

# Non-monotonic Disclosure in Policy Advice\*

Anna Denisenko<sup>†</sup>      Catherine Hafer<sup>‡</sup>      Dimitri Landa<sup>§</sup>

## Abstract

The strategic context of bureaucratic advice to policymakers often takes the form of a disclosure game in which the relevant bureaucracy has an ideal policy interior to the policymaker’s action space. We characterize conditions under which this game has sequential equilibria in which the sender adopts a non-monotonic disclosure strategy, implying partial disclosure. Further, multiple sequential equilibria exist under some conditions, including a fully revealing equilibrium and multiple partially revealing equilibria that vary in extent of disclosure. We show that these equilibria are strictly rankable both by actors’ welfare and by their robustness to belief perturbations. Further, for sender preferences that are sufficiently close to the expected value of the state, (1) the most robust equilibrium is partially revealing and (2) set of states that the sender discloses becomes larger as the divergence in sender’s and receiver’s ex ante preferences increases.

## 1 Introduction

Effective policy-making requires expert information, for which policymakers must often rely on bureaucratic agencies. Because agencies have their own policy preferences, the problem of strategic disclosure – how to effectively motivate agencies to disclose information available to them without undermining policy-making – is a first-order concern in understanding incentives in policy-making.

While the literature on disclosure has developed important insights that shed light on the strategic logic of disclosure, existing models of disclosure assume away a crucial feature present in many settings: experts/agents/bureaucrats have an ideal action they would like to see implemented. Indeed, in many instances, these ideal actions are state-independent – for example, bureaucrats are often described as being strongly biased in favor of the status quo (not the least, because they are often residual claimants on costs of policy changes – see [Kaufman \(1981\)](#); [McCarty \(2004\)](#)). The presence of ideal actions belies the standard

---

\*We thank Steve Callander, Wioletta Dziuda, Anthony Fowler, Amanda Friedenberg, Kun Heo, Roger Myerson, Keith Schnakenberg, and participants in the Virtual Formal Theory Workshop, Stanford GSB PE Workshop, NYU PE Workshop, the Political Economy Winter Workshop at Bocconi University, the Harris PE Workshop and Harris Political Economy of American Democracy Conference for valuable feedback on earlier versions of this paper.

<sup>†</sup>Post-Doctoral Researcher, Harris School of Public Policy, University of Chicago, e-mail: ad4205@nyu.edu

<sup>‡</sup>Associate Professor, Wilf Family Department of Politics, NYU, e-mail: catherine.hafer@nyu.edu

<sup>§</sup>Professor, Wilf Family Department of Politics, NYU, e-mail: dimitri.landa@nyu.edu

motivating examples of the disclosure literature, in which the sender always wants a higher action (e.g., selling more cars).

We study a model of disclosure that departs from the [Milgrom \(1981\)](#) canonical setting in focusing on senders with such preferences. We show that it generates a host of predictions that substantially depart from the conventional prediction of the “unraveling” logic of disclosure whereby an informed agent has an incentive to disclose her information to avoid the decision-maker inferring the worst possible state from non-disclosure ([Grossman, 1981](#); [Milgrom, 1981](#)). While the studies of disclosure have consistently assumed monotonicity of sender preferences, we show that what the unraveling result needs is a sender with a most-preferred action that is sufficiently far from the expected value of the distribution of states. Monotonicity is not necessary for the existence of the full-disclosure equilibrium, but given non-monotonicity, if the sender’s most-preferred action is within a specified interval around the expected value of the state, that can fundamentally change the nature of disclosure. Specifically, it can give rise to equilibria in which unraveling stops before being complete. We delineate two types of partial disclosure equilibria that can be sustained in such a case: the *less expansive partial-disclosure equilibrium* (henceforth, “less expansive equilibrium”), where the sender reveals relatively less information, and the *more expansive partial-disclosure equilibrium* (henceforth, “more expansive equilibrium”), where it reveals relatively more. We show that only the less expansive and the full disclosure equilibria are *belief-stable* in the sense of robustness to small perturbations in players’ beliefs, and that, as the ideal action of the sender moves close to the expected value, the extent of such robustness is highest in the less expansive equilibrium. Starkly, our analysis demonstrates that in this equilibrium, disclosure increases with ex ante preference divergence between the sender and the receiver, contrasting sharply with the canonical prediction on the effects of preference divergence on communication in the cheap-talk signaling context.

The rest of the paper is organized as follows. In Section 2, we briefly discuss the connection to the existing literature. In Section 3, we describe a simplified version of our general environment and provide its initial analysis. Section 4 introduces the concept of belief-stable equilibria, which we use to refine our equilibrium predictions, and develops its implications for the model in Section 3. In Section 5, we generalize our formal framework and provide our main analysis. Section 6 shows robustness of our key results in contexts with partial verifiability and with possibility of vague statements by the sender. Section 7 provides discussion of our results, comparing them to predictions from cheap-talk models of communication and models of delegation and optimal choice of agency.

## 2 Connection to the Literature

A key result in the communication games with verifiable messaging is that all private information is revealed in equilibrium ([Grossman \(1981\)](#), [Milgrom \(1981\)](#), [Milgrom \(2008\)](#)). Following the initial formulation of this result, subsequent works have studied conditions under which the unraveling logic of full disclosure remains intact in a variety of formal and substantive environments. An important review of the disclosure games literature to-date in [Milgrom \(2008\)](#). A key lesson from this literature has been the general robustness the unraveling prediction.

To create the possibility for equilibrium non-disclosure, the sender must not be fully informed (Dye (1985), Jung and Kwon (1988), Shin (1994)). Alternatively, the receiver must be uncertain about sender’s preferences (Wolinsky (2003), Dziuda (2011)). Callander, Lambert and Matouschek (2021) find that incomplete disclosure can also be supported in a (stylized multi-dimensional) setting in which the sender can not only provide a direct recommendation but also referential information that can influence the decision-maker’s beliefs about other options.

The analysis of the incentives to disclose has focused on settings where the sender’s revelation is monotonic in the state. In contrast to the preference for always “higher” or always “lower” policies (e.g., selling more cars in Milgrom’s seminal example), preference satiation, which corresponds to interior ideal points on the range of possible policy alternatives, makes it possible to have interior boundaries on revelation intervals and non-monotonic revelation preferences, which we show to be supportable in equilibrium. Preference satiation is a standard feature of preferences in political economy contexts, which often work with the spatial model of preferences, where they are represented as ideal points. Seidmann and Winter (1997) provide a generalization of Milgrom’s classic setting to environments with objective functions that are concave in actions. The key assumption in their study is that sender’s utility is more state-dependent than the receiver’s – the opposite of the assumption we maintain in this paper. Denisenko, Hafer and Landa (2024) study the transmission of verifiable information from an unbiased sender with a fixed and known ideal point, and focus on the effects of sender competence on information transmission.

Delegation and communication within hierarchies is a focus of a substantial body of political economy scholarship (for reviews, see Gailmard and Patty (2012) and Sobel (2013)). The dominant approach to modeling communication in this literature has been as “cheap talk” in which bureaucrats’ potential messages are not directly constrained by their information (Crawford and Sobel (1982); Gilligan and Krehbiel (1989); Austen-Smith (1990); Austen-Smith (1993)). A key comparative static result that is at the core of this literature is the opposite of what we provide below for the setting with verifiable messaging – viz., that divergence in the actors’ preferences curtails communication, and successful communication at all occurs only when the advisor’s and the policymaker’s preferences are sufficiently aligned. An important exception is Callander (2008), which studies an expert bureaucrat’s advice to a legislator in an environment in which the bureaucrat’s expertise is endogenously acquired and the legislator may not be able to fully recover the bureaucrat’s private information from the advice. Callander shows that, in the absence of an institutionalized commitment to implement the received advice, greater divergence in primitive preferences between bureaucrat and legislator sometimes induces greater voluntary delegation of policy-making powers from the legislator to the bureaucrat. This suggests a certain affinity with our result that greater preference divergence spurs more information disclosure. The mechanisms producing these results are, however, very different. (Battaglini (2002) and Aybas and Callander (2023) identify policy-relevant settings in which cheap-talk communication fully favors, respectively, the receiver, and the sender.)

### 3 Simplified Model: Uniform Prior and State-independent Preferences

We begin with a relatively simple setting to illustrate the key features of equilibrium results. We generalize the model in subsequent sections with respect to the distribution of the state, the functional form of utilities, and the dependence of the Agency’s utility on the realized state.

Informally, the basic setting is the interaction between an Agency and a Policymaker, where the Agency possesses information relevant to the Policymaker’s current political agenda and has discretion over whether to share this information with the Policymaker, who then chooses a policy, given the received message.

Formally: The Agency observes  $\omega$  and decides whether to disclose it to the Policymaker. The Agency’s information is verifiable and it can send one of two messages:  $m \in \{\omega, \emptyset\}$ . The message  $m = \emptyset$  is commonly understood to be not intrinsically informative.<sup>1</sup>

The Agency’s utility function is

$$u_A(p) = -(i - p)^2,$$

where  $p \in \mathbb{R}$  is the policy choice of the Policymaker. The Agency’s utility is maximized when the implemented policy matches its most preferred policy, denoted by  $i$ . We later generalize to state-dependent Agency preferences.

The Policymaker’s utility function is

$$u_P(p) = -(\omega - p)^2.$$

Thus, the Policymaker aims to set policy  $p = \omega$ .

The state of the world,  $\omega$ , is drawn from a continuous distribution with cumulative distribution function (cdf)  $F(\cdot)$  and probability density function (pdf)  $f(\cdot)$  over a support  $\Omega$ . For initial tractability and clarity of exposition, in the special case we assume  $\omega$  is uniformly distributed on  $\Omega := [-1, 1]$ , i.e.,  $f(\omega) = 1/2$  for  $\omega \in [-1, 1]$  and zero otherwise.

Note that the absolute value  $|i|$  measures the divergence in the two actors’ ex ante preferences. The Policymaker’s ex ante preferred policy is  $p = \mathbb{E}[\omega]$ , while the Agency’s is  $p = i$ , with the distance between these reflecting the alignment of their preferences.

The timing of the game is as follows:

1. Nature draws the state of the world  $\omega$ .
2. The Agency observes  $\omega$  and chooses a message  $m$ .
3. The Policymaker observes  $m$  and selects a policy  $p$ .

Let  $\mu : \Omega \rightarrow [0, 1]$  represent the Agency’s disclosure strategy, where  $\mu(\omega)$  is the probability the Agency discloses state  $\omega$ . The Policymaker’s beliefs, conditional on observing message

---

<sup>1</sup>The Agency’s message space is restricted, excluding vague but truthful messages, which provides the most challenging setting for our results, as we discuss later.

$m$ , are represented by the pdf  $\beta$ . Consistent with verifiable information,  $\beta(\omega|m = \omega) = 1$  and  $\beta(\omega|m = \hat{\omega} \neq \omega) = 0$ . From Bayes' Rule,

$$\beta(\omega|\emptyset; \mu) = \frac{(1 - \mu(\omega))f(\omega)}{\int_{\Omega}(1 - \mu(\hat{\omega}))f(\hat{\omega})d\hat{\omega}}.$$

The Policymaker's optimal strategy  $p^* : \Omega \cup \emptyset \rightarrow \mathbb{R}$  maximizes the Policymaker's utility given her beliefs conditional on  $m$ .

The following lemma summarizes the Policymaker's optimal policy-implementation strategy:

**Lemma 1.** *The Policymaker's optimal policy  $p^*(m)$  is*

$$p^*(m) = \begin{cases} m & \text{if } m \neq \emptyset \\ \mathbb{E}[\omega|\emptyset; \mu^*(\cdot)] & \text{if } m = \emptyset, \end{cases} \quad (1)$$

where  $\mu^*(\cdot)$  denotes the Policymaker's conjecture about the Agency's disclosure strategy.

The Agency discloses its information to the Policymaker when withholding it would result in a policy farther from the Agency's preferred policy than the one the Policymaker would implement if fully informed about the state. The following lemma describes the Agency's optimal disclosure strategy, given the Agency conjectures that the Policymaker's strategy is of the form specified in Lemma 1, with  $x := p^*(\emptyset)$ .

**Lemma 2.** *The Agency's optimal disclosure strategy is*

$$\mu^*(\omega) = \begin{cases} 1 & \text{if } \omega \in M(x, i) \\ 0 & \text{if } \omega \in N(x, i), \end{cases} \quad (2)$$

where  $M(x, i) := [i - \sqrt{(x - i)^2}, i + \sqrt{(x - i)^2}] \cap \Omega$  and  $N(x, i) := \Omega \setminus M(x, i)$ .<sup>2</sup>

Although some signals are never disclosed, the Policymaker infers in equilibrium that when she receives  $m = \emptyset$ , the state fell outside the disclosure interval, and she updates her beliefs about the state absent disclosure accordingly. In every equilibrium, after observing  $\omega$ , the Agency follows an optimal disclosure strategy  $\mu^*(\omega)$ , anticipating that absent disclosure the Policymaker will select the optimal policy, denoted

$$x^* := \mathbb{E}[\omega|\emptyset; \mu^*(\cdot)].$$

In the next proposition, we characterize all disclosure strategies supported in Sequential Equilibrium (SE); for the remainder of the paper, "equilibrium" will mean Sequential Equilibrium.

**Proposition 1.** *1. For all  $i \in \Omega$ , a full disclosure strategy can be supported in equilibrium, with  $x^* = x_F$  and the disclosure interval  $M(x^*, i) = \Omega$ , where*

$$x_F := \begin{cases} 1 & \text{if } i \leq 0 \\ -1 & \text{if } i \geq 0. \end{cases}$$

---

<sup>2</sup>Both  $m = \omega$  and  $m = \emptyset$  are optimal if  $\omega = i - \sqrt{(x^* - i)^2}$  or  $\omega = i + \sqrt{(x^* - i)^2}$ . Henceforth, we will assume the Agency discloses when indifferent, with no substantive effect on our results.

2. For  $i \in [-\frac{1}{4}, \frac{1}{4}]$ , two partial disclosure strategies can be supported in equilibrium, with  $x^* \in \{x_M, x_L\}$  and the disclosure interval  $M(x^*, i) \subset \Omega$ , where

$$x_M := \frac{1}{2} \left( 2 \cdot i - \text{sign}(i) - \text{sign}(i) \cdot \sqrt{1 - 4 \cdot |i|} \right);$$

$$x_L := \frac{1}{2} \left( 2 \cdot i - \text{sign}(i) + \text{sign}(i) \cdot \sqrt{1 - 4 \cdot |i|} \right).$$

*Proof.* See Appendix. □

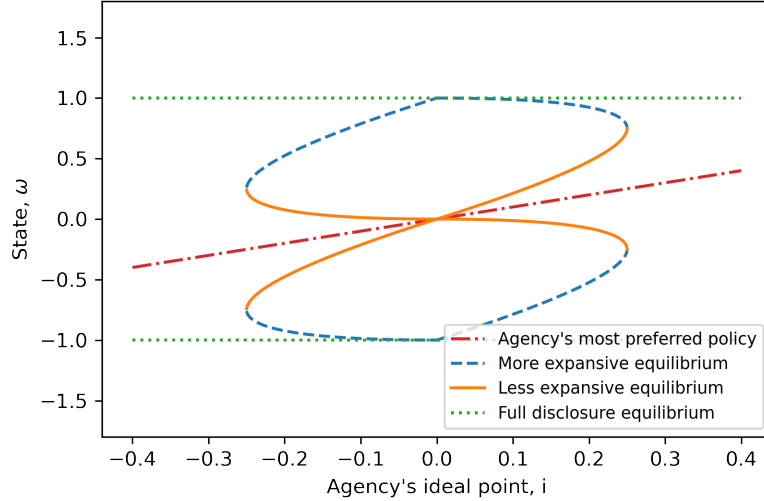


Figure 1: Agency’s disclosure boundaries in more expansive, less expansive, and fully revealing equilibria as a function of Agency’s ideal point  $i$ .

We will, for convenience, adopt the following terminology for this special case. We refer to the disclosure strategy characterized by  $x_M$  as the *more expansive* partial-disclosure strategy, and the disclosure strategy characterized by  $x_L$  as the *less expansive* partial-disclosure strategy. We term the equilibria in which these strategies are employed as the *more expansive equilibrium* and the *less expansive equilibrium*, correspondingly. This nomenclature is motivated by the difference in the extent of information revelation across equilibria: the Agency discloses a strictly broader set of states in the more expansive equilibrium than in the less expansive equilibrium. This feature is visually represented in Figure 1, which juxtaposes all equilibrium disclosure boundaries. As the figure demonstrates, the less expansive disclosure interval is nested within the more expansive disclosure interval, which, in turn, lies within the full disclosure interval for all values of  $i$ .<sup>3</sup>

The nestedness of the disclosure strategies has a direct implication for the Policymaker’s welfare, summarized in the following corollary.

<sup>3</sup>While the nestedness of equilibrium disclosure sets is maintained across different prior distributions, the existence of at most two partial disclosure equilibria is a consequence of the uniform prior assumption.

**Corollary 1.** *For any given Agency’s ideal point  $i$ , the Policymaker’s ex ante expected utility is highest when there is full disclosure in equilibrium and lowest in the (partial disclosure) less expansive equilibrium.*

The Policymaker trivially prefers more disclosure to less, and the full disclosure equips the Policymaker with complete information about the state, letting her select policies tailored best to her preferences. In contrast, partial disclosure leaves the Policymaker with residual uncertainty, leading to suboptimal policy choices and reduced utility. The welfare loss is larger when disclosure is smaller and, therefore, Policymaker’s utility is lowest under the less expansive equilibrium.

Proposition 1 establishes that for any given ideal point  $i$ , there are at most three pure-strategy equilibrium profiles. We now argue that mixed strategies are not sustainable in any equilibrium. For any state realization  $\omega$ , and conditional on the policy implemented in the absence of disclosure, the Agency has a strict preference for disclosing states that produce a policy closer to its ideal point. Probabilistic disclosure is therefore never incentive compatible (except for the knife-edge conditions at the boundaries of  $M(x, i)$ ). Similarly, the Policymaker’s optimal policy choice is also restricted to pure strategies. Were the Policymaker to randomize across policies following non-disclosure, this would induce a change in the agency’s optimal disclosure interval. However, given any disclosure interval chosen by the Agency, there exists a unique optimal policy for the Policymaker in the absence of disclosure that constitutes a best response. Thus, neither player employs mixed strategies in equilibrium.

As Figure 1 illustrates, partial disclosure is non-monotonic in realized states of the world. It is this non-monotonicity that prevents full unraveling: The Agency’s unwillingness to disclose signals both too high and too low, relative to  $i$ , ensures that policy absent disclosure is not too extreme. Consequently, since the expected state conditional on non-disclosure is not too extreme, the Agency optimally chooses to withhold some states on either side of its ideal point.

The unique full-disclosure equilibrium is the only equilibrium that is monotonic in realized states. Proposition 1 establishes necessary and sufficient conditions for full disclosure to be the unique equilibrium of the game. Specifically, when Agency’s ideal point  $i$  is sufficiently far from the mean of the prior distribution (or, alternatively, when ex ante preference divergence  $|i|$  is sufficiently high), the unique equilibrium involves full disclosure. Conversely, when  $i$  lies within the interval  $[-\frac{1}{4}, \frac{1}{4}]$  (when  $|i| < \frac{1}{4}$ ), multiple equilibria exist, including those with partial disclosure.

As  $|i|$  approaches zero, the Agency’s incentives to disclose information approach symmetry around the mean of the distribution of states. The Agency, being indifferent between right-leaning and left-leaning policies, does not prioritize the disclosure of one over the other. Thus, all else equal, as  $|i|$  converges zero, the expected state in the absence of disclosure should converge to the mean of the prior distribution (zero), following the Agency’s optimal strategy as established in Lemma 2 (with the Policymaker’s strategy,  $x$ , held constant). (Full disclosure becomes possible at  $i = 0$  because the Policymaker forms extreme beliefs, leading her to implement an extreme policy ( $-1$  or  $1$ ) absent disclosure, thereby encouraging the Agency to fully disclose information, reinforcing the Policymaker’s beliefs.) While this anticipated state convergence is sustained in the less expansive equilibrium, it is not in the

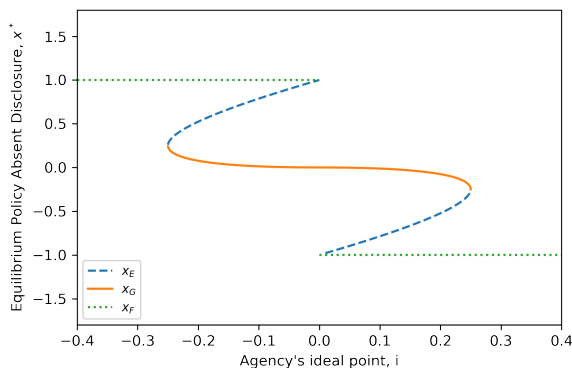


Figure 2: Policymaker's equilibrium policy selection in the absence of disclosure.

equilibria characterized by  $x_F$  or  $x_M$ . In these latter cases, as the Policymaker continuously updates her beliefs about the state expecting the Agency to disclose (weakly) more information as preference misalignment  $|i|$  decreases, this belief becomes self-reinforcing, resulting in greater disclosure from the Agency.

### 3.1 Comparative Statics

In this section, we examine how the game's parameters affect the players' decisions and the resulting outcomes. Our first result details how the Agency's ideal point,  $i$ , impacts the Policymaker's beliefs and choices in the absence of disclosure:

**Proposition 2.** *Increasing  $i$ , the difference between the Agency's ideal point and the Policymaker's ex ante expected ideal point,*

1. *has no effect on  $x^* = x_F$  for  $i \in (-\frac{1}{4}, 0) \cup (0, \frac{1}{4})$  but discontinuously decreases at  $i = 0$ , in the full-disclosure equilibrium;*
2. *decreases  $x^* = x_L$  on  $i \in [-\frac{1}{4}, \frac{1}{4}]$  in the less expansive equilibrium; and*
3. *increases  $x^* = x_M$  on  $i \in (-\frac{1}{4}, 0) \cup (0, \frac{1}{4})$  but discontinuously decreases it at  $i = 0$  in the more expansive equilibrium.*

*Proof.* See Appendix. □

Figure 2 illustrates the policy in the absence of disclosure, as a function of the Agency's ideal point  $i$ , for each of the three equilibria. Notably, the comparative statics for the less expansive and more expansive equilibria present a stark contrast: the equilibrium policy absent disclosure decreases in the Agency's ideal point  $i$  in the former, but increases with  $i$  (for  $i \neq 0$ ) in the latter.

We next characterize the impact of ex ante preference divergence  $|i|$  on disclosure.

**Proposition 3.** *Increasing the magnitude of the ex ante preference divergence between the Agency and the Policymaker,  $|i|$ ,*

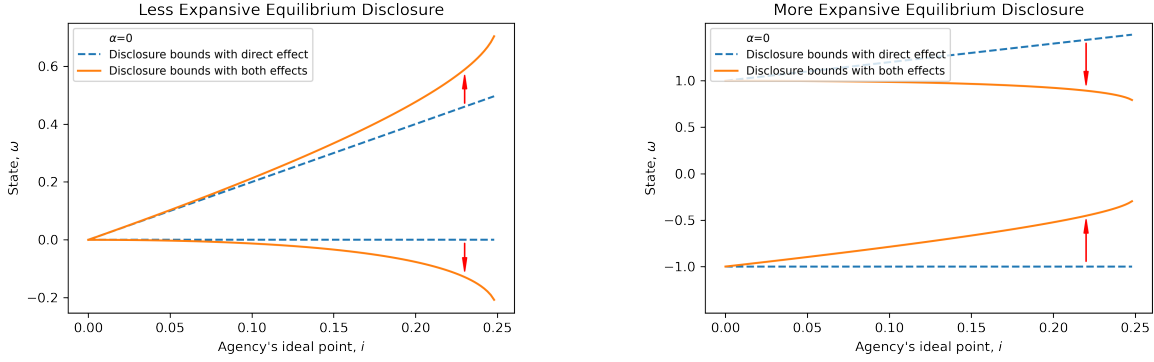


Figure 3: Direct and indirect effects of ex ante preference divergence on Agency's equilibrium disclosure in the less expansive (left panel) and more expansive (right panel) equilibria for  $i \geq 0$ .

1. has no effect on disclosure in the full-disclosure equilibrium;
2. (weakly) decreases disclosure in the more expansive equilibrium; and
3. increases disclosure in the less expansive equilibrium.

*Proof.* See Appendix. □

The ex ante preference divergence has two effects on the Agency's disclosure strategy. First, there is a *direct effect*: Policymaker's strategy being fixed, the Agency discloses more information as divergence increases to obtain policy closer to its own, relatively more extreme tastes. However, there is a secondary *indirect effect* of the ex ante divergence outlined in Proposition 2: the Agency needs to take into account how its changing disclosure strategy will affect the Policymaker's beliefs, and hence the policy she implements, after non-disclosure. Disclosure depends on both the direct and indirect effects of preference divergence.

Figure 3 shows direct and indirect effects in the less expansive equilibrium in the left panel, and in the more expansive equilibrium in the right panel. In the less expansive equilibrium, the two effects reinforce each other, and so as ex ante preference divergence increases, communication becomes more informative. Greater divergence implies that the Agency's preferences become increasingly asymmetric relative to the mean of the prior distribution. This asymmetry shapes the Agency's incentives, encouraging it to prioritize disclosure of states on one side relative to the other, allowing the Policymaker to infer more about the state when information is withheld. Consequently, the policy choice in the absence of disclosure shifts away from the prior mean, further encouraging the Agency to disclose information.

In the more expansive equilibrium, however, the direct and indirect effects are opposed. Thus, greater divergence in the Agency's preferences encourages the Agency to disclose more, which results in the disclosure interval being more symmetric, and making the policy following non-disclosure more centrist. Because the Policymaker, in the absence of disclosure, chooses policy closer to the prior mean, the Agency has less incentive to disclose, and so the direct and indirect effects of communication counteract each other. In this equilibrium, the indirect effect dominates the direct effect, leading to a decline in disclosure as preferences diverge.

## 4 Belief-Stable Equilibria

The difference in comparative statics across equilibria gives the refinement of the set of equilibria greater salience. As we detail above, all three equilibria – the full-disclosure, less expansive, and more expansive equilibria – are sequential equilibria, and all three satisfy standard action-perturbation refinement conditions. With this in mind, in this section, we introduce a novel refinement, *belief-stability*, which employs perturbations in beliefs to identify robust equilibria.

We begin with the following definition:

**Definition 1.** Consider an equilibrium strategy profile and system of beliefs  $(\sigma, \beta)$  and a perturbed system of beliefs  $\beta_j^\varepsilon$ . Let  $\sigma^\varepsilon$  be sequentially rational given the beliefs  $(\beta_j^\varepsilon, \beta_{-j})$ , and let  $\hat{\beta}_j^\varepsilon$  be consistent with  $\sigma^\varepsilon$ . If there exists an  $\varepsilon > 0$  such that, for every  $\beta_j^\varepsilon$  that satisfies  $|\beta_j^\varepsilon(y) - \beta_j(y)| < \varepsilon$ , condition  $|\hat{\beta}_j^\varepsilon(y) - \beta_j(y)| \leq |\beta_j^\varepsilon(y) - \beta_j(y)|$  is satisfied for all decision nodes  $y$  assigned to  $j$ , then we say that equilibrium  $(\sigma, \beta)$  is **belief-stable for player  $j$** . Equilibrium  $(\sigma, \beta)$  is **belief-stable** if it is belief-stable for every player  $j$ , and is **belief-unstable** otherwise.

**Definition 2.** Let  $\varepsilon_j^*$  be the largest value  $\varepsilon > 0$  such that, for every  $\beta_j^\varepsilon$  that satisfies  $|\beta_j^\varepsilon(y) - \beta_j(y)| < \varepsilon$ , condition  $|\hat{\beta}_j^\varepsilon(y) - \beta_j(y)| \leq |\beta_j^\varepsilon(y) - \beta_j(y)|$  is satisfied for all decision nodes  $y$  assigned to  $j$ . We say  $\varepsilon_j^*$  is the **extent of belief-stability of  $(\sigma, \beta)$  for player  $j$** .

Intuitively *belief-stability* ensures that small perturbations in players' beliefs do not result in large deviations in their optimal strategies: in belief-stable equilibria, if players' beliefs depart slightly from equilibrium beliefs, the feedback they receive as they play the game reinforces the equilibrium beliefs. In contrast, if the equilibrium is belief-unstable, feedback will provoke larger and larger deviations from equilibrium beliefs.

The next proposition applies the concept of belief-stability to the sequential equilibria in the game.

**Proposition 4.**

1. The less expansive equilibrium is belief-stable if  $|i| \neq \frac{1}{4}$ .
2. The more expansive equilibrium is belief-unstable.
3. The full-disclosure equilibrium is belief-stable for all  $|i| \neq 0$  and is belief-unstable at  $i = 0$ .
4. For  $|i| \leq \frac{1}{4}$ , the extent of belief-stability of the full-disclosure equilibrium increases and of the less expansive equilibrium decreases in  $|i|$ .

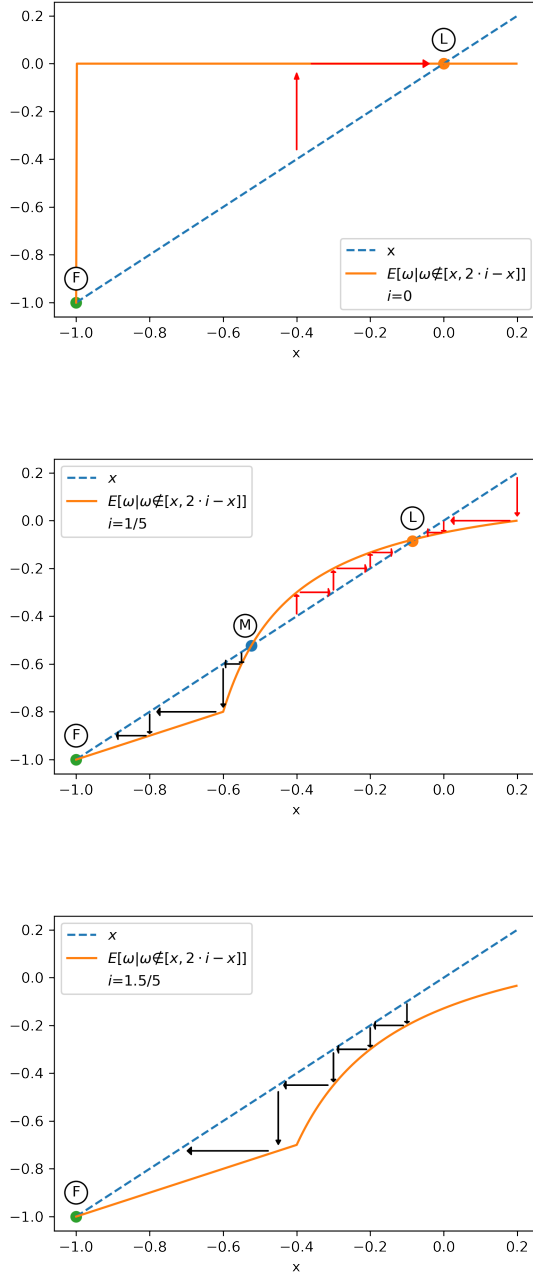


Figure 4: Policymaker's beliefs absent disclosure,  $E[\omega | \omega \notin [x, 2i - x]]$ , as a function of the policy adopted in the absence of disclosure,  $x$

Figure 4 illustrates three substantively different scenarios in this game: when  $i = 0$ ,  $i \in (0, 1/4]$ , and  $i > 1/4$ . The policy adopted in the absence of disclosure,  $x$ , is on the horizontal axis, and the blue dashed line is the 45° line, i.e.,  $y = x$ . The Agency's optimal disclosure interval for  $i \geq 0$ , parameterized by  $x$ , is  $M(x, i) = [x, 2 \cdot i - x] \cap \Omega$ . Using  $E[\omega | \cdot]$  as a summary statistic for beliefs, the orange solid line represents the Policymaker's beliefs given non-disclosure, assuming the Agency uses this disclosure interval for every given

value  $x$ . Because the Policymaker’s equilibrium policy choice  $x^*$  is such that  $x^* = E[\omega | \omega \notin M(x^*, i)]$  in equilibrium, intersections of the blue dashed line and the orange solid line identify equilibrium values  $x^*$ .

Arrows in Figures 4 (a)-(c) show the direction of the best-response updating following an initial perturbation in the Policymaker’s beliefs. For example, in the more expansive equilibrium, if the Policymaker’s  $E[\omega | m^*(\omega) = \emptyset]$  shifts from  $x = x_M$  to  $x = x_M + \varepsilon$ , the Agency’s best-response disclosure strategy shifts rightward, revealing less information. In turn, the Policymaker updates her beliefs, resulting in  $E[\omega | \omega \notin M(x_M + \varepsilon, i)] > x_M + \varepsilon$ . In fact, the more expansive equilibrium is belief-unstable and any perturbation of beliefs will cause adjustments that move behavior and beliefs farther from this equilibrium. Depending on the nature of the initial deviation, the game will converge to an equilibrium where the Policymaker’s expectation of the state given non-disclosure is either  $x_L$  or  $x_F$ .<sup>4</sup> In contrast, small perturbations in the Policymaker’s beliefs after non-disclosure in the less expansive equilibrium give rise, by analogous argument, to adjustments that lead back to the less expansive equilibrium.

While both the full-disclosure equilibrium and the less expansive equilibrium are belief-stable for  $|i| > 0$ , their robustness to belief perturbations move in opposing directions as ex ante preferences converge. As  $|i|$  decreases, the extent of belief-stability decreases for the full-disclosure equilibrium, but increases for the less expansive equilibrium (See Figure 5). Consequently, for sufficiently aligned preferences, partial disclosure not only remains a possibility but emerges as the more belief-stable outcome.

Note that at the boundary case of perfect ex ante preference alignment,  $i = 0$ , the full-disclosure equilibrium is not belief-stable. Its support relies on an extreme off-path threat by the Policymaker, but with  $i = 0$  any marginal perturbation results in the Agency concealing a symmetric neighborhood of 0. Given such non-disclosure, the Policymaker’s posterior jumps discontinuously away from the initial belief. The unique belief-stable equilibrium in case of perfect ex ante preference alignment is, thus, the less expansive equilibrium in which the disclosure set collapses to  $M = \{0\}$ .

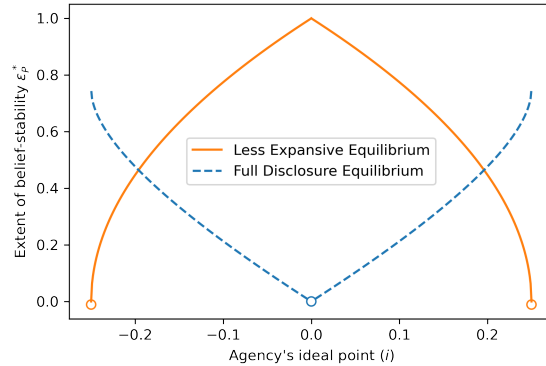


Figure 5: Extent of belief-stability for the less expansive equilibrium and for the equilibrium with full disclosure as a function of the Agency’s most preferred policy  $i$ .

<sup>4</sup>Given that  $|i| \notin \{0, 1/4\}$ .

## 5 A (More) General Model

In this section we study a more general environment, using the equilibrium stability concept we delineated above to get a sharper characterization and comparative statics.

Let the state space  $\Omega \subseteq \mathbb{R}$  be a compact set, with the convex hull denoted by  $Conv(\Omega) = [\underline{\omega}, \bar{\omega}]$ .<sup>5</sup> The state of the world,  $\omega \in \Omega$ , is drawn from a continuous cumulative distribution function  $F$  with a corresponding probability density function  $f$ , where  $E[\omega] = 0$ .

The Policymaker's von Neumann-Morgenstern (vN-M) utility function  $u_P(p; \omega)$  is continuously differentiable and strictly concave in  $p$ , and the Policymaker's ideal policy is  $p^P(\omega) := \arg \max_p u_P(p; \omega) = \omega$ .<sup>6</sup> The Agency's vN-M utility function  $u_A(p; \omega, \alpha, i)$  is also continuously differentiable and strictly concave in  $p$ , and the Agency's ideal policy given by  $p^A(\omega, \alpha, i) := \arg \max_p u_A(p; \omega, \alpha, i) = \alpha \cdot p^P(\omega) + (1 - \alpha) \cdot i$ , where  $i \in \mathbb{R}$  is the Agency's *preference (additive) bias* and  $\alpha \in [0, 1]$  is the Agency's *preference state dependence*. A higher  $\alpha$  indicates that the Agency's preferences are more closely aligned with those of the Policymaker, while a lower  $\alpha$  indicates a stronger weight for the Agency's bias  $i$ . Let  $p_0^P := \arg \max_p \mathbb{E}[u_P(p; \omega)]$  be the Policymaker's ex ante optimal policy. Thus, a game  $G := (\{\omega, \emptyset\}_{\omega \in \Omega}, \Omega; u_A(p; \omega, \alpha, i), u_P(p; \omega); F)$ . Let  $\mathcal{G}$  be the set of all such games satisfying the above conditions.

We analyze three categories of equilibria based on the equilibrium disclosure set,  $M \subseteq \Omega$ . In a full-disclosure equilibrium, the Agency reveals the state for all  $\omega \in \Omega$ . In a *partial-disclosure equilibrium*, the Agency withholds information for a non-empty set of states, so that  $\emptyset \neq M \subset \Omega$ . Finally, we distinguish a substantively important limiting case of partial disclosure, which we will refer to as a *non-disclosure equilibrium*. In a non-disclosure equilibrium, the disclosure set is the singleton  $M = \{p_0^P\}$ . This outcome is substantively equivalent to complete non-disclosure from the Policymaker's perspective, as her posterior belief upon observing non-disclosure remains her prior.

This section first provides specific conditions for the existence and belief-stability of the full-disclosure equilibrium and the non-disclosure equilibrium. It then characterizes the necessary and sufficient conditions for the existence of partial-disclosure equilibria and provides an indirect means of determining the belief-stability of any partial-disclosure equilibria: The disclosure sets of equilibria are nested, providing a natural ordering of all equilibria, and belief-stability alternates on this ordering. Thus knowing the belief-stability of one equilibrium is sufficient to determine the belief-stability of all others. The section concludes with comparative statics, examining how the outcomes in belief-stable equilibria change in response to model parameters.

Define  $\hat{x}(x)$  to be the Policymaker's optimal policy choice absent disclosure given the Agency discloses if and only if  $\omega \in M(x, \alpha, i)$ . Formally,

$$\hat{x}(x) := \arg \max_p \int_{\omega \notin M(x, \alpha, i)} u_P(p; \omega) dF(\omega). \quad (3)$$

---

<sup>5</sup>While compactness of  $\Omega$  is not strictly required for the existence of partial disclosure equilibria, it is important to ensure that the concept of full-disclosure equilibrium is well-defined. Without compactness, full disclosure may not be attainable.

<sup>6</sup>We assume here that the Policymaker is unbiased and show that all the results hold for biased Policymaker in the appendix.

In equilibrium, the Agency's conjecture  $x$ , on which it bases its disclosure interval  $M(x, \alpha, i)$  must be correct; hence, for any equilibrium value  $x^*$ , it must be that  $x^* = \hat{x}(x^*)$ .

The following proposition establishes conditions under which a full-disclosure equilibrium can be sustained and when it is belief-stable.

**Proposition 5.**

1. (a) If  $u_A(p^P(\bar{\omega}); \bar{\omega}, \alpha, i) > u_A(p^P(\underline{\omega}); \bar{\omega}, \alpha, i)$ , then for all conditions on primitives such that there exists a full-disclosure equilibrium s.t.  $x^* = p^P(\underline{\omega})$ , that equilibrium is belief-stable;
- (b) If  $u_A(p^P(\bar{\omega}); \bar{\omega}, \alpha, i) = u_A(p^P(\underline{\omega}); \bar{\omega}, \alpha, i)$ , then for all conditions on primitives, such that there exists a full-disclosure equilibrium s.t.  $x^* = p^P(\underline{\omega})$ , that equilibrium is belief-unstable;
- (c) If  $u_A(p^P(\bar{\omega}); \bar{\omega}, \alpha, i) < u_A(p^P(\underline{\omega}); \bar{\omega}, \alpha, i)$ , there does not exist a full-disclosure equilibrium such that  $x^* = p^P(\underline{\omega})$ .<sup>7</sup>
2. (Seidmann and Winter, 1997) There exists a full-disclosure equilibrium if the Agency's utility,  $u_A(\cdot)$ , satisfies single-crossing.

*Proof.* See Appendix. □

Part 1 of Proposition 5 addresses belief-stability of full-disclosure equilibrium, establishing that for a full-disclosure equilibrium sustained by a boundary policy to be stable, the Agency must *strictly* prefer disclosure at the opposite boundary. If the Agency is indifferent, the equilibrium is not robust to belief perturbations. At this point of indifference, an arbitrarily small perturbation to the Policymaker's belief causes the Agency's non-disclosure set to include states from neighborhoods of both  $\underline{\omega}$  and  $\bar{\omega}$ . The Policymaker's best choice of policy absent disclosure then jumps discontinuously to a policy strictly greater than  $p^P(\underline{\omega})$ , violating the condition for belief-stability.

Part 1 also provides a necessary condition for a full-disclosure equilibrium to be sustained. If the Agency has a strict incentive to conceal a boundary state, full-disclosure equilibrium is not incentive compatible, rendering belief-stability analysis moot. Finally, part (2), established by Seidmann and Winter (1997), provides a sufficient condition for the existence of a full-disclosure equilibrium.

Having established the belief-stability conditions for a full-disclosure equilibrium, we now turn to the conditions for the existence and belief-stability of a non-disclosure equilibrium.

**Proposition 6.**

1. There exists a unique threshold  $\alpha^* \in (0, 1)$  such that a non-disclosure equilibrium exists if and only if the Agency's bias  $i = p_0^P$  and preference state-dependence  $\alpha \leq \alpha^*$ .
2. A non-disclosure equilibrium is belief-stable.

*Proof.* See Appendix. □

---

<sup>7</sup>Results symmetric to 1.(a)-1.(c) hold for the full-disclosure equilibrium at  $x^* = p^P(\bar{\omega})$ .

Proposition 6 identifies the non-disclosure equilibrium as a belief-stable outcome in the case in which there is no ex-ante preference divergence. We now characterize the conditions under which partial disclosure equilibria, the central focus of our analysis, can be sustained. The following proposition provides the necessary and sufficient conditions for the existence of such equilibria, linking them to the state-dependence of the Agency's preference,  $\alpha$ , and the magnitude of its additive bias,  $i$ .

**Proposition 7.** *There exists a threshold  $\alpha^{**} \in [\alpha^*, 1)$  and,  $\forall \alpha \leq \alpha^{**}$ , an interval  $I^*(\alpha) \subset \Omega$  containing  $p_0^P$  such that  $G \in \mathcal{G}$  has a partial-disclosure equilibrium if and only if  $\alpha \leq \alpha^{**}$  and  $i \in I^*(\alpha)$ .*

*Proof.* See Appendix. □

Proposition 7 provides the necessary and sufficient conditions for the existence of partial disclosure equilibrium. It confirms that partial disclosure is a robust possibility sustainable when the Agency's preferences are not excessively state-dependent (low  $\alpha$ ) and its intrinsic bias is moderate ( $i \in I^*(\alpha)$ ). Further, proposition 7 emphasizes that when non-disclosure equilibrium exists, so does partial disclosure equilibrium, while the converse does not hold.

Note that Seidmann and Winter's condition – single-crossing – is a restriction on preferences that is not required for our analysis of partial disclosure. With this in mind, observe that the parameter spaces supporting partial-disclosure equilibrium and full-disclosure equilibrium are not mutually exclusive: if the Agency's utility,  $u_A(\cdot)$ , satisfies single-crossing, then for any parameter profile  $(\alpha, i)$  that supports a partial disclosure equilibrium ( $\alpha \leq \alpha^{**}$  and  $i \in I^*(\alpha)$ ), the game admits multiple equilibria.

Define  $X^*(\alpha, i)$  such that  $x^* \in X^*(\alpha, i)$  if and only if  $x$  is the policy absent disclosure in some equilibrium of the game with preference parameters  $(\alpha, i)$ . The following proposition establishes that these equilibria exhibit a highly structured pattern, which may be leveraged to determine, in conjunction with Proposition 6 or 7, the belief-stability of every equilibrium.

**Proposition 8.**

1. For any distinct  $x_j^*$  and  $x_k^*$ ,

$$M(x_k^*, \alpha, i) \subseteq M(x_j^*, \alpha, i) \Leftrightarrow u_A(x_k^*; \omega, \alpha, i) \geq u_A(x_j^*; \omega, \alpha, i).$$

2. Index set  $X^*(\alpha, i)$  s.t. if  $j < k, x_j^* < x_k^*$ . Then the belief-stability of equilibria alternates along this ordering, i.e., if the equilibrium corresponding to  $x_j^*$  is belief-stable, then the equilibrium corresponding to  $x_{j+1}^*$  (if it exists) is not belief-stable and the equilibrium corresponding to  $x_{j+2}^*$  (if it exists) is.

*Proof.* See Appendix. □

Part 1 of this proposition establishes that the Agency's utility absent disclosure is higher in an equilibrium in which less is disclosed. A direct corollary is that all equilibria can also be ranked by the Policymaker's ex-ante expected welfare, and that ranking is the reverse of the ranking by the Agency's utility absent disclosure. This inverse relationship is starkest in the limiting cases. The least disclosing equilibrium (the equilibrium with the smallest

disclosure set) is most preferred by the Agency, as it reveals the least information. The most disclosing equilibrium maximizes the Policymaker’s ex-ante welfare and minimizes that of the Agency.<sup>8</sup>

Part (2) of Proposition 8 shows that for any given preference profile, belief-stable and belief-unstable equilibria must alternate along the set of equilibria ordered by policies absent disclosure. This property means we do not need to analyze the belief-stability of every equilibrium. Instead, if we can determine the belief-stability of a single one (for instance a full-disclosure equilibrium) we can infer the belief-stability of all others.

**Corollary 2.** *For any given  $(\alpha, i)$ , knowing the belief-stability of one equilibrium is sufficient to determine the belief-stability of all equilibria.*

The following propositions describe how policy and disclosure vary with the Agency’s preferences in belief-stable equilibria. Furthermore, the comparative statics are the opposite in equilibria that are not belief-stable.

**Proposition 9.** *The equilibrium policy absent disclosure,  $x^*$ , is*

1. *decreasing in the Agency’s bias,  $i$ ; and*
2. *increasing in the Agency’s preference state-dependence,  $\alpha$ , when  $x^* < i$ , and decreasing in  $\alpha$  otherwise,*

*if and only if the equilibrium is belief-stable.*

*Proof.* See Appendix. □

Proposition 9 characterizes how, in belief-stable equilibria, the policy following nondisclosure  $x^*$  responds to the Agency’s preferences, as captured by parameters  $i$  and  $\alpha$ . An increase in  $i$  corresponds to a greater ex ante difference in preferences between the Agency and the Policymaker, which results in a decrease in the Policymaker’s choice absent disclosure. An increase in  $\alpha$ , on the other hand, indicates a reduction in their ex post preference divergence, which results in policy absent disclosure moving closer to  $i$ . The mechanism underlying this result is a product of the Policymaker’s recognition of the difference between her own interests and the Agency’s and its implications for disclosure. In canonical disclosure models (e.g., Milgrom (1981); Milgrom and Roberts (1986)), a sophisticated receiver recognizes the sender’s incentive to conceal unfavorable information, and so the receiver adopts a skeptical posture regarding the content of undisclosed information. In our belief-stable partial disclosure equilibria, the underlying sophisticated behavior is analogous: an increase in the Agency’s additive bias drives down the policy chosen absent disclosure.

**Proposition 10.** *The equilibrium disclosure interval,  $M(x^*, \alpha, i)$ , is*

1. *expanding in the Agency’s bias,  $i$ , when  $x^* \leq i$ , and contracting in  $i$  otherwise; and*
2. *expanding in the Agency’s preference state-dependence,  $\alpha$*

---

<sup>8</sup>If non-disclosure equilibrium and full-disclosure equilibrium exist, they are the least disclosing and the most disclosing equilibria correspondingly.

if and only if the equilibrium is belief-stable.

*Proof.* See Appendix. □

Proposition 10 characterizes the net effect of parameters on the information disclosed. The comparative statics with respect to the Agency’s preference state-dependence,  $\alpha$ , is straightforward. A higher  $\alpha$  implies that the Agency’s preferences are more closely aligned with the state-contingent goals of the Policymaker, reducing the conflict of interest. In any belief-stable equilibrium, this greater alignment unambiguously leads to a larger disclosure interval and, thus, to more information transmission.

The effect of the Agency’s additive bias,  $i$ , is more complex. First, the disclosure interval  $M(x^*, \alpha, i)$  is a function of both  $x^*$  and  $i$ ; a change in the Agency’s bias  $i$  exerts both a direct influence on the boundaries of the disclosure interval and an indirect influence mediated through the adjustment of the equilibrium policy  $x^*$ . Second, an increase in  $i$  can represent either a convergence of ex-ante preferences (when  $i < p_0^P$ ) or a divergence (when  $i \geq p_0^P$ ). In standard applications with sufficient symmetry (such as in the case with uniform prior and quadratic objective functions we studied above), the Policymaker’s ex-ante optimum,  $p_0^P$ , and the equilibrium non-disclosure policy,  $x^*$ , both lie on the same side of the Agency’s bias.<sup>9</sup> According to Proposition 10, this implies that greater ex-ante preference misalignment will result in *more* information disclosure. If  $i$  is sufficiently high or sufficiently low, the Agency discloses all states in equilibrium.

As established in Proposition 9,  $x^*$  always shifts to oppose the Agency’s bias. When  $x^*$  and  $p_0^P$  are on the same side of  $i$ , preference divergence makes the non-disclosure outcome more punishing for the Agency, leading to greater disclosure. This counterintuitive result is reversed, however, if the equilibrium geometry is such that the non-disclosure policy,  $x^*$ , and the ex-ante optimum,  $p_0^P$ , lie on opposite sides of the Agency’s bias  $i$  (i.e.  $(x^* - i) \cdot (p_0^P - i) < 0$ ). The next section establishes that the counterintuitive result is, indeed, the one that corresponds to many standard environments in the applied literature.

## 5.1 Preference Divergence and Disclosure

In this section, we first identify a class of games that share the same equilibrium geometry as the game in Section 3; then establish equilibrium properties of those games, including the response of the disclosure interval to changes in preference divergence; and, finally, provide a condition on game primitives that serves as a test for membership in this class that is readily evaluated for standard environments in the applied literature.

We first define a class of games  $\hat{\mathcal{G}} \subset \mathcal{G}$  that share the same equilibrium geometry of the game in Section 3. In doing so, we leverage the observation that equilibrium properties in games with no ex ante preference misalignment ( $i = p_0^P$ ) correspond systematically to equilibrium properties in otherwise equivalent games with preference misalignment. For convenience, we will now write a generic member of the set  $\mathcal{G}$  as  $G(i)$  to emphasize the dependence on  $i$ . Let  $\mathbb{G}_0 := \{G(i = p_0^P) \in \mathcal{G} : \text{the non-disclosure equilibrium is the unique}$

---

<sup>9</sup>E.g., if the Agency’s ex-ante preference is to the left of the Policymaker’s, the non-disclosure set is typically concentrated to the left of  $p_0^P$ , and the resulting equilibrium non-disclosure policy lies to the right of the Agency’s bias ( $x^* > i$ ).

belief-stable equilibrium of  $G(i = p_0^P)$ . Let  $\mathbb{G}(G'(i')) := \{G(i) : G(i') = G'(i')\}$  for  $G'(i') \in \mathcal{G}$ . Finally, let  $\hat{\mathcal{G}} := \bigcup_{\hat{G}(i) \in \mathbb{G}_0} \mathbb{G}(\hat{G}(i)) \subset \mathcal{G}$ .

A game's membership in  $\hat{\mathcal{G}}$  pins down both the geometry of all of its partial disclosure equilibria and, consequently, the relationship between preference divergence and disclosure in any belief-stable equilibrium.

**Proposition 11.** *For any game  $G(i) \in \hat{\mathcal{G}}$ ,*

1. *for any partial-disclosure equilibrium,  $(x^* - i) \cdot (p_0^P - i) \geq 0$ ;*
2. *disclosure interval  $M(x^*, \alpha, i)$  expands as the magnitude of ex-ante preference divergence,  $|i - p_0^P|$ , increases if and only if the equilibrium is belief-stable.*

*Proof.* See Appendix. □

For games in  $\hat{\mathcal{G}}$ , policy absent disclosure  $x^*$  and the ex-ante optimum policy  $p_0^P$  lie on the same side of the Agency's additive bias  $i$ . Consequently, for such games, belief-stability selects equilibria where, as preferences diverge,  $x^*$  becomes less attractive to the Agency, encouraging it to reveal a strictly larger set of outcomes. It implies that the regularity condition imposed by  $\hat{\mathcal{G}}$  is sufficient to generate unambiguously the prediction that greater preference conflict enhances, rather than undermines, verifiable information disclosure.

While a full-disclosure equilibrium may exist in  $G(i = p_0^P) \in \hat{\mathcal{G}}$ ,<sup>10</sup> by definition of  $\hat{\mathcal{G}}$ , it cannot be belief-stable. Recall that for the game analyzed in Section 3, at  $i = 0$  the non-disclosure equilibrium is belief-stable while the full-disclosure equilibrium is not (see Proposition 4). The following proposition demonstrates that this is not a knife-edge result, but, instead, the limit point of a continuous dynamic for all games in  $\hat{\mathcal{G}}$ .

**Proposition 12.** *In all  $G(i) \in \hat{\mathcal{G}}$ , there exists an  $\varepsilon$  neighborhood of  $p_0^P$  such that for all  $i$  in that neighborhood, the relationship between preference alignment and belief-stability is as follows:*

1. *For the least expansive partial-disclosure equilibrium, the extent of belief-stability ( $\varepsilon^*$ ) increases as the magnitude of ex-ante preference divergence  $|i - p_0^P|$  decreases.*
2. *For any other belief-stable equilibrium (including full-disclosure equilibrium), there exists a threshold  $\hat{\Delta}$  such that for all  $i$  such that  $|i - p_0^P| < \hat{\Delta}$ , the extent of belief-stability ( $\varepsilon^*$ ) decreases as the magnitude of ex-ante preference divergence  $|i - p_0^P|$  decreases.*

An immediate implication of Proposition 12 is that greater ex-ante preference alignment enhances the belief-stability of the *least* expansive equilibrium while decreasing the extent of belief-stability of any more expansive belief-stable equilibrium (where one exists). Thus, the equilibrium that provides least disclosure is a robust prediction in the belief-stability sense. Furthermore, it becomes the most belief-stable prediction as the Agency's additive bias  $i$  converges to the mean of the state distribution  $p_0^P$ . From Proposition 11, in this robust prediction, better ex-ante preference alignment in the sense of smaller  $|i - p_0^P|$  implies that the Agency discloses less information. This finding stands in sharp contrast to the

---

<sup>10</sup>Observe that  $G(i) \in \hat{\mathcal{G}}$  does not preclude  $G(i)$ 's satisfying single-crossing.

canonical predictions of cheap-talk models, where greater preference divergence further limits communication.<sup>11</sup>

The following remark, which focuses on distance-based utility functions, shows that the condition defining  $\hat{\mathcal{G}}$  admits standard environments familiar from applied literature (THIS WANTS CITES).

**Remark 1.** *Consider a game  $G(i)$  such that  $\alpha = 0$  and utilities are functions of  $|p - i|$  and  $|p - \omega|$ , respectively. Then  $G(i) \in \hat{\mathcal{G}}$  if and only if for all potential disclosure boundaries  $x > 0$  and all states  $\omega > x$ , the prior density  $f$  satisfies*

$$\frac{\omega - x}{\omega + x} < \frac{f(-\omega)}{f(\omega)} < \frac{\omega + x}{\omega - x}. \quad (4)$$

*Proof.* See Appendix. □

Condition 4 is easily interpretable— it restricts the density from being too skewed in the tails of the distribution (and so is satisfied by any prior that is symmetric around its mean, including the normal and the uniform distribution, which is used in the analysis in Section 3). Further, the continuity of the Agency’s best response implies that  $\hat{\mathcal{G}}$  will include games in which the Agency places positive weight on matching the state.

## 6 Robustness

In this section, we test the robustness of our core findings by relaxing two key assumptions: perfect message verifiability and the restriction to a binary disclosure choice. To maintain analytical tractability and isolate the impact of each modification, we conduct this analysis within the framework of the uniform prior and state-independent preferences model from Section 3. Finally, we assume, that  $i \geq 0$ ; the case of  $i \leq 0$  is symmetric.

### 6.1 Partial Verifiability

Throughout the preceding analysis, all messages observed by the Policymaker are assumed to be perfectly verifiable. That is, whenever the Agency discloses a state, the Policymaker can infer with certainty that the message accurately reflects the Agency’s observation. Formally, the Agency can send a message  $m = \omega$  if and only if it observes state  $\omega$ .

In this section, we relax this assumption by introducing partial verifiability, allowing the Agency to distort information. Specifically, the Agency may send a point message  $m = \tilde{\omega} \in \Omega$  that differs from the true state  $\omega$ . The Policymaker, in turn, has the ability to verify whether the received message accurately reflects the underlying state.

Verification occurs probabilistically: with probability  $q \in [0, 1]$ , the Policymaker receives a signal  $s(m)$  that indicates whether the reported message is truthful. If the message is truthful ( $m = \omega$ ), the Policymaker observes  $s(m) = \text{True}$ ; if the message is distorted ( $m \neq \omega$ ),

---

<sup>11</sup>In particular, in a cheap-talk communication game in which the Sender’s and Receiver’s preferences are like those we assume above, there exist informative equilibria if and only if  $|i|$  is small enough, and the precision of the messages is unequal and grows more so as  $|i|$  increases; and if the Sender’s additive bias  $i$  is too great, the unique equilibrium policy is  $p = p_0^P = 0$  for all states. See Supplemental Appendix.

the Policymaker observes  $s(m) = \text{False}$ . With probability  $1 - q$ , the verification mechanism is inconclusive, and the Policymaker receives no additional information, observing  $s(m) = \emptyset$ . When  $q = 1$ , all messages are perfectly verifiable, meaning that any distortion by the Agency is immediately detected by the Policymaker. Conversely, when  $q = 0$ , messages are never verifiable, and the Policymaker receives no information beyond the reported message itself.

Observe that in this game with partial verifiability, a full-disclosure equilibrium cannot exist. Consider an Agency with  $i \geq 0$  that observes the state realization  $\omega = \underline{\omega}$ . The Agency will refrain from distorting its message only if the expected policy implemented following a distortion is at least as far from its ideal point as the policy implemented when the true state is disclosed. This condition implies that the Policymaker must implement at least  $p = \underline{\omega}$  whenever verification is inconclusive.

However, since the probability of verification is independent of whether the message corresponds to the true state or is strategically distorted: in full-disclosure equilibrium all types including  $\omega \neq \underline{\omega}$  choose to disclose their state by sending  $m = \omega$ . The Policymaker, thus, must incorporate this uncertainty into their policy decision. Following an inconclusive verification outcome,  $s(m) = \emptyset$ , the Policymaker's sequentially optimal policy must be closer to  $i$  than to  $\underline{\omega}$ . This, in turn, implies that type  $\omega = \underline{\omega}$  strictly prefers to distort its message rather than disclose truthfully, contradicting the existence of a full-disclosure equilibrium.

The following proposition characterizes an equilibrium of this game:

**Proposition 13.** *There is a belief-stable equilibrium such that*

$$p^* = \begin{cases} m & \text{if } s(m) = T \\ \frac{i(i-y^*)}{i-y^*-1} & \text{if } s(m) = F \\ z^* & \text{if } s(m) = \emptyset \end{cases}$$

$$m^*(\omega) = \begin{cases} \omega & \text{if } \omega \in [y^*, 2 \cdot i - y^*] \cap [-1, 1] \\ \tilde{\omega} & \text{else} \end{cases}$$

where

$$y^* = \frac{i(1+q) - 1 + \sqrt{(1-i(1+q))^2 - 4i^2q}}{2}, \quad z^* = 0, \quad \text{and} \quad \tilde{\omega} \sim U[[y^*, 2 \cdot i - y^*] \cap [-1, 1]].$$

*Proof.* See Appendix. □

In this equilibrium, the Agency fully discloses its information when the realized state falls within the interval  $[i - \sqrt{(y-i)^2}, i + \sqrt{(y-i)^2}] \cap [-1, 1]$ . When the realized state lies outside this region, the Agency distorts its information, instead sending a message drawn from a uniform distribution over this interval, effectively mimicking the prior distribution.

Figure 6 illustrates the disclosure boundaries as a function of the Agency's ideal point  $i$  for different values of the verifiability parameter  $q$ .

The following proposition characterizes the comparative statics of the disclosure intervals with respect to the Agency's ideal point  $i$  and the verifiability parameter  $q$ .

**Proposition 14.** *The disclosure interval within which the Agency fully reveals the state to the Policymaker,  $[y^*, 2 \cdot i - y^*]$ , is expanding in  $i$  and in  $q$ .*

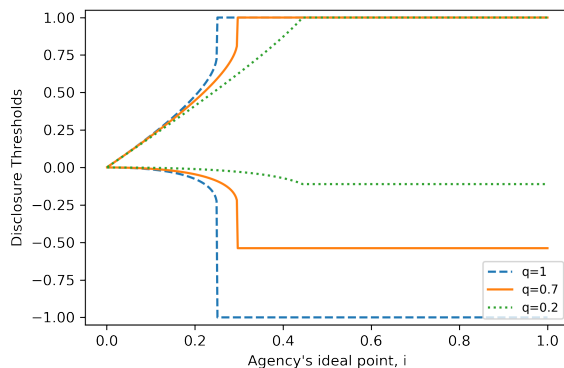


Figure 6: Disclosure intervals as a function of the Agency’s ideal point  $i$  for different levels of partial verifiability  $q$ .

*Proof.* See Appendix. □

As the Agency’s ideal point moves farther from the prior mean, the interval of disclosed states expands. Higher verifiability reduces the incentive to engage in misreporting, leading to a contraction in the set of states that remain undisclosed.

## 6.2 Agency’s Vagueness

Suppose next that the Agency can choose the precision of its message: Instead of disclosing  $\omega$  exactly, as in the main model, the Agency chooses message of the form “the true state is in  $T$ ,” where  $T \subseteq \Omega$  and  $\omega \in T$ . Assume that message  $T$  is verifiable, meaning that the Policymaker learns with certainty that the true state  $\omega$  is an element of  $T$ . The Agency’s message space is thus a set  $\{T \subseteq \Omega \mid \omega \in T\}$ . A message strategy for the Agency is a function  $m : \Omega \rightarrow \mathcal{P}(\Omega)$ , where  $\mathcal{P}(\Omega)$  is the power set of  $\Omega$ , such that for every  $\omega \in \Omega$ ,  $\omega \in m(\omega)$ . We continue to work with the assumptions of the special case:  $\Omega = [-1, 1]$ ,  $f(\cdot)$  is the uniform density on  $\Omega$ , and utility functions are  $u_P(p) = -(\omega - p)^2$  and  $u_A(p) = -(i - p)^2$ .

If the Agency sends a singleton message  $m(\omega) = T = \{\omega\}$ , we say the state  $\omega$  is fully disclosed. If  $m(\omega) = T \neq \{\omega\}$ , the message is vague. Upon receiving a message  $T$ , the Policymaker updates her beliefs via Bayes’ rule when applicable, conditional on her conjecture of the Agency’s strategy  $m(\cdot)$  and the prior  $f(\cdot)$ . If  $T = \{\omega\}$ , the Policymaker’s posterior belief assigns probability one to state  $\omega$ .

The following proposition characterizes a sequential equilibrium that corresponds to the belief-stable (less expansive) partial disclosure equilibrium of the game in Section 3:

**Proposition 15.** *The following strategy profile and system of beliefs can be supported in a Sequential Equilibrium:*

$$m^*(\omega) = \begin{cases} \{\omega\} & \text{if } \omega \in M_L \\ N_L & \text{if } \omega \notin M_L, \end{cases}$$

where  $M_L = \{\omega' \in \Omega \mid -(i - \omega')^2 \geq -(i - x_L)^2\}$ ,  $x_L$  is the policy absent disclosure in the less expansive equilibrium (Proposition 1), and  $N_L = \Omega \setminus M_L$ ;

$$p^*(T) = \mathbb{E}[\omega \mid T; m^*(\cdot), \beta(\cdot|T)],$$

where  $\beta(\cdot|T)$  are Policymaker's beliefs given by

$$\beta(\omega|T, m(\cdot)) = \begin{cases} 1 & T = \{\omega\} \\ \mathbb{1}_{\omega \in N_L} \cdot \frac{f(\omega)}{\int_{N_L} f(\hat{\omega}) d\hat{\omega}} & T = N_L \\ \mathbb{1}_{\omega = \hat{\omega}(T)} & T \in \mathcal{T}_{off}(\omega). \end{cases}$$

with  $\hat{\omega}(T) := \arg \max_{\omega \in T} |i - \omega|$ <sup>12</sup> and  $\mathcal{T}_{off}(\omega) := \{T : \omega \in T, T \neq \{\omega\}, T \neq N_L\}$ .

*Proof.* See Appendix. □

Note that the system of Policymaker's beliefs described in the proposition is skeptical in that any deviation is attributed to the type  $\hat{\omega}(T)$ , which is the attribution that is least favorable to the Agency. The construction of such beliefs is crucial for deterring deviations to arbitrary vague messages not specified on the equilibrium path.

Proposition 15 demonstrates that the equilibrium outcome characterized in Proposition 1 – the unique belief-stable partial disclosure equilibrium – is robust to a significant enlargement of the Agency's message space. The equilibrium assessment  $(m^*, p^*, \beta)$  retains the same partition of types into disclosure and non-disclosure sets, inducing the same policy absent disclosure. Consequently, the sequential equilibrium of the model with vague communication retains the key features and comparative statics of the belief-stable equilibrium, emphasizing that the mere availability of arbitrarily precise messages does not inherently improve communication.

## 7 Discussion

### 7.1 Policymaker's Optimal Choice of Agency

Policymakers in many institutional settings often have discretion over the selection of agents or advisors from whom they receive policy advice. The biases of these agents significantly influence the quality and nature of the information transmitted.

The standard intuition, borne of the extensive cheap-talk signaling literature, is that information-revelation increases with the preference proximity between the sender and receiver. A related idea is the benchmark ‘‘ally principle’’ from the delegation literature, which posits that principals should delegate authority to agents with co-aligned preferences (Bendor and Meirowitz, 2004). A substantial body of literature has shown that this expectation might fail, and a principal might prefer an agent with preferences divergent from her own, when agents' information is endogenous. Che and Kartik (2009) argue that differences in preferences between advisors and decision-makers create incentives for advisors to acquire

---

<sup>12</sup>If multiple such states exist, ties can be broken arbitrarily.

information, which can benefit decision-makers when the advisors' biases are moderate. Similarly, [Krishna and Morgan \(2001\)](#) demonstrate that the presence of another biased expert can enhance the decision-maker's ability to extract information from a biased advisor. [Gailmard and Patty \(2007\)](#) characterize scenarios where greater agent bias encourages the acquisition of expertise, as biased agents are more motivated to influence outcomes. [Prendergast \(2007\)](#) also shows that higher agent bias may motivate agents to exert more effort beyond what might be achieved via monetary incentives.

In contrast to the previous literature, we have shown that even when information is exogenous and readily available to the Agency, strategic policy implementation by the Policymaker is sometimes – depending on the nature of the equilibrium played by the Policymaker and the Agency – not enough to guarantee full disclosure. Our analysis suggests that, assuming that the Policymaker is not in a position to effect the selection of the full-disclosure equilibrium when partial-disclosure equilibria are possible, she may be better off selecting an Agency with sufficiently divergent ex ante policy preference. Doing so may guarantee that the only equilibrium possible is full-disclosure, and helps increase disclosure in the belief-stable partial-disclosure equilibria otherwise.

## 7.2 More Discussion TBA

# Appendix

## Lemma 1

Because information is verifiable, in case of disclosure, subsequent beliefs are independent of Agency strategy:

$$\Pr(\omega|m = \omega') = \begin{cases} 1 & \text{if } \omega = \omega' \\ 0 & \text{if } \omega \neq \omega'. \end{cases} \quad (5)$$

In contrast,  $p(\omega|\emptyset)$  is determined by Bayes Rule and depends on the Agency's disclosure strategy  $m(\omega)$ :

$$p(\omega|m = \emptyset) = \frac{\mathbb{1}(m(\omega) = \emptyset)p(\omega)}{\int_{-1}^1 \mathbb{1}(m(\omega') = \emptyset)p(\omega')d\omega'}, \quad (6)$$

where  $\mathbb{1}$  is an indicator function. For the quadratic utility function, the Policymaker's expected utility is maximized at  $p = \mathbb{E}[\omega|m]$ , thus equation 1 follows.

## Lemma 2

The Agency strictly prefers to disclose its information  $\omega$  to the Policymaker when  $p = \omega$  yields higher utility than  $p = x$ :

$$\begin{aligned} & -(i - \omega)^2 > -(i - x)^2, \\ \Leftrightarrow \omega \in & \begin{cases} (x, 2 \cdot i - x) & \text{if } x \leq i \\ (2 \cdot i - x, x) & \text{if } x \geq i. \end{cases} \end{aligned}$$

Because the support of the distribution if  $\omega$  is  $[-1, 1]$ , the Agency discloses  $\omega \in (x, 2 \cdot i - x) \cap [-1, 1]$  for  $i > x$  and does not disclose  $\omega \in [-1, 1] \setminus [x, 2 \cdot i - x]$ . The Agency is indifferent for  $\omega = x$  and  $\omega = 2 \cdot i - x$ . The argument for  $i \leq x$  case is symmetric.

Finally, note that because  $x = \mathbb{E}[\omega|\emptyset; \mu^*(\cdot)]$ , when  $i \geq x$ ,  $x < 0$ .

## Proposition 1

1. Suppose the Agency discloses all states it observes, and after non-disclosure ( $m = \emptyset$ ), the Policymaker selects policy  $x^* = -1$  if  $i > 0$ , and  $x^* = +1$  if  $i < 0$ , and selects either  $x^* = -1$  or  $x^* = +1$  if  $i = 0$ . Then, given  $x^*$ , Agency prefers  $p^* = \omega$  to  $p^* = x^* \forall \omega$ .

To establish that this strategy profile is supported in a sequential equilibrium (SE), it remains to show that beliefs supporting  $p^*(\emptyset) = x^*$  as an optimal choice for the Policymaker are fully consistent with the strategy profile. A sequence of completely mixed disclosure strategies  $\{\mu^k(\omega)\}_1^\infty$  can be constructed with

$$\mu^k(\omega) = \begin{cases} 1 - \varepsilon^k(\omega) & \text{if } \omega \in M(x^*, i) \\ \varepsilon^k(\omega) & \text{if } \omega \in [-1, 1] \setminus M(x^*, i). \end{cases} \quad (7)$$

If  $\{\varepsilon^k(\omega)\}_1^\infty \rightarrow 0$  for every  $\omega$ , then  $\{\mu^k(\omega)\}_1^\infty \rightarrow \mu^*(\omega)$  for every  $\omega \in \Omega$ . In particular, let  $\varepsilon^k(\omega')$  converge faster than  $\varepsilon^k(\omega'')$  for every  $\omega', \omega''$  s.t.  $\omega' > \omega''$ ; then  $\{x^{*k}\}_{k=1}^\infty := \{\mathbb{E}[\omega|m = \emptyset; \mu^k]\}_{k=1}^\infty$  converges to -1. Likewise, if  $\varepsilon^k(\omega')$  converges faster than  $\varepsilon^k(\omega'')$  for every  $\omega', \omega''$  s.t.  $\omega' < \omega''$ , it converges to +1. Thus, the full-disclosure equilibria are SE,

$$x^F := \begin{cases} 1, & \text{if } i \leq 0, \\ -1, & \text{if } i > 0 \end{cases}$$

2. Lemma 1 and Lemma 2 characterize Policymaker's and Agency's equilibrium behavior respectively.

Suppose  $i = 0$ . When  $x^* = 0$ , the Agency is indifferent between disclosing and not disclosing  $\omega = 0$ , and strictly prefers not disclosing ( $m = \emptyset$ ) for all  $\omega \neq 0$ . If the Agency does not disclose for all  $\omega \neq 0$ , then, regardless of  $\mu(0)$ ,  $x^* = \mathbb{E}[\omega|m = \emptyset, \mu^*] = 0$ . It remains to show that the Policymaker's beliefs are fully consistent. Because for all  $\omega$  beliefs following  $m = \omega$  are determined by the verifiability of information, and because  $m = \emptyset$  occurs with positive probability in equilibrium, there are no off-path-of-play information sets, and so beliefs are fully consistent with  $\mu^*(\cdot)$ , constituting SE.

Suppose  $i > 0$  (the case  $i < 0$  is symmetric). We assume the disclosure interval  $M(x, i) = [x, 2i - x]$  is a proper, non-empty subset of  $[-1, 1]$ . The equilibrium policy  $x^*$  must be a fixed point where:  $x^* = \mathbb{E}[\omega|m = \emptyset, \mu^*(\cdot)]$ . Assuming a uniform distribution where  $f(\omega) = 1/2$  and  $F(\omega) = \frac{\omega+1}{2}$ :

$$\begin{aligned} x &= \mathbb{E}[\omega|m = \emptyset, \mu^*(\cdot)] = \frac{\int_{-1}^x \omega \cdot f(\omega) d\omega + \int_{2i-x}^{+1} \omega \cdot f(\omega) d\omega}{F(\infty) - F(2 \cdot i - x) + F(x) - F(-1)} \\ &= -\frac{\int_x^{2i-x} \omega \cdot f(\omega) d\omega}{1 - F(2 \cdot i - x) + F(x)} = -\frac{\int_x^{2i-x} \omega/2 d\omega}{1 - (2 \cdot i - x + 1)/2 + (x + 1)/2} \\ &= \frac{(i - x) \cdot i}{i - x - 1} \end{aligned}$$

Thus,  $x^*$  solves

$$\begin{aligned} x &= \frac{(i - x) \cdot i}{i - x - 1} \\ &\Leftrightarrow \\ i \cdot x - x^2 - x &= i^2 - i \cdot x \\ &\Leftrightarrow \\ x^2 - x \cdot (-1 + 2 \cdot i) + i^2 &= 0 \end{aligned}$$

Thus,  $x^* \in \{x_M, x_L\}$  where

$$x_M := \frac{2 \cdot i - 1 - \sqrt{1 - 4 \cdot i}}{2}$$

and

$$x_L := \frac{2 \cdot i - 1 + \sqrt{1 - 4 \cdot i}}{2}.$$

These two solutions represent potential partial disclosure equilibria. One where the Policymaker chooses  $x_M$  absent disclosure and the other one where the Policymaker chooses  $x_L$ . These solutions are valid if and only if the underlying assumption — that the disclosure interval is a subset of  $[-1, 1]$  — holds. This requires  $x^* \in (-1, 0)$ , or, equivalently,  $\frac{2 \cdot i - 1 - \sqrt{1 - 4 \cdot i}}{2} > -1$  and  $\frac{2 \cdot i - 1 + \sqrt{1 - 4 \cdot i}}{2} < 0$ . These inequalities are satisfied when  $i \in [0, 1/4]$ . When  $i > 1/4$  neither less expansive nor more expansive equilibrium exists.

Because information is verifiable, and  $m = \emptyset$  occurs on the path of play, beliefs are fully consistent and hence these equilibria are sequential equilibria.

## Corollary 1

Suppose  $i \geq 0$ , the case of  $i \leq 0$  is omitted. Note that  $x_M \leq x_L \leq 0$ , and thus the disclosure intervals corresponding to these two equilibria are nested:

$$[x_L, 2 \cdot i - x_L] \subseteq [x_M, 2 \cdot i - x_M] \subseteq [-1, 1].$$

The Policymaker's ex ante expected utility increases in disclosure, thus, it is the highest with full-disclosure equilibrium, lowest with less expansive equilibrium.

## Proposition 2

1. From Proposition 1,  $x_F$  is a step function defined by

$$x_F = \begin{cases} 1, & i \leq 0; \\ -1, & i \geq 0. \end{cases} \quad (8)$$

Thus,  $x_F$  is constant for all  $i \neq 0$  and has a downward jump discontinuity at  $i = 0$  from 1 to  $-1$ .

2. Differentiating  $x_L$ , defined in Proposition 1, wrt  $i$

$$\frac{dx_L}{di} = 1 - \frac{1}{\sqrt{1 - \text{sign}(i) \cdot 4 \cdot i}} \leq 0 \text{ for } i \in (-1/4, 1/4), \quad (9)$$

$$\lim_{i \rightarrow \pm 1/4} \frac{dx_L}{di} = \lim_{i \rightarrow \pm 1/4} 1 - \frac{1}{\sqrt{1 - \text{sign}(i) \cdot 4 \cdot i}} < 0, \quad (10)$$

hence  $x_L$  is decreasing in  $i \in [-1/4, 1/4]$ .

3. Differentiating  $x_M$ , defined in Proposition 1, wrt  $i$

$$\frac{dx_M}{di} = 1 + \frac{1}{\sqrt{1 - 4 \cdot i}} \geq 0, \text{ for } i \in (0, 1/4), \quad (11)$$

$$\frac{dx_M}{di} = 1 + \frac{1}{\sqrt{1+4 \cdot i}} \geq 0, \text{ for } i \in (-1/4, 0),$$

and

$$\lim_{i \rightarrow \pm 1/4} \frac{dx_M}{di} = \lim_{i \rightarrow \pm 1/4} 1 + \frac{1}{\sqrt{1+4 \cdot i}} > 0.$$

Note that  $\lim_{i \rightarrow 0^+} x_M = -1$ , whereas  $\lim_{i \rightarrow 0^-} x_M = 1$ . Thus,  $x_M$  has a downward discontinuity at  $i = 0$ .

### Proposition 3

1. From Proposition 1, the disclosure interval in the full-disclosure equilibrium is  $M(x^*, i) = \Omega$  and does not depend on  $i$ .
2. Suppose  $i \geq 0$ , the case of  $i \leq 0$  is omitted. Then, the upper bounds of the disclosure intervals depend on  $i$  both directly and indirectly (via  $x$ ), and the lower bounds on  $i$  only indirectly (via  $x$ ).

Differentiating the upper bound,

$$\frac{d(2 \cdot i - x^*)}{di} = \underbrace{2}_{\text{direct effect}} \underbrace{-\frac{dx^*}{di}}_{\text{indirect effect}}.$$

From (11), in the more expansive equilibrium, the indirect effect of increasing  $i$  dominates the direct effect, lowering the upper bound of the disclosure interval.

$$\frac{d(2 \cdot i - x_M)}{di} = 2 - 1 - \frac{1}{\sqrt{1-4 \cdot i}} \leq 0.$$

From Proposition 2, the lower bound of the disclosure interval is (weakly) increasing; thus the Agency discloses less in the more expansive equilibrium as  $i$  increases.

3. In the less expansive equilibrium, both the direct and indirect effects of increasing  $i$  are aligned, resulting in an expansion of the disclosure interval. From Lemma 2 and (10), the derivative of the upper bound of the disclosure interval in the less expansive equilibrium is

$$\frac{d(2 \cdot i - x_L)}{di} = 2 - 1 + \frac{1}{\sqrt{1-4 \cdot i}} > 0.$$

This positive derivative indicates that as  $i$  increases, the upper threshold  $2i - x_L$  increases. From Proposition 2, the lower threshold  $x_L$  decreases with  $i$ , and so the disclosure interval expands as  $i$  increases in the less expansive equilibrium.

### Proposition 4

1. Suppose  $i \in (-1/4, 1/4)$ . Consider the less expansive SE. Policymaker chooses  $x_L = (2 \cdot i - 1 + \sqrt{1-4 \cdot i})/2$  absent disclosure. Note that Agency's best response to Policymaker's selection of policy  $x_0 \in [x_L - \varepsilon, x_L + \varepsilon]$  where  $\varepsilon = \sqrt{1-4 \cdot i}$  will be to disclose

states  $\omega$  such that  $\omega \in [x_0, 2 \cdot i - x_0]$ . Importantly, Policymaker's best response to this disclosure strategy is to select policy  $x_1$  such that

$$x_1 = -\frac{\int_{x_0}^{2 \cdot i - x_0} \omega \cdot f(\omega) d\omega}{1 - (2 \cdot i - x_0 + 1)/2 + (x_0 + 1)/2} = \frac{(i - x_0) \cdot i}{i - x_0 - 1}.$$

Note that  $x_0 \leq x_1$  when

$$x_0 \leq \frac{(i - x_0) \cdot i}{i - x_0 - 1} \Leftrightarrow x_0 \in \left[ \frac{2 \cdot i - 1 - \sqrt{1 - 4 \cdot i}}{2}, \frac{2 \cdot i - 1 + \sqrt{1 - 4 \cdot i}}{2} \right]. \quad (12)$$

and  $x_0$  exceeds  $x_1$  otherwise. Therefore, when  $\varepsilon = \sqrt{1 - 4 \cdot i}$  and  $x_0 \in [x_L - \varepsilon, x_L]$  policy  $x_0 \leq x_1$ . Further, when  $x_0 \in [x_L - \varepsilon, x_L]$  policy  $x_1$  cannot exceed  $x_L$ . Assume instead that  $x_1$  exceeds  $x_L$ :

$$\begin{aligned} x_1 &= \frac{(i - x_0) \cdot i}{i - x_0 - 1} > x_L = \frac{2 \cdot i - 1 + \sqrt{1 - 4 \cdot i}}{2} \\ x_0 &> \frac{2 \cdot i - 1 + \sqrt{1 - 4 \cdot i}}{2}, \end{aligned}$$

which contradicts expression (12). Therefore, when  $x_0 \in [x_L - \varepsilon, x_L]$ ,  $|x_L - x_0| < |x_L - x_1|$ . By similar logic, when  $x_0 \in [x_L, x_L + \varepsilon]$ ,  $|x_L - x_0| < |x_L - x_1|$ . Therefore,

$$\forall x_0 : x_0 \in [x_L - \varepsilon, x_L + \varepsilon], \varepsilon = \sqrt{1 - 4 \cdot i}, |x_L - x_0| < |x_L - x_1|, \quad (13)$$

and the less expansive equilibrium is belief-stable for all  $i \in (-1/4, 1/4)$ .

2. The more expansive equilibrium, in contrast, is belief-unstable. Note that for any policy  $x_0$  chosen absent disclosure from the interval  $[x_M, x_M + \varepsilon]$ , where  $\varepsilon = \sqrt{1 - 4 \cdot i}$ , the Agency responds by disclosing states  $\omega \in [x_0, 2 \cdot i - x_0]$ . The Policymaker's response to this is

$$x_1 = -\frac{\int_{x_0}^{2 \cdot i - x_0} \omega \cdot f(\omega) d\omega}{1 - (2 \cdot i - x_0 + 1)/2 + (x_0 + 1)/2} = \frac{(i - x_0) \cdot i}{i - x_0 - 1}.$$

By statement (13),  $x_1 > x_0$ . Now, let us consider  $x_0 \in [-1, x_M]$ . Because

$$x_1 = \frac{(i - x_0) \cdot i}{i - x_0 - 1}$$

and  $x_0 \leq x_1$  if and only if  $x_0 \in \left[ \frac{2 \cdot i - 1 - \sqrt{1 - 4 \cdot i}}{2}, \frac{2 \cdot i - 1 + \sqrt{1 - 4 \cdot i}}{2} \right]$ , when  $x_0 < x_M$  policy  $x_1 < x_0$ . Therefore, less expansive equilibrium is never belief-stable.

3. Finally, when equilibrium is fully revealing and  $i > 0$ , the Agency's best response to any policy  $x_0 \in [-1, x_M]$  is  $x_1 = \frac{(i - x_0) \cdot i}{i - x_0 - 1}$  s.t.  $x_1 \leq x_0$ . Therefore, fully revealing equilibrium will be belief-stable. When  $i = 0$ , policy  $x_M$  converges to  $-1$  and fully revealing equilibrium is belief-unstable.

4. Given definition, the extent of belief stability is the radius of the largest neighborhood around an equilibrium policy  $x^*$  such that for any policy  $x_0$  within that neighborhood that was produced by perturbed actors' beliefs, the best-response update  $x_1$  is closer to  $x^*$  than  $x_0$  was.

This region of convergence is bounded by the nearest unstable equilibrium. Let's analyze the case for  $i \in (0, 1/4)$ ; the case for  $i \in (-1/4, 0)$  is symmetric. The three equilibria are ordered on  $[-1, 0)$  as  $x_F = -1 < x_M < x_L < 0$ . The equilibrium at  $x_M$  is unstable.

For the stable equilibrium  $x_L$ , any initial perturbation in the interval  $(x_M, 0)$ <sup>13</sup> will lead to a sequence of updates converging to  $x_L$ . Thus, the extent of belief-stability for less expansive equilibrium,  $\varepsilon_L^*$ , is the distance between  $x_L$  and  $x_M$ .

$$\begin{aligned}\varepsilon_L^* &= \left| \frac{2i - 1 + \sqrt{1 - 4i}}{2} - \frac{2i - 1 - \sqrt{1 - 4i}}{2} \right| \\ &= \left| \frac{2\sqrt{1 - 4i}}{2} \right| = \sqrt{1 - 4i}.\end{aligned}$$

The extent of belief-stability for the less expansive equilibrium strictly increases as preference divergence  $|i|$  decreases.

In contrast, perturbation in beliefs leading to policy  $x_0$  will lead to convergence to  $x_F$  if and only if  $x_0 \in [-1, x_M)$ . Thus, the extent of belief-stability for full-disclosure equilibrium,  $\varepsilon_F^*$ , is

$$\begin{aligned}\varepsilon_F^* &= 1 + \frac{2i - 1 - \sqrt{1 - 4i}}{2} \\ &= \frac{2 + 2i - 1 - \sqrt{1 - 4i}}{2} = \frac{1}{2} \left( 1 + 2i - \sqrt{1 - 4i} \right).\end{aligned}$$

The extent of belief-stability for the full-disclosure equilibrium decreases as preference divergence  $|i|$  decreases.

Thus, as the Agency's and Policymaker's ex-ante preferences become more aligned ( $i \rightarrow 0$ ), the robustness of the less expansive equilibrium increases, while the robustness to belief perturbations of the full-disclosure equilibrium decreases.

## Proposition 5

1. We analyze the belief-stability of a full-disclosure equilibrium characterized by  $x^* = p^P(\underline{\omega})$ ; the analysis for  $x^* = p^P(\bar{\omega})$  is symmetric.

A full-disclosure equilibrium is belief-stable if small perturbations to the Policymaker's beliefs about the off-path policy do not lead to divergent best responses. Formally, it

<sup>13</sup>The interval  $(x_M, 0)$  is the basin of attraction for the stable equilibrium  $x_L$ .

is belief-stable iff

$$\lim_{\varepsilon \rightarrow 0^+} \frac{\hat{x}(p^P(\underline{\omega}) + \varepsilon) - \hat{x}(p^P(\underline{\omega}))}{\varepsilon} \leq 1. \quad (14)$$

When  $\hat{x}(x)$  is right-differentiable at  $p^P(\underline{\omega})$ , this condition simplifies to

$$\lim_{x \rightarrow p^P(\underline{\omega})^+} \frac{d\hat{x}}{dx} \leq 1.$$

The differentiability of  $\hat{x}(x)$  at  $p^P(\underline{\omega})$  is not guaranteed. If the non-disclosure set,  $N(x, \alpha, i)$ , is disconnected for  $x$  in a right-neighborhood of  $p^P(\underline{\omega})$ ,  $\hat{x}(x)$  will be discontinuous,  $\lim_{x \rightarrow p^P(\underline{\omega})^+} \hat{x}(x) \neq p^P(\underline{\omega})$ . Such discontinuity implies belief-instability. We, thus, focus on the case where  $\hat{x}(x)$  is right-differentiable, which requires that  $\exists \delta > 0$ : the non-disclosure interval  $\forall x \in [p^P(\underline{\omega}), p^P(\underline{\omega}) + \delta)$ , is a single interval  $N(x) = [\underline{\omega}, \omega_b(x)]$ , where  $\omega_b(x)$  is such that  $u_A(p^P(\omega_b(x)); \omega_b(x), \alpha, i) = u_A(x; \omega_b(x), \alpha, i)$ .

Define

$$K(x, y; \alpha, i) := \int_{\Omega} \frac{\partial}{\partial p} u_P(p; \omega) \Big|_{p=y} dF(\omega | \omega \in N(x, \alpha, i)). \quad (15)$$

The best response  $\hat{x}(x)$  is implicitly defined by  $K(x, \hat{x}(x); \alpha, i) = 0$ . Applying the Implicit Function Theorem yields

$$\frac{d\hat{x}}{dx} = - \frac{\partial K(x, y; \alpha, i) / \partial x}{\partial K(x, y; \alpha, i) / \partial y} \Big|_{y=\hat{x}(x)}. \quad (16)$$

We evaluate Equation 16 as  $x \rightarrow p^P(\underline{\omega})^+$ . The strict concavity of  $u_P$  ensures that the denominator,  $\partial K / \partial y$ , is always strictly negative.

$$\lim_{x \rightarrow p^P(\underline{\omega})^+} \frac{\partial K}{\partial y} \Big|_{y=\hat{x}(x)} = \frac{\partial^2 u_P(p; \omega)}{\partial p^2} \Big|_{p=p^P(\underline{\omega}), \omega=\underline{\omega}} < 0. \quad (17)$$

Next, we evaluate the numerator. Define

$$\begin{aligned} A(x, y, \alpha, i) &:= \int_{\omega \in N(x, \alpha, i)} \frac{\partial}{\partial p} u_P(p; \omega) \Big|_{p=y} f(\omega) d\omega \\ B(x, \alpha, i) &:= \int_{\omega \in N(x, \alpha, i)} f(\omega) d\omega. \end{aligned} \quad (18)$$

The equilibrium condition  $K = 0$  implies  $A(x = x^*, y = \hat{x}(x)) = 0$ . Therefore, the numerator of Equation 16 simplifies to

$$\frac{\partial K(x, y; \alpha, i)}{\partial x} \Big|_{y=\hat{x}(x)} = \frac{dA/dx \cdot B - dB/dx \cdot A}{B^2} \Big|_{y=\hat{x}(x)} = \frac{dA/dx}{B} \Big|_{y=\hat{x}(x)}. \quad (19)$$

As  $x \rightarrow p^P(\underline{\omega})$ , both numerator and denominator converge to zero. We apply L'Hôpital's Rule to determine  $\lim_{x \rightarrow p^P(\underline{\omega})^+} \frac{\partial K(x, y; \alpha, i)}{\partial x} \Big|_{y=\hat{x}(x)}$ .

$$\lim_{x \rightarrow p^P(\underline{\omega})^+} \frac{dB}{dx} = \lim_{x \rightarrow p^P(\underline{\omega})^+} f(\omega_b(x)) \cdot \frac{d\omega_b(x)}{dx}. \quad (20)$$

We analyze components of  $\frac{dA}{dx}$  separately, denoting

$$\begin{aligned} \lim_{x \rightarrow p^P(\underline{\omega})^+} \frac{dA}{dx} &= \lim_{x \rightarrow p^P(\underline{\omega})^+} \underbrace{\frac{\partial u_P}{\partial p} \Big|_{p = \hat{x}(x), \omega = \omega_b(x)} \cdot f(\omega_b(x)) \cdot \frac{d\omega_b}{dx}}_{\text{Term 1}} \\ &+ \lim_{x \rightarrow p^P(\underline{\omega})^+} \underbrace{\frac{d\hat{x}}{dx} \cdot \int_{\underline{\omega}}^{\omega_b(x)} \frac{\partial^2 u_P}{\partial p^2} \Big|_{p = \hat{x}(x)} f(\omega) d\omega}_{\text{Term 2}}. \end{aligned} \quad (21)$$

Given  $\partial u_p / \partial p$  approaches 0 as  $x$  approaches  $p^P(\underline{\omega})$ ,

$$\begin{aligned} \lim_{x \rightarrow p^P(\underline{\omega})^+} \frac{d}{dx} (\text{Term 1}) &= \lim_{x \rightarrow p^P(\underline{\omega})^+} \frac{d}{dx} \left( \frac{\partial u_P}{\partial p} \Big|_{p = \hat{x}(x), \omega = \omega_b(x)} \cdot f(\omega_b(x)) \cdot \frac{d\omega_b}{dx} \right) \\ &= \left( \frac{\partial^2 u_P}{\partial p \partial \omega} \cdot \frac{d\omega_b}{dx} + \frac{\partial^2 u_P}{\partial p^2} \cdot \frac{d\hat{x}}{dx} \right) \Big|_{p^P(\underline{\omega}), \underline{\omega}} \cdot f(\underline{\omega}) \cdot \frac{d\omega_b}{dx} \Big|_{p^P(\underline{\omega})}. \end{aligned} \quad (22)$$

$$\begin{aligned} \lim_{x \rightarrow p^P(\underline{\omega})^+} \frac{d}{dx} (\text{Term 2}) &= \lim_{x \rightarrow p^P(\underline{\omega})^+} \left( \frac{d}{dx} \left( \int_{\underline{\omega}}^{\omega_b(x)} \frac{\partial^2 u_P}{\partial p^2} \Big|_{p = \hat{x}(x)} \cdot f(\omega) d\omega \right) \right) \cdot \frac{d\hat{x}}{dx} \\ &+ \lim_{x \rightarrow p^P(\underline{\omega})^+} \left( \int_{\underline{\omega}}^{\omega_b(x)} \frac{\partial^2 u_P}{\partial p^2} \Big|_{p = \hat{x}(x)} \cdot f(\omega) d\omega \right) \cdot \frac{d^2 \hat{x}}{dx^2} \\ &= \lim_{x \rightarrow p^P(\underline{\omega})^+} \frac{\partial^2 u_P}{\partial p^2} \Big|_{\hat{x}(x), \omega_b} \cdot f(\omega_b) \cdot \frac{d\omega_b}{dx} \cdot \frac{d\hat{x}}{dx} + 0 \\ &= \frac{\partial^2 u_P}{\partial p^2} \Big|_{p^P(\underline{\omega}), \underline{\omega}} \cdot f(\underline{\omega}) \cdot \frac{d\omega_b}{dx} \Big|_{p^P(\underline{\omega})} \cdot \frac{d\hat{x}}{dx} \Big|_{p^P(\underline{\omega})}. \end{aligned} \quad (23)$$

Therefore,

$$\begin{aligned} \lim_{x \rightarrow p^P(\underline{\omega})^+} \frac{\partial K}{\partial x} \Big|_{y = \hat{x}(x)} &= \frac{\left( \frac{\partial^2 u_P}{\partial p \partial \omega} \cdot \frac{d\omega_b}{dx} + 2 \cdot \frac{\partial^2 u_P}{\partial p^2} \cdot \frac{d\hat{x}}{dx} \right) \Big|_{p^P(\underline{\omega}), \underline{\omega}} \cdot f(\underline{\omega}) \cdot \frac{d\omega_b}{dx} \Big|_{p^P(\underline{\omega})}}{f(\underline{\omega}) \cdot \frac{d\omega_b}{dx} \Big|_{p^P(\underline{\omega})}} \\ &= \frac{\partial^2 u_P}{\partial p \partial \omega} \Big|_{p^P(\underline{\omega}), \underline{\omega}} \cdot \frac{d\omega_b}{dx} \Big|_{p^P(\underline{\omega})} + 2 \cdot \frac{\partial^2 u_P}{\partial p^2} \Big|_{p^P(\underline{\omega}), \underline{\omega}} \cdot \frac{d\hat{x}}{dx} \Big|_{p^P(\underline{\omega})}. \end{aligned} \quad (24)$$

Substituting (24) into the implicit function formula gives an equation for the slope  $\frac{d\hat{x}}{dx} \Big|_{p^P(\underline{\omega})}$ :

$$\begin{aligned} \frac{d\hat{x}}{dx} \Big|_{p^P(\underline{\omega})} &= - \frac{\left( \frac{\partial^2 u_P}{\partial p \partial \omega} \cdot \frac{d\omega_b}{dx} + 2 \cdot \frac{\partial^2 u_P}{\partial p^2} \cdot \frac{d\hat{x}}{dx} \right) \Big|_{p^P(\underline{\omega}), \underline{\omega}}}{\frac{\partial^2 u_P}{\partial p^2} \Big|_{p^P(\underline{\omega}), \underline{\omega}}} \\ &= \left( - \frac{\frac{\partial^2 u_P}{\partial p \partial \omega} \cdot \frac{d\omega_b}{dx}}{\frac{\partial^2 u_P}{\partial p^2}} - 2 \cdot \frac{d\hat{x}}{dx} \right) \Big|_{p^P(\underline{\omega}), \underline{\omega}}. \end{aligned} \quad (25)$$

Solving for the derivative yields the expression for the slope of the best-response function at the boundary:

$$\left. \frac{d\hat{x}}{dx} \right|_{p^P(\underline{\omega})} = -\frac{1}{3} \frac{\frac{\partial^2 u_P}{\partial p \partial \omega}}{\frac{\partial^2 u_P}{\partial p^2}} \cdot \left. \frac{d\omega_b}{dx} \right|_{p^P(\underline{\omega}), \underline{\omega}},$$

where  $\omega_b(x)$  is defined by the Agency's indifference  $u_A(p = x; \omega = \omega_b, \alpha, i) = u_A(p = p^P(\omega_b); \omega = \omega_b, \alpha, i)$ . Then

$$\begin{aligned} \frac{d\omega_b(x)}{dx} &= -\frac{\partial(u_A(x; \omega_b, \alpha, i) - u_A(p^P(\omega_b); \omega_b, \alpha, i))/\partial x}{\partial(u_A(x; \omega_b, \alpha, i) - u_A(p^P(\omega_b); \omega_b, \alpha, i))/\partial \omega_b} \\ &= -\frac{\partial u_A/\partial p|_{p=x, \omega=\omega_b}}{\partial u_A/\partial \omega|_{p=x, \omega=\omega_b} - (\partial u_A/\partial p \cdot \frac{dp^P(\omega_b)}{d\omega_b} + \partial u_A/\partial \omega)|_{p=p^P(\omega_b), \omega=\omega_b}} \end{aligned} \quad (26)$$

As  $x$  approaches  $p^P(\underline{\omega})$ ,  $\omega_b$  approaches  $\underline{\omega}$ , thus

$$\left. \frac{d\omega_b(x)}{dx} \right|_{p^P(\underline{\omega})} = -\frac{\partial u_A/\partial p|_{p=p^P(\omega_b), \omega=\omega_b}}{-\partial u_A/\partial p|_{p=p^P(\omega_b), \omega=\omega_b} \cdot \frac{dp^P(\underline{\omega})}{d\omega} + \frac{dp^P(\underline{\omega})}{d\omega}} = \frac{1}{\frac{dp^P(\underline{\omega})}{d\omega}} \quad (27)$$

Given  $p^P(\underline{\omega}) : \frac{\partial u_P(p; \omega)}{\partial p} = 0$ , we have  $\frac{d}{d\omega} \frac{\partial u_P(p; \omega)}{\partial p} = \frac{\partial^2 u_P(p; \omega)}{\partial p^2} \cdot \frac{dp^P(\underline{\omega})}{d\omega} + \frac{\partial^2 u_P(p; \omega)}{\partial \omega \partial p} \cdot \frac{d\omega}{d\omega} = 0$ . Therefore,

$$\left. \frac{d\hat{x}}{dx} \right|_{p^P(\underline{\omega})} = 1/3 \leq 1 \quad (28)$$

and full-disclosure equilibrium is belief-stable at  $x^* = p^P(\underline{\omega})$  if  $\hat{x}(\cdot)$  is differentiable.

Finally, assume, contrary to the proposition, that a full-disclosure equilibrium exists and that both of the following conditions hold

$$u_A(p^P(\bar{\omega}); \underline{\omega}, \alpha, i) > u_A(p^P(\underline{\omega}); \underline{\omega}, \alpha, i), \quad (29)$$

$$u_A(p^P(\underline{\omega}); \bar{\omega}, \alpha, i) > u_A(p^P(\bar{\omega}); \bar{\omega}, \alpha, i). \quad (30)$$

In a full-disclosure equilibrium, the Agency discloses the state  $\omega$  for all  $\omega \in \Omega$ . Since the Agency's utility  $u_A$  is strictly concave in the policy, for any interior policy absent disclosure  $x \in (p^P(\underline{\omega}), p^P(\bar{\omega}))$ , the set of states  $\omega$  for which the Agency's ideal point  $p^A(p^P(\omega), \alpha, i)$  is closer to  $x$  than to  $p^P(\omega)$  would be a non-empty set. Therefore, any policy  $x^*$  that sustains a full-disclosure equilibrium must be  $x^* \in \{p^P(\underline{\omega}), p^P(\bar{\omega})\}$ .

Suppose  $x^* = p^P(\underline{\omega})$ . Note that the Agency observing the state  $\bar{\omega}$  prefers to conceal the state given Inequality 30. The case of  $x^* = p^P(\bar{\omega})$  is symmetric, thus no full-disclosure equilibrium can exist.

2. The proof of sufficiency for the single-crossing condition is provided by Seidmann and Winter (1997) and is omitted.

## Proposition 6

1. If the Agency never discloses, the Policymaker's posterior belief remains her prior. By definition, the Policymaker's ex-ante optimal policy is  $p_0^P = \arg \max_p \mathbb{E}[u_P(p; \omega)]$ .

For the Agency's strategy to be a best response, it must prefer the non-disclosure outcome,  $p_0^P$ , to the disclosure outcomes,  $p = p^P(\omega)$ , for all possible states  $\omega$ . The incentive compatibility condition is

$$u_A(p_0^P; \omega, \alpha, i) \geq u_A(p^P(\omega); \omega, \alpha, i) \quad \forall \omega \in \Omega.$$

Since the Agency's utility  $u_A$  is strictly concave in the policy  $p$  and maximized at its ideal point  $p^A(\omega, \alpha, i) = \alpha p^P(\omega) + (1 - \alpha)i$ , unless  $i = p_0^P$  the IC condition is violated for all  $\omega \in [\min\{p_0^P, i\}, \max\{i, p_0^P\}]$ . Thus,  $i = p_0^P$ .

Finally, note that as  $u_A(p^P(\omega); \omega, \alpha, i)$  increases in  $\alpha$ , thus there exists  $\alpha^*$  such that the IC holds if and only if  $\alpha \leq \alpha^*$  and is violated otherwise.

2. For an interior equilibrium, the belief-stability condition  $\frac{d\hat{x}}{dx}|_{x=x^*} \leq 1$  is equivalent to the condition derived from the Implicit Function Theorem

$$\frac{\partial K(x, y; \alpha, i = p_0^P)}{\partial x} \Big|_{x=p_0^P, y=p_0^P} + \frac{\partial K(x, y; \alpha, i = p_0^P)}{\partial y} \Big|_{x=p_0^P, y=p_0^P} \leq 0.$$

We evaluate this condition for the non-disclosure equilibrium, where  $x = p_0^P, i = p_0^P$ . First, consider the partial derivative with respect to  $x$ . From expression 19

$$\frac{\partial K(x, y; \alpha, i)}{\partial x} \Big|_{y=\hat{x}(x)} = \frac{dA/dx \cdot B - dB/dx \cdot A}{B^2} \Big|_{y=\hat{x}(x)} = \frac{dA/dx}{B} \Big|_{y=\hat{x}(x)}, \quad (31)$$

where  $A$  and  $B$  are defined in 18. Given  $B$  is positive, the sign of this derivative is characterized by the derivative of the expression  $A$ . By the Leibniz Integral Rule

$$\frac{\partial A}{\partial x} \Big|_{x=p_0^P} = \pm \frac{\partial}{\partial p} u_P(p; \omega = p_0^P) \Big|_{p=p_0^P} \cdot f(p_0^P) \cdot \frac{\partial}{\partial x} p_0^P. \quad (32)$$

Second, consider the partial derivative with respect to  $y$ . Given the strict concavity of  $u_P(\cdot)$  in  $p$ , this term is strictly negative.

$$\frac{\partial K}{\partial y} \Big|_{x=p_0^P, y=p_0^P} = \mathbb{E} \left[ \frac{\partial^2 u_P(p; \omega)}{\partial p^2} \Big|_{p=p_0^P} \Big| \omega \in N(p_0^P) \right] < 0.$$

By the model's definition, the Policymaker's ideal policy is  $p^P(\omega) = \omega$ . This implies that the Policymaker's marginal utility is zero whenever the policy matches the state. Therefore,  $\frac{\partial A}{\partial x} \Big|_{x=p_0^P} = 0$  and the Non-Disclosure Equilibrium, if it exists, is belief-stable.

## Proposition 7

The proof proceeds in two main steps. First, we establish the existence and uniqueness of the threshold  $\alpha^{**}$ . Second, for any  $\alpha \leq \alpha^{**}$ , we establish the existence of a bounded interval  $I^*(\alpha)$  containing  $p_0^P$ .

A partial disclosure equilibrium exists if there is a non-disclosure policy  $x \in \Omega$  and a non-empty, proper subset of states  $M \subset \Omega$  such that: (i)  $u_A(p^P(\omega); \omega, \alpha, i) \geq u_A(x; \omega, \alpha, i) \forall \omega \in M$ ,  $u_A(p^P(\omega); \omega, \alpha, i) \leq u_A(x; \omega, \alpha, i) \forall \omega \notin M$  and (ii)  $x = \arg \max_p \mathbb{E}[u_P(p; \omega) | \omega \notin M]$ . The Agency's decision to disclose depends on the sign of the net gain from full disclosure of state  $\omega$ , defined as  $\Delta(x, \omega; \alpha, i) := u_A(p^P(\omega); \omega, \alpha, i) - u_A(x; \omega, \alpha, i)$ . A partial disclosure equilibrium is possible only if there exists an  $x$  such that the set  $M = \{\omega \in \Omega \mid \Delta(x, \omega; \alpha, i) \geq 0\}$  is nonempty and is not equal to  $\Omega$ .

We first show that  $M$  expands monotonically in  $\alpha$ . Note that because  $u_A$  is strictly concave, for all  $\omega$  and  $x$  such that  $(p^A - p^P(\omega)) \cdot (p^A - p^P(x)) > 0$ , the Agency discloses a fixed state  $\omega$  over a fixed induced policy  $x$  when  $p^P(\omega)$  is close to  $p^A$  than  $x$ . As  $\alpha$  increases, the Agency's ideal point  $p^A(p^P(\omega), \alpha, i)$  moves closer to  $p^P(\omega)$ . Thus, if a partial disclosure fails to exist for some  $(\alpha_0, i)$ , it must also fail to exist for all  $(\alpha, i)$  such that  $\alpha > \alpha_0$ . Similarly, if a partial disclosure exists for some  $(\alpha_0, i)$ , it must also exist for all  $(\alpha, i)$  such that  $\alpha \leq \alpha_0$ .

Now consider the limiting cases. At  $\alpha = 1$ , the Agency's ideal point is  $p^A = p^P(\omega)$ . For any  $x \neq p^P(\omega)$ ,  $\Delta_A(x, \omega; 1, i) > 0$ . The Agency strictly prefers to disclose every state rather than have any policy  $x \neq p^P(\omega)$  implemented. The partial disclosure is not sustainable. Conversely, at  $\alpha = 0$  there exists  $i$  for which the set of non-disclosure is non-empty. In particular, if  $i = p_0^P$ , a partial disclosure equilibrium can be sustained by the non-disclosure policy  $x = p_0^P$ . Since a partial disclosure equilibrium exists for  $\alpha = 0$  (for some  $i$ ) but not for  $\alpha = 1$  (for all  $i$ ), and  $M$  is monotonically expanding in  $\alpha$ , there must exist a unique threshold  $\alpha^{**} \in (0, 1)$  such that a partial disclosure equilibrium exists for some  $i$  if and only if  $\alpha \leq \alpha^{**}$ .

Fix any  $\alpha \in [0, \alpha^{**}]$ . By definition of  $\alpha^{**}$ , there exist at least one  $i$  that supports a partial disclosure equilibrium. Finally, monotonicity of  $M$  in  $\alpha$  also implies that if the condition for the existence of NDE is met, then the condition for the existence of PDE must also be met. Thus,  $\alpha^{**} > \alpha^*$ .

Let  $I^*(\alpha)$  be the set of all such  $i$ . We must show that this non-empty (by definition) interval is bounded and  $p_0^P \in I^*(\alpha)$ . Given Agency's utility function is concave, for every  $\alpha$  there exist unique ideal points  $\underline{I}(\alpha)$  and  $\bar{I}(\alpha)$  such that an Agency with ideal point  $\underline{I}(\alpha)$  (respectively,  $\bar{I}(\alpha)$ ) is indifferent between inducing the policy  $p_0^P$  (through non-disclosure) and inducing policy  $p^P(\underline{\omega})$  (respectively,  $p^P(\bar{\omega})$ ) through disclosure of the boundary state. That is,  $u_A(p_0^P; \underline{\omega}, \alpha, \underline{I}(\alpha)) = u_A(p^P(\underline{\omega}); \underline{\omega}, \alpha, \underline{I}(\alpha))$  and  $u_A(p_0^P; \bar{\omega}, 0, \bar{I}) = u_A(p^P(\bar{\omega}); \bar{\omega}, 0, \bar{I})$ .

These critical ideal points define an interval  $I^*(\alpha) \subseteq [\underline{I}(\alpha), \bar{I}(\alpha)] \subset \Omega$ . For any Agency's ideal point  $i \in I^*(\alpha)$ , a partial disclosure equilibrium can be sustained. Finally, while policy  $p_0^P$  is not necessarily the optimal policy absent disclosure, if the Agency cannot conceal boundary states even when induced policy absent disclosure is  $p_0^P$ , it never conceals states in equilibrium.

## Proposition 8

1. Let  $(x_j^*, M_j^*)$  and  $(x_k^*, M_k^*)$  constitute two distinct partial disclosure equilibria, where  $M_j^* = M(x_j^*, \alpha, i)$  and  $M_k^* = M(x_k^*, \alpha, i)$ .

Assume that  $u_A(x_j^*; \omega_0, \alpha, i) \leq u_A(x_k^*; \omega_0, \alpha, i)$  but  $M_k^* \not\subseteq M_j^*$ . Then there exists  $\omega_0$  s.t.  $\omega_0 \in M_k^* \setminus M_j^*$ . This implies:  $u_A(p^P(\omega_0); \omega_0, \alpha, i) \geq u_A(x_k^*; \omega_0, \alpha, i)$  and  $u_A(p^P(\omega_0); \omega_0, \alpha, i) < u_A(x_j^*; \omega_0, \alpha, i)$ . Thus,  $u_A(x_j^*; \omega_0, \alpha, i) > u_A(p^P(\omega_0); \omega_0, \alpha, i) \geq u_A(x_k^*; \omega_0, \alpha, i)$ . This simplifies to  $u_A(x_j^*; \omega_0, \alpha, i) > u_A(x_k^*; \omega_0, \alpha, i)$ , presenting a contradiction.

The proof for the converse implication follows a symmetric argument and is omitted.

2. Since the Agency's utility function  $u_A$  is continuous in all its arguments, the boundaries of the set  $N(x, \alpha, i)$  are continuous functions of the hypothetical policy  $x \in (\underline{\omega}, \bar{\omega})$ . Given  $\omega$  is drawn from a continuous distribution  $F$ , the integral defining  $N(\cdot)$  is a continuous function of its (continuously moving) boundaries. Therefore,  $\hat{x}(x)$  is continuous for all  $x \in (\underline{\omega}, \bar{\omega})$ .

Let the set of all equilibrium non-disclosure policies, including any boundary full-disclosure equilibria ( $X^*$ ) be strictly ordered. An equilibrium  $x_j^*$  is belief-stable if, in a neighborhood of  $x_j^*$ , the graph of  $\hat{x}(x)$  crosses the 45-degree line from above. It is belief-unstable if it crosses from below. Given continuity of  $\hat{x}(x)$  for all  $x \in (\underline{\omega}, \bar{\omega})$ , the elements of  $X^*$  must alternate in their belief-stability properties.

## Proposition 9

1. Consider an equilibrium characterized by a policy  $x^*$  absent disclosure. By the implicit function theorem and given equation 15,

$$\frac{\partial x^*}{\partial i} = - \frac{\partial K(x, y; \alpha, i) / \partial i |_{x=x^*, y=x^*}}{\partial K(x, y; \alpha, i) / \partial x |_{x=x^*, y=x^*}}. \quad (33)$$

We first determine the sign of the numerator,  $\partial K(x, y; \alpha, i) / \partial i |_{x=x^*, y=x^*}$ .

$$\begin{aligned} \partial K(x, y; \alpha, i) / \partial i |_{x=x^*, y=x^*} &= \frac{\partial}{\partial i} \int_{\Omega} \frac{\partial}{\partial p} u_P(p; \omega) \Big|_{p=x^*} dF(\omega | \omega \in N(x^*, \alpha, i)) = \\ &= \frac{\partial}{\partial i} \frac{\int_{\omega \in N(x^*, \alpha, i)} \frac{\partial}{\partial p} u_P(p; \omega) |_{p=x^*} dF(\omega)}{\int_{\omega \in N(x^*, \alpha, i)} dF(\omega)} \end{aligned} \quad (34)$$

Given Equations 18

$$\frac{\partial}{\partial i} K(x, y; \alpha, i) \Big|_{x=x^*, y=x^*} = \frac{\partial}{\partial i} \frac{A}{B} = \frac{\partial A / \partial i \cdot B - \partial B / \partial i \cdot A}{B^2}, \quad (35)$$

where expressions for  $A$  and  $B$  are defined in 18. Because  $A(x = x^*, y = \hat{x}(x)) = 0$ , the sign of  $\partial K(x, y; \alpha, i) / \partial i$  at equilibrium is determined by the sign of  $\partial A / \partial i$ . By Leibniz Integral Rule, when  $M(x^*, \alpha, i) = [x^*, \bar{M}(x^*, \alpha, i)]$

$$\partial A/\partial i = -\frac{\partial}{\partial p}u_P(p; \omega = \overline{M}(x^*, \alpha, i))|_{p=x^*} \cdot f(\overline{M}(x^*, \alpha, i)) \cdot \frac{\partial}{\partial i}\overline{M}(x^*, \alpha, i) < 0, \quad (36)$$

where  $\frac{\partial}{\partial p}u_P(x^*, \overline{M}(x^*, \alpha, i)) > 0$  follows  $u_P(\cdot)$  concavity; and given concavity of  $u_A(\cdot)$  holds, an increase in the Agency's ideal point  $i$  shifts its indifference points outwards, so  $\partial \overline{M}/\partial i > 0$ . Alternatively, if the disclosure interval is  $M(x^*, \alpha, i) = [\underline{M}(x^*, \alpha, i), x^*]$ ,

$$\partial A/\partial i = \frac{\partial}{\partial p}u_P(p; \omega = \underline{M}(x^*, \alpha, i))|_{p=x^*} \cdot f(\underline{M}(x^*, \alpha, i)) \cdot \frac{\partial}{\partial i}\underline{M}(x^*, \alpha, i) < 0, \quad (37)$$

where  $\frac{\partial}{\partial p}u_P(x^*, \underline{M}(x^*, \alpha, i)) < 0$  follows  $u_P(\cdot)$  concavity and  $\frac{\partial}{\partial i}\underline{M}(x^*, \alpha, i) > 0$  given Agency's objective function satisfies concavity. Therefore

$$\partial K(x, y; \alpha, i)/\partial i|_{x=x^*, y=x^*} < 0 \quad (38)$$

and

$$\text{sign} \frac{\partial x^*}{\partial i} = \text{sign} \left. \frac{\partial K(x, y; \alpha, i)}{\partial x} \right|_{x=x^*, y=x^*}. \quad (39)$$

By the Implicit Function Theorem

$$\frac{d\hat{x}(x)}{dx} = -\frac{\left. \frac{\partial K(x, y; \alpha, i)}{\partial x} \right|_{y=\hat{x}(x)}}{\left. \frac{\partial K(x, y; \alpha, i)}{\partial y} \right|_{y=\hat{x}(x)}}. \quad (40)$$

Given  $\hat{x}(x)$  is optimal given no disclosure,  $\left. \frac{\partial K(x, y; \alpha, i)}{\partial y} \right|_{y=\hat{x}(x)} < 0$ . Therefore,

$$\begin{cases} \frac{d\hat{x}(x)}{dx} \leq 1, & \frac{\partial K(x, y=\hat{x}(x); \alpha, i)}{\partial x} + \frac{\partial K(x, y=\hat{x}(x); \alpha, i)}{\partial y} \leq 0, \\ \frac{d\hat{x}(x)}{dx} > 1, & \frac{\partial K(x, y=\hat{x}(x); \alpha, i)}{\partial x} + \frac{\partial K(x, y=\hat{x}(x); \alpha, i)}{\partial y} > 0. \end{cases} \quad (41)$$

Finally, by the chain rule

$$\partial K(y = x^*, x = x^*; \alpha, i)/\partial x^* = \left. \frac{\partial K(x, y; \alpha, i)}{\partial y} \right|_{y=x^*, x=x^*} + \left. \frac{\partial K(x, y; \alpha, i)}{\partial x} \right|_{y=x^*, x=x^*}. \quad (42)$$

Combining equations 42, 41, and 39, belief-stable equilibria exhibit  $\partial x^*/\partial i \leq 0$  (non-increasing  $x^*$  as  $i$  increases), while belief-unstable equilibria exhibit  $\partial x^*/\partial i \geq 0$ .

2. Consider function  $K(\cdot)$  defined at 15. By the implicit function theorem

$$\frac{\partial x^*}{\partial \alpha} = -\frac{\left. \frac{\partial K(x, y; \alpha, i)}{\partial \alpha} \right|_{x=x^*, y=x^*}}{\left. \frac{\partial K(x, y; \alpha, i)}{\partial x} \right|_{x=x^*, y=x^*}}. \quad (43)$$

Following logic equivalent to that in part 1 of Proposition 9, the sign of  $\frac{\partial x^*}{\partial \alpha}$  in belief-stable equilibria is determined by the sign of  $\partial A/\partial \alpha$  (for the definition of A see expression 18). If the disclosure interval is  $M(x^*, \alpha, i) = [x^*, \bar{M}(x^*, \alpha, i)]$ ,

$$\partial A/\partial \alpha = -\frac{\partial}{\partial p} u_P(p; \omega = \bar{M}(x^*, \alpha, i)) \Big|_{p=x^*} \cdot f(\bar{M}(x^*, \alpha, i)) \cdot \frac{\partial}{\partial \alpha} \bar{M}(x^*, \alpha, i) < 0, \quad (44)$$

where  $\frac{\partial}{\partial \alpha} \bar{M}(x^*, \alpha, i) > 0$  follows monotonicity argument from Proposition 7. If  $M(x^*, \alpha, i) = [\underline{M}(x^*, \alpha, i), x^*]$ ,

$$\partial A/\partial \alpha = \frac{\partial}{\partial p} u_P(p; \omega = \underline{M}(x^*, \alpha, i)) \Big|_{p=x^*} \cdot f(\underline{M}(x^*, \alpha, i)) \cdot \frac{\partial}{\partial \alpha} \underline{M}(x^*, \alpha, i) > 0, \quad (45)$$

where  $\frac{\partial}{\partial \alpha} \underline{M}(x^*, \alpha, i) < 0$  follows monotonicity argument from Proposition 7.

## Proposition 10

1. Consider the case where the equilibrium non-disclosure policy satisfies  $x^* \leq i$ ; the argument for  $x^* > i$  is symmetric. In this case, the disclosure interval is  $M(i) = [x^*, \bar{M}(x^*, \alpha, i)]$ , where the upper boundary  $\bar{M}$  is defined by the Agency's indifference. From Proposition 9, in any belief-stable partial disclosure equilibrium, the equilibrium non-disclosure policy  $x^*$  is a strictly decreasing function of the Agency's bias  $i$ . Therefore,  $\frac{dx^*}{di} < 0$ . The lower boundary strictly decreases.

The upper boundary  $\bar{M}$  is a function of both the equilibrium policy  $x^*$  and the bias  $i$ . Its total derivative with respect to  $i$  is given by the chain rule:

$$\frac{d\bar{M}}{di} = \underbrace{\frac{\partial \bar{M}}{\partial i}}_{\text{Direct Effect}} + \underbrace{\frac{\partial \bar{M}}{\partial x^*} \cdot \frac{dx^*}{di}}_{\text{Indirect Effect}}.$$

Holding  $x^*$  constant, an increase in bias  $i$  shifts the Agency's ideal point further from  $x^*$ , making it more willing to disclose states far from  $x^*$ . Therefore,  $\frac{\partial \bar{M}}{\partial i} > 0$ . Holding  $i$  constant, a decrease in the non-disclosure policy  $x^*$  makes non-disclosure a worse outside option for the Agency, strengthening its incentive to disclose. Thus,  $\frac{\partial \bar{M}}{\partial x^*} < 0$ .

The direct and indirect effects are mutually reinforcing, causing the upper boundary to strictly increase:  $\frac{d\bar{M}}{di} > 0$ . Since the lower boundary  $x^*$  strictly decreases and the upper boundary  $\bar{M}$  strictly increases with  $i$ , it follows that for any  $i_2 > i_1$ , we have  $M(i_1) \subset M(i_2)$ . The disclosure interval is strictly expanding in  $i$ .

2. Consider the disclosure interval  $M = [x^*, \bar{M}(x^*(\alpha), \alpha, i)]$ . From Proposition 9, the equilibrium policy  $x^*$  is a strictly decreasing function of state-dependence  $\alpha$ . Thus,  $\frac{dx^*}{d\alpha} < 0$ .

The total derivative of the upper boundary with respect to  $\alpha$  is:

$$\frac{d\bar{M}}{d\alpha} = \frac{\partial \bar{M}}{\partial \alpha} + \frac{\partial \bar{M}}{\partial x^*} \cdot \frac{dx^*}{d\alpha},$$

where  $\frac{\partial \bar{M}}{\partial \alpha} > 0$  and  $\frac{\partial \bar{M}}{\partial x^*} < 0$ . The total derivative is positive.

Since the lower boundary strictly decreases and the upper boundary strictly increases with  $\alpha$ , the disclosure interval  $M$  is strictly expanding in the Agency's preference state-dependence.

## Proposition 11

1. The proof relies on a key property of the Policymaker's best-response function,  $\hat{x}(x; i)$ . For any fixed interior policy  $x$ , the function  $\hat{x}(x; i)$  is strictly decreasing in the Agency's bias,  $i$ .

Suppose that at  $i = p_0^P$ , there exists an equilibrium  $x^* > p_0^P$ . Now, consider a small increase in the Agency's bias to  $i' = p_0^P + \epsilon$  for some small  $\epsilon > 0$ . For  $\epsilon$  small enough and given continuity of  $\hat{x}(\cdot)$  there will be an equilibrium  $x^* > i$ .

By definition, for any  $G \in \hat{\mathcal{G}}$ , for  $i = p_0^P$ , there is a NDE and it is the unique belief-stable equilibrium. This (together with continuity of the Policymaker's best response function) implies that for all  $x > p_0^P$ ,  $\hat{x}(x; i = p_0^P) < x$ , and for all  $x < p_0^P$ ,  $\hat{x}(x; i = p_0^P) > x$ . Consider any  $i > p_0^P$ . The best-response function for this new bias, let's call it  $\hat{x}_i(x)$ , must lie at or below the original function for  $i = p_0^P$ , i.e.,  $\hat{x}_i(x) \leq \hat{x}_{p_0^P}(x)$  for all  $x$ . This rules out the existence of any new equilibrium at a policy greater than  $p_0^P$ . Thus, any belief-stable equilibrium  $x^*$  for an agent with bias  $i > p_0^P$  must satisfy  $x^* \leq p_0^P$ .

2. Given any belief-stable partial-disclosure equilibrium satisfies  $(x^* - i) \cdot (p_0^P - i) \geq 0$ , and the direct and indirect effects of  $i$  on disclosure are co-aligned, as  $i$  departs from  $p_0^P$ , disclosure expands. For any belief-unstable equilibrium, as  $i$  increases, optimal policy absent disclosure – one of the disclosure boundaries – converges to Agency's ideal point.

## Proposition 12

Consider a game  $G(i) \in \hat{\mathcal{G}}$ . By definition of  $\hat{\mathcal{G}}$ , when there is no ex-ante preference misalignment ( $i = p_0^P$ ), the non-disclosure equilibrium (NDE) is the unique belief-stable equilibrium. All other equilibria that exist at  $i = p_0^P$  must be belief-unstable. Without loss of generality, let  $p_0^P = 0$ .

The set of equilibrium policies,  $X^*(i)$ , varies with the parameter  $i$ . At  $i = 0$ , we have the stable NDE at  $x^* = 0$ . Let's analyze the case for  $i > 0$  in a neighborhood of 0 (the case for  $i < 0$  is symmetric). Let  $x_S^*(i)$  be the policy of the stable, least expansive partial-disclosure equilibrium that emerges from the NDE as  $i$  moves away from 0, such that  $\lim_{i \rightarrow 0^+} x_S^*(i) = 0$ .

Due to the alternating stability property (Proposition 8), the equilibrium nearest to 0 is belief-unstable and acts as a boundary for the basins of attraction of the belief-stable equilibria. Let  $x_U^*(i)$  be the policy of the nearest unstable equilibrium, such that  $\lim_{i \rightarrow 0^+} x_U^*(i) = x_U^*(0)$ . From continuity and the configuration at  $i = 0$ , for small  $i > 0$ , these equilibria will be ordered  $x_U^*(i) < x_S^*(i)$ .

1. The extent of belief-stability for the equilibrium at  $x_S^*(i)$  is determined by its distance to  $x_U^*(i)$ :

$$\varepsilon_S^*(i) = |x_S^*(i) - x_U^*(i)| = x_S^*(i) - x_U^*(i)$$

From Proposition 9, the equilibrium policy  $x^*$  is decreasing in  $i$  if and only if the equilibrium is belief-stable. This implies that for a belief-unstable equilibrium,  $x^*$  is increasing in  $i$ . Therefore  $\frac{d\varepsilon_S^*}{di} = (\text{negative}) - (\text{positive}) < 0$ .

Since  $i > 0$ ,  $|i - p_0^P| = i$ . A negative derivative means that as  $i$  increases,  $\varepsilon_S^*$  decreases. Conversely, as the magnitude of preference divergence  $|i - p_0^P|$  decreases, the extent of belief-stability  $\varepsilon_S^*$  increases.

2. Let  $x_{S2}^*(i)$  be the policy of any other belief-stable equilibrium (including a potential full-disclosure equilibrium). Given continuity and alternation of belief-stability, there is a belief-unstable equilibrium with policy  $x_{U2}^*(i)$  closest to  $x_{S2}^*(i)$  such that  $x_{U2}^*(i) > x_{S2}^*(i)$  (if this condition is not satisfied,  $x_{S2}^*(i)$  should have been the least expansive equilibrium). The extent of belief-stability of the equilibrium with policy  $x_{S2}^*(i)$  is the distance between it and the nearest belief-unstable equilibrium. As  $i$  converges to  $p_0^P$ ,  $x_{S2}^*(i)$  increases and  $x_{U2}^*(i)$  decreases. Thus, there exists a threshold  $\hat{\Delta}$  such that  $x_{U2}^*(i)$  is the closest belief-unstable equilibrium policy to  $x_{S2}^*(i)$ . Thus, as  $i$  converges to  $p_0^P = 0$  for all  $|i| < \hat{\Delta}$  the extent of belief-stability will decrease of the equilibrium with  $x_{S2}^*(i)$  will decrease.

## Remark 1

Without loss of generality, assume that the support  $\Omega$  is normalized to  $[-1, 1]$ , and the prior distribution is normalized to have mean  $E[\omega] = 0$ . Given utilities are distance-based. At  $i = 0$ , the agent's utility is  $u_A(p) = -(p - 0)^2 = -p^2$ . Note if there is no equilibrium such that  $(x^* - i) \cdot (p_0^P - i) < 0$  at  $i = 0$ , there is no such equilibrium for all  $|i| > 0$ . Let  $x$  be the policy implemented by the policymaker in the absence of disclosure. The agent, observing state  $\omega$ , will disclose it if their utility from disclosure,  $-(\omega)^2$ , is greater than their utility from non-disclosure,  $-(x)^2$ . This is equivalent to  $\omega^2 < x^2$ , or  $|\omega| < |x|$ . Therefore, the non-disclosure set is  $N(x) = \{\omega \in [-1, 1] \mid |\omega| \geq |x|\}$ .

Assume that  $x \in (0, 1)$ . Let  $a = x$ , where  $a \in (0, 1)$ .<sup>14</sup> The non-disclosure set is  $N(a) = [-1, -a] \cup [a, 1]$ . No equilibria such that  $(x^* - i) \cdot (p_0^P - i) < 0$  exist (given that NDE is belief-stable), if and only if for all  $a \in (0, 1)$  we have  $\mathbb{E}[\omega \mid \omega \in N(a)] < a$ .

---

<sup>14</sup>This notation becomes relevant when we consider  $x \in (-1, 0)$ .

$$\begin{aligned}
& \frac{\int_{-1}^{-a} \omega f(\omega) d\omega + \int_a^1 \omega f(\omega) d\omega}{\int_{-1}^{-a} f(\omega) d\omega + \int_a^1 f(\omega) d\omega} < a \\
& \int_{-1}^{-a} \omega f(\omega) d\omega + \int_a^1 \omega f(\omega) d\omega < a \left( \int_{-1}^{-a} f(\omega) d\omega + \int_a^1 f(\omega) d\omega \right) \quad (46) \\
& \int_a^1 (\omega - a) f(\omega) d\omega + \int_{-1}^{-a} (\omega - a) f(\omega) d\omega < 0
\end{aligned}$$

Let  $u = -\omega$ , which implies  $\omega = -u$  and  $d\omega = -du$ . Limits become  $u = 1$  (from  $\omega = -1$ ) and  $u = a$  (from  $\omega = -a$ ).

$$\begin{aligned}
\int_{\omega=-1}^{\omega=-a} (\omega - a) f(\omega) d\omega &= \int_{u=1}^{u=a} (-u - a) f(-u) (-du) \\
&= \int_1^a (u + a) f(-u) du \\
&= - \int_a^1 (u + a) f(-u) du
\end{aligned} \quad (47)$$

Substituting this back

$$\int_a^1 (\omega - a) f(\omega) d\omega - \int_a^1 (u + a) f(-u) du < 0 \quad (48)$$

Let  $\omega = v$  and  $u = v$ , then

$$\int_a^1 (v - a) f(v) dv < \int_a^1 (v + a) f(-v) dv \quad (49)$$

For this to hold for any  $a \in (0, 1)$ ,  $\forall v \in (a, 1)$

$$\begin{aligned}
(v - a) f(v) &< (v + a) f(-v) \\
\frac{f(-v)}{f(v)} &> \frac{v - a}{v + a}
\end{aligned} \quad (50)$$

Assume  $x \in (-1, 0)$ . Let  $a = -x$ , where  $a \in (0, 1)$ . The non-disclosure set is  $N(-a) = [-1, -a] \cup [a, 1]$ . For no equilibria such that  $(x^* - i) \cdot (p_0^P - i) < 0$  to exist  $\mathbb{E}[\omega \mid \omega \in N(-a)] > -a$

$$\int_{-1}^{-a} (\omega + a) f(\omega) d\omega + \int_a^1 (\omega + a) f(\omega) d\omega > 0 \quad (51)$$

Let  $u = -\omega$ , then

$$\begin{aligned}
\int_{\omega=-1}^{\omega=-a} (\omega + a) f(\omega) d\omega &= \int_{u=1}^{u=a} (-u + a) f(-u) (-du) \\
&= \int_1^a (u - a) f(-u) du \\
&= - \int_a^1 (u - a) f(-u) du
\end{aligned} \quad (52)$$

Substitute back

$$-\int_a^1 (u-a)f(-u)du + \int_a^1 (u+a)f(u)du > 0 \quad (53)$$

Let  $\omega = v$  and  $u = v$ , then

$$\int_a^1 (v+a)f(v)du > \int_a^1 (v-a)f(-v)du \quad (54)$$

Then, for all  $v \in (a, 1)$

$$\begin{aligned} (v+a)f(v) &> (v-a)f(-v) \\ \frac{f(-v)}{f(v)} &< \frac{v+a}{v-a} \end{aligned} \quad (55)$$

Combining Equations 55 and 50 we get

$$\frac{v-a}{v+a} < \frac{f(-v)}{f(v)} < \frac{v+a}{v-a}.$$

For any boundary  $a > 0$ , for all states in the tail of the distribution ( $v \in (a, 1)$ ) the ratio of the probability density at points  $-v$  and  $v$  is bounded from above and from below.

It implies that no partial disclosure equilibrium such that  $(x^* - i) \cdot (p_0^P - i) < 0$  exists if and only if the prior distribution is not too skewed. Note that for any symmetric prior distribution this inequality will be satisfied.

Finally, note that the boundaries of the disclosure set are continuous functions of  $\alpha$  for all  $\alpha \in [0, 1]$ . As a result, all interim values in inequalities 46 and 51 are continuous. By continuity, strict inequality in condition 55 guarantees that if the inequalities holds, Thus, if the conditions of Remark 1 are satisfied for  $\alpha = 0$ , there necessarily exists an  $\varepsilon > 0$  such that for any  $\alpha \in [0, \varepsilon)$ , the non-disclosure equilibrium remains the unique belief-stable equilibrium, ensuring the game is still in  $\hat{\mathcal{G}}$ .

### Proposition 13

When the Policymaker receives message  $m$  and signal  $s(m) = T$ , the Policymaker implements policy equal to the message observed. Next, signal  $s(m) = \emptyset$  and signal  $s(m) = F$  both should produce state-independent policies in equilibrium. This implies that Agency's distortion must replicate prior distribution on the disclosure interval. Denote policy the Policymaker implements following signal  $s(m) = \emptyset$  as  $z$  and policy the Policymaker implements following signal  $s(m) = F$  as  $x$ .

Next, consider an Agency with an ideal point  $i$ . Note that if an Agency with state's realization  $\tilde{\omega}$  prefers to disclose its information to the Policymaker instead of distorting it, any Agency with state  $\omega : |\omega - i| < |\tilde{\omega} - i|$  will disclose its state instead of concealing it. In this case, disclosing state produces policy  $q \cdot \omega + (1 - q) \cdot z$  while distorting it leads to policy  $q \cdot x + (1 - q) \cdot z$ . Thus, there exists a threshold  $y$  such that the Agency discloses states  $\omega \in [y, 2 \cdot i - y] \cap [-1, 1]$  and distorts states otherwise.

It immediately follows from the previous paragraph that when the Policymaker observes message  $m$  and signal  $s(m) = F$ , she implements policy  $x^* = E[\omega | \omega \notin [y, 2 \cdot i - y]] = \frac{i \cdot (i - y)}{-1 + i - y}$  when  $2 \cdot i - y < 1$  and  $x^* = \frac{y-1}{2}$  otherwise.

If the Policymaker observes  $s(m) = \emptyset$ , she implements equilibrium policy

$$z^* = E[\omega | s(m) = \emptyset] = E[Pr[\omega \in [y, 2 \cdot i - y]] \cdot m + Pr[\omega \notin [y, 2 \cdot i - y]] \cdot x] = 0. \quad (56)$$

Because both Agency that discloses information and the Agency that does not have equal probability to generate not-informative message, regardless of  $y$ , observing  $s(m) = \emptyset$  conveys no information beyond prior about state's realization.

In any equilibrium, the following holds

$$y = x^* \cdot q + z^* \cdot (1 - q) = \frac{i \cdot (i - y) \cdot q}{-1 + i - y}.$$

We only focus on belief-stable equilibria, thus,

$$y^* = \frac{i \cdot (1 + q) - 1 + \sqrt{(1 - i(1 + q))^2 - 4 \cdot i^2 \cdot q}}{2}.$$

### Proposition 14

1.

$$\frac{\partial y^*}{\partial i} = \frac{(1 + q) \cdot \sqrt{-4 \cdot i^2 \cdot q + (1 - i \cdot (1 - q))^2} - (1 - i \cdot (1 - q)^2 + q)}{2 \cdot \sqrt{-4 \cdot i^2 \cdot q + (1 - i \cdot (1 + q))^2}}$$

Note that numerator is always negative when  $q \in [0, 1]$ ,  $i \in [0, 1]$ . We will show that

$$(1 + q) \cdot \sqrt{-4 \cdot i^2 \cdot q + (1 - i \cdot (1 + q))^2} < (1 - i \cdot (1 + q)) \cdot (1 + q) + 4 \cdot i \cdot q$$

Squaring both sides, we need to show

$$(1 + q)^2((1 - i \cdot (1 + q))^2 - 4 \cdot i^2 \cdot q) < ((1 - i \cdot (1 + q)) \cdot (1 + q) + 4 \cdot i \cdot q)^2$$

which simplifies to

$$i \cdot (1 - q)^2 < 2 + 2 \cdot q.$$

Since  $i \in [0, 1]$  and  $q \in [0, 1]$ , we have  $i \cdot (1 - q)^2 \leq (1 - q)^2 \leq 1$  and  $2 + 2 \cdot q \geq 2$ . Since  $1 < 2$ , the inequality  $i \cdot (1 - q)^2 < 2 + 2 \cdot q$  holds. Therefore, the numerator is less than 0, and hence  $\frac{\partial y^*}{\partial i} < 0$ .

Because  $\frac{\partial y^*}{\partial i} < 0$ , the lower boundary of the disclosure interval decreases in  $i$  while the upper boundary increases in  $i$ .

2.

$$\frac{\partial y^*}{\partial q} = i \cdot \frac{\sqrt{-4 \cdot i^2 \cdot q + (1 - i \cdot (1 - q))^2} - (1 + i \cdot (1 - q))}{2 \cdot \sqrt{-4 \cdot i^2 \cdot q + (1 - i \cdot (1 + q))^2}}$$

We aim to show that  $\sqrt{-4 \cdot i^2 \cdot q + (1 - i \cdot (1 - q))^2} \leq (1 + i \cdot (1 - q))$ . Since both sides are non-negative, it is sufficient to show that their squares satisfy the inequality:

$$-4 \cdot i^2 \cdot q + (1 - i \cdot (1 - q))^2 \leq (1 + i \cdot (1 - q))^2$$

Rearranging terms yields:

$$(1 - i \cdot (1 - q))^2 - (1 + i \cdot (1 - q))^2 \leq 4 \cdot i^2 \cdot q$$

Factoring the difference of squares, we have:

$$\begin{aligned} [(1 - i \cdot (1 - q)) - (1 + i \cdot (1 - q))] \cdot [(1 - i \cdot (1 - q)) + (1 + i \cdot (1 - q))] &\leq 4 \cdot i^2 \cdot q \\ -4 \cdot i \cdot (1 - q) &\leq 4 \cdot i^2 \cdot q \end{aligned}$$

For  $i \geq 0$ , dividing both sides by  $4i$  (when  $i > 0$ ) or observing directly (when  $i = 0$ ), we require

$$q(1 - i) \leq 1.$$

Since  $0 \leq q \leq 1$  and  $0 \leq i \leq 1$ , it follows that  $0 \leq 1 - i \leq 1$ , and thus  $0 \leq q(1 - i) \leq 1$ . The inequality  $q(1 - i) \leq 1$  always holds. Therefore, numerator is non-positive.

Because of that,  $\frac{\partial y^*}{\partial q} \leq 0$  and the disclosure interval expands as  $q$  increases.

## Proposition 15

Given the beliefs  $\beta(\cdot|T, m(\cdot))$ , the Policymaker's policy choice  $p^*(T)$  is optimal by definition for on-path messages. For an off-path message  $T_{off}$ , the belief is concentrated on a single state  $\hat{\omega}(T_{off}) = \arg \max_{\tilde{\omega} \in T_{off}} |i - \tilde{\omega}|$ , so  $p^*(T_{off}) = \hat{\omega}(T_{off})$  is optimal.

If  $\omega \in M_L$ : By sending  $m^*(\omega) = \{\omega\}$ , the Agency's utility is  $-(i - \omega)^2$ . If it deviates to send  $N_L$ , its utility is  $-(i - x_L)^2$ . By definition of  $x_L$ ,  $-(i - \omega)^2 \geq -(i - x_L)^2$ . If it deviates to some  $T_{off}$  (where  $\omega \in T_{off}$ ), the policy will be  $\hat{\omega}(T_{off})$ . The utility is  $-(i - \hat{\omega}(T_{off}))^2$ . Since  $\hat{\omega}(T_{off})$  is the state in  $T_{off}$  furthest from  $i$ , and  $\omega \in T_{off}$ , it must be that  $|i - \hat{\omega}(T_{off})| \geq |i - \omega|$ . Thus,  $-(i - \hat{\omega}(T_{off}))^2 \leq -(i - \omega)^2$ . So, no profitable deviation to  $T_{off}$ .

If  $\omega \in N_L$ : By sending  $m^*(\omega) = N_L$ , the Agency's utility is  $-(i - x_L)^2$ . If it deviates to send  $\{\omega\}$ , its utility is  $-(i - \omega)^2$ . By definition of  $x_L$ ,  $-(i - x_L)^2 > -(i - \omega)^2$ . Deviations to  $T_{off}$  are deterred as above, as  $-(i - \hat{\omega}(T_{off}))^2 \leq -(i - \omega)^2 < -(i - x_L)^2$ .

Finally, to show that proposed beliefs are consistent, we construct a sequence of strictly mixed strategy profiles  $(m^n(\cdot), p^n(\cdot))$  that converges to  $(m^*(\cdot), p^*(\cdot))$  and corresponding sequence of Bayesian beliefs  $\beta^n$  that converges to  $\beta$ . Denote the family of all off-path messages available for the Agency observing  $\omega$  for which  $\omega$  is the furthest from the Agency's ideal point  $i$  as  $\overline{\mathcal{T}}_{off}(\omega) := \{T : \omega = \hat{\omega}(T), \omega \in T, T \neq \{\omega\}, T \neq N_L\}$ .

We construct  $(m^n(\cdot), p^n(\cdot))$  as follows. Let  $P^n(T|\omega)$  denote the probability type  $\omega$  sends a message  $T$ . Suppose the Agency observing realization  $\omega$  sends its equilibrium message  $m^*(\omega)$  with probability  $P^n(m^*(\omega)|\omega) = 1 - 1/n - 1/n^2$ . With total probability  $1/n$ , the Agency sends an off-path message  $T_{off}$  selected uniformly from the set  $\overline{\mathcal{T}}_{off}(\omega)$ . With the remaining total probability  $1/n^2$ , the Agency sends an off-path message  $T_{off}$  selected uniformly from the set  $\mathcal{T}_{off}(\omega) \setminus \overline{\mathcal{T}}_{off}(\omega)$ .

As  $n \rightarrow \infty$ , for an off-path  $T_{off}$ , Bayes' rule for  $\beta^n(\tilde{\omega}|T_{off}, m(\cdot))$  requires

$$\beta^n(\tilde{\omega}|T_{off}, m(\cdot)) = \frac{P^n(T_{off}|\tilde{\omega})f(\tilde{\omega}|\tilde{\omega} \in T_{off})}{\int_{\omega \in T_{off}} P^n(T_{off}|\omega)dF(\omega|\omega \in T_{off})}.$$

If  $\tilde{\omega} = \hat{\omega}(T_{off})$ , the numerator term  $P^n(T_{off}|\tilde{\omega})$  is  $O(1/n)$ . For any other  $\omega' \in T_{off}$ ,  $P_n(T_{off}|\omega')$  is  $O(1/n^2)$ . Thus, in the limit, the probability mass concentrates on  $\hat{\omega}(T_{off})$ . This ensures that  $\lim_{n \rightarrow \infty} \beta^n(\cdot|T_{off}, m(\cdot)) = \beta(\cdot|T_{off}, m(\cdot))$  as specified. The on-path beliefs are similarly consistent.

## Supplemental Appendix: Cheap Talk

Suppose the Sender's and Receiver's utility functions are  $u_S(p, \omega; \alpha, i) = -((\alpha\omega + (1-\alpha)i - p)^2)$  and  $u_R(p, \omega) = -((\omega - p)^2)$ , respectively. Assume  $\omega \sim U[-1, 1]$  and let  $i \geq 0$ . (The case  $i \leq 0$  is symmetric.) After the Sender observes  $\omega$ , she sends a message from a rich message space  $M$ ; the Receiver observes the message  $m \in M$  and then updates her beliefs and chooses  $p \in \mathbb{R}$ .

Suppose that for some  $(\alpha, i)$  there exists an informative equilibrium in which two distinct messages  $m_1, m_2 \in M$ ,  $m_1 \neq m_2$ , provoke two distinct policies  $p_1, p_2 \in \mathbb{R}$ ,  $p_1 \neq p_2$ , in response. Suppose wlog  $p_1 < p_2$ . Then, Sender prefers  $p_1$  to  $p_2$ , and hence  $m_1$  to  $m_2$ , if

$$\alpha\omega + (1 - \alpha)i \leq \frac{p_1 + p_2}{2}. \quad (57)$$

Then there exists a threshold  $\omega^*$  such that Sender prefers  $m_1$  to  $m_2$  for all  $\omega \leq \omega^*$  and  $m_2$  to  $m_1$  otherwise. Given this monotonicity of the Sender's induced preferences over messages, we may henceforth restrict attention to an equilibrium in which two messages are informative.

Because  $p_1$  and  $p_2$  are optimal Sender responses,  $p_1 = \mathbb{E}[\omega|m_1]$  and  $p_2 = \mathbb{E}[\omega|m_2]$ ; thus

$$p_1 + p_2 = \int_{-1}^{\omega^*} \frac{p(\omega)}{P(\omega^*)} \omega d\omega + \int_{\omega^*}^1 \frac{p(\omega)}{P(\omega^*)} \omega d\omega.$$

Because  $p(\omega) = \frac{1}{2}$  and  $P(\omega) = \frac{\omega^*+1}{2} \forall \omega \in [-1, 1]$ , integrating yields  $2\omega^*$ . Recognizing that Inequality 57 holds at equality for  $\omega = \omega^*$  and substituting  $\omega^*$  for  $p_1 + p_2$ ,

$$\omega^* = \frac{2(1 - \alpha)}{1 - 2\alpha} i \quad (58)$$

for  $\alpha \neq \frac{1}{2}$ .

From Equality 58, if  $\alpha < \frac{1}{2}$ ,  $\omega^* \geq 0$  and increasing in  $i$  for  $i \in [0, \frac{1-2\alpha}{2(1-\alpha)})$ , where the upper bound on  $i$ ,  $\frac{1-2\alpha}{2(1-\alpha)}$ , decreases from  $\frac{1}{2}$  to 0 as  $\alpha$  increases from 0 to  $\frac{1}{2}$ . Also from (58), if  $\alpha > \frac{1}{2}$ ,  $\omega^* \leq 0$  and decreasing in  $i$  for  $i \in [0, \frac{2\alpha-1}{2(1-\alpha)})$ , where the upper bound on  $i$ ,  $\frac{2\alpha-1}{2(1-\alpha)}$ , increases from 0 to infinity as  $\alpha$  increases from  $\frac{1}{2}$  to 1.

Thus,  $\forall \alpha \neq \frac{1}{2}$ , as  $i$  increases,  $|\omega^* - p_0^P|$  increases until  $\omega^*$  reaches a boundary of the support of the state space. For higher  $i$ , the premise that both messages are played in equilibrium is violated, that is, it is no longer possible to support informative cheap-talk communication in equilibrium.

## References

Austen-Smith, David. 1990. "Information transmission in debate." *American Journal of political science* pp. 124–152.

- Austen-Smith, David. 1993. "Interested experts and policy advice: Multiple referrals under open rule." *Games and Economic Behavior* 5(1):3–43.
- Aybas, Yunus, C. and Steven Callander. 2023. "Efficient Cheap Talk in Complex Environments." *Working paper* .
- Battaglini, Marco. 2002. "Multiple referrals and multidimensional cheap talk." *Econometrica* 70(4):1379–1401.
- Bendor, Jonathan and Adam Meirowitz. 2004. "Spatial models of delegation." *American Political Science Review* 98(2):293–310.
- Callander, Steven. 2008. "A theory of policy expertise." *Quarterly Journal of Political Science* 3(2):123–140.
- Callander, Steven, Nicolas S Lambert and Niko Matouschek. 2021. "The power of referential advice." *Journal of Political Economy* 129(11):3073–3140.
- Che, Yeon-Koo and Navin Kartik. 2009. "Opinions as incentives." *Journal of Political Economy* 117(5):815–860.
- Crawford, Vincent P and Joel Sobel. 1982. "Strategic information transmission." *Econometrica* pp. 1431–1451.
- Denisenko, Anna, Catherine Hafer and Dimitri Landa. 2024. "Competence and Advice."
- Dye, Ronald A. 1985. "Disclosure of nonproprietary information." *Journal of accounting research* pp. 123–145.
- Dziuda, Wioletta. 2011. "Strategic argumentation." *Journal of Economic Theory* 146(4):1362–1397.
- Gailmard, Sean and John W Patty. 2007. "Slackers and zealots: Civil service, policy discretion, and bureaucratic expertise." *American Journal of Political Science* 51(4):873–889.
- Gailmard, Sean and John W Patty. 2012. "Formal models of bureaucracy." *Annual Review of Political Science* 15(1):353–377.
- Gilligan, Thomas W and Keith Krehbiel. 1989. "Asymmetric information and legislative rules with a heterogeneous committee." *American journal of political science* pp. 459–490.
- Grossman, Sanford J. 1981. "The informational role of warranties and private disclosure about product quality." *The Journal of law and Economics* 24(3):461–483.
- Jung, Woon-Oh and Young K Kwon. 1988. "Disclosure when the market is unsure of information endowment of managers." *Journal of Accounting research* pp. 146–153.
- Kaufman, Herbert. 1981. "Fear of bureaucracy: A raging pandemic." *Public Administration Review* pp. 1–9.

- Krishna, Vijay and John Morgan. 2001. "A model of expertise." *The Quarterly Journal of Economics* 116(2):747–775.
- McCarty, Nolan. 2004. "The appointments dilemma." *American Journal of Political Science* 48(3):413–428.
- Milgrom, Paul. 2008. "What the seller won't tell you: Persuasion and disclosure in markets." *Journal of Economic Perspectives* 22(2):115–131.
- Milgrom, Paul and John Roberts. 1986. "Relying on the information of interested parties." *The RAND Journal of Economics* pp. 18–32.
- Milgrom, Paul R. 1981. "Good news and bad news: Representation theorems and applications." *The Bell Journal of Economics* pp. 380–391.
- Prendergast, Canice. 2007. "The motivation and bias of bureaucrats." *American Economic Review* 97(1):180–196.
- Seidmann, Daniel J and Eyal Winter. 1997. "Strategic information transmission with verifiable messages." *Econometrica: Journal of the Econometric Society* pp. 163–169.
- Shin, Hyun Song. 1994. "The burden of proof in a game of persuasion." *Journal of Economic Theory* 64(1):253–264.
- Sobel, Joel. 2013. "Giving and receiving advice." *Advances in economics and econometrics* 1:305–341.
- Wolinsky, Asher. 2003. "Information transmission when the sender's preferences are uncertain." *Games and Economic Behavior* 42(2):319–326.