Persuasion in Evidentiary Mechanisms

Sam Kapon*

[Click here for the latest version]

January 8, 2023

Abstract

A primary method of collecting evidence against a group of misbehaving agents (e.g., a cartel) is to offer *whistleblower* rewards to agents who come forward with evidence. But even before a whistleblower comes forward, the regulator typically possesses her own private evidence, though it may not guarantee conviction in court. This paper studies how revealing this private evidence to group members can facilitate whistleblowing. The more damning the regulator's evidence, the more incentivized, all else equal, a whistleblower is to come forward. I formalize this environment using information design in games. A principal (sender) informs two agents (receivers), about a state (the *evidence* state) that affects the payoffs to whistleblowing. I first ask, if agents can communicate amongst themselves, can the principal improve over public communication with the group? I answer the question in the affirmative, and provide conditions under which the principal's value is independent of whether agents can communicate or not. Second, interpreting the evidence 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. I show that the principal can facilitate more whistleblowing in groups with more asymmetrically distributed gains from misbehavior across agents. Finally, I demonstrate a class of simple information structures that, though potentially suboptimal, robustly improve over public communication.

*Princeton University: kapon@nyu.edu. I thank Sylvain Chassang, Guillaume Fréchette, Alessandro Lizzeri, Dilip Abreu, Navin Kartik, Paula Onuchic, Elliot Lipnowski, and Mauricio Ribeiro for valuable discussions and comments.

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

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. Does communication between group members—an inherent feature in these settings—mean the principal can do no better than publicly communicating with the group and if not, how harmful is communication between group members? What do optimal whistleblowing outcomes look like, and how do they vary with the underlying features of the misbehaving group? Finally, are there simple information structures that improve over public communication alone that can be implemented 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 on the group's

[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.

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]

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.

I provide three sets of results, corresponding to the three sets of questions I posed. In the first set of results, I provide answers to the questions, can the principal improve over public communication and if so, how harmful is communication amongst the agents to the principal? In general, public information design is insufficient to implement the principal's value. Communication amongst agents can therefore be harmful if it means the principal can only achieve public information design outcomes. My starting point is the class of *unraveling information structures*, which are information structures satisfying the following property; each type of an agent (i.e., signal agent receives from the principal) can be assigned a position in a list such that informing is optimal if all those in higher positions inform, and agents are never assigned the same position. These information structures guarantee that informing is the unique BNE outcome: for agents at the top of the list, informing is dominant, for

[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.

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

agents just below them informing is (iteratively) dominant given the behavior of those at the top, and so on, exhausting all possible positions. Information structures of this form have been deployed extensively throughout a variety of interrelated literatures (discussed in the literature section below)—as described in Halac, Lipnowski, and Rappoport (2022) (in a general environment in which the principal can also make allocations), such information structures comprise essentially all those that implement the principal's desired action as the unique rationalizable outcome.[6]

In general, unraveling information structures *are not* immune to communication, which I show by example: they may yield high payoffs for the principal when agents *cannot* communicate but low payoffs when agents *can* communicate. I then show that unraveling information structures satisfying a property I call *lower-rank uniqueness* are immune to communication. In particular, I establish that unraveling information structures satisfying lower-rank uniqueness are communication-proof; they induce the same outcomes whether agents can communicate or not.[7] An unraveling information structure is lower-rank unique if, after each agent observes their position in the list, they place positive probability on at most one type of their partner with a lower position on the list. The driving force underlying communication-proofness is the assumption that each agent prefers his partner not inform. As a result of this assumption, an agent who plans to inform sends whatever message leads his partner to be most likely *not* to inform, so his partner is skeptical of any attempt to convince him not to inform. In lower-rank unique unraveling information structures, this effect is so severe that communication between agents is impossible.

Unraveling information structures are not sufficiently rich to be without loss of value for the principal in all possible environments; if they were, then the principal could always achieve her first best. Indeed, if strong evidence against the group is sufficiently unlikely—so that prior to communication from the principal, informing is unattractive to agents—unraveling information structures need not exist. To solve the principal's problem, I therefore define *partially* unraveling information structures, which append to an unraveling information structure the possibility of a public signal whose observation prompts both agents *not* to inform. I show that if agents' payoffs are supermodular, it is without loss of value for the principal to restrict to partially unraveling information structures satisfying

[6] Unraveling information structures are a version of *strict ranking schemes* in Halac, Lipnowski, and Rappoport (2022).

[7] 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).

lower-rank uniqueness (Proposition 2). Supermodular payoffs can be translated into the requirement that an agent's gain when switching from *not* informing to informing is larger when his partner is informing than when his partner is not. To prove this result, I use the information structures in Morris, Oyama, and Takahashi (2022), which are shown in that paper to be sufficient to implement any outcome (in BNE). In particular, I show that one can restrict without loss of value to information structures implementing *perfectly coordinated* outcomes—those in which either both agents inform or neither does in the principal's worst equilibrium. A (slightly modified) version of the subset of information structures that implement perfectly coordinated outcomes are partially unraveling information structures satisfying lower-rank uniqueness, and hence are communication-proof.

In my second question, I 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 real numbers and agents' preferences are affine in the evidence state. I show that, as long as informing is dominant for both agents at the highest evidence state, there exists an optimal outcome in which the likelihood that agents inform is *monotonically increasing* in the evidence state. More precisely, the result states that an optimal outcome for the principal takes a simple form: there exists a threshold evidence state, both agents inform with certainty when the state is above the threshold, and neither agent informs when the state is below the threshold. The result implies that, unless the principal can achieve her first best, it is optimal to forgo inducing agents to inform in low evidence states (i.e., low likelihood of proving misbehavior without an informant) because it is too costly in terms of agents' incentives. Linear environments are natural when the evidence state is viewed as the probability that the principal can successfully prove and stop misbehavior *without* an informant.

When an agent's payoff from informing increases, holding all else fixed, the principal's optimal value increases. I also study how the principal's optimal value varies across groups with more *asymmetric* payoffs. In particular, consider any environment which is symmetric across agents except for payoffs in the event that neither agent informs. I 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.

The information structures required to implement the principal's optimal value can be complex. In my final set of questions, I ask, are there simple information structures the principal can use that improve over public communication and if so, can they be implemented without detailed knowledge of the primitives? To answer this question, I propose a class of information structures that only require private communication in which the principal fully discloses the state or discloses nothing—call these *simple discriminatory* information structures. In simple discriminatory information structures, the principal reveals a public signal to both agents, and excludes one agent from private communication. To the other agent the principal privately reveals the evidence state if it is a dominant strategy for that agent to inform, but reveals nothing otherwise. In practice, the principal may already be employing a set of public signals about her evidence e.g., press releases of an initiated investigation, costly raids on company offices, and so on. As a result, I consider the choice of simple discriminatory information structures taking as given the public signal the principal is endowed with. I show that for any public information structure, if agents' payoffs are supermodular, then there is always an improving simple discriminatory information structure.

However, given a public information structure, the details of an optimal simple discriminatory information structure generally depend finely on primitives. I show that the simple discriminatory information structure that privately communicates with the agent who is *least tempted* to inform improves over public communication alone, and requires the principal only be able to express limited knowledge; the principal must be able to identify an agent who is least tempted to inform after each public signal, and must be able to identify the states in which it is dominant to inform for each agent (or at least a subset of such states). Given a public signal realization, an agent is said to be the least tempted to inform if either informing is not dominant for him, or informing is dominant for both agents.

A stronger statement can be made if there exists an agent who is *unambiguously least tempted* to inform. An agent is unambigously least tempted to inform if in every state, the payoff from informing minus the payoff from not informing, given that one's partner is not informing, is smaller for that agent than for his partner. If such an agent exists, then the simple discriminatory information structure that privately communicates with the agent who is unambiguously least tempted to inform improves over public communication alone *for any prior belief* over evidence states.

The paper proceeds as follows: after describing the literature, I describe the model in Section 2 and describe an example in Section 3. I define (partially) unraveling information structures and their key property without communication in Section 4, and establish that they are communication-proof if they satisfy lower-rank uniqueness in Section 5. I establish that partially unraveling information structures satisfying lower-rank uniqueness are rich enough to solve the principal's problem in supermodular environments in Section 6, study linear environments in Section 7, describe simple discriminatory information structures in Section 8, and discuss the model's assumptions and limitations in Section 9.

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).[8] 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.[9]

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

[9] 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).

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 (2022), Camboni and Porcellacchia (2022), 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), Mathevet, Perego, and Taneva (2020), Ziegler (2020), Sandmann (2021), Li, Song, and Zhao (2022), Hoshino (2022), Morris, Oyama, and Takahashi (2022), and Inostroza and Pavan (2022).[10] 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. The stark difference between these solution concepts in the games considered here is illustrated in Section 3. I use a result in Morris, Oyama, and Takahashi (2022), 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

[10] 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).

communication in this paper.[11]

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 (2021) and Angelucci and Russo (2022). For instance, Pei and Strulovici (2021) 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.

2 Model

States. There is a finite set of states Θ, with arbitrary element denoted θ. The principal and agents share a full support 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 $a_i \in A \equiv \{w, n\}$.[12,13] 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.

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

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

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

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}.

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 $\mathcal{I}$.[14] Let $C(\mathcal{I})$ denote the set of communication equilibria given an information structure $\mathcal{I}$.

Principal. The principal chooses the information structure, $\mathcal{I}$, 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. The principal's problem is:

$$V^* \equiv \sup_{\mathcal{I}} \inf_{\sigma \in C(\mathcal{I})} \mathbb{E}_{\sigma, \mathcal{I}} \left(v(\mathbf{a}, \theta)\right)$$

I will also call this the principal's problem *with group communication.*[15]

It is useful to define another problem in which agents are *not* allowed to communicate. Let BNE($\mathcal{I}$) be the set of BNE in the game induced by information structure $\mathcal{I}$. Then, the

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

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

principal's problem *without group communication* is

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

An information structure $\mathcal{I}$ is called *communication-proof* if:

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

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, information design would be unnecessary.*

Microfoundation. To provide intuition for the model, consider the following microfoundation: $\Theta \subset (0,1)$ and θ, the evidence state, is the likelihood that the principal can prove and stop misbehavior *without* either agent informing.[16] Suppose that the principal can prove and stop misbehavior *with certainty* if at least one agent informs. The principal's value is the likelihood of proving and stopping misbehavior in each state.

Let $g_i > 0$ be agent i's *profit from misbehavior* when neither agent informs and the principal is unable to prove misbehavior. Let $\ell_i > 0$ be agent i's *punishment* if neither agent informs but the principal is able to prove misbehavior. Suppose that if agent $-i$ informs while i does not, then i receives the full punishment, ℓ_i, while $-i$ receives a reduced punishment $\underline{\ell}_{-i} < \ell_{-i}$. Further, suppose that if both agents inform, each agent i has a chance $p_i \in (0,1)$ to be punished at the reduced level, $\underline{\ell}_i$, and chance $1 - p_i$ to be punished at the full level, ℓ_i. Agent i's payoffs are then given by,

	n	w
n	$g_i(1-\theta) - \ell_i\theta$	$-\ell_i$
w	$-\underline{\ell}_i$	$-p_i\ell_i - (1-p_i)\underline{\ell_i}$

where i is the row player and $-i$ the column player.

Under this parameterization, the assumptions of the model are satisfied: the principal prefers (w,w) to (n,n) (and is indifferent as long as at least one agent informs), w is a strict

[16] This is a reduction of the private evidence the principal possesses to a single number, the probability of conviction given that evidence. Implicitly, I assume that any collection of evidence the principal has that leads to the same probability of conviction enters payoffs in the same way (and so can be combined into a single evidence state).

best-response to w for each agent i (Assumption 2), and whatever is i's action, i's payoff is strictly higher if $-i$ chooses n (Assumption 1). Observe that in this setting, preferences are *affine functions* of θ for both agents and the principal; I return to this special but natural case in Section 7.

3 Example

Consider the following setting. Nature flips two fair coins, coin 1 and coin 2. Each coin can come up heads (H) or tails (T), so $\Theta = \{HH, TT, HT, TH\}$. Payoffs are:

	n	w
n	x, x	$-2, 1$
w	$1, -2$	$0, 0$

where $x = -1$ if $\theta = HH$ and $x = 2$ otherwise. In Appendix D, I show that with public information design only, the principal cannot achieve first best. Below, I first demonstrate an information structure which does achieve first best in the principal's problem without group communication, but performs poorly with group communication. Then, I demonstrate a simple communication-proof information structure which achieves first best in the principal's problem with group communication.

3.1 A Private Information Structure

Consider the following information structure: the principal privately shows coin 1 to agent 1 and coin 2 to agent 2. Say that agent i has type H (T) if he observes that coin i comes up H (T).

Without communication. Without communication, the unique BNE is for both agents to choose w. Assuming that $-i$ chooses n, if i of type H chooses w then he receives payoff 1, while if he chooses n he receives $2(\frac{1}{2}) - 1(\frac{1}{2}) = \frac{1}{2}$; so, w is strictly preferred to n. All other distributions of $-i$'s action preserve this strict preference, so that an agent of type H always chooses w. For agent i of type T, the payoff to choosing n is 0, under the assumptions that $-i$ of type H chooses w and that $-i$ of type T chooses n. The payoff to choosing w under the same assumptions is $\frac{1}{2}$. As a result, w is strictly preferred over n for an agent of type T, given that $-i$ of type H chooses w and $-i$ of type T chooses n. Any other distribution of $-i$'s actions consistent with $-i$ of type H choosing w preserves this strict preference for w.

The unique BNE is therefore (w, w), and the principal's value under this information structure is $\frac{1}{4} \sum\limits_{\theta \in \Theta} v((w, w), \theta)$, the highest possible.

With communication. Consider now a communication mechanism, σ, that elicits messages $m_i \in \{H, T\}$ from each agent and recommends the action profile that maximizes the total payoff: if $m_i = H$ for each $i \in I$, the recommendation is w to both agents and otherwise the recommendation is n to both agents.

If agent i is type T, then reporting T guarantees the highest possible payoff (assuming $-i$ obeys), so truth-telling and obedience are satisfied. If i is type H, then conditional on truthfully reporting, obeying is optimal; obeying recommendation n leads to payoff 2 while disobeying leads to payoff 1, and obeying recommendation w leads to payoff 0 while disoebying leads to payoff -2. If i is type H and he misreports T, he receives recommendation n. If he obeys and chooses n, his payoff is $\frac{1}{2}(2) - \frac{1}{2}(1) = \frac{1}{2}$. If he disobeys and chooses w, his payoff is 1, so the only relevant deviation is the double deviation of misreporting T and choosing w (leading to payoff 1). But, by reporting truthfully and obeying, he receives $\frac{1}{2}(2) + \frac{1}{2}(0) = 1$, so truth-telling and obeying is a best-response for an agent of type H (to truth-telling and obedience by his partner).

In this case, the principal's payoff is

$$\frac{1}{4}v((w, w), HH) + \sum_{\theta \in \Theta \setminus \{HH\}} \frac{1}{4}v((n, n), \theta).$$

This is the same payoff the principal would achieve if she had simply revealed both coins to both agents, i.e., publicly communicated all of her information.

3.2 Communication-Proof Private Information Design

I now provide an information structure that implements the principal's first best outcome (both agents always choose w) in the unique communication equilibrium outcome.

Suppose that the principal shows agent 1 coin 1, but shows agent 2 nothing. Again, I refer to agent 1 as type H (T) if coin 1 comes up H (T). By the same argument given in Section 3.1, w is a strict best-response for agent 1 of type H, for any belief about agent 2's action. Consider agent 2's incentives. From the argument in Section 3.1, even if agent 2 could observe that coin 2 is T (the most optimistic he can be about coin 2), w is a strict best-response for any belief about agent 1's choice consistent with agent 1 of type H choosing

w. As a result, agent 2 must choose w in any BNE. Finally, since agent 2 chooses w with certainty, Assumption 2 implies agent 1 of type T must choose w. As a result, the unique BNE is (w, w).

This information structure is communication-proof. To see why, fix an arbitrary communication equilibrium. Without communication amongst agents, it is strictly dominant for agent 1 of type H to choose w, and so the only way agent 1 of type H would choose n is if his belief about coin 2 changed after a recommendation from the communication mechanism. But agent 2 has no information about coin 2, and so agent 1's belief about the state after observing the recommendation from the communication mechanism cannot change. As a result, to maintain obedience, the communication equilibrium must recommend w with probability 1 to agent 1 who reports H.

Observe next that if agent 2 is recommended to choose n with positive probability, it must be that the communication equilibrium recommends n to agent 2 *more often* after agent 1 reports H than T. To see why, recall that agents strictly prefer their partner choose n over w (Assumption 1). As a result, truth-telling by agent 1 of type H requires that agent 2 is recommended to choose n with (weakly) higher probability after agent 1 reports H than T, otherwise it would be profitable for agent 1 of type H to deviate to reporting T, inducing a greater likelihood that agent 2 chooses n. As a result, when agent 2 observes recommendation n, agent 2's belief that agent 1 is type H (weakly) *increases*. But since absent the recommendation it would be a strict best-response for agent 2 to choose w for any belief consistent with agent 1 of type H choosing w, the recommendation to agent 2 to choose n *cannot* be obedient. To maintain obedience then, the communication equilibrium must recommend w to agent 2 with probability 1.

Finally, agent 1 of type T knows that agent 2 is being recommended w with probability 1. To maintain obedience, agent 1 of type T must therefore receive recommendation w with probability 1. Any communication equilibrium therefore recommends w to all agents with probability 1.

Using the information structure described here, the principal is able to implement the first best using a communication-proof information structure (hence improving over public communication).

4 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 6, 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 5).

For any information structure (T, π), let π_i denote the marginal distribution of π along dimension i (i.e., the marginal distribution of agent i's type). Fix an information structure (T, π), an agent $i \in I$ and a type $t \in \operatorname{supp}(\pi_i) \subseteq T_i$. 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 $(t_{-i}, a_{-i}) \sim \beta$. In words, an interim strict-BR is a strict-BR if the solution concept for the agents' game were BNE rather than communication equilibrium.

Given an information structure (T, π), a *ranking function* is a pair of functions $\lambda = (\lambda_1, \lambda_2)$ with $\lambda_i : T_i \to \mathbb{N} \cup \{\infty\}$, such that $\lambda_i(t_i) \neq \lambda_{-i}(t_{-i})$ for any $(t_1, t_2) \in \operatorname{supp}(\pi)$ whenever $\max\limits_{i \in I}\{\lambda_i(t_i)\} < \infty$. Call $\lambda_i(t_i)$ the λ-*rank* of type t_i. Finally, for any $i \in I$ and type $t \in T_i$, let

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

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

Definition 1. *An information structure (T, π) is a* ***unraveling information structure*** *if there exists a ranking function λ such that $\lambda_i(T_i) \subseteq \mathbb{N}$ for each $i \in I$, and for any $i \in I$ and $t_i \in T_i$, w is an interim strict-BR for any belief $\beta \in \Phi_i(t_i, \lambda)$ about $-i$'s action.*

A function λ satisfying the conditions in the definition is called **admissible** with respect to (T, π). Note that admissible ranking functions cannot take value ∞—once I introduce *partially* unraveling information structures, the role of a λ-rank of ∞ will become clear. The key property of unraveling information structures is that 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*

- T_i *can be partitioned into disjoint sets* $\widetilde{T}_i$ *and* $\{\infty\}$
- *If* $\pi(\boldsymbol{t}) > 0$*, then* $\boldsymbol{t} \in \widetilde{T}$ *or* $\boldsymbol{t} = (\infty, \infty)$*, where* $\widetilde{T} = \widetilde{T}_1 \times \widetilde{T}_2$
- $(\widetilde{T}, \pi_{\widetilde{T}})$ *is an unraveling information structure given prior* $\mu_{\widetilde{T}}$ *over* Θ
- *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 second requirement states that there is 0 probability that one agent's type is ∞ while another agent's type is *not* ∞. The third requirement states that on $\widetilde{T}$, agents face an unraveling information structure. The fourth requirement is that if there is a positive probability that both types are ∞ (under π), 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 $i \in I$.[17]

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. Any ranking function λ that is an admissible ranking function when restricted to $(\widetilde{T}, \pi_{\widetilde{T}})$ and has $\lambda_i(\infty) = \infty$ for each i is called **admissible**.[18]

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* $t_i = \infty$*.*

[17] 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.

[18] Recycling the term admissible is convenient, though it represents a slight abuse of terminology.

5 Information Structures: With Communication

The information structure in the example in Section 3.1, in which the principal shows each agent one coin, is an unraveling information structure. For instance, let $\lambda_1(H) = 1, \lambda_2(H) = 2$, $\lambda_1(T) = 3$ and $\lambda_2(T) = 4$. Then, λ is an admissible ranking function. But, I have shown that communication ruins the principal's value. Observe that agent 1 with type H (λ-rank 1) faces *two* types of agent 2 with larger λ-ranks, both H (λ-rank 2) and T (λ-rank 4). Indeed, in the information structure of the two coin example in Section 3.1, any admissible ranking function λ will have this property for at least one agent. In this section, I show that unraveling information structures *are* communication-proof if they satisfy an additional restriction that rules out this possibility.

Recall that for any information structure (T, π), π_i denotes the marginal distribution of agent i's type. For any information structure (T, π), let $\pi^i_{t_i}$ denote the distribution of $-i$'s type, t_{-i}, conditional on i having type $t_i \in \operatorname{supp}(\pi_i) \subset T_i$.

Fix a ranking function λ and for any $t_i \in \operatorname{supp}(\pi_i)$, let

$$Lower_i(t_i, \lambda) \equiv \operatorname{supp}(\pi^i_{t_i}) \cap \left\{ t_{-i} \in T_{-i} \middle| \lambda_{-i}(t_{-i}) > \lambda_i(t_i) \right\}.$$

and

$$Upper_i(t_i, \lambda) \equiv \operatorname{supp}(\pi^i_{t_i}) \cap \left\{ t_{-i} \in T_{-i} \middle| \lambda_{-i}(t_{-i}) < \lambda_i(t_i) \right\}.$$

$Lower(t_i, \lambda)$ is the set of types t_{-i} with λ-rank *larger* than the λ-rank of t_i. $Upper(t_i, \lambda)$ is the set of types t_{-i} with λ-rank *smaller* than the λ-rank of t_i. When it is clear, I suppress the dependence of $Lower(t_i, \lambda)$ and $Upper(t_i, \lambda)$ on λ.

Definition 3. *Fix a partially unraveling information structure (T, π). An admissible ranking function λ is* ***lower-rank unique*** *if for any $t_i \in \operatorname{supp}(\pi_i)$,*

$$|Lower_i(t_i, \lambda)| \leq 1$$

and is ***upper-rank unique*** *if for any $t_i \in \operatorname{supp}(\pi_i)$,*

$$|Upper_i(t_i, \lambda)| \leq 1$$

and is ***rank unique*** *if it is both upper and lower-rank unique. The information structure (T, π) is lower-rank unique (upper-rank unique, rank unique) if there exists an admissible* *ranking function that is lower-rank unique (upper-rank unique, rank unique).*

An unraveling information structure that is both upper and lower-rank unique is a version of an electronic mail game information structure (Rubinstein, 1989): each type t_i believes that there is at most one type of $-i$ with a lower λ-rank and at most one type of $-i$ with a higher λ-rank. As I show in Proposition 1 below, lower-rank uniqueness is sufficient for communication-proofness of a partially unraveling information structure. In Section 6, partially unraveling information structures satisfying both upper and lower-rank uniqueness are shown to be sufficient to implement the principal's optimal value.[19]

I provide intuition for why lower-rank unique unraveling information structures are communication-proof—afterwards, I comment on how this intuition this can be easily applied to show the same for *partially* unraveling information structure. Fix any lower-rank unique unraveling information structure (T, π) and corresponding lower-rank unique admissible ranking function λ.

First, consider an agent i of type t_i with the smallest λ-rank, i.e., an agent for whom $\lambda_i(t_i) = \min\limits_{i \in I, t_i \in \text{supp}\,(T_i)} \{\lambda_i(t_i)\}$. The definition of an unraveling information structure implies that, if the group could not communicate, this type of agent chooses w for *any belief* about his partner's action. As a result, the only way to convince this type of agent not to choose w is to change his belief about θ. But under lower-rank uniqueness, he is *certain* of his partner's type, and as a result no message he receives in a communication equilibrium with positive probability can change his belief about θ. Hence, an agent with the smallest possible λ-rank must be recommended w with certainty in any communication equilibrium.

The argument for larger λ-ranks proceeds by induction. Suppose that in any communication equilibrium, agents with λ-rank strictly smaller than m are recommended w with certainty. Then, I claim there can be no communication equilibrium in which t_i with λ-rank m is recommended n with positive probability. To understand why, fix any t_i with λ-rank m and first observe that for any types $t_{-i} \in Upper_i(t_i)$, lower-rank uniqueness guarantees that $Lower_{-i}(t_{-i}) = \{t_i\}$. The inductive hypothesis implies that any type $t_{-i} \in Upper_i(t_i)$ believes t_i is the only type in $\text{supp}\,(\pi^{-i}_{t_{-i}})$ that may choose n in equilibrium. Assumption 1 implies that t_{-i} would deviate from truth-telling if an alternative report convinced t_i to choose n with higher probability. To ensure truth-telling then, all reports by $-i$ of types $t_{-i} \in Upper_i(t_i)$ must lead to *the same* probability that t_i chooses n. The key implica-

[19] However, it may be that in other settings, upper-rank uniqueness is restrictive, and the principal may prefer to use information structures in which agent i believes there are multiple possible types of $-i$ with smaller λ-rank. For instance, when the principal is able to make allocations, or if the agents can side-contract, it may be optimal to violate upper-rank uniqueness.

tion of this is that upon observing recommendation n, i's belief about t_{-i} conditional on $t_{-i} \in Upper_i(t_i)$ is *unchanged.*

Similarly, observe that by lower-rank uniqueness, $Lower(t_i)$ has at most one element, say $\underline{t}_{-i}$.[20] Any type $t_{-i} \in Upper_i(t_i)$ can misreport $\underline{t}_{-i}$ and so, to ensure truth-telling of types $t_{-i} \in Upper_i(t_i)$, reporting any $t_{-i} \in Upper_i(t_i)$ must lead to a weakly *higher* likelihood that t_i chooses n than reporting $\underline{t}_{-i}$. The key implication of this is that upon observing recommendation n, i's belief that $t_{-i} = \underline{t}_{-i}$ must *decrease* or, equivalently, i's belief that $t_{-i} \in Upper_i(t_i)$ must *increase.*

From the above observations, I argue that no communication equilibrium can recommend n to type t_i with positive probability. Suppose towards contradiction that t_i is recommended n with positive probability. The definition of an unraveling information structure and the inductive hypothesis imply that, if i's belief about $-i$'s type does not change after observing recommendation n, only w is an obedient recommendation to t_i. Then to convince t_i to obey recommendation n, one of the following must happen to i after observing the recommendation: (a) his belief that $t_{-i} \in Upper_i(t_i)$ (and hence that $-i$ necessarily chooses w by the inductive hypothesis) strictly decreases, or (b) his belief about θ conditional on $t_{-i} \in Upper_i(t_i)$ must change in a way to make w less attractive. But as I have argued, after observing recommendation n, (a) t_i's belief that $t_{-i} \in Upper_i(t_i)$ must weakly *increase*, and (b) t_i's belief about t_{-i} conditional on $t_{-i} \in Upper_i(t_i)$ *must not change*, and hence t_i's belief about θ conditional on $t_{-i} \in Upper_i(t_i)$ must not change. As a result, t_i can only become *more* pessimistic about n after receiving recommendation n, relative to the interim stage (i.e., after learning type but before receiving recommendation n), and so violates obedience by choosing w. Hence, no communication equilibrium can involve a recommendation to any type of any agent to choose n with positive probability.

For a *partially unraveling information structure* (T, π), if $t_i \neq \infty$ for each i, then the intuition for communication-proofness is exactly the same as for unraveling information structures, and if $\boldsymbol{t} = (\infty, \infty)$, then both agents choose (n, n) in the principal's worst BNE and communication equilibrium. By the definition of a partially unraveling information structure, these are the only two possibilities.

Proposition 1. *Any lower-rank unique partially unraveling information structure, (T, π), is communication-proof, and the principal's worst BNE outcome is identical to the principal's worst communication equilibrium outcome.*

[20] If $Lower(t_i, \lambda)$ is empty, the result is immediate: the inductive hypothesis implies that $-i$ chooses w with certainty, so t_i must be recommended w with certainty.

Note that if (T, π) is an unraveling information structure, the result implies that (w, w) is the unique and hence 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.*

6 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 rank unique partially unraveling information structures.

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.[21]

Definition 4. *A set of information structures, B,* ***implements*** *V^* if*

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

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

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

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 rank unique partially unraveling information structures implements V^* and V^0.

[21] 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.

The proof is given in the appendix and follows by combining Morris et al. (2022)'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. It is useful here to recall the characterization in Morris et al. (2022)'s Theorem 1. To facilitate the explanation, suppose there is some state, $\hat{\theta}$, such that informing is strictly dominant for both agents at $\hat{\theta}$, and in no other state is informing strictly dominant. A few preliminary definitions are required.

A *outcome* is an element $\nu \in \Delta(A^I \times \Theta)$, and is *consistent* if $\sum_{\mathbf{a} \in A^I} \nu(\mathbf{a}, \theta) = \mu(\theta)$ for each $\theta \in \Theta$, i.e., it has marginal over states equal to the prior over states. An outcome ν is *implemented* by an information structure if it is the principal's worst BNE under that information structure. Define $\Gamma \equiv \{\emptyset, 1, 2, 12, 21\}$ and $\Gamma_i \equiv \Gamma \backslash \{-i, \emptyset\}$. Think of each $\gamma \in \Gamma$ as representing the set of agents who will choose w, and the order of those agents in γ as a ranking; if i comes before j in $\gamma \in \Gamma$, then i is said to be ranked higher than j. For any $\gamma \in \Gamma$, define

$$a_{-i}(\gamma) \equiv \begin{cases} w & \text{if } -i \text{ is ranked higher than } i \text{ in } \gamma \text{ (or } i \text{ is not in } \gamma \text{ while } -i \text{ is)} \\ n & \text{otherwise} \end{cases}$$

An *ordered outcome* is a distribution $\nu_\Gamma \in \Delta(\Gamma \times \Theta)$, and satisfies *sequential obedience* if

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

for any i with $v_\Gamma(\Gamma_i \times \Theta) > 0$. An outcome ν is n-obedient if

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

Finally, an ordered outcome ν_Γ *implements* outcome ν if

$$\begin{aligned} \nu((w,w),\theta) &= \nu_\Gamma(12,\theta) + \nu_\Gamma(21,\theta), & \nu((n,n),\theta) &= \nu_\Gamma(\emptyset,\theta) \\ \nu((w,n),\theta) &= \nu_\Gamma(1,\theta), & \nu((n,w),\theta) &= \nu_\Gamma(2,\theta) \end{aligned}$$

Consider the set of outcomes ν such that $\nu((w,w),\hat{\theta}) > 0$. Then, under the assumptions I have made, an immediate implication of Morris et al. (2022)'s Theorem 1 is that an outcome ν is implementable as the principal's worst BNE for some information structure if and only if

ν is n-obedient, consistent, and there exists a sequentially obedient ordered outcome ν_Γ that implements ν. I describe now an information structure that can be used to generate such an outcome ν as the principal's worst BNE. Fix a sequentially obedient ordered outcome ν_Γ that implements ν. Since $\nu((w,w),\hat{\theta}) > 0$, $\nu_\Gamma(12,\hat{\theta}) + \nu_\Gamma(21,\hat{\theta}) > 0$. Suppose then that $\nu_\Gamma(12,\hat{\theta}) > 0$. Set $T_i = \{0,1,2,...\} \cup \{\infty\}$. Fix $\eta, \epsilon > 0$ s.t. $\eta << \epsilon < \nu_\Gamma(12,\hat{\theta})$. Set

$$\pi(\boldsymbol{t},\theta) = \begin{cases} \eta(1-\eta)^{m-1}\left(\nu_\Gamma(12,\theta) - \mathbf{1}_{\theta=\hat{\theta}}\epsilon\right) & \text{if } \boldsymbol{t} = (n, n+1) \text{ and } m \geq 1 \\ \eta(1-\eta)^{m-1}\nu_\Gamma(21,\theta) & \text{if } \boldsymbol{t} = (n+1, n) \text{ and } m \geq 1 \\ \eta(1-\eta)^{m-1}\nu_\Gamma(1,\theta) & \text{if } \boldsymbol{t} = (n, \infty) \text{ and } m \geq 1 \\ \eta(1-\eta)^{m-1}\nu_\Gamma(2,\theta) & \text{if } \boldsymbol{t} = (\infty, n) \text{ and } m \geq 1 \\ \nu_\Gamma(\emptyset,\theta) & \text{if } \boldsymbol{t} = (\infty, \infty) \\ \mathbf{1}_{\theta=\hat{\theta}}\frac{\epsilon}{2} & \text{if } \boldsymbol{t} = (0,1) \\ \mathbf{1}_{\theta=\hat{\theta}}\frac{\epsilon}{2} & \text{if } \boldsymbol{t} = (1,0) \end{cases}$$

For ϵ and η sufficiently small, sequential obedience ensures that any type $t_i < \infty$ chooses w in any equilibrium. To see why, observe that type $t_i = 0$ knows $\theta = \hat{\theta}$ and hence it is strictly dominant to inform. For ϵ sufficiently large relative to η, type $t_i = 1$ believes it is very likely that $\theta = \hat{\theta}$ and so it is also strictly dominant to inform. For type $t_i = 2$ and for η, ϵ sufficiently small, sequential obedience ensures that it is optimal to inform as long as $t_i = 1$ informs, which he does. Proceeding inductively through the list of finite types shows that in any BNE, an agent with a finite type chooses to inform. Finally, n-obedience guarantees that in the principal's worst equilibrium, agents with type ∞ choose not to inform.

This information structure is not, in general, a partially unraveling information structure. If $t_{-i} < \infty$ but $t_i = \infty$ with positive probability (i.e., if $\nu((w,n),\theta) > 0$ or $\nu((n,w),\theta) > 0$), the principal's worst BNE involves one agent choosing w while the other chooses n, a property absent from partially unraveling information structures. However, a consequence of Assumption 2 and the assumption that the principal prefers more informing is that the principal's value can always be (weakly) improved by modifying the information structure to reveal to type $t_i = \infty$ that $t_{-i} < \infty$ (and therefore $-i$ chooses w), which (all else equal) incentivizes i to choose w as well. This additional revelation does not affect the actions of any other agents in the resulting worst equilibrium, and does not harm the principal's value. As a result, it is possible to restrict to outcomes ν such that $\nu((w,n),\theta) = \nu((n,w),\theta) = 0$. The information structures implementing these outcomes are partially unraveling information structure satisfying lower-rank uniqueness and hence, are communication-proof by Proposition 1.[22]

Remark 3 (Linear Programming Formulation)**.** *A useful implication of this characterization is that the principal's problem can be stated as a linear program: the principal chooses any ordered outcome that satisfies sequential obedience (with weak inequality) and implements a consistent outcome ν (n-obedience can be dropped, as it is necessarily satisfied at optimality), and the principal's objective for any ordered outcome is $\mathbb{E}_\nu\left(v(\boldsymbol{a}, \theta)\right)$. This linear program is applied to the study of linear environments in Section 7.*

7 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. (2022). 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 characterization of an optimal policy below is proved by examining the dual of the linear program and applying complementary slackness. The proof is given in Appendix B.

Recall, 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) $V^* = \mathbb{E}_\nu\left(v(\mathbf{a}, \theta)\right)$ 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 $\underline{\theta} \equiv \min\{\Theta\}$.[23] Finally, say that an action $a \in A$ is *dominant at* $\theta \in \Theta$ if

$$u_i(a, a_{-i}, \theta) \geq u_i(a', a_{-i}, \theta)$$

[22] They are also upper-rank unique and hence, rank unique.

[23] Both exist because Θ is finite.

for any $a', a_{-i} \in A$.

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*

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

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. Indeed, the microfoundation of the model offered in Section 2 satisfies these conditions. 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.

7.1 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 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., $\sum\limits_{i \in I} u_i(\mathbf{a}, \theta; \mathcal{G}) = \sum\limits_{i \in I} u_i(\mathbf{a}, \theta; \mathcal{G}^{\epsilon,\epsilon})$, 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 using the microfoundation in Section 2, in particular 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, using the microfoundation in Section 2. Recall that in the microfoundation described in Section 2, θ is interpreted as the probability that the principal can prove misbehavior *without* an informant, and $u_i((n, n), \theta) = g_i \times (1 - \theta) - \ell_i \times (\theta)$, where 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 other's markets.[24] 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 in principle re-allocate buyers or geographies in such a way 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

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

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.

8 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 propose a simple information structure that only requires private communication by fully disclosing the state or disclosing nothing.

Before defining the simple information structures I will investigate, I formally define a public information structure. An information structure (T, π) is called *public* if whenever $\pi(t_i, t_{-i}) > 0$ for some pair $(t_i, t_{-i}) \in T$, $\pi(t_i, t'_{-i}) = 0$ for any $t'_{-i} \neq t_{-i}$. Any pair (P, π^P) with P countable and $\pi^P \in \Delta(P \times \Theta)$ induces a public information structure given by (T, π) with $T = P \times P$ and $\pi((p, p'), \theta) = \mathbf{1}_{p=p'} \pi^P(p, \theta)$. As a result, I refer to any such pair (P, π^P) as a public information structure.

Let $\textsc{Dom}_i \subset \Theta$, be the set of states θ such that w is strictly dominant for i in state θ.

Definition 5. *An information structure* (T, π) *is called* **simple discriminatory** *if*

$$T_i = P \times \left(\{\emptyset\} \cup \overline{\Theta}_i\right)$$

for some set P*, and there exists a function*

$$i^*(.) : P \to I$$

called the favored agent *function, s.t. for any* $\boldsymbol{t} = ((p, \theta_1), (p, \theta_2)) \in T$*,* $\pi(\boldsymbol{t}) > 0$ *implies*

$$\begin{aligned} \theta_{-i^*(p)} &= \emptyset, \\ \theta \notin \overline{\Theta}_{i^*(p)} \implies \theta_{i^*(p)} &= \emptyset, \\ \theta \in \overline{\Theta}_{i^*(p)} \implies \theta_{i^*(p)} &= \theta \end{aligned}$$

and for any $\boldsymbol{t} = ((p_1, \theta_1), (p_2, \theta_2)) \in T$ *with* $p_1 \neq p_2$*,* $\pi(\boldsymbol{t}) = 0$*. The pair* (P, π^P) *with*

$\pi^P(p,\theta) = \sum_{t \in T} \pi((p,t_1),(p,t_2),\theta)$ *for each* $p \in P$ *is called the* public signal structure of (T,π)*, and for each type* $t_i = (p,\theta)$*,* p *is called the* public signal *of* t_i*.*

Each simple discriminatory information structure can be represented as the triple (P, π^P, i^*) consisting of public signal space P, distribution over public signal space π^P, and favored agent mapping i^*. Denote the set of all such triples by $\mathcal{D}$.

A simple discriminatory information structure can be interpreted as the following two-step procedure: the principal sends a public signal from P, and then privately reveals the state to *one agent*, $i^*(p)$, if it is a dominant strategy to inform in that state, and otherwise reveals nothing.

Definition 6. *A simple discriminatory information structure* (P, π^P, i^*) ***targets the agents least (most)*** **tempted** ***to inform after public signal*** $p \in P$ *if,*

$$\mathbb{E}_p\left(u_{i^*(p)}(w,n,\theta) - u_{i^*(p)}(n,n,\theta)\right) > 0$$
$$\implies (\impliedby)\ \mathbb{E}_p\left(u_{-i^*(p)}(w,n,\theta) - u_{-i^*(p)}(n,n,\theta)\right) > 0.$$

where $\mathbb{E}_p(\cdot)$ *is the conditional expectation given public signal* p*. Denote by* $\textsc{least}(p; P, \pi^P)$ *the set of agents who are least tempted inform after* $p \in P$*.*

A simple discriminatory information structure (P, π^P, i^*) ***targets the agent least (most) tempted to inform*** *if* $i^*(p) \in \textsc{least}(p; P, \pi^P) \quad \forall p \in P$*. Denote by* $\textsc{least}(P, \pi^P)$ *the set of favored agent functions* i^* *with* $i^*(p) \in \textsc{least}(p; P, \pi^P)$ *for each* $p \in P$*.*

When clear, I suppress the dependence of $\textsc{least}(p; P, \pi^P)$ on (P, π^P). If a simple discriminatory information structure is an unraveling information structure, then it is lower-rank unique and hence, is communication-proof. If it is not an unraveling information structure, it must be because in the principal's worst BNE, there is some $p \in P$ such that any type who observes p but does not observe the private signal revealing the state chooses n. In that case, the principal's worst communication equilibrium involves exactly the same outcome, and hence, is communication-proof. A key benefit of simple discriminatory information structures is that they require relatively simple private communication from the principal: if evidence is verifiable, the principal need only reveal evidence.

Observe now that for any simple discriminatory information structure, if $i^*(p) = i$ then the principal's worst BNE given public signal p is the same for any simple discriminatory information structure in which $i^*(p) = i$. As a result, for any public signal (P, π^P), it is possible to define equilibrium objects after observation of signal realization p given a favored agent mapping evaluated at p without any reference to the favored agent mapping elsewhere.

For any simple discriminatory information structure (P, π^P, i^*), let

$$P^L(P, \pi^P, i^*) \equiv \left\{ p \in P \middle| \text{in any BNE, any type with public signal } p \text{ chooses } w \right\}$$

$$I^L(p; P, \pi^P) \equiv \left\{ i \in I \middle| \exists \left(P, \pi^P, i^*\right) \in \mathcal{D} \text{ such that } i^*(p) = i \text{ and } p \in P^L(P, \pi^P, i^*) \right\}$$

Thus, $P^L(P, \pi^P, i^*)$ is the set of public signals after which (w, w) is the unique equilibrium outcome in the simple discriminatory information structure (P, π^P, i^*). $I^L(p, P, \pi^P)$ is the set of all agents i such that there exists some simple discriminatory information structure in which $i^*(p) = i$ and (w, w) is the unique equilibrium outcome after public signal p.

Define the principal's *value for an information structure* $\mathcal{I}$ as

$$V(\mathcal{I}) \equiv \inf_{\sigma \in C(\mathcal{I})} \mathbb{E}_{\sigma, \mathcal{I}} \left(v(\mathbf{a}, \theta) \right).$$

Let $\mathbf{a}^i \in A^I$ be defined by $a^i_i = w$ and $a^i_{-i} = n$.

Proposition 6. *Fix any public information structure (P, π^P) and suppose that agents' payoffs are supermodular. There exists an optimal simple discriminatory information structure (P, π^P, i^*) satisfying:*

1. *If $I^L(p; P, \pi^p) \neq \emptyset, \quad i^*(p) \in I^L(p; P, \pi^p)$*
2. *If $I^L(p; P, \pi^p) = \emptyset, \quad i^*(p) = \arg\max_{i \in I} \left\{ \sum_{\theta \in \overline{\Theta}_i} \left(v(\boldsymbol{a}^i, \theta) - v((n, n), \theta) \right) \mathbb{P}(\theta | p) \right\}$ where $\mathbb{P}(\theta | p)$ is the probability of state θ conditional on public signal p*
3. *The principal's value is (weakly) larger under simple discriminatory information structure (P, π^p, i^*) than under public information structure (P, π^P).*

The existence of an optimal simple discriminatory information structure with public signal (P, π^P) is immediate, as there are only finitely many. The existence of an optimal simple discriminatory information structure with public signal (P, π^P) satisfying the first two requirements is straightforward. If after a public signal p, privately informing some agent leads both agents to choose w, then set $i^*(p)$ to be any such agent (point 1). Otherwise, set $i^*(p)$ such that after public signal p, if only the type who observes the state chooses w, the principal's value at that public signal is maximized. The final part of the proposition follows from the next proposition, where I show that there always exists a simple discriminatory information structure that improves over public information structure (P, π^P).

Proposition 7. *Fix a public information structure* (P, π^p) *and suppose that agents' payoffs are supermodular. Then, any simple discriminatory information structure* (P, π^p, i^*) *with* $i^* \in \textsc{least}(P, \pi^p)$ *(weakly) improves the principal's value over public information structure* (P, π^p)*.*

The result is not immediate since it may be that simple discriminatory information structures involve strictly more information: in principle, this could lead to lower payoffs. Indeed, choosing $i^*(\cdot)$ poorly can lead to strictly lower payoffs for the principal.

The intuition behind the proposition is the following. The principal's worst equilibrium after a public signal (and without any further private signals) is in pure strategies, and is either (w, w) or (n, n), a consequence of Assumptions 2 and 3. If after some public signal p, the principal's worst equilibrium is (n, n), then this is the worst possible outcome, so no further private information can make matters worse for the principal.

If after some public signal p, the principal's worst equilibrium is (w, w), then there must be some agent, say i, who strictly prefers w even if $-i$ chooses n, otherwise (n, n) would be an equilibrium. Then, $\textsc{least}(p; P, \pi^P)$ is non-empty, since it at least contains $-i$. Then, set $i^*(p) = -i$: agent i is left completely uninformed, and, since the information structure is communication-proof, i still chooses w (a consequence of supermodularity), while Assumption 2 implies that $-i$ chooses w as well. Hence, the principal is made no worse off by the private revelation.

Robustness to μ.

Define a *public experiment* to be a pair $(P, (\pi_\theta)_{\theta\in\Theta})$ composed of a public signal space P and a distribution over P for each $\theta \in \Theta$. The public information structure *generated* by a public experiment $(P, (\pi_\theta)_{\theta\in\Theta})$ and a prior μ is the pair (P, π^P), where $\pi^P(p, \theta) = \mu(\theta) \times \pi_\theta(p)$. Given a favored agent function, i^*, and a public experiment $(P, (\pi_\theta)_{\theta\in\Theta})$, define a *simple discriminatory experiment* as the triple $(P, (\pi_\theta)_{\theta\in\Theta}, i^*)$. The simple discriminatory information structure *generated* by a simple discriminatory experiment $(P, (\pi_\theta)_{\theta\in\Theta}, i^*)$ and prior μ is the triple (P, π^P, i^*), where $\pi^P(p, \theta) = \mu(\theta) \times \pi_\theta(p)$. Given a prior μ, define the principal's value for a public (simple discriminatory) experiment to be the principal's value for the public (simple discriminatory) information structure generated by the prior.

Say that agent i is *unambiguously least tempted* if for all $\theta \in \Theta$,

$$u_i(w, n, \theta) - u_i(n, n, \theta) \leq u_{-i}(w, n, \theta) - u_{-i}(n, n, \theta).$$

Then observe that i is least tempted to inform after any public signal in any public information structure, i.e., $i \in \textsc{least}(p; P, \pi^P)$ for any (P, π^P) and $p \in P$. An implication of Proposition 7 is the following proposition.

Proposition 8. *Suppose agents' payoffs are supermodular. Fix any public experiment* $(P, (\pi_\theta)_{\theta\in\Theta})$ *and suppose that* $i \in I$ *is unambiguously least tempted. The simple discriminatory experiment* $(P, (\pi_\theta)_{\theta\in\Theta}, i^*)$ *with* $i^*(p) = i$ *for each* $p \in P$ *(weakly) improves the principal's value over public experiment* $(P, (\pi_\theta)_{\theta\in\Theta})$ ***for any prior*** μ.

The result states that if some agent can be identified as having smaller incentives to inform relative to his partner in each state, then any public experiment with any prior can be improved by the simple discriminatory experiment that targets that agent.

8.1 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 discriminatory information structure, the regulator needs to *commit* to two things: a *public* signal and releasing *private* information to only one agent.

On the public signal, the problem is that, once the regulator commits to a signal that informs cartel members that its evidence is strong, it will tempted to send that signal excessively. Sauvagnat (2015) discusses microfoundations for commitment of the regulator to an informative public signal. In particular, initiating an investigation is costly, and so serves as a potentially credible signal of the regulator's evidence—the investigation would be initiated, and cost incurred, only if the regulator has sufficiently strong evidence or is sufficiently confident of the existence of a cartel. Transparency can facilitate the signaling value of initiating an investigation if the regulator is forced to publish statistics about the share of initiated investigations that result in successful prosecutions.

On private information, it may be that after some public signals, revealing the state to one agent (say, agent 1) fails 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 (beyond the public signal), she may be tempted to reveal to player 2 that $\theta \in \overline{\Theta}_1 \cup \overline{\Theta}_2$, which may spur player 2 to report when he otherwise may not have.[25] If player 2 anticipates this, interpreting no signal from the regulator as indication that $\theta \notin \overline{\Theta}_1 \cup \overline{\Theta}_2$, then he and player 1 may be able to avoid informing when $\theta \notin \overline{\Theta}_1 \cup \overline{\Theta}_2$.

If the regulator's communications to agents are observable to the courts, then one way to create commitment is by creating a regulatory rule that for some fixed amount of time after the regulator has communicated its evidence to agent 1, it is prohibited from communicating with agent 2. If courts reduce fines when the regulator violates this rule, then it seems possible for the regulator to make this commitment, especially if courts are observed to enforce such reductions in fines when the regulator does violate the rule.

Another way to create some commitment is to implement a "first-in" rule that only the first agent to inform is granted leniency, and thus committing the regulator to only extract evidence voluntarily from the first reporting agent.[26] In that case, if $\theta \in \overline{\Theta}_1$, 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 $\theta \in \overline{\Theta}_2 \cap \overline{\Theta}_1^C$, 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.[27]

For reasons outside the model—for instance, the regulator is unsure if $\theta \in \overline{\Theta}_1$—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. For this, the regulatory rule prohibiting communication with agent 2 after communication with agent 1 for some fixed amount of time may be useful as well, to

[25] Recall $\overline{\Theta}_i$ is the subset of states Θ for which w is strictly dominant for agent i.

[26] 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.

[27] 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.

convince agent 2 that by the time the regulator communicates with him, agent 1 will have had enough time to act.

Choosing i^*. To generate improving simple discriminatory information structures given a public signal p, one needs to choose $i^*(p)$ carefully. If (w, w) is the principal's worst outcome at signal p, there is guaranteed to be an agent, i, for whom w is a strict best-response even if $-i$ chooses n. Setting $i^*(p) = -i$ preserves the principal's worst case outcome as (w, w). But how can one identify $i^*(p)$, without fine payoff information? One possibility 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 or informed on 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, leaving the recently shrunk firm (privately) uninformed appears to be a good choice.

8.2 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 discriminatory 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 may 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 prosecute 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.

To implement a simple discriminatory information structure, 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 discriminatory 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.[28] 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 discriminatory 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).

9 Discussion

Principal-Worst Equilibrium. Instead of evaluating the principal's payoff against her worst communication equilibrium, one could evaluate against the principal's worst communication equilibrium among the set of Pareto efficient equilibria for the agents. This does not change results: partially unraveling information structures induce a principal worst equilibrium that is efficient for agents (among equilibria).

Supermodularity and Rationalizability. When the game agents play is supermodular, the principal's worst rationalizable outcome coincides with the principal's worst BNE (Milgrom and Roberts (1990)). This result is applied in Morris et al. (2022), which I combine with Proposition 1 to prove Proposition 2 in this paper.

In contrast, if the game is not supermodular, it is possible that the principal's worst rationalizable outcome and BNE do not coincide. An example in which the principal's worst rationalizable outcome and BNE do not coincide is given in Appendix E.[29]

[28] 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).

[29] Of course, it is easy to generate such examples without satisfying the payoff assumptions, i.e., Assumptions 1 and 2. The example in the appendix provides an example that satisfies the payoff assumptions.

Bounded Rationality. A benefit of the simple discriminatory information structures considered in Section 8 is that they require relatively little sophisticated reasoning from the agents to implement, unlike (in general) the information structures required to achieve the principal's optimal value. The motivating environments involve games between agents in which the gains and losses from informing or being informed on are experienced infrequently, and the sophisticated reasoning required to reach the outcomes generated by unraveling information structures through introspection in theory may be too demanding in practice. Experimental evidence suggests, at least in static settings, that higher-order reasoning will be severely limited—a potential problem even with the basic leniency policy in the economics literature, which draws value from an unraveling argument (Motchenkova, 2004). Experience may lead to more higher-order reasoning, but in the settings motivating this paper—especially in the case of antitrust—many agents will be first time offenders. On the other hand, the high stakes involved may strain the external validity of lab experiments, and cartelists may engage in more higher-order.

10 Conclusion

In this paper, I studied the problem of a principal facing a group of misbehaving agents. Each member of the group could inform on the group's misbehavior, and the principal could reveal information to the group in order to spur whistleblowing. Novel to this setting, I allowed the group of misbehaving agents to communicate amongst themselves, after the principal communicates with them, and showed that such communication is *not harmful* to the principal's value in supermodular environment. I studied linear environments, and showed that under a dominance condition, there exists a solution to the principal's problem in which the likelihood of informing is monotonically increasing in the evidence state. I then provided a simple information structure, requiring only public communication and full disclosure of the evidence, that improves over public communication alone, and argued for the feasibility of its implementation in practice.

A number of questions, on both theoretical and practical matters, have been left open. First, what happens with more than two misbehaving agents? Partially unraveling information structures naturally generalize, but their communication-proofness does not; except in special cases, the techniques used to prove communication-proofness for two agents does not generalize to more than two agents. In that case, does communication-proofness place real restrictions on the principal's value, or are there alternative information structures that can still implement the principal's value without group communication? Second, what are optimal information structures when agents can side-contract? It would be natural to allow agents to re-balance the gains from misbehavior in order to make the group less susceptible to the principal's information revelation. In this case, information structures that are communication-proof without side-contracting may no longer be with side-contracting. Third, what happens if the principal can jointly design the payoff structure of the game as well as the information agents have? With sufficiently generous rewards, any cartel can be dismantled, but these rewards are costly, so a principal must balance the loss from rewards and the operation of the cartel. Fourth, a natural extension would allow for more than two actions—for instance, agents may be able to take action to further conceal their misbehavior. In this case the regulator must take care not to inform agents if it will lead them not to whistleblowing but to concealing their misbehavior. Finally, what happens if evidence arrives slowly over time, as would be the case when investigating a cartel? Does the principal slowly reveal information, or gather evidence for a long period of time before revealing, and how are these questions affected by the principal's ability to commit?

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 (2022): “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. (2022): “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 (2022): "Adversarial Coordination and Public Information Design," *Working Paper*.

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.

Milgrom, P. and J. Roberts (1990): "Rationalizability, learning, and equilibrium in games with strategic complementarities," *Econometrica: Journal of the Econometric Society*, 1255–1277.

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 (2022): "Implementation via information design in binary-action supermodular games," *Available at SSRN 3697335*.

Motchenkova, E. (2004): "Effects of leniency programs on cartel stability," .

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 (2021): "Crime Aggregation, Deterrence, and Witness Credibility," *arXiv preprint arXiv:2009.06470.*

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 5

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 a lower-rank unique unraveling information structure, (T, π) and let $\lambda = (\lambda_i)_{i \in I}$ be an admissible ranking function satisfying lower-rank uniqueness. 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 λ-rank $\in \{0, 1, ...\}$.

Base Case: If $\mathbb{P}_{\pi}(\lambda_i(t_i) = 0) = 0$ for each i and t_i, then there is nothing to show. Otherwise, fix any $i \in I$ and $t_i \in T_i$ such that $\lambda(t_i) = 0$ 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 lower-rank uniqueness, $|Lower_i(t_i)| = 1$.[30] 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 $\lambda(m_i) < k$. I will prove the statement for any $\mathbf{m}$ such that $\lambda(m_i) = k$. To this end, fix any type profile $\mathbf{t}$ such that $\lambda(t_i) = n$ such that $\pi(\mathbf{t}) > 0$. If no such profile exists, we are done. Otherwise, let $t \equiv t_i$. By lower-rank uniqueness, $|Lower_i(t)| \in \{0, 1\}$.

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

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 $\mathbb{P}_\pi(\lambda(t_i) = \lambda(t_{-i})) = 0$.

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 $n, n' \in Upper_i(t)$,*

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

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:

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

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

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

Then, since $u_{-i}(w, n, \theta) > u_i(w, w, \theta)$ by Assumption 1, the expression is maximized by reporting $\hat{m} \in \mathit{Upper}_i(t) \cup \mathit{Lower}_i(t)$ 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 $t_{-i} \in \mathit{Lower}_i$ must weakly decrease, (ii) conditional on $t_{-i} \in \mathit{Upper}_i$, i's belief about θ is unchanged related to her interim belief and hence, her expected payoff from any action profile conditional on $t_{-i} \in \mathit{Upper}_i$ 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:

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

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:

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

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 Sections 6

Before proving results in Sections 6 and 7, it is necessary to define a number of preliminaries in order to modify results in Morris et al. (2022). Some of these are repeats of definitions offered in Section 6.

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 $\sum\limits_{\mathbf{a}\in A^I} \nu(\mathbf{a},\theta) = \mu(\theta)$. 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[31]

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

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. (2022), so that only asymmetric "grain of dominance" (defined in Morris et al. (2022), 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.

[31] Note that S-implementability corresponds to "smallest" equilibrium implementation in Morris et al. (2022). Labeling w as 1 and n as 0, principal-worst equilibrium is the same as S-implementable.

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. (2022)'s Theorem 1(2): Since payoffs are supermodular for agents, Morris et al. (2022)'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.

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. (2022), 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. (2022): 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. (2022) 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. (2022) 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. (2022), that π is consistent. I state a modified version of the claim A.1.

Modified Claim A.1 (from Morris et al. (2022)'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. (2022)'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. (2022)'s Theorem 1(2) proof). *For any $i \in I$, any $\tau \in \{2, 3, ...\}$, and any $S \subset I \backslash i$,*

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

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

Claim A.3 (from Morris et al. (2022)'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. (2022) 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 with ranking function $\lambda_i(t_i) = t_i$ (a consequence of Claim A.1, and the argument after Claim A.3 concluding the proof in Morris et al. (2022)).

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 $\nexists\ \overline{\Theta}$ 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 maximum 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 7: Fix any public information structure and denote by s^p the public signal. For any public signal s^p, 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). Can a mixed strategy equilibrium 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 $\mathbb{E}^0(.) \equiv \mathbb{E}(.|s^p)$, 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, given s^p. But then (n, n) is an equilibrium.

Therefore, given any public signal 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 to the public signal s^p when worst-equilibrium behavior is (w,w) or (n,n).

If under s^p, both players choose n in the principal's worst equilibrium, the introduction of private information cannot lower the principal's value. Suppose instead s^p 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.[32] Then, under the simple discriminatory information structure for agent 1, agent 2 still chooses w by supermodularity. But then, player 1 chooses w by Assumption 2.

To see that a simple discriminatory information structure with public signal space P and favored agent mapping i^* is communication-proof, fix some public signal realization s^p. In any communication equilibrium, if $i^*(s^p)$ observes $\theta \in \overline{\Theta}_{i^*(s^p)}$ (i.e., dominant to choose w) then he must receive recommendation w with probability 1. Let $i^* \equiv i^*(s^p)$. Denote by t_i^j the type of agent $i \in I$ who observes signal $j \in \overline{\Theta}_i \cup \{\emptyset\}$. There are then two cases to

[32] 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.

consider:

Case 1: Suppose that both agents choose w with probability 1 in the principal's worst BNE. Then, conditional on s^p, the information structure is an unraveling information structure satisfying lower-rank uniqueness: and hence, by Proposition 1, is communication-proof.

Case 2: Suppose that the principal's worst BNE involves $t^{\emptyset}_{-i^*}$ and $t^{\emptyset}_{i^*}$ choosing n with positive probability. Then, by supermodularity, the principal's worst BNE is $t^{\theta}_{i^*}$ choosing w for any $\theta \in \overline{\Theta}_{i^*}$ and $t^{\emptyset}_{i}$ choosing n for each $i \in I$. Then, the only way a communication equilibrium can lower the principal's value is if $t^{\theta}_{i^*}$ chooses n with positive probability for some $\theta \in \overline{\Theta}_{i^*}$, 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 discriminatory information structures are communication-proof. □

C Proofs of Section 7

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., $\nu((w,n),\theta) = \nu((n,w),\theta) = 0$.[33] As in Morris et al. (2022), the characterization of S-implementable outcomes in the proof of Proposition 2, implies that the principal's optimal value is the

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

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:

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

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

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

Given $(\gamma_i)_{i\in I}$, an optimal $\lambda(\theta)$ is $\lambda(\theta) = \max\{0, \max\limits_{i\in I}\{\alpha_i(\theta)\}\}$. 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 $\lambda(\theta) = \max\{0, \max\limits_{i\in I}\{\alpha_i(\theta)\}\} > 0 \implies w_1(\theta) + w_2(\theta) = \mu(\theta)$ 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 $\lambda(\theta) = c > 0$ 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., $V^* = \sum\limits_{\theta\in\Theta}\mu(\theta)v((w,w),\theta)$. 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, $V^* = \sum\limits_{\theta \in \Theta} \mu(\theta) v((w, w), \theta)$. 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 $\theta > \min\limits_{i \in I}\{\theta_i^*\}$, $\lambda(\theta) > 0$ 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: The following chain of implications holds regarding asymmetric grain of dominance (AGD):

$$\text{AGD at } \mathcal{G}^{\epsilon,\epsilon} \implies \text{AGD at } \mathcal{G}^{\epsilon,\delta} \implies \text{AGD at } \mathcal{G}^{\delta,\delta} \implies \text{AGD at } \mathcal{G}.$$

If AGD fails in some environment, then the principal's value in that environment is the smallest feasible (and of course is independent of ϵ, δ). I prove the result assuming that AGD is satisfied at $\mathcal{G}$, but the proof is identical if AGD is satisfied at some but not all of the environments $\mathcal{G}$, $\mathcal{G}^{\delta,\delta}$, $\mathcal{G}^{\epsilon,\delta}$, and $\mathcal{G}^{\epsilon,\epsilon}$).

I prove first the relationship

$$V^*(\mathcal{G}) \leq V^*(\mathcal{G}^{\epsilon,\epsilon}) \leq V^*(\mathcal{G}^{\delta,\delta}) \tag{7}$$

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

$$d_i^n(\theta; \mathcal{G}) = g_i^n(\mathcal{G})(1 - \theta) + \ell_i^n(\mathcal{G})\theta \tag{8}$$

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

$$\begin{aligned} d_1^n(\theta;\mathcal{G}^{\gamma,\gamma}) &= g_1^n(\mathcal{G}^{\gamma,\gamma})(1-\theta)+\ell_1^n(\mathcal{G}^{\gamma,\gamma})\theta \qquad (9)\\ &= (g^n(\mathcal{G})+\epsilon)(1-\theta)+\ell^n(\mathcal{G})\theta \qquad (10)\end{aligned}$$

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

$$d_2^n(\theta;\mathcal{G}^{\gamma,\gamma}) = (g^n(\mathcal{G})-\epsilon)(1-\theta)+\ell^n(\mathcal{G})\theta \tag{11}$$

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

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

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 3. 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 $\overline{\theta}^{*\prime}, x^{*\prime} \leq \mu(\overline{\theta}^{*\prime})$ such that

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

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 $p \in [0,1]$

$$\Phi_1(p;\gamma') = \Phi_1(p;0) + \gamma' \sum_{\theta\in\Theta} w_1^p(\theta)(1-\theta) \geq \Phi_1(p;\gamma) \tag{12}$$

$$\Phi_2(p;\gamma') = \Phi_2(p;0) - \gamma' \sum_{\theta\in\Theta} w_2^p(\theta)(1-\theta) \leq \Phi_2(p;\gamma) \tag{13}$$

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

$$\Phi_1(p^*;\gamma) \geq 0 \tag{14}$$

$$\Phi_2(p^*;\gamma) \geq 0 \tag{15}$$

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

$$\Phi_1(p;0) - \Phi_1(p';0) = \Phi_2(p';0) - \Phi_2(p;0) \tag{16}$$

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

$$\Phi_1({p^*}'; \gamma') \geq 0 \tag{17}$$

$$\Phi_2({p^*}'; \gamma') \geq 0 \tag{18}$$

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

  $$\Phi_1(p_1; \gamma') - \Phi_1(p^*; \gamma) = 0 \tag{19}$$

  which exists because $\Phi_1(p; \gamma)$ is continuous in p, $\Phi_1(p^*; \gamma') \geq \Phi_1(p^*; \gamma)$ by (12), $\Phi_1(p^*; \gamma) \geq 0$ by (14), 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 (15). To see why, observe that for any $p' \in [p^*, 1]$,

  $$\Phi_2(p'; \gamma') - \Phi_2(p^*; \gamma') = \Phi_1(p^*; \gamma') - \Phi_1(p'; \gamma') \geq 0 \tag{20}$$

  by (16) and the definition of w_i^p. Further, by Lemma 3 (proof given below),

  $$0 \geq \Phi_2(p^*; \gamma') - \Phi_2(p^*; \gamma) \geq \Phi_1(p^*; \gamma) - \Phi_1(p^*; \gamma') \tag{21}$$

  for $\gamma' \geq \gamma$. Combining (20) and (21),

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

  where the second lines follows by (20) and (21) 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).

$\square$

Lemma 3. *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*

$$\sum_{\theta \in \Theta} (1-\theta) w_1(\theta) \geq \sum_{\theta \in \Theta} (1-\theta) w_2(\theta) \tag{22}$$

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):

$$\begin{aligned} &\Phi\left((w_i)_{i \in I}; \gamma\right) \geq 0, \quad i \in I \\ &w_i(\theta) \geq 0, \quad i \in I, \theta \in \Theta \\ &\textstyle\sum_{i \in I} w_i(\theta) \leq \mu(\theta), \quad \theta \in \Theta \end{aligned} \tag{23}$$

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

$$\sum_{\theta \in \Theta} w_1(\theta)(1-\theta) < \sum_{\theta \in \Theta} w_2(\theta)(1-\theta) \tag{24}$$

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

$$\Phi_1(\mathbf{w}; \gamma) = \Phi_1(\mathbf{w}; 0) + \gamma \sum_{\theta \in \Theta} w_1(\theta)(1-\theta) \geq 0 \tag{25}$$

$$\Phi_2(\mathbf{w}; \gamma) = \Phi_2(\mathbf{w}; 0) - \gamma \sum_{\theta \in \Theta} w_2(\theta)(1-\theta) \geq 0 \tag{26}$$

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,

$$\begin{aligned}\Phi_1(\mathbf{v};\gamma) - \Phi_1(\mathbf{w};\gamma) &= \Phi_2(\mathbf{w};\gamma) - \Phi_2(\mathbf{v};\gamma) + 2\gamma \sum_{\theta\in\Theta} w_2(\theta)(1-\theta) - w_1(\theta)(1-\theta) \\ &\geq \Phi_2(\mathbf{w};\gamma) - \Phi_2(\mathbf{v};\gamma) \qquad (27)\end{aligned}$$

By (26), $\Phi_1(\mathbf{v};\gamma) \geq 0$. 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 (26). 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,

$$\Phi_1(\mathbf{w}^{p^*};\gamma) = \Phi_1(\mathbf{w};\gamma) \qquad (28)$$

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 (25), and at $p = 1$, $\Phi_1(\mathbf{w}^p;\gamma) < 0$. Then by definition,

$$\Phi_1(\mathbf{v};\gamma) - \Phi_1(\mathbf{w}^{p^*;\gamma}) \leq \Phi_2(\mathbf{w}^{p^*;\gamma}) - \Phi_2(\mathbf{v};\gamma) \qquad (29)$$

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 (27) and (29) 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 $[p^v, 1]$. It is omitted for concision.

□

D Public Information Design in Section 3

In this section I demonstrate that public information design cannot achieve the principal's first best in the example of Section 3. For simplicity, suppose that $v(\mathbf{a}, \theta) = v(\mathbf{a}, \theta')$ for any pair $\theta, \theta' \in \{TT, HT, TH\}$. An information structure (T, π) is called *public* if whenever $\pi(t_i, t'_{-i}) > 0$ for some pair $(t_i, t_{-i}) \in T$, $\pi(t_i, t'_{-i}) = 0$ for any $t'_{-i} \neq t_{-i}$. After any signal, the principal's worst communication equilibrium is the same as the principal's worst BNE, is in pure strategies, and is either (w, w) or (n, n).[34] The principal's value is achieved as the limit of values of the following public information structures as $\epsilon \to 0$:[35]

1. The principal sends a binary public signal, $p \in \{n, w\}$
2. Signal $p = n$ is sent if $\theta \neq HH$ with prob. $\frac{1}{3} + \epsilon$ and if $\theta = HH$ with prob. 0
3. Signal $p = w$ is sent after $\theta \neq HH$ with prob. $\frac{2}{3} - \epsilon$ and if $\theta = HH$ with prob. 1

After $p = w$, w is strictly dominant for each i, so (w, w) is the unique equilibrium. After $p = n$, the principal's worst equilibrium is (n, n). As a result, as long as $v((w, w), \theta) > v((n, n), \theta)$ for each θ, the principal cannot achieve her first best using public information design alone.

E Rationalizability and BNE

In this section, I present an example of a crime game which does not have supermodular payoffs in each state, along with an information structure in which the principal's worst BNE is strictly better than the principal's worst rationalizable outcome.

Suppose that $T_1 = \{0\}$ and $T_2 = \{0, 1\}$, so that player 1 has a single type while player 2 has two types. Let $\Theta = \{0, 1\}$. The information structure is $\pi((0, t_2), \theta_2) = \mathbf{1}_{t_2 = \theta_2})$, i.e., player 2 is fully informed of the state while player 1 is not informed beyond the prior. Suppose that the prior is $\mathbb{P}_F(\theta = 1) = \mathbb{P}_F(\theta = 0) = \frac{1}{2}$.

When $\theta = 0$, payoffs are

[34] It is straightforward to show that, in equilibrium, if one agent mixes between actions then the other agent must as well. Assumption 2—(w, w) is an equilibrium—then implies that (n, n) is an equilibrium.

[35] The ϵ perturbation is required to guarantee strict incentives.

$$\begin{array}{ccc} & n & w \\ n & 5,3 & -\frac{1}{10},\frac{5}{2} \\ w & 100,0 & 0,2 \end{array}$$

and when $\theta = 1$, payoffs are

$$\begin{array}{ccc} & n & w \\ n & 3,3 & 0,2 \\ w & 2,0 & 1,1 \end{array}$$

When $\theta = 0$, the payoffs of player 1 are as in a prisoner's dilemma while the payoffs of player 2 are as in a coordination game. When $\theta = 1$, the game is a coordination game. Observe that this environment satisfies Assumptions 2 and 1, (i) if $-i$ chooses w, w is always a strict best-response and (ii) i's payoff is higher when $-i$ chooses n than when $-i$ chooses w.

Bayesian Nash Equilibrium. Under the information structure described above, there is no equilibrium in which all types choose n: if $t_2 = 0$ chooses n with probability 1, then $t_1 = 0$ chooses w. Thus, either $t_2 = 0$ chooses n with probability less than 1 or $t_1 = 0$ chooses w.

Rationalizability. If type $t_2 = 0$ chooses n, type $t_1 = 0$ has strict best-response w. If type $t_2 = 0$ chooses w and type $t_2 = 1$ chooses n, type $t_1 = 0$ has strict best-response n. If type $t_1 = 0$ chooses n, type $t_2 = 1$ has strict best-response w. If $t_1 = 0$ chooses w, type $t_2 = 1$ has strict best-response w. As a result, rationalizability does not eliminate any actions, and the principal's worst rationalizable action profile is that all agents choose n.