\quad + \\
&\quad \mathbb{P}(r_i = b|m_i \in Lower_{-i}(s), m_{-i} = \hat{m})\mathbb{E}\left(u_{-i}(w, w, \theta)|t_i \in Lower_{-i}(s), t_{-i} = s\right) \Big)
\end{aligned}$$

where the first line follows from the inductive hypothesis. By assumption, $Lower_{-i}(s) = \{t\}$, so the expression becomes

$$\begin{aligned}
&\mathbb{P}(t_i \in Upper_{-i}(s)|t_{-i} = s) \times \mathbb{E}\Big(u_{-i}(w, w, \theta)|t_{-i} = s, t_i \in Upper_{-i}(s)\Big) \\
&+ \\
&\mathbb{P}(t_i = t|t_{-i} = s) \times \Big(\\
&\quad \mathbb{P}(r_i = s|m_i = t, m_{-i} = \hat{m})\mathbb{E}\left(u_{-i}(w, n, \theta)|t_i = t, t_{-i} = s\right) \\
&\quad + \\
&\quad \mathbb{P}(r_i = b|m_i = t, m_{-i} = \hat{m})\mathbb{E}\left(u_{-i}(w, w, \theta)|t_i = t, t_{-i} = s\right) \Big)
\end{aligned}$$

Then, since $u_{-i}(w,n,\theta) > u_i(w,w,\theta)$ by Assumption 1, the expression is maximized by reporting $\hat{m} \in Upper_i(t) \cup Lower_i(t)$ that maximizes $\mathbb{P}(r_i = s | m_i = t, m_{-i} = \hat{m} = \mathbb{P}(r_i^{\sigma(\mathbf{m})})$ for $\mathbf{m}$ with $m_i = t, m_{-i} = \hat{m}$. □

Observe then that agent i's posterior after observing recommendation n has two properties: (i) i's belief that $t_{-i} \in Lower_i$ must weakly decrease, (ii) conditional on $t_{-i} \in Upper_i$, i's belief about θ is unchanged related to her interim belief and hence, her expected payoff from any action profile conditional on $t_{-i} \in Upper_i$ is unchanged.

Suppose now towards contradiction that $\mathbb{P}(r_i = n | m_i = t_i = t) > 0$. Consider then, the payoff to *obeying* the recommendation, choosing n, less the payoff to disobeying, choosing w:

$$\begin{aligned}
U^{obey}(t) \equiv \quad & \mathbb{P}(t_{-i} = Lower_i(t) | r_i = n, m_i = t = t_i) \\
& \times \Big(\mathbb{P}(r_{-i} = n | r_i = n, m_i = t, m_{-i} = Lower_i(t)) \mathbb{E}\left(u_i(n,n,\theta) - u_i(w,n,\theta) | t_i = t, t_{-i} = Lower_i(t)\right) \\
& \quad + \mathbb{P}(r_{-i} = w | r_i = n, m_i = t, m_{-i} = Lower_i(t)) \mathbb{E}\left(u_i(n,w,\theta) - u_i(w,w,\theta) | t_i = t, t_{-i} = Lower_i(t)\right) \\
& + \quad \mathbb{P}(t_{-i} \in Upper_i(t) | r_i = n, m_i = t = t_i) \times \Big(\\
& \qquad \mathbb{E}\left(u_i(n,w,\theta) - u_i(w,w,\theta) | t_i = t, r_i = n, t_{-i} \in Upper_i(t)\right) \Big)
\end{aligned}$$

If obedience is to hold, it must be that $U^{obey}(t) \geq 0$. By Assumption 2, it must be that sum of the terms in the second and third line is weakly positive. But then, by the claim:

$$\begin{aligned}
U^{obey}(t) \leq \mathbb{P}_\pi(t_{-i} = Lower_i(t) | m_i = t = t_i) \times \Big(& \\
& \mathbb{P}(r_{-i} = n | r_i = n, m_i = t, m_{-i} = Lower_i(t)) \mathbb{E}\left(u_i(n,n,\theta) - u_i(w,n,\theta) | t_i = t, t_{-i} = Lower_i(t)\right) \\
& + \mathbb{P}(r_{-i} = w | r_i = n, m_i = t, m_{-i} = Lower_i(t)) \mathbb{E}\left(u_i(n,w,\theta) - u_i(w,w,\theta) | t_i = t, t_{-i} = Lower_i(t)\right) \\
& \Big) \\
& + \mathbb{P}(t_{-i} \in Upper_i(t) | m_i = t = t_i) \times \Big(\mathbb{E}\left(u_i(n,w,\theta) - u_i(w,w,\theta) | t_i = t, t_{-i} \in Upper_i(t)\right) \Big) \\
< \quad & 0
\end{aligned}$$

where the last line follows by the definition of an unraveling information structure. This

contradicts obedience, and so we conclude that $\mathbb{P}(r_i = n | m_i = t_i = t) = 0$, and the result follows. □

B Proofs of Section 5

Before proving results in Sections 5 and 6, it is necessary to define a number of preliminaries in order to modify results in Morris et al. (2024).

Preliminaries. Let $d_i(a_{-i}, \theta) \equiv u_i(b, a_{-i}, \theta) - u_i(s, a_{-i}, \theta)$. An *outcome* is a distribution $\nu \in \Delta(A^I \times \Theta)$. Let $\Gamma \equiv \{\emptyset, (1), (2), (1,2), (2,1)\}$ and $\Gamma_i \equiv \Gamma \backslash \{i, \emptyset\}$. $\Gamma_2 \equiv \{(2), (1,2), (2,1)\}$. An outcome is *consistent* if $\sum_{\mathbf{a} \in A^I} \nu(\mathbf{a}, \theta) = \mu(\theta)$. An outcome is *obedient* if for each $i \in I, a_i \in \{w, n\}, a' \in \{w, n\}$:

$$\sum_{\theta \in \Theta} u_i(a_i, a_{-i}) \nu(a_i, a_{-i}) \geq \sum_{\theta \in \Theta} u_i(a', a_{-i}) \nu(a_i, a_{-i})$$

An *ordered outcome* is a distribution $\nu_\Gamma \in \Delta(\Gamma \times \Theta)$. Given $\gamma \in \Gamma$, let $a_{-i}(\gamma)$ denote the action for $-i$ equal to w if $-i$ comes before i in γ or if i is not in γ while $-i$ is, and n otherwise. An ordered outcome satisfies *sequentual obedience* if

$$\sum_{\gamma \in \Gamma_i, \theta \in \Theta} \nu_\Gamma(\gamma, \theta) d_i(a_{-i}(\gamma), \theta) > 0$$

for any i with $v_\Gamma(\Gamma_i \times \Theta) > 0$. Let $\overline{a}(\gamma)$ denote the strategy profile in which agents appearing in γ choose w and otherwise choose n. An outcome ν is *induced by* and ordered outcome ν_Γ if

$$\nu(\mathbf{a}, \theta) = \sum_{\gamma : \overline{a}(\gamma) = a} \nu_\Gamma(\gamma, \theta).$$

An outcome ν is said to satisfy sequential obedience if there exists an ordered outcome $\nu_\Gamma \in \Delta(\Gamma \times \Theta)$ that satisfies sequential obedience and induces ν.

Finally, an outcome ν satisfies *asymmetric grain of dominance* if there exists i and $\overline{\theta}$, such that $d_i(a_{-i}, \theta) > 0$ for any a_{-i} i.e., w is strictly dominant, and $\nu((w, w), \overline{\theta}) > 0$.

An outcome ν is said to be *S-implementable* if there exists an information structure $\mathcal{I}$

such that[22]

$$\inf_{\sigma \in C(\mathcal{I})} \mathbb{E}_{\sigma,I}(v(\mathbf{a},\theta) = \mathbb{E}_{\nu}(v(\mathbf{a},\theta)).$$

Finally, since payoffs are supermodular, for any information structure there will exist a principal's worst equilibrium—this coincides with the smallest equilibium when action w is labeled 1 and action n is labeled 0.

Proof of Proposition 2: To prove this result, I will proceed in four steps:

1. Modify the statement and proof of Theorem 1(2) in Morris et al. (2024), so that only asymmetric "grain of dominance" (defined in Morris et al. (2024), with asymmetric version defined below) is necessary and the information structure used in the proof never involves both agents having the same types if those types are finite.

2. Show that if the principal constraints herself to information structures that exhibit asymmetric grain of dominance, it is without loss of generality for the principal's value to ignore information structures that do not generate *perfect coordination* in the principal's worst equilibrium—agents either both choose w or both choose n.

3. Show that information structures in (1) that satisfy the perfect coordination property are partially unraveling information structures.

4. If there exists no state $\theta \in \Theta$ such that at least one agent finds w dominant, then the worst equilibrium under any information structure is the pure strategy profile in which neither agent informs. Otherwise, I show that requiring asymmetric grain of dominance is without loss of value for the principal.

Step 1: Modying Morris et al. (2024)'s Theorem 1(2): This modification is straightforward, though tedious.

Since payoffs are supermodular for agents, Morris et al. (2024)'s Theorem 1(1) applies, so that obedience, consistency, and sequential obedience are necessary conditions for an outcome to be n-implementable. Further, it is easy to see that if an outcome ν fails to satisfy asymmetric grain of dominance, then it is not implementable; indeed, if asymmetric grain of dominance fails then each player choosing n is an equilibrium. As a result, an additional necessary condition for ν to be n-implementable is that ν satisfies asymmetric grain of dominance.

[22]Note that S-implementability corresponds to "smallest" equilibrium implementation in Morris et al. (2024). Labeling w as 1 and n as 0, principal-worst equilibrium is the same as S-implementable.

The restatement of Theorem 1(2) that I will prove is

> *If an outcome satisfies consistency, obedience, sequential obedience and asymmetric grain of dominance, then it is S-implementable.*

So, fix an outcome $\nu \in \Delta(A \times \Theta)$ and suppose that it satisfied *asymmetric grain of dominance*. For the proof, suppose that asymmetric grain of dominance is satisfied for player 1: there is $\overline{\theta}$ such that $d_1(a_2, \overline{\theta}) > 0$ and $v((w,w), \overline{\theta}) > 0$. The proof will work exactly the same if it is $i = 2$ who has the dominant action, so I will only proceed with the case in which $i = 1$ has the dominant action.

I will now follow the steps of Morris et al. (2024), pointing out where small modifications must be made to the information structure. I will purposely stay as close as possible to their proof, so the modification becomes clear.

Since ν satisfies sequential obedience, there exists an ordered outcome $v_\Gamma \in \Delta(\Gamma \times \Theta)$ that induces ν and satisfies sequential obedience. Since $\nu((w,w), \overline{\theta}) > 0$ by asymmetric grain of dominance, there is $\overline{\gamma} \in \Gamma$ containing all players with $\nu_\Gamma(\overline{\gamma}, \overline{\theta}) > 0$. Pick any $\epsilon > 0$ so that $\epsilon < \nu_\Gamma(\overline{\gamma}, \overline{\theta})$ and define

$$\tilde{v}_\Gamma(\gamma, \theta) \equiv \frac{v_\Gamma(\gamma, \theta)}{1 - \epsilon} - \left(\mathbf{1}_{(\gamma,\theta)=(\overline{\gamma},\overline{\theta})}\right) \frac{\epsilon}{1 - \epsilon}$$

where ϵ is sufficiently small that $\tilde{\nu}_\Gamma$ satisfies sequential obedience (possible because ν does). Since $d_1(a_2, \overline{\theta}) > 0$, there exists $\overline{q}_1 < 1$ such that

$$\overline{q} d_1(s, \overline{\theta}) + (1 - \overline{q}) \min_{\theta \neq \overline{\theta}} d_1(s, \theta) > 0. \tag{4}$$

By assumption 2, $d_2(b, \overline{\theta}) > 0$, so there exists $\overline{q}_2 < 1$ such that

$$\overline{q} d_2(b, \overline{\theta}) + (1 - \overline{q}) \min_{\theta \neq \overline{\theta}} d_2(s, \theta) > 0. \tag{5}$$

Let $\overline{q} = \max\{\overline{q}_1, \overline{q}_2\}$. This is the first minor difference from Morris et al. (2024): assumption 2 allows for a slightly less constrained dominance state assumption, here called asymmetric grain of dominance.

Now, let $\eta > 0$ be such that

$$\frac{\frac{\epsilon}{2}}{\frac{\epsilon}{2} + \eta} > \overline{q}$$

and

$$\sum_{\gamma\in\Gamma_i,\theta\in\Theta}(1-\eta)^{1-n(a_{-i}(\gamma))}\tilde{\nu}_\Gamma(\gamma,\theta)d_i(a_{-i}(\gamma),\theta)>0.$$

for all i, where $n(a_{-i}(\gamma))$ is and indicator equal to 1 if $a_{-i}(\gamma)=b$. Let type space T be defined as follows:

$$T_1=\begin{cases}\{0,1,2,...\} & \text{if } \tilde{\nu}_\Gamma(\Gamma_i\times\Theta)=1\\ \{0,1,2,...\}\cup\{\infty\} & \text{otherwise}\end{cases}$$

and

$$T_2=\begin{cases}\{1,2,...\} & \text{if } \tilde{\nu}_\Gamma(\Gamma_i\times\Theta)=1\\ \{1,2,...\}\cup\{\infty\} & \text{otherwise}\end{cases}$$

The only difference from Morris et al. (2024) is that T_1 now contains 0. Let

$$\ell(i,\gamma)\equiv\begin{cases}\ell & \text{if there exists } \ell\in\{1,...,k\} \text{ such that } i_\ell=i\\ \infty & \text{otherwise}\end{cases}$$

for each $i\in I$ and $\gamma=(i_1,...,i_k)\in\Gamma$. Then, let $\pi\in\Delta(T\times\Theta)$:

$$\pi(\mathbf{t},\theta)\equiv\begin{cases}(1-\epsilon)\eta(1-\eta)^m\tilde{\nu}_\Gamma(\gamma,\theta) & \text{if } t_i<\infty \text{ for some } i \text{ and there exists } m\geq 0,\\ & \text{such that for all } i, t_i=m+\ell(i,\gamma)\\ \frac{\epsilon}{2} & \text{if } t_1=0,t_2=1,\theta=\overline{\theta}\\ \frac{\epsilon}{2} & \text{if } t_1=1,t_2=2,\theta=\overline{\theta}\\ (1-\epsilon)\tilde{v}_\Gamma(\emptyset,\theta) & \text{if } t_1=t_2=\infty\\ 0 & \text{otherwise}\end{cases}$$

The only difference between this information structure and the one in Morris et al. (2024) is that the mass that was originally on $t_1=1$ has been split between the new type $t_1=0$ and $t_1=1$. It follows from Morris et al. (2024), that π is consistent. I state a modified version of the claim A.1.

Modified Claim A.1 (from Morris et al. (2024)'s Theorem 1(2) proof). *For any t_i with $\mathbb{P}_\pi(t_i)>0$,*

$$\pi(\overline{\theta}|t_i=0)\geq\overline{q}.$$

To see this, observe that $\pi(\overline{\theta}|t_1 = 0) = 1$ by definition, and for each $i \in I$

$$\pi(\overline{\theta}|t_i = 1) \geq \frac{\frac{\epsilon}{2}}{\frac{\epsilon}{2} + \eta} \geq \overline{q} \tag{6}$$

Claims A.2 and A.3 in Morris et al. (2024)'s Theorem 1(2) proof do not need to be restated and the proofs follow in exactly the same way. They are stated here for completeness (with $|I| = 2$ plugged in)

Claim A.2 (from Morris et al. (2024)'s Theorem 1(2) proof). *For any $i \in I$, any $\tau \in \{2, 3, ...\}$, and any $S \subset I \backslash i$,*

$$\pi(\{j \neq i|t_j\} = S, \theta|t_i = \tau) = \frac{(1-\eta)^{1-|S|}\tilde{\nu}_\Gamma(\{\gamma \in \Gamma_i | a_{-i}(\gamma) = b_S\} \times \theta)}{\sum_{\ell=1}^{2}(1-\eta)^{2-\ell}\tilde{\nu}_\Gamma(\{\gamma = (i_1, ..., i_k) \in \Gamma_i | i_\ell = i\} \times \Theta)}$$

where b_S equals w if $-i$ is in S and n otherwise.

Claim A.3 (from Morris et al. (2024)'s Theorem 1(2) proof). *For any $i \in I$ such that $\tilde{\nu}_\Gamma(\Gamma_i \times \Theta) < 1$, $\pi(\{j \neq i|t_j < \infty\} = S|t_i = \infty) = \frac{\nu(b_S, \theta)}{(1-\epsilon)(1-\tilde{\nu}_\Gamma(\Gamma_i \times \Theta)}$ for all $S \subset I \backslash \{i\}$.*

Now, we can complete Step 1. First, observe that action w is strictly dominant for $t_1 = 0$ and $t_2 = 1$ by Claim A.1 and conditions 4 and 5. For $\infty > \tau \geq 2$, the same exact steps can be made as in Morris et al. (2024) to show that for each type $t_i < \infty$, it is a strict-BR to choose w as long as types $t_{-i} < t_i$ do so (definition of an unraveling information structure). So, the unique rationalizable outcome is w for any $t_i < \infty$, and the principal's worst rationalizable outcome is for all agents with finite type to choose w and all agents with time $t_i = \infty$ to choose n (and this is also the principal's worst BNE).

Step 2. I show that if an outcome ν satisfies asymmetric grain of dominance, it is without loss for the principal's value to use information structures that satisfy *perfect coordination*—in the principal's worst equilibrium, both choose w or both choose n.

From Step 1, we know that if an outcome ν can be generated as the principal's worst equilibrium from some information structure, then there exists a $\pi(t, \theta)$ defined on $T \times \Theta$ with $T \in (\mathbb{N} \cup \{\infty\})^2$ such that, in the principal's worst equilibrium, agent i chooses w if and only if $t_i = \infty$. If $\mathbb{P}_\pi(t_i = \infty, t_{-i} < \infty) = 0$ for each i, then we are done. Otherwise, consider the modification $(\tilde{T}, \tilde{\pi})$ defined by:

- $\tilde{T}_i = (\mathbb{N} \cup \infty)^2$
- $\tilde{\pi}((t_1, 0), (t_2, 0)) = \tilde{\pi}(t_1, t_2)$ if $t_1, t_2 < \infty$ or $t_1 = t_2 = \infty$
- $\tilde{\pi}((t_1, t_2), (t_2, 0)) = \tilde{\pi}(t_1, t_2)$ if $t_1 = \infty$ and $t_2 < \infty$
- $\tilde{\pi}((t_1, 0), (t_2, t_1)) = \tilde{\pi}(t_1, t_2)$ if $t_2 = \infty$ and $t_1 < \infty$
- $\tilde{\pi}((t_1, t_1'), (t_2, t_2')) = 0$ otherwise

Under this information structure, the principal's worst equilibrium involves players choosing w when their type is $(t_i, 0)$ with $t_i < \infty$ but *also* if their type (t_i, x) with $t_i = \infty$ and $x < \infty$, consequences of Assumption 2 and supermodularity. Hence, the principal's value under under $(\tilde{T}, \tilde{\pi})$ is higher than under (T, π).

As a result, it is without loss of generality for the principal's value to constrain to the subset of information structures described in step (1) with the property that $\pi(\boldsymbol{t}) > 0$ only if either (i) $t_1 = t_2 = \infty$ or (ii) $t_1, t_2 < \infty$.

Step 3. Observe that any information structure (T, π) from step (1) which satisfies $\pi(\boldsymbol{t}, \theta) > 0$ only if either $t_1 = t_2 = \infty$ or $t_1, t_2 < \infty$ has the following properties (i) $\pi(\boldsymbol{t}, \theta) = 0$ unless $|t_1 - t_2| = 1$, (ii) $\pi((m, m), \theta) = 0$ for any $m \neq \infty$, (iii) (T, π) is a partially unraveling information structure (a consequence of Claim A.1, and the argument after Claim A.3 concluding the proof in Morris et al. (2024)).

These are the only properties required for an information structure to be a partially unraveling information structures, and so combining steps 1 and 2 shows that if an outcome satisfies asymmetric grain of dominance, it is without loss of value for the principal to restrict to partially unraveling information structures.

Step 4. Finally, I show that as long as there exists a state in which w is strictly dominant for at least one agent, requiring asymmetric grain of dominance is without loss of value for the principal. If there is no state such that w is strictly dominant for one agent, then the principal's worst equilibrium under any information structure is for both agents to choose n with probability 1.

Otherwise, there exists a set $\overline{\Theta}$ and some i, say $i = 1$, such that $d_1(a_2, \theta) > 0$ for each $a_2 \in \{w, n\}$ and any $\theta \in \overline{\Theta}$. Fix an implementable outcome $\nu \in \Delta(A^I \times \Theta)$ that fails to satisfy asymmetric grain of dominance, i.e., for every $\theta \in \overline{\Theta}$, $v((w, w), \theta) = 0$.

Let $\mu(\theta)$ be the prior probability of θ and consider the modification

$$\tilde{\nu}(\mathbf{a},\theta) = (1-\epsilon)\nu(\mathbf{a},\theta) + \epsilon \mathbf{1}_{\mathbf{a}=(w,w)}\mu(\theta).$$

Notice that $\tilde{v}(\mathbf{a},\theta)$ satisfies asymmetric grain of dominance, as well as consistency. So, if it satisfies obedience and sequential obedience, that will conclude the proof. Obedience for r is maintained, by the dominance assumption for agent 1 and subsequently assumption 2 (Jointly Informing) for agent 2. The obedience constraint for n is unchanged, and obedience for $\tilde{v}$ follows from obedience of
nu.

Since ν satisfies sequential obedience, there exists $\nu_\Gamma \in \Delta(\Gamma \times \Theta)$ such that for each i with $\nu_\Gamma(\Gamma_i \times \Theta) > 0$ we have

$$\sum_{\gamma \in \Gamma_i, \theta \in \Theta} \nu_\Gamma(\gamma,\theta) d_i(a_{-i}(\gamma),\theta) > 0$$

and $\nu(a,\theta) = \sum_{\gamma | a = \overline{a}(\gamma)} \nu_\Gamma(\gamma,\theta)$. Then, consider

$$\tilde{\nu}_\Gamma(\gamma,\theta) = \mathbf{1}_{\theta \neq \overline{\theta}} \nu_\Gamma(\gamma,\theta) + \mathbf{1}_{\theta=\overline{\theta}} \left((1-\epsilon)\nu_\Gamma(\gamma,\theta) + \epsilon \mu_{\overline{\theta}} \mathbf{1}_{\gamma=(1,2)} \right).$$

For ϵ sufficiently small, sequential obedience holds. As a result, $\tilde{v}(\gamma,\theta)$ is implementable or ϵ sufficiently small. Finally, the change in the principal's value moving from ν to $\tilde{\nu}$ is $O(\epsilon)$, so the principal's value can be approximated arbitrarily well by taking ϵ small. This concludes the proof. □

Proof of Proposition 6: Fix any prior μ. The principal's worst BNE is in pure strategies and is either (w,w) or (n,n). To see this, observe that there is no worst equilibrium in which one player chooses w and the other chooses n, a result of the assumption 2 (Jointly Informing). To see that a mixed strategy equilibrium cannot be the principal's worst equilibrium, suppose that some agent, say player 1, is mixing with strictly positive probability on both w and n. Then, player 2 must be choosing n with strictly positive probability, otherwise player 1 has a strict best-response to choose w. Then I claim that (n, n) is an equilibrium, which is weakly worse for the principal. To see this, let p_i be the probability that player i places on choosing w. Then, letting $\mathbb{E}^0(.) \equiv \mathbb{E}_\mu(.)$, i.e., the expectation given the prior, best-response

requires:

$$p_{-i}\mathbb{E}^0(u(w,w,\theta)) + (1-p_{-i})\mathbb{E}^0(u(w,n,\theta)) \leq p_{-i}\mathbb{E}^0(u(n,w,\theta)) + (1-p_{-i})\mathbb{E}^0(u(n,n,\theta))$$

for each i, where the lhs is payoff to n and rhs is payoff to w (and with equality for $i = 1$, who is strictly mixing). Rearranging yields

$$p_{-i}\mathbb{E}^0(u(w,w,\theta) - u(n,w,\theta)) \leq (1-p_{-i})\mathbb{E}^0(u(n,n,\theta) - u(w,n,\theta))$$

Observe that the left hand-side is positive, and so the right-hand side must be as well. Since $1 - p_{-i} > 0$ for each i, then choosing n is a best-response to $-i$ choosing n. But then (n, n) is an equilibrium.

Therefore, given any prior, the principal's worst equilibrium is either (w, w) or (n, n). As a result, we need only consider the effect of introducing private signals when worst-equilibrium behavior is (w, w) or (n, n).

If both players choose n in the principal's worst equilibrium, the introduction of private information cannot lower the principal's value. Suppose instead μ is such that both players choose w in the principal's worst equilibrium. Then, there must exist some player, say player 2, for whom w is strictly dominant.[23] By the definition of least tempted, it must be true that w is dominant for $-i^{least}$. Then, under the simple information structure with $i^* = i^{least}$, agent $-i*$ still chooses w, by supermodularity. But then, player i^* chooses w by Assumption 2.

To see that a simple information structure with $i^* = i^{least}$ is communication-proof, observe that in any communication equilibrium, if i^* observes $\theta \in \text{DOM}_{i^*}$ (i.e., dominant to choose w) then he must receive recommendation w with probability 1. Denote by t_i^j the type of agent $i \in I$ who observes signal $j \in \text{DOM}_{i*} \cup \{\emptyset\}$. There are then two cases to consider:

Case 1: Suppose that both agents choose w with probability 1 in the principal's worst BNE. Then, the information structure is an unraveling information structure satisfying the assumption of Proposition 1 and hence, is communication-proof.

Case 2: Suppose that the principal's worst BNE involves $t_{-i^*}^{\emptyset}$ and $t_{i^*}^{\emptyset}$ choosing n with positive probability. Then, by supermodularity, the principal's worst BNE is $t_{i^*}^{\theta}$ choosing w

[23]Formally, suppose this is not true. Then, n is a (weak) best-response for each agent to some choice of his partner. But, w is a strict best-response to w, so n must be a (weak) best-reponse to n. Hence, both choosing n is the principal's worst equilibrium.

for any $\theta \in \text{DOM}_{i^*}$ and $t_i^{\emptyset}$ choosing n for each $i \in I$. Then, the only way a communication equilibrium can lower the principal's value is if $t_{i^*}^{\theta}$ chooses n with positive probability for some $\theta \in \text{DOM}_{i^*}$, but this is impossible.

By Assumption 2, it is not possible that only one player chooses n with positive probability in the principal's worst BNE. Hence, these two cases are exhaustive and show that simple information structures are communication-proof. □

C Proof of Proposition 7

Let $\Delta u_i(\boldsymbol{a}_{-i}, \theta) \equiv u_i(n, \boldsymbol{a}_{-i}, \theta) - u_i(w, \boldsymbol{a}_{-i}, \theta)$. Let

$$\underline{s}_1 \equiv \min_{i \in I, \theta \in \Theta, \boldsymbol{a_{-i}} \in \{(n,w),(w,n)\}} \{\Delta u_i(\boldsymbol{a}_{-i}, \theta) - \Delta u_i((w,w), \theta)\}$$

and

$$\underline{s}_2 \equiv \min_{i \in I, \theta \in \Theta, \boldsymbol{a_{-i}} \in \{(n,w),(w,n),(w,w)\}} \{\Delta u_i((n,n), \theta) - \Delta u_i(\boldsymbol{a}_{-i}, \theta)\}.$$

Observe that, as a consequence of assumptions 4 and 5, $\underline{s}_1, \underline{s}_2 > 0$.

Lemma 3. *There exists $\phi, \psi, \bar{\epsilon} > 0$ s.t. $\forall \epsilon \leq \bar{\epsilon}$ and $\nu_\Gamma \in \Delta(\Gamma \cup \{\emptyset\} \times \Theta)$ with $\max_i |s_i(\nu_\Gamma)| < \epsilon$, then for all $i \in I$,*

$$\sum_{j \neq k \neq i, \theta \in \Theta} \nu_\Gamma(ijk, \theta) \in (\phi, 1 - \phi)$$

and

$$\sum_{j \neq k \neq i, \theta \in \Theta} \Delta u_i((n,n), \theta) \nu_\Gamma(ijk, \theta) \geq \psi > 0.$$

Proof. Recall now that

$$s_i(\nu_\Gamma) = \sum_{j \neq k, \theta \in \Theta} \nu_\Gamma(ijk, \theta) \Delta u_i((n,n), \theta) + \sum_{\theta \in \Theta, j \neq k \neq i, \gamma \notin \{ijk, ikj\}} \nu_\Gamma(\gamma, \theta) \Delta u_i(a_{-i}(\gamma), \theta).$$

The result then follows from the fact that w is a strict best-response to $a_{-i} \in \{(w,w), (n,w), (w,n)\}$ in every state θ. □

Proof of Proposition 7: In an arbitrary communication mechanism, denote a reported type profile by $\tilde{\boldsymbol{t}} = (\tilde{t}_1, \tilde{t}_2, \tilde{t}_3)$, and denote by $\tilde{\boldsymbol{t}}^{reord}$ the triple with type reports ordered from smallest to largest (with ties broken arbitrarily).

Consider the following communication mechanism, $\sigma^{\gamma,\delta}$, for arbitrary $\gamma, \delta \in (0,1)$:

- If $\tilde{\boldsymbol{t}}^{reord} = (k, k+1, k+2)$ for $k \geq 5$, $\sigma = (n,n,n)$ w.p. 1
- If $\tilde{\boldsymbol{t}}^{reord} = (0,1,2)$, $\sigma = (w,w,w)$ w.p. 1
- If $\tilde{\boldsymbol{t}}^{reord} = (1,2,3)$, $\sigma_i = n$ w.p. $1-\delta$ for $\tilde{t}_i = 3$, and $\sigma_i = w$ w.p. 1 otherwise
- If $\tilde{\boldsymbol{t}}^{reord} = (2,3,4)$, $\sigma_i = n$ w.p. $1-\delta$ for $\tilde{t}_i = 3$, $\sigma_i = n$ w.p. 1 for $\tilde{t}_i = 4$, and $\sigma_i = w$ w.p. 1 for $\tilde{t}_i = 2$.
- If $\tilde{\boldsymbol{t}}^{reord} = (3,4,5)$, $\sigma = (n,n,n)$ w.p. $1-\gamma\delta$, $\sigma_i = w$ w.p $\gamma\delta$ for $\tilde{t}_i \in \{3,4\}$, and $\sigma_i = n$ w.p. 1 for $\tilde{t}_i = 5$
- If $\tilde{\boldsymbol{t}}^{reord} = (4,5,6)$, $\sigma = (n,n,n)$ w.p. $1-\gamma\delta$, $\sigma_i = n$ w.p 1 if $\tilde{t}_i = 6$, and $\sigma_i = w$ w.p. $\gamma\delta$ for $\tilde{t}_i \in \{4,5\}$
- If $\tilde{\boldsymbol{t}}^{reord} \in \{(0,2,4),(1,3,5),(0,1,5),(2,4,6),(1,2,6)\}$, types with $\sigma_i = n$ w.p. 1 for $\tilde{t}_i \in \{4,5,6\}$ and $\sigma_i = w$ w.p. 1 for $\tilde{t}_i \in \{0,1,2,3\}$
- Otherwise, $\sigma_i = w$ for all i

To prove the proposition, I will show that there exists $\bar{\epsilon}$ sufficiently small such that, if $\min_{i \in I} |s_i(\nu_\Gamma)| \leq \bar{\epsilon}$, then there exists γ, δ sufficiently small that for any η sufficiently close to 1, $\sigma^{\gamma,\delta}$ is a communication equilibrium. The result then follows immediately. The proof is in a number of steps, each step corresponding to showing that the incentive constraints of a type are satisfied for properly chosen $\bar{\epsilon}, \gamma, \delta$, and η. To start, fix some $\phi, \psi > 0$ such that Lemma 3 is true for some common $\bar{\epsilon}_{init} > 0$.

1. **Obedience and Truth-telling for** $t_i = 0$: An agent with type $t_i = 0$ finds w dominant, and upon truthtelling receives recommendation w w.p. 1 and hence, learns nothing from the communication mechanism $\sigma^{\gamma,\delta}$ that changes the dominance of w. As a result, conditional on truth-telling, obedience is guaranteed. For any misreport of $t_i = 0$, $\sigma_{-i} = (w,w)$ w.p. 1, to which w is a best-response. Hence, truth-telling and obedience is optimal.

2. **Obedience and Truth-telling for** $t_i = 1$: An agent with type $t_i = 1$ who reports $\tilde{t}_i = 1$ receives $\sigma_i = w$ w.p. 1. Since w is a strict best-response to w and $p > 0$, there exists $\eta_1 < 1$ s.t. for all $\eta \in [\eta_1, 1]$, it is optimal for i to obey. An agent with type $t_i = 1$ could misreport, in which case:

- If i misreports $\tilde{t}_i = 4$, no other agent's recommendation changes, and $\sigma_i = n$ w.p. 1. Hence, misreporting $\tilde{t}_i = 4$ is not profitable, under the assumption that other agents report the truth and obey their recommendations.
- If i misreports $\tilde{t}_i \notin \{4, 1\}$, all agents receive recommendation $\sigma_i = w$ w.p. 1. Since by assumption 2 w is a best-response to w, misreporting $\tilde{t}_i \notin \{4, 1\}$ is not profitable.

3. **Obedience and Truth-telling for** $t_i = 2$: I argue that for any $\delta > 0$, there exists $\bar{\gamma}_2(\delta) > 0$ s.t. $\forall \gamma \geq \bar{\gamma}_2(\delta)$, and $\epsilon \leq \bar{\epsilon}_{init}$, truth-telling and obedience are satisfied.

   To see this, observe that for any $\delta > 0$, for $\gamma = 1$, $\tilde{t}_i = 2$ strictly prefers truth-telling to mis-reporting $\tilde{t}_i = 5$, by assumption 4. By continuity of payoffs in δ and γ (assuming others report the truth and obey), there exists a $\bar{\gamma}_2(\delta)$ s.t. for all $\gamma \in (\bar{\gamma}_2(\delta), 1)$, this remains true. All other mis-reports lead to $\sigma = (w, w, w)$ with probability 1. By the definition of sequential obedience, an agent with type $\tilde{t}_i = 2$ who reports truthfully obeys the recommendation w w.p. 1.

4. **Obedience and Truth-telling for** $t_i = 3$**:** I will show that for all $\gamma, \delta > 0$, there exists $\bar{\epsilon}_3 > 0$ s.t. $\forall \epsilon \leq \bar{\epsilon}_3$, $t_i = 3$ prefers truth-telling and obedience for η sufficiently close to 1.

   Fix any $\gamma, \delta > 0$. For $\epsilon \leq \bar{\epsilon}_{init}$,

$$\sum_{i \neq j \neq k, \theta \in \Theta} \nu_\Gamma(ijk, \theta) \Delta u_i((n, n), \theta) \geq \psi > 0 \tag{7}$$

   Recall now that

$$s_i(\nu_\Gamma) = \sum_{i \neq j \neq k, \theta \in \Theta} \nu_\Gamma(ijk, \theta) \Delta u_i((n, n), \theta) + \sum_{\theta \in \Theta, \gamma \notin \{ijk, ikj\}, i \neq j \neq k} \nu_\Gamma(\gamma), \theta) \Delta u_i(a_{-i}(\gamma), \theta).$$

   Then, there is $\bar{\epsilon}_3$ sufficiently small s.t. if $|s_i(\nu_\Gamma)| < \bar{\epsilon}_3$, the definition of $s_i(\nu_\Gamma)$ and

inequality 7 imply that

$$\begin{aligned}
&\sum_{\theta\in\Theta,\gamma\notin\{ijk,ikj\}} \nu_\Gamma(\gamma),\theta)\Delta u_i(a_{-i}(\gamma),\theta) \\
&= \sum_{\theta\in\Theta,j\neq k\neq i,\gamma\in\{jii,kij\}} \nu_\Gamma(\gamma,\theta)\Delta u_i(a_{-i}(\gamma),\theta) + \sum_{\theta\in\Theta,j\neq k\neq i\gamma\in\{jki,kji\}} \nu_\Gamma(\gamma,\theta)\Delta u_i(a_{-i}(\gamma),\theta) \\
&< \quad 0 \qquad (8)
\end{aligned}$$

If $t_i = 3$ reports truthfully, then recommendation n from the communication mechanism is obedient as long as:

$$\begin{aligned}
&\eta^2\left(\frac{1-\gamma\delta}{1-\delta}\right)\sum_{j\neq k,\theta\in\Theta}\nu_\Gamma(ijk,\theta)\Delta u_i((n,n),\theta) \\
&+\eta\sum_{\theta\in\Theta,j\neq k\neq i,\gamma\in\{jii,kij\}}\nu_\Gamma(\gamma),\theta)\Delta u_i(a_{-i}(\gamma),\theta) \\
&+\sum_{\theta\in\Theta,j\neq k\neq i,\gamma\in\{jki,kji\}}\nu_\Gamma(\gamma),\theta)\Delta u_i(a_{-i}(\gamma),\theta)\geq 0.
\end{aligned}$$

Observe that the left-hand side of this inequality equals $s_i(\nu_\Gamma)$ at $\nu = 1$ and $\delta = 0$. From sequential obedience of ν_Γ, we know that $s_i(\nu_\Gamma) > 0$. Therefore, for any $\delta, \gamma \in (0,1)$ and for $\epsilon \leq \bar{\epsilon}_3$, there exists η_4 sufficiently close to 1 such that recommendation n is obedient. Let $\bar{\epsilon}_3(\gamma,\delta)$ and $\bar{\eta}(\gamma,\delta)$ be such that, for any $\epsilon \leq \bar{\epsilon}_3$ and $\eta \in (\bar{\eta}(\gamma,\delta),1)$, obedience holds.

Finally, if $t_i = 3$ misreports $\tilde{t}_i = 6$, she receives $\sigma_i = n$ w.p. 1, is worse off when $\tilde{t}_i \in \{(4,5),(5,4)\}$ and induces $\sigma_{-i} = (w,w)$ otherwise. Hence, mis-reporting $\tilde{t}_i = 6$ is not profitable. All other misreports lead to $\sigma = (w,w,w)$, and hence are not profitable.

5. **Obedience and Truth-telling for $t_i = 4$:** I show that there exists $\bar{\delta}_4, \bar{\eta}_4$ s.t. $\forall \delta \leq \bar{\delta}_4$, $\epsilon \leq \bar{\epsilon}_{init}$, $\gamma \in [0,1]$, and $\eta \in (\bar{\eta}_4, 1)$, $t_i = 4$ prefers truth-telling and obedience. To see this, observe that if $t_i = 4$ reports the truth and

   - receives recommendation w, then $\sigma_{-i} = (w,w)$ w.p. 1, and hence obedience holds by assumption 2

- receives recommendation n, then obedience holds if

$$\overbrace{\eta^2 \sum_{j\neq k,\theta\in\Theta} \nu_\Gamma(ijk,\theta)\Delta u_i((n,n),\theta)}^{A_\eta} + \overbrace{\eta \sum_{\theta\in\Theta,\gamma\in\{jik,kij\}} \nu_\Gamma(\gamma,\theta)\Delta u_i((n,n),\theta)}^{B^1_{\eta,\gamma,\delta}}$$

$$+ \overbrace{\frac{(1-\delta)}{1-\gamma\delta} \sum_{\theta\in\Theta} \nu_\Gamma(kji,\theta)\Delta u_i((a_j=n,a_k=w),\theta) + \nu_\Gamma(jki,\theta)\Delta u_i((a_j=w,a_k=n),\theta)}^{B^{2a}_{\eta,\gamma,\delta}}$$

$$+ \overbrace{\frac{\delta}{1-\gamma\delta} \sum_{\theta\in\Theta,\gamma\in\{jki,kji\}} \nu_\Gamma(\gamma,\theta)\Delta u_i((w,w),\theta)}^{B^{2b}_{\eta,\gamma,\delta}} \tag{9}$$

$$\geq 0 \tag{10}$$

By assumption, we know that

$$\overbrace{\sum_{j\neq k,\theta\in\Theta,\gamma\in\{ijk,ikj\}} \nu_\Gamma(ijk,\theta)\Delta u_i((n,n),\theta)}^{A} + \overbrace{\sum_{\theta\in\Theta,\gamma\in\{jik,kij\}} \nu_\Gamma(\gamma,\theta)\Delta u_i(a_{-i}(\gamma),\theta)}^{B^1}$$

$$+ \overbrace{\sum_{\theta\in\Theta,\gamma\in\{jki,kji\}} \nu_\Gamma(\gamma,\theta)\Delta u_i((w,w),\theta)}^{B^2}$$

$$> 0 \tag{11}$$

If $\epsilon \leq \bar{\epsilon}_{init}$, then by lemma 3,

$$\sum_{\theta\in\Theta,\gamma\in\{jik,kij\}} \nu_\Gamma(\gamma,\theta) + \sum_{\theta\in\Theta,\gamma\in\{jki,kji\}} \nu_\Gamma(\gamma,\theta) \geq \phi > 0$$

Since $\underline{s}_1, \underline{s}_2 > 0$, there exists $\delta_4 \in (0,1), \bar{\eta}_4 < 1$ s.t. for any $\gamma \in [0,1], \delta \leq \bar{\delta}_4, \eta \in (\bar{\eta}_4, 1)$, $|A_\eta - A| \leq \min\{\bar{s}_1, \bar{s}_2\}\frac{\phi}{2}$ and $B^1_{\eta,\gamma,\delta} + B^{2a}_{\eta,\gamma,\delta} + B^{2b}_{\eta,\gamma,\delta} - (B^1 + B^2) > \phi \min\{\bar{s}_1, \bar{s}_2\}$. Hence, obedience holds for any such δ and η.

If $t_i = 4$ misreports $\tilde{t}_i = 1$, then i receives an uninformative recommendation of w, and the recommendations to other agents do not improve i's payoffs. All other mis-reports lead to $\sigma = (w, w, w)$ w.p. 1. Hence, truthtelling and obedience is optimal for $t_i = 4$.

6. **Obedience and Truth-telling for $t_i = 5$:** All agents truthfully reporting type $t_i = 5$ receive $\sigma_i = n$ w.p. approaching 1 as $\delta \to 0$, and conditional on receiving n, $-i$ receive $\sigma_j = n$. Conditional on $t_i = 5$ reporting truthfully and receiving recommendation w, i knows that $\boldsymbol{t} = (4, 5, 6)$ and $t_i = 4$ receives recommendation w as well, hence w is obedient. For any mis-report and for $\epsilon \leq \bar{\epsilon}_{init}$, there exists a probability at least ϕ that partners will receive recommendation (w, w). Since $\underline{s}_2 > 0$, then, for $\epsilon \leq \bar{\epsilon}_{init}$, there exists $\bar{\eta}_5$ and $\bar{\delta}_5 \in (0, 1)$ s.t. $\forall \eta \in (\bar{\eta}_5, 1)$, $\delta \leq \bar{\delta}_5$, $t_i = 5$ prefers truth-telling and obedience.

7. **Obedience and Truth-telling for $t_i = 6$:** All agents truthfully reporting type $t_i = 6$ receive $\sigma_i = n$ w.p. 1. As δ shrinks to 0 and η converges to 1, $-i$ receive $\sigma_i = n$ w.p. converging to 1. For any mis-report, there exists a probability at least ϕ that partners will receive recommendation (w, w). Since $\underline{s}_2 > 0$ then, for $\epsilon \leq \bar{\epsilon}_{init}$, there exists $\bar{\eta}_6$ and $\bar{\delta}_6$ s.t. $\forall \eta \in (\bar{\eta}_6, 1)$, $\delta \leq \bar{\delta}_6$, $t_i = 6$ prefers obedience and truth-telling.

8. **Obedience and Truth-telling for $t_i \geq 7$:** All agents truthfully reporting type $t_i \geq 7$ receive $\sigma_i = n$ w.p. 1, as do their partners. There exists $\bar{\eta}_7$ s.t. $\forall \eta \in (\bar{\eta}_7, 1)$, n is obedient. Hence, $t_i = 7$ prefers truth-telling and obedience.

Taking $\bar{\epsilon} = \min\{\bar{\epsilon}_{init}, \bar{\epsilon}_3\}$, for any $\delta \leq \min\{\bar{\delta}_4, \bar{\delta}_5, \bar{\delta}_6\}$, $\gamma \leq \bar{\gamma}_2(\delta)$, and $\bar{\eta} = \max\{\eta_1, \bar{\eta}(\gamma, \delta), \bar{\eta}_4, \bar{\eta}_5, \bar{\eta}_6, \bar{\eta}_7\}$, $\sigma^{\gamma,\delta}$ is a communication equilibrium for any $\epsilon \leq \bar{\epsilon}$ and $\eta \in (\bar{\eta}, 1)$. The results follows immediately. □

D Proofs of Section 6

Proof of Proposition 3: Observe first that it is without loss of generality to suppose that $v((n, n), \theta) = 0$, so I will proceed under this assumption. Observe also that since w is dominant in state $\overline{\theta}$ for both agents, I assume without loss of value for the principal that the principal implements w after $\overline{\theta}$ with probability 1.

Suppose that there exists no state in which some agent finds it strictly dominant to whistleblow. Then, the principal's worst equilibrium is (n, n) with certainty, independent of the information structure. Then setting, $\underline{\theta}^* = \underline{\theta}$ and $\overline{\theta}^* = \overline{\theta}$ delivers the result.

Suppose instead that there exists some state in which some agent finds it strictly dominant to whistleblow (so there exist outcomes satisfying asymmetric grain of dominance). From Proposition 2, rank unique partially unraveling information structures implement V^*. As

a result, there exists an optimal outcome ν that is *perfectly coordinated*, i.e., $\nu((w,n),\theta) = \nu((n,w),\theta) = 0$.[24] As in Morris et al. (2024), the characterization of S-implementable outcomes in the proof of Proposition 2, implies that the principal's optimal value is the solution to the linear program

$$\begin{aligned} V^* = \quad & \max \sum_{\theta\in\Theta}\sum_{i\in I} \left(v((w,w),\theta)\right) w_i(\theta) \\ \text{s.t.} \quad & \sum_{\theta\in\Theta} w_i(\theta) d_i(n,\theta) + w_{-i}(\theta) d_i(w,\theta) \geq 0, \quad i \in I \\ & w_i(\theta) \geq 0, \quad i \in I, \theta \in \Theta \\ & \textstyle\sum_{i\in I} w_i(\theta) \leq \mu(\theta), \quad \theta \in \Theta \end{aligned} \tag{P}$$

and, if $(w_i^*(\theta))_{i\in I,\theta\in\Theta}$ is an optimal solution to this problem, then an optimal outcome is:

$$\nu(\mathbf{a},\theta) = \begin{cases} \sum_{i\in I} w_i^*(\theta) & \mathbf{a} = (w,w) \\ \mu(\theta) - \sum_{i\in I} w_i^*(\theta) & \mathbf{a} = (n,n) \\ 0 & \text{otherwise} \end{cases}$$

It is convenient to formulate the dual of (P), and apply strong duality to find:

$$\begin{aligned} V^* = \quad & \min \sum_{\theta\in\Theta} \lambda(\theta)\mu(\theta) \\ \text{s.t.} \quad & \overbrace{v((w,w),\theta) + \gamma_i d_i(n,\theta) + \gamma_{-i} d_{-i}(w,\theta)}^{\alpha_i(\theta)} - \lambda(\theta) \leq 0, \quad i \in I, \theta \in \Theta \\ & \lambda(\theta) \geq 0, \qquad \theta \in \Theta \\ & \gamma_i \geq 0, \qquad i \in I \end{aligned} \tag{D}$$

Given $(\gamma_i)_{i\in I}$, an optimal $\lambda(\theta)$ is $\lambda(\theta) = \max\{0, \max_{i\in I}\{\alpha_i(\theta)\}\}$. Complementary slackness implies that if $(\lambda(\theta),\gamma_i)_{i\in I,\theta\in\Theta}$ is an optimal solution to (D), then there is an optimal solution $(w_i(\theta))_{i\in I,\theta\in\Theta}$ to (P) such that $\lambda(\theta) = \max\{0, \max_{i\in I}\{\alpha_i(\theta)\}\} > 0 \implies w_1(\theta) + w_2(\theta) = \mu(\theta)$ and $\max_{i\in I}\alpha_i(\theta) < 0 \implies w_1(\theta) = w_2(\theta) = 0$.

Observe that because the environment is linear, $\alpha(\theta)$ is affine in θ. Then, consider a possible solution in which $\alpha_i(\theta)$ is constant in θ for both i, in which case either $\lambda(\theta) = c > 0$

[24] The closure of any set of perfectly coordinated outcomes includes only perfectly coordinated outcome.

for all θ, or $\lambda(\theta) = 0$ for all θ. In the former case, complementary slackness implies that the principal achieves first best i.e., $V^* = \sum_{\theta \in \Theta} \mu(\theta) v((w,w),\theta)$. Then, the result follows by setting $\theta^* = \underline{\theta}^*$. Instead, suppose $\lambda(\theta) = 0$. But then, observe that since w is dominant at $\overline{\theta}$ for both agents and $\Delta v(\theta) > 0$ (since $d_i(n,\theta) \geq 0$ and $d_i(w,\theta) > 0$), the only way $\lambda(\overline{\theta}) = 0$ is if $\gamma_i = \gamma_{-i} = 0$. But in that case, if $\lambda(\theta) = 0$ for all $\theta \in \Theta$, it must be that $v((w,w),\theta) = 0$ for all $\theta \in \Theta$. In that case, the principal is indifferent whether agents choose w or n; Then, an optimal policy is to provide no information, in which case agents either choose w or choose n independent of θ. Then setting $\theta^* = \overline{\theta}$ or $\theta^* = \underline{\theta}$ leads to the result.

Suppose next that $\alpha_i(\theta)$ is constant for some i but is non-constant for $-i$, say $i = 1$. Again since w is dominant at $\theta = \overline{\theta}$, it cannot be that $\alpha_i(\theta) < 0$. If $\alpha_1(\theta) > 0$, then the same argument as above applies. Suppose instead that $\alpha_1(\theta) = 0$ for all θ. As before, it must be that $D_i(\overline{\theta}) > 0$ for each i, and so it must be that $\gamma_i = 0$ for each i. But then, as above, $\Delta v(\theta) = 0$, and the principal can set $\theta^* = \overline{\theta}$.

So, I proceed now under the assumption that $\alpha_i(\theta)$ is non-constant for each i. First, suppose that $\alpha_i(\overline{\theta}) = 0$ for some i. In that case, $\gamma_i = 0$ for each i and so $\lambda(\theta) = v((w,w),\theta)$. Then, $V^* = \sum_{\theta \in \Theta} \mu(\theta) v((w,w),\theta)$. If $v((w,w),\theta) \neq 0$ for each θ, then the principal necessarily implements w in each state, and setting $\theta^* = \underline{\theta}$ leads to the result. Suppose instead that $v((w,w),\theta) = 0$ for some θ. Since $v((w,w),\theta)$ is linear, there are three possibilities:

- $v((w,w),\theta) = 0$ for all θ: in this case the principal is indifferent whether agents choose w or n, in which the result follows as above.

- $v((w,w),\overline{\theta}) = 0$ and $v((w,w),\theta) > 0$ otherwise: In this case, the principal must induce w with probability 1 in every state other than $\overline{\theta}$. But, I assumed that the principal implements w after $\overline{\theta}$ with probability 1, and so the principal in fact implements w with probability 1 after each state. Setting $\theta^* = \underline{\theta}$ implies the result.

- $v((w,w),\underline{\theta}) = 0$ and $v((w,w),\overline{\theta})) > 0$ otherwise:

So, I proceed under the assumption that $\alpha_i(\overline{\theta}) \neq 0$ for each i. As already stated, it must be that $\alpha_i(\overline{\theta}) > 0$ for each i.

Case 1: $\alpha_i(\theta)$ is strictly decreasing in θ for some i. In this case, since $\alpha_i(\overline{\theta}) > 0$, it must be that $\alpha_i(\theta) > 0$ for all $\theta \in \Theta$. But then by complementary slackness, an optimal outcome for the principal is w with probability 1, independent of θ. Setting $\theta^* = \underline{\theta}$ implies the result.

Case 2: $\alpha_i(\theta)$ is strictly increasing in θ for each i. In this case, let $\theta_i^* = \min\{\theta|\alpha_i(\theta) \geq 0\}$. Then, for any $\theta > \min_{i\in I}\{\theta_i^*\}$, $\lambda(\theta) > 0$ and complementary slackness implies the principal implements w with certainty after any such θ. Since $\alpha_i(\theta)$ is strictly increasing, for any $\theta < \min_{i\in I}\{\theta_i^*\}$, $\alpha_i(\theta) < 0$ for each i, in which case $\lambda(\theta) = 0$, and the first constraint in (D) constraint is slack; as a result, it must be that the principal implements w with probability 0 after any such θ. Setting $\theta^* = \min_{i\in I}\{\theta_i^*\}$ implies the result. □

Proof of Proposition 5: I prove first the relationship

$$V^*(\mathcal{G}) \leq V^*(\mathcal{G}^{\epsilon,\epsilon}) \leq V^*(\mathcal{G}^{\delta,\delta}) \tag{12}$$

for any $\epsilon, \delta \in \mathbb{R}_+$. The relationship of $V^*(\mathcal{G}^{\epsilon,\delta})$ to the others follows immediately from this and Proposition 4.

For any environment $\mathcal{G}$, let

$$d_i^n(\theta;\mathcal{G}) \equiv u_i(w,n,\theta) - u_i(n,n,\theta).$$

Plugging in the linear, symmetric preferences, we have

$$d_i^n(\theta;\mathcal{G}) = (g_i(n,n;\mathcal{G}) - g_i(n,w;\mathcal{G}))(1-\theta) + (\ell_i(n,n;\mathcal{G}) - \ell_i(n,w;\mathcal{G}))\theta.$$

Letting $g_i^n(\mathcal{G}) \equiv g_i(n,n;\mathcal{G}) - g_i(n,w;\mathcal{G})$ and $\ell_i^n(\mathcal{G}) \equiv \ell_i(n,n;\mathcal{G}) - \ell_i(n,w;\mathcal{G})$, then

$$d_i^n(\theta;\mathcal{G}) = g_i^n(\mathcal{G})(1-\theta) + \ell_i^n(\mathcal{G})\theta \tag{13}$$

For a symmetric environment $\mathcal{G}$, and $\gamma \in \mathbb{R}_+$, then:

$$\begin{aligned} d_1^n(\theta;\mathcal{G}^{\gamma,\gamma}) &= g_1^n(\mathcal{G}^{\gamma,\gamma})(1-\theta) + \ell_1^n(\mathcal{G}^{\gamma,\gamma})\theta & (14)\\ &= (g^n(\mathcal{G}) + \epsilon)(1-\theta) + \ell^n(\mathcal{G})\theta & (15)\end{aligned}$$

where I drop the dependence of $g_1^n(\mathcal{G})$ on i since $\mathcal{G}$ is symmetric. Similarly,

$$d_2^n(\theta;\mathcal{G}^{\gamma,\gamma}) = (g^n(\mathcal{G}) - \epsilon)(1-\theta) + \ell^n(\mathcal{G})\theta \tag{16}$$

From the proof of Proposition 3, for any $\gamma \geq 0$,

$$
\begin{aligned}
V^*(\mathcal{G}^{\gamma,\gamma}) = \quad & \max \sum_{\theta\in\Theta}\sum_{i\in I} v((w,w),\theta) w_i(\theta) \\
\text{s.t.} \quad & \overbrace{\sum_{\theta\in\Theta} w_1(\theta)\left((g^n(\mathcal{G})+\gamma)(1-\theta)+\ell^n(\mathcal{G})\theta\right) + w_2(\theta)(u_1(w,w,\theta)-u_1(n,w,\theta))}^{\Phi_1((w_i)_{i\in I};\gamma)} \geq 0 \\
& \overbrace{\sum_{\theta\in\Theta} w_2(\theta)\left((g^n(\mathcal{G})-\gamma)(1-\theta)+\ell^n(\mathcal{G})\theta\right) + w_1(\theta)(u_2(w,w,\theta)-u_2(n,w,\theta))}^{\Phi_2((w_i)_{i\in I};\gamma)} \geq 0 \\
& w_i(\theta) \geq 0, \quad i \in I, \theta \in \Theta \\
& \textstyle\sum_{i\in I} w_i(\theta) \leq \mu(\theta), \quad \theta \in \Theta
\end{aligned}
\tag{P^γ}
$$

From the proof of Proposition 3, there exists $\theta^*, \overline{\theta}^* \in \Theta, x^*, z^* \in \mathbb{R}_+$ with $x^* \leq \mu(\overline{\theta}^*)$ and $z^* \leq \mu(\theta^*)$, and $i^* \in I$ such that an optimal solution to this linear program is $(w_i^*)_{i\in\mathcal{I}}$ defined, for each $\theta \in \Theta$, as:

$$
\begin{aligned}
w_{i^*}(\theta) &\equiv \mathbf{1}_{\theta>\overline{\theta}^*}\mu(\theta) + \mathbf{1}_{\theta=\overline{\theta}^*}x^* \\
w_{-i^*}(\theta) &\equiv \mathbf{1}_{\theta^*<\theta<\overline{\theta^*}}\mu(\theta) + \mathbf{1}_{\theta=\overline{\theta}^*}(\mu(\theta^*)-x^*) + \mathbf{1}_{\theta=\theta^*}z^*
\end{aligned}
$$

Call any solution of this form a *monotone partition* solution. Now, observe that:

$$
\begin{aligned}
\frac{\partial\Phi_1(w_1,w_2)}{\partial\gamma} &= \sum_{\theta\in\Theta} w_1(\theta)(1-\theta) \\
\frac{\partial\Phi_1(w_1,w_2)}{\partial\gamma} &= -\sum_{\theta\in\Theta} w_2(\theta)(1-\theta)
\end{aligned}
$$

I claim that there exists an optimal solution of the form described above with the property $\frac{\partial\Phi_1((w_i)_{i\in I})}{\partial\gamma} \geq -\frac{\partial\Phi_2((w_i)_{i\in I})}{\partial\gamma}$, and the proof is relegated to Lemma 4. Without loss then, suppose that $\frac{\partial\Phi_1((w_i)_{i\in I})}{\partial\gamma} \geq -\frac{\partial\Phi_2((w_i)_{i\in I})}{\partial\gamma}$.

To complete the proof, I show that for any $\gamma' > \gamma$ such that $\mathcal{G}^{\gamma',\gamma'}$ is admissible, there exists $\overline{\theta}^{*'}, x^{*'} \leq \mu(\overline{\theta}^{*'})$ such that

$$
\begin{aligned}
w_{i^*}^{(x^{*'},\theta^{*'})}(\theta) &\equiv \mathbf{1}_{\theta>\overline{\theta}^{*'}}\mu(\theta) + \mathbf{1}_{\theta=\overline{\theta}^{*'}}x^{*'} \\
w_{-i^*}^{(x^{*'},\theta^{*'})}(\theta^{*'}) &\equiv \mathbf{1}_{\theta^*<\theta<\overline{\theta^*}}\mu(\theta) + \mathbf{1}_{\theta=\overline{\theta}^{*'}}(\mu(\overline{\theta}^{*})-x^{*'}) + \mathbf{1}_{\theta=\theta^*}z^*
\end{aligned}
$$

is feasible in problem $P^{\gamma'}$. Then, since the principal's value under $(w_i')_{i\in I}$ is identical to the principal's value under $(w_i)_{i\in I}$, the proof will be complete.

To show this, fix $\gamma' > \gamma$. Let $Q(p)$ be the quantile function of $\mu(\theta)$ (where recall, $\Theta \subset \mathbb{R}$, and consider the function $\hat{Q}(p) = \left(\sum_{\theta<Q(p)} \mu(\theta) - p, Q(p) \right)$. Define

$$\mathbf{w}^p \equiv (w_i^p)_{i\in I} \equiv \left(w_i^{\hat{Q}(p)} \right)_{i\in I}.$$

Let p^* be such that $\hat{Q}(p^*) = (x^*, \overline{\theta}^*)$. Finally, let

$$\Phi_i\left(p;\gamma'\right) \equiv \Phi_i\left(\left(w_i^{\hat{Q}(p)}\right)_{i\in I};\gamma'\right).$$

Now, observe that for any $p \in [0,1]$

$$\Phi_1(p;\gamma') = \Phi_1(p;0) + \gamma' \sum_{\theta\in\Theta} w_1^p(\theta)(1-\theta) \geq \Phi_1(p;\gamma) \tag{17}$$

$$\Phi_2(p;\gamma') = \Phi_2(p;0) - \gamma' \sum_{\theta\in\Theta} w_2^p(\theta)(1-\theta) \leq \Phi_2(p;\gamma) \tag{18}$$

and since $\hat{Q}(p^*)$ is feasible in $\mathcal{G}^\gamma$, is must be that

$$\Phi_1(p^*;\gamma) \geq 0 \tag{19}$$

$$\Phi_2(p^*;\gamma) \geq 0 \tag{20}$$

Further, for any $p, p' \in [0,1]$

$$\Phi_1(p;0) - \Phi_1(p';0) = \Phi_2(p';0) - \Phi_2(p;0) \tag{21}$$

The goal is to find $p^{*'}$ such that

$$\Phi_1(p^{*'};\gamma') \geq 0 \tag{22}$$

$$\Phi_2(p^{*'};\gamma') \geq 0 \tag{23}$$

There are two cases to consider:

- $i^* = 2$: In this case, it is straightforward to see that $\Phi_1(p,\gamma')$ is decreasing in p and $\Phi_2(p,\gamma')$ is increasing in p. Also, $w_2^p(\theta)$ is decreasing in p and $w_1^p(\theta)$ is increasing in p.

Let p_1 be a solution in $[p^*, 1]$ to

$$\Phi_1(p_1; \gamma') - \Phi_1(p^*; \gamma) = 0 \tag{24}$$

which exists because $\Phi_1(p; \gamma)$ is continuous in p, $\Phi_1(p^*; \gamma') \geq \Phi_1(p^*; \gamma)$ by (17), $\Phi_1(p^*; \gamma) \geq 0$ by (19), and $\Phi_1(1; \gamma') \leq 0$ by Assumption 1. Then, I claim that

$$\Phi_2(p_1; \gamma') - \Phi_2(p^*; \gamma) \geq 0$$

from which the result follows by (20). To see why, observe that for any $p' \in [p^*, 1]$,

$$\Phi_2(p'; \gamma') - \Phi_2(p^*; \gamma') = \Phi_1(p^*; \gamma') - \Phi_1(p'; \gamma') \geq 0 \tag{25}$$

by (21) and the definition of w_i^p. Further, by Lemma 4 (proof given below),

$$0 \geq \Phi_2(p^*; \gamma') - \Phi_2(p^*; \gamma) \geq \Phi_1(p^*; \gamma) - \Phi_1(p^*; \gamma') \tag{26}$$

for $\gamma' \geq \gamma$. Combining (25) and (26),

$$\begin{aligned}
\Phi_2(p_1; \gamma') - \Phi_2(p^*; \gamma) &= \Phi_2(p^*; \gamma') - \Phi_2(p^*; \gamma) + \Phi_2(p_1; \gamma') - \Phi_2(p^*; \gamma') \\
&= \Phi_2(p^*; \gamma') - \Phi_2(p^*; \gamma) + \Phi_1(p^*; \gamma') - \Phi_1(p_1; \gamma') \\
&\geq \Phi_1(p_1; \gamma) - \Phi_1(p^*; \gamma') \\
&= 0
\end{aligned}$$

where the second lines follows by (25) and (26) and the last line follows by the definition of p_1.

- $i^* = 1$: The proof is essentially the same, except that Φ_1 is increasing in p rather than decreasing and Φ_2 is decreasing in p rather than increasing (so one must reverse all of the relevant equations).

□

Lemma 4. *Fix a symmetric linear environment $\mathcal{G}$, in which payoffs are supermodular for agents, and w is dominant for both agents at $\overline{\theta}$. Then, for any $\gamma \geq 0$ and admissible*

perturbation $\mathcal{G}^{\gamma,\gamma}$, there exists a monotone partition solution of (P^γ), $w = (w_1, w_2)$ in which

$$\sum_{\theta\in\Theta}(1-\theta)w_1(\theta) \geq \sum_{\theta\in\Theta}(1-\theta)w_2(\theta) \tag{27}$$

Proof. From Proposition 3, there exists a monotone partition solution to the principal's problem, denoted by $\mathbf{w} = (w_1, w_2)$. Recall the constraints in (P^γ):

$$\begin{aligned} &\Phi\left((w_i)_{i\in I};\gamma\right) \geq 0, \quad i \in I \\ &w_i(\theta) \geq 0, \quad i \in I, \theta \in \Theta \\ &\textstyle\sum_{i\in I} w_i(\theta) \leq \mu(\theta), \quad \theta \in \Theta \end{aligned} \tag{28}$$

Suppose that at $\mathbf{w}$ the result in the lemma statement holds, then the proof is complete. Otherwise, suppose that

$$\sum_{\theta\in\Theta} w_1(\theta)(1-\theta) < \sum_{\theta\in\Theta} w_2(\theta)(1-\theta) \tag{29}$$

Since $\mathbf{w}$ is a solution, it is feasible, and hence (using the definition of Φ)

$$\Phi_1(\mathbf{w};\gamma) = \Phi_1(\mathbf{w};0) + \gamma\sum_{\theta\in\Theta} w_1(\theta)(1-\theta) \geq 0 \tag{30}$$

$$\Phi_2(\mathbf{w};\gamma) = \Phi_2(\mathbf{w};0) - \gamma\sum_{\theta\in\Theta} w_2(\theta)(1-\theta) \geq 0 \tag{31}$$

From the definition of a monotone partition solution, there exists $i^* \in I$, $\theta^*, \overline{\theta}^* \in \Theta$, $x^* \leq \mu(\overline{\theta}^*)$ and $z^* \leq \mu(\theta^*)$ such that

$$\begin{aligned} w_{i^*}(\theta) &\equiv \mathbf{1}_{\theta>\overline{\theta}^*}\mu(\theta) + \mathbf{1}_{\theta=\overline{\theta}^*}x^* \\ w_{-i^*}(\theta) &\equiv \mathbf{1}_{\theta^*<\theta<\overline{\theta^*}}\mu(\theta) + \mathbf{1}_{\theta=\overline{\theta}^*}(\mu(\theta^*)-x^*) + \mathbf{1}_{\theta=\theta^*}z^* \end{aligned}$$

There are two cases to consider:

- $i^* = 2$: Consider now the *reverse* policy $\mathbf{v} = (w_2, w_1)$. Then, observe that by definition,

$$\begin{aligned} \Phi_1(\mathbf{v};\gamma) - \Phi_1(\mathbf{w};\gamma) &= \Phi_2(\mathbf{w};\gamma) - \Phi_2(\mathbf{v};\gamma) + 2\gamma\sum_{\theta\in\Theta} w_2(\theta)(1-\theta) - w_1(\theta)(1-\theta) \\ &\geq \Phi_2(\mathbf{w};\gamma) - \Phi_2(\mathbf{v};\gamma) \end{aligned} \tag{32}$$

By (31), $\Phi_1(\mathbf{v};\gamma) \geq 0$. If $\Phi_1(\mathbf{v};\gamma) \leq \Phi_1(\mathbf{w};\gamma)$, then it must be that $\Phi_2(\mathbf{w};\gamma) - \Phi_2(\mathbf{v};\gamma) \leq 0$, and the result follows from (31). Otherwise, suppose $\Phi_1(\mathbf{v};\gamma) > \Phi_1(\mathbf{w};\gamma)$. Then, let p^v be defined such that $(\mathbf{w}_i^{p^v})_{i\in I} = \mathbf{v}$ and let $p^* \in [p^v, 1]$ be such that,

$$\Phi_1(\mathbf{w}^{p^*};\gamma) = \Phi_1(\mathbf{w};\gamma) \tag{33}$$

which exists because at $p = p^v$, $\Phi_1(\mathbf{w}^p;\gamma) = \Phi_1(\mathbf{v};\gamma) > \Phi_1(\mathbf{w};\gamma) \geq 0$ by assumption and (30), and at $p = 1$, $\Phi_1(\mathbf{w}^p;\gamma) < 0$. Then by definition,

$$\Phi_1(\mathbf{v};\gamma) - \Phi_1(\mathbf{w}^{p^*;\gamma}) \leq \Phi_2(\mathbf{w}^{p^*;\gamma}) - \Phi_2(\mathbf{v};\gamma) \tag{34}$$

Then,

$$\begin{aligned}
\Phi_2(\mathbf{w}^{p^*};\gamma) - \Phi_2(\mathbf{w};\gamma) &= \Phi_2(\mathbf{w}^{p^*};\gamma) - \Phi_2(\mathbf{v};\gamma) + \Phi_2(\mathbf{v};\gamma) - \Phi_2(\mathbf{w};\gamma) \\
&\geq \Phi_1(\mathbf{v};\gamma) - \Phi_2(\mathbf{w}^{p^*};\gamma) + \Phi_2(\mathbf{w};\gamma) - \Phi_2(\mathbf{v};\gamma) \\
&= \Phi_1(\mathbf{w};\gamma) - \Phi_1(\mathbf{w}^{p^*};\gamma) \\
&= 0
\end{aligned}$$

where the second line follows from (32) and (34) and the last line follows from the definition of p^*. Thus, $\mathbf{w}^{p^*}$ is also feasible for the principal, and delivers the same value. To conclude, observe that by definition and the fact that $p^* \geq p^v$ and the assumption that $\sum_{\theta\in\Theta} w_1(\theta)(1-\theta) < \sum_{\theta\in\Theta} w_2(\theta)(1-\theta)$,

$$\sum_{\theta\in\Theta} w_1^{p^*}(\theta)(1-\theta) \geq \sum_{\theta\in\Theta} w_1^{p^v}(\theta)(1-\theta) \geq \sum_{\theta\in\Theta} w_2^{p^v}(\theta)(1-\theta) \geq \sum_{\theta\in\Theta} w_2^{p^*}(\theta)(1-\theta).$$

- $i^* = 1$: The proof is identical, except that $\Phi_1(\mathbf{w}^p;\gamma)$ ($\Phi_2(\mathbf{w}^p;\gamma)$) is increasing (decreasing) in p and p^* is chosen in the set $[0, p^v]$ rather than $[p^v, 1]$.

□