ISSN 1936-5349 (print)
ISSN 1936-5357 (online)

HARVARD

JOHN M. OLIN CENTER FOR LAW, ECONOMICS, AND BUSINESS
FELLOWS' DISCUSSION PAPER SERIES

TRANSPARENCY AND POWER IN RULEMAKING

(Formerly "Transparency in Industrial Regulation")

Laurence Tai

Discussion Paper No. 34

07/2010

Harvard Law School
Cambridge, MA 02138

Contributors to this series are John M. Olin Fellows or Terence M. Considine Fellows in Law and Economics at Harvard University.

This paper can be downloaded without charge from:

The Harvard John M. Olin Fellow's Discussion Paper Series:
http://www.law.harvard.edu/programs/olin_center/

Transparency and Power in Rulemaking

Laurence Tai*

Abstract

This paper develops a model analyzing the impact of information transparency in administrative rulemaking. The key items of information are the regulated party's communication and the agent's signal based on that communication. Information disclosure not only allows the principal to observe the information, but it may also increase her power in the decision, measured by the probability that she can to select the final policy. A key result is that, even without mandating any disclosures, the principal can have the same knowledge as the agent does about what level of regulation would be optimal. Instead of increasing knowledge, transparency primarily benefits the principal when it increases her power through the disclosures. However, it may also discourage the regulated party or agent from generating information in the first place. There are empirical implications to determine the model's applicability and institutional design implications to the extent it is applicable.

JEL Classes: D72, D73, K23

---

*Ph.D. Candidate in Public Policy, Harvard Kennedy School and J.D. Candidate, Harvard Law School. The author is a Terence M. Considine Fellow in Law and Economics at Harvard Law School and acknowledges support from the Considine Family Foundation and the School's John M. Olin Center for Law, Economics, and Business. This paper has received valuable feedback from Daniel Carpenter, Ken Shepsle, Matthew Stephenson, Yuki Takagi, Craig Volden, and Richard Zeckhauser. Prior versions of this paper have been presented at the Political Economy Workshop at Harvard, the 2010 MPSA Annual Conference, and the 2010 APSA Annual Meeting.

Transparency and Power in Rulemaking

> "A popular Government, without popular information, or the means of acquiring it, is but a Prologue to a Farce or a Tragedy; or, perhaps both. Knowledge will forever govern ignorance: And a people who mean to be their own Governors, must arm themselves with the power which knowledge gives."
>
> – James Madison, Letter to William T. Barry, August 4, 1822 (quoted in Madison 1999, 790).

1 Introduction

The above quote, which often appears in works about transparency (see Fenster 2006, 895), is perhaps the earliest expression by an American statesperson of the notion that access to government information can help ensure that the government serves the popular will, rather than the narrow interest of its officers or some other organized group. With the administrative state has emerged regulatory capture, a theory according to which executive branch agencies cater to the interests of the entities that Congress has charged them with regulating (see Levine and Forrence 1990, 169). Whether these entities unduly influence agency policymaking is an open question (Carpenter 2013); but there is at least a perception that through their influence regulated parties are able to reap gains at the expense of the beneficiaries of regulation.

In the face of potential capture, public interest advocates have made access to government information an important element of their efforts to reduce the extent to which agencies bias their policies in favor of regulated interests (see Wagner 2010, 1323–24). One of the more

prominent laws in this area is the Freedom of Information Act (FOIA), which obligates agencies to release nonexempt documents to anyone upon request. The trend toward greater transparency has continued in the Obama Administration, which has made this value a theme (Coglianese 2009). In particular, it has called upon agencies to adopt a presumption of disclosure for FOIA requests and to release more documents proactively (Obama 2009).

After all of these initiatives to increase government transparency, there are two conflicting results, which can be treated as stylized facts. On one hand, many agencies continue to withhold or delay the release of information requested through the FOIA and have responded in limited fashion to President Obama's memorandum on FOIA implementation (Hicks 2013). On the other hand, agencies are often willing to make of their information transparent, even when the law does not require them to do so (Moffitt 2010). In the rulemaking process, agencies typically present large amounts of supporting information in their notices of proposed rulemaking (NPRM) and of their final rules, and they make additional information accessible in regulatory dockets (Kerwin and Furlong 2011, 64–65).

These patterns cannot be fully reconciled with an assertion that the agencies that usually voluntarily disclose information are different from those that frequently withhold it, unless any agency that provides information in rulemaking also rarely resists in responding to FOIA requests.[1] These patterns also cannot obviously be explained with the idea that agencies release only uninformative documents while keeping truly informative documents secret, since courts are sometimes asked to review supporting information and are often satisfied as to its validity.

[1]Although agencies may disclose the information supporting their proposed rules because they believe that doing so is necessary to secure court approval, disclosure is still voluntary in the sense that courts are not formally compelling the disclosure of any particular item of information in the way they require agencies to comply with the FOIA.

An alternative explanation is that an agency's willingness to disclose some documents and not others derives not from the information that concerned citizens would learn from those documents about a policy question, but instead from the power or influence they might gain vis-à-vis the agency. The potential of a released document to increase citizens' power, apart from its potential to increase their knowledge, is arguably embedded in discussions about transparency. Stiglitz (2002) has asserted that "secrecy gives those in government exclusive control over certain areas of knowledge, and thereby increases their power" (29-30); as a particular example, he notes that the International Monetary Fund has argued that "open discussions . . . may feed the opposition (38). In the context of rulemaking, this logic can be derived from the idea that "transparency . . . enables better public participation" (Coglianese, Kilmartin, and Mendelson 2009, 928), which implies that "all interested citizens have the ability to participate and to have an agency consider their interests even-handedly" (id., 927).

More generally, Madison's reference to "the power that knowledge gives" is ambiguous enough to allow this notion that document disclosure increases power. A number of mechanisms can be imagined: In a political context, an agency might release information it received about the industry with which a media report could cast that industry in a negative light so that public interest groups have more influence in the rest of the policymaking process. Alternatively, such information might make it easier for them to activate the "fire-alarm" form of congressional oversight described in McCubbins and Schwartz (1984). In a legal setting, a plaintiff representing beneficiaries of regulation might be willing to challenge a regulation and increase her chances of success in judicial review if she has more the agency's information not because she knows more about what policy would be optimal, but because she can better respond to the agency's arguments defending its proposed rule. This logic is

consistent with the reported belief of agency officials that information in dockets can be a "source of ammunition for lawsuits" (West 2004, 70).

This logic is different from the standard notion that transparency increases citizens' knowledge via the information that released documents convey. Discussion of this knowledge function may appear together empowerment function in work that discusses transparency. In the electoral context, Stiglitz (2002) argues that, "if democratic oversight is to be achieved, then the voters have to be informed" (31). For rulemaking, Coglianese, Kilmartin, and Mendelson (2009) states that"transparency . . . mak[es] information more readily available to more people" (928). These works on transparency mention both rationales, they do not appear to have analyzed them.

This paper presents a simple model exploring the relationship between transparency and power in a common regulatory setting, that of rulemaking. It offers three results that support increased power rather than increased knowledge as the rationale for transparency. First, even with no transparency, the agent will disclose enough information for citizens to know as much about the regulated party as he does. Second, if information disclosure can directly empower public interest groups, transparency can yield benefits, but not because they are able to learn more about the policy question from release documents. Instead, their gains derive from their increased power or from the knowledge they based on from whether a document exists. Finally, if no type of information disclosure can increase their power, then the agent has no reason to withhold information even when (and possibly because) there are no transparency requirements.

The rest of this paper proceeds as follows: Section 2 sets up a formal model designed to match the rulemaking process, and Section 3 describes the results that follow from it. Section 4 the importance of power in understanding transparency beyond the confines of the

model. Section 5 discusses policy implications for the model, and Section 6 concludes.

2 The Model

The game, which is structured to capture some of the salient features of notice-and-comment rulemaking, features three players: a principal (P, she), whose preferences are assumed to be synonymous with those of the general public; a regulated party or target (R, it), which has information relevant to the policy decision, and an agent or agency (A, he), who, unlike the principal, can process information that the regulated party generates and communicates and can make policy commitments based on this information. The principal has a use for an agent because he has these two special abilities, for a policy result that may be better than directly choosing policy herself. The goal of the analysis is to understand how transparency and power affect the principal's payoff. Throughout, public interest groups and the principal will be used interchangeably, as if the former represents society's interest. If one does not believe that these interests are congruent, then the model provides a positive account for how citizens in favor of stricter regulation than either the agent or regulated party could benefit from transparency.

2.1 Policies and Payoffs

A policy or regulation $x \in \mathbb{R}_+$, such as the permissible emission levels of a pollutant or the stringency of standards for workplace safety, is to be set at a particular level. The costs of this regulatory policy will fall on the regulated party and are represented by $rc(x)$, where $c(\cdot)$ is a continuous function with $c(0) = c'(0) = 0$, $c'(x) > 0$, $\forall x > 0$, and $c''(x) > 0$, $\forall x \geq 0$, and where $r > 0$ is a parameter reflecting how costly the regulation is for that party. Its

cost parameter can take one of two values h, or l, with $h > l$. The probability of each of these cost levels t is τ_t. All players know the cost function, the possible cost parameters, and the probability of each parameter. However, only the target knows its type $T \in \{H, L\}$, reflecting that its costs for each level of regulation are high or low. The idea that the key item of unknown information is the regulated party's type is a common feature of games of informational lobbying by interest groups (Potters and Van Winden 1992, Sloof 1998).

The following features about the other players' payoffs are common knowledge: The public benefit of regulation takes the functional form $b(x)$, a continuous function with $b(0) = 0$, and $b'(x) > 0$, and $b''(x) < 0$, $\forall x \geq 0$. The principal's utility is simply social welfare $b(x) - rc(x)$. The agency's utility, however, is $b(x) - (1 + a)rc(x)$, with $a > 0$ a divergence parameter.[2] Because $a > 0$, he weighs the costs more greatly than the principal and will thus tend to act more favorably toward the regulated party than she would.

2.2 Policymaking Steps

A series of steps involving the regulated party, the agency, and possibly the principal, are involved in arriving at the final regulation. In general, actions by the agent and the target lead to a proposal, which the principal has some chance of amending.

The Regulated Party's Message to the Agency The regulated party's single action is whether to communicate with the agency. This decision can be understood as a stylized version of submitting additional policy-relevant material to the agency. Formally, it decides to send a message m, or not to send one, $\emptyset$. Both target types are able to send messages that do not allow anyone else to distinguish them at all without some additional action by

[2]Because there will be no side transfers among parties, the scale of benefits is not important.

the agent, so the target types are treated as though each can transmit the same message. Conveying a message does not directly cost the type, but doing so can indirectly cost the target depending on what the agent does with the information.

Agency Interpretation of the Target's Message The agent, like any player in a game of imperfect information, can attempt to infer the target's type from its transmissions. However, he has the unique ability to generate a signal about the type if he processes a message that the target has sent. He can choose to interpret the message, n, or not to do so, $\emptyset$. If he interprets the message, the signal is denoted by $s \in \{\tilde{H}, \tilde{L}\}$, corresponding to the target's type, with $\Pr(s = \tilde{H}|T = H) = \Pr(s = \tilde{L}|T = L) = \alpha \in (1/2, 1)$. The agent does not incur any direct costs through his act of interpretation. Thus, if he chooses not to activate his interpretive abilities, it is not because he has decided that doing so is not worth the effort. Interpretation requires a message, so if the target decides not to communicate (chooses $\emptyset$), the agent cannot generate a signal. Lack of a signal, whether by choice or due to lack of a message, will be denoted by $s = \emptyset$.

Some quantities can be derived from the case in which both target types provide a message and the agent interprets it. First,

$$i \equiv \frac{\alpha\tau_h h + (1-\alpha)\tau_l l}{\alpha\tau_h + (1-\alpha)\tau_l} \text{ , } \tau_i \equiv \alpha\tau_h + (1-\alpha)\tau_l \text{ , } k \equiv \frac{\alpha\tau_l l + (1-\alpha)\tau_h h}{\alpha\tau_l + (1-\alpha)\tau_h}, \text{ and } \tau_k \equiv \alpha\tau_l + (1-\alpha)\tau_h$$

can represent respectively the expected cost level when the signal is high, the probability of a high signal, the expected cost level when the signal is low, and the probability of a low signal. The agent's imperfect sorting makes the target's message fall somewhere between soft and hard information. Next, $j \equiv \tau_h h + \tau_l l$ can denote the expected cost parameter according to the prior beliefs, and which may apply when $s = \emptyset$ because both types have sent a message but he

has not interpreted. Then $h > i > j > k > l$, and t can denote one of these five values. Then, for the principal, $EU_t^P(x) \equiv b(x) - tc(x)$ and $x_t^P \equiv \arg\max_x b(x) - tc(x)$ can respectively denote her expected utility from a given level of regulation given an expected cost parameter t and her optimal policy level for that cost parameter. $EU_t^A(x) \equiv b(x) - (1+a)tc(x)$ and $x_t^A \equiv \arg\max_x b(x) - (1+a)tc(x)$ denote the analogous terms for the agent. Thus, $EU_t^P(x_t^P)$ and $EU_t^A(x_t^A)$ are these players' respective optimal payoffs for a given expected cost level t.

The Agent's Policy Proposal In this stage, the agent chooses the policy that will obtain if the principal does not override it by making a proposal x^A. This stage of the game is supposed to be equivalent to an agency's notice of proposed rule-making (NPRM), since the content of the rule typically does not change much after the proposal apart from unusual political pressures, described in the next paragraph. Reasons are that changing the rule significantly may trigger the need for another notice-and-comment period (West 2004, 73), that agency officials are psychologically committed to a given policy and reluctant to change (*id.*, 72–73), and that it has made costly investments in orienting itself toward the proposed policy and away from others (Ting 2011). Although rules can subtly change even under routine circumstances, the simplifying assumption that the agent's proposal is binding limits the scope of alterations to those that result from more deliberate intervention by political leaders or a court.

The Agent's Disclosures to the Principal Along with the proposal, the agent also makes certain disclosures to the principal. This timing for the disclosures is consistent with the perceived tendency of agencies to communicate with preferred interest groups and to formulate a proposal before the NPRM (see Coglianese, Kilmartin, and Mendelson 2009, 931–32). Based on the previous three steps, the agent has up to three items that he can

disclose: the target's message, his signal, and his policy proposal. The focus on these specified pieces can be rationalized in part by representing a message from the target as a study, a signal about the cost as a report by the agency, and the proposal as a memo detailing the agency's plans. For these categories of information, the analysis assumes that the agent has a way of credibly disclosing any information in his possession, such as a high-cost signal $\tilde{H}$ or a policy proposal $x^A = 10$. Making transparent information clearly observable for the principal is a standard feature for transparency models (e.g., Prat 2005).

Two other clarifications about the nature of disclosures are important: First, the agent has no independent way of conveying the lack of a certain kind of information. For example, if the agent has actually received a message but chooses not to display it, his nondisclosure decision is observationally equivalent to not receiving a message if he is not required to release all information. Thus, a transparency requirement can help the principal distinguish between the nondisclosure and nonexistence of information. Second, even when the agent is not required to convey some item of information, he is still permitted to do so. In actual policymaking, certain types of records, such as those relating to intelligence and trade secrets, are not only exempted by the FOIA from disclosure, but also typically prohibited altogether from release. Another exemption in the FOIA renders the statute consistent with laws that prohibit disclosure of other kinds of information. Statutes that preclude the fulfillment of FOIA requests imply that some other value, such as national security or innovation, is supported by some level of secrecy. This model is limited to cases in which withholding information does not directly confer some policy benefit. Thus, the analysis is testing the impacts of transparency, in which information that the agent can transmit to the principal, he must send to her; rather than of mere observability, which refers only to whether she actually receives and comprehends the item of information.

Policy Change by the Principal Occasionally, the content of a rule does significantly change after its proposal. When this occurs after the agency has proposed the rule, it may be due to intervention from political players outside the agency, either higher up in the executive branch or in Congress (West 2004, 72). This possible step in policymaking is represented by a probability, $\pi \in (0, 1]$, that the principal will be able to select the final policy, x^P, according to her preferences. This is the simplest way of modeling the principal's receipt of any agency disclosures, followed by her response. The random chance π is what represents the prinicpal's baseline level of power in the model. It can represent the likelihood that public interest groups will attract the attention and support of political insiders or the ease with which standing rules allow them to challenge regulations in court.

The two items of information that can result in the principal's gaining power when the agent discloses them are the target's message and the low-cost signal. More power from the target's message can be rationalized by the notion that, if concerned citizens can access a regulated party's information earlier in the regulatory process, they can marshal better arguments against it and increase their chance of changing policy. This belief is consistent with the tendency of participants in notice-and-comment ruling to submit their comments as late as possible "to have the last word" (see Coglianese, Kilmartin, and Mendelson 2009, 947) Meanwhile, more power from a low-cost signal can be justified with the idea that citizens may have a better case for political intervention, such as congressional oversight. Since interest groups might conceivably pull fire alarms whenever they can, possessing hard information that supports intervention helps legislators determine which alarms are worth responding to. The increases in the power parameter due to target's message and the low-cost signal will be denoted respectively by $\Delta\pi_m$ and $\Delta\pi_{\tilde{L}}$, each of which is nonnegative, and which together are constrained so that $\Delta\pi_m + \Delta\pi_{\tilde{L}} \leq 1 - \pi$.

For the sake of completeness, it is worth noting that the principal only has the potential to select the policy; in particular, he cannot interpret any message from the regulated party on her own and can only read signals that the agent generates. However, like the agent, she can try to infer the target's type based on the totality of the information she receives.

Summary of Stages The order of gameplay can be listed as follows:

1. Nature selects the regulated party's type, $T \in \{H, L\}$.
2. The target of regulation decides whether to send a message, m, or to stay silent, $\emptyset$.
3. The agent decides whether to process the target's message (if he has one) and generate a signal s.
4. The agent makes a policy proposal, x^A, and decides on his disclosures to the principal.
5. Either the principal selects the policy or the agent's proposal stands. The probability that the principal substitutes her choice depends on her baseline level of power and on what items of information the agent discloses.

2.3 Strategies and Beliefs

The players' strategies can be expressed as follows: The simplest strategy to notate is that of the regulated party, each type of which has a single component: $\sigma^T \in \{m, \emptyset\} \equiv M$, with $T \in \{H, L\}$.

The agent has the largest number of actions. First, he chooses whether to interpret the target's message, if he can. This decision can be represented as $\sigma_n^A : M \to \{n, \emptyset\} \equiv N$, with $\sigma_n^A(\emptyset) = \emptyset$, since he can only interpret a message if the target has provided one.

The possible signals he may have after interpretation are $S \equiv \{\tilde{H}, \tilde{L}, \emptyset\}$. Then his proposal is $x^A : M \times S \to \mathbb{R}_+$, where some ordered pairs in the arguments are logically precluded (e.g., $(\emptyset, \tilde{L})$). His final move involves his disclosures to the principal. With the restrictions above on information transmissions and δ ($\emptyset$) representing (non)disclosure, the agent's strategy for what to convey to the principal can be denoted by the ordered triple $\sigma_d^A \equiv (d_m, d_s, d_x) : M \times S \times \mathbb{R}_+ \to \{\delta, \emptyset\}^3$. The possibilities for disclosure may be constrained by a transparency requirement. Overall, the agent's strategy can be more concisely notated as $\sigma^A \equiv (\sigma_n^A, x^A, \sigma_d^A)$.

Finally, the principal's strategy depends on the disclosures she has about the target's message, the agent's signal, and his proposal. With $\mathring{M} \equiv M$, $\mathring{S} \equiv S$, and $\mathring{x}^A \equiv \emptyset \cup \mathbb{R}_+$ representing the set of possibilities for each category of information, her strategy is $\sigma^P \equiv x^P : \mathring{M} \times \mathring{S} \times \mathring{x}^A \to \mathbb{R}_+$. For convenience, $\mathring{d} \equiv (\mathring{d}_m, \mathring{d}_s, \mathring{d}_x)$ can represent an ordered triple of information she receives. Overall, the strategy profile for the game be notated as $\sigma \equiv (\sigma^H, \sigma^L, \sigma^A, \sigma^P)$.

Beliefs for the agent and principal center on the target's type. Let β_L^A and β_L^P represent their respective beliefs that $T = L$. Then $\beta_L^A : M \times S \to [0, 1]$. Although the agent's beliefs can change twice during the game, her belief between the target's communication and his interpretation or lack thereof is sufficiently represented by $\beta_L^A(\cdot, \emptyset)$. The principal's beliefs change only once, so her posteriors are $\beta_L^P : \mathring{M} \times \mathring{S} \times \mathring{x}^A \to [0, 1]$. With $\beta \equiv \beta_L^A, \beta_L^P$, strategy-belief profiles can be denoted as (σ, β).

2.4 Preliminary Comparisons of Payoffs

Determining the value of transparency requires a comparison of different equilibrium payoffs for the principal. Meanwhile, the same comparison for the agent helps identify how he

might respond to mandated disclosure. For each player $q \in \{P, A\}$, the maximum possible payoff, which would entail selecting the optimal level of regulation for each cost parameter, is $\tau_h EU_h^q(x_h^q) + \tau_l EU_l^q(x_l^q)$. A kind of second-best payoff if both target types message, a signal is generated, and a player chooses optimally for each signal is $\tau_i EU_i^q(x_i^q) + \tau_k EU_k^q(x_k^q)$. If, however, no signal is generated, then choosing optimally in ignorance yields $EU_j^q(x_j^q)$. Unsurprisingly, the principal and agent each prefer partial information about the target's type to none, and full information to partial when each has authority.

Lemma 1. *For $q \in \{P, A\}$, the following inequality holds:*

$$EU_j^q(x_j^q) < \tau_i EU_i^q(x_i^q) + \tau_k EU_k^q(x_k^q) < \tau_h EU_h^q(x_h^q)) + \tau_l EU_l^q(x_l^q). \tag{1}$$

Proof. Proofs of all numbered results except Corollary 3 are in Appendix B. ■

For the principal, the lowest utility in Inequality 1 can be understood as her default payoff, in that she benefits from granting power to an agent whose preferences diverge from hers only if she improves upon this payoff.

3 Equilibrium Results

The equilibrium concept is perfect Bayesian equilibrium in pure strategies, except for the low-cost target, which can randomize between transmitting and not transmitting a message. As is the case in many messaging games, many equilibria exist, and a challenge is to rule out implausible equilibria. In particular, it is important to prevent the principal from always believing that the regulated party has low costs off the equilibrium path, even when the agent's disclosures suggest a posterior probability $\beta_L^P < 1$. For example, if the principal observes

the target's message in a deviation from an equilibrium in which both types communicate a message, her most pessimistic belief should be that the agent has the low-cost signal, which means the target might still have high costs.[3] Because there is an agent in between the sender (the target) and the receiver (the principal) and because the regulation to be chosen is from the real line rather than from a finite set, standard refinements like the intuitive criterion (Cho and Kreps 1987) and universal divinity (Banks and Sobel 1987) cannot readily be applied. Instead, two refinements are developed in Appendix A. Though they operate differently, they identify the same set of plausible equilibria in the results. Thus, a *natural* equilibrium will be one that satisfies a given refinement, and the results in this section can be read with either refinement in mind.

Among natural equilibria, one type that will receive special focus is one in which both target types always message the agent, who then analyzes the message to generate a signal. Formally, a *message-signal* equilibrium is one in which $\sigma^H = \sigma^L = m$ and $\sigma_n^A(m) = n$. If the principal is able to discern the agent's signal with or without seeing it, then she can benefit from his ability to scrutinize the regulated party's communications.

3.1 Policy Choices and Power Levels

The most general result involves the policy choices and power levels of the principal and agent in a natural message-signal equilibrium. Although there are many possibilities for message-signal equilbria in general, the refinements lead to a single set of regulation and power levels.

Theorem 2. *In any natural message-signal equilibrium,* $x^{P*} = x_i^P$ *and* $x^{A*} = x_i^A$ *following*

[3]If anything, appealing to this logic provides additional support for mandated disclosure, since it prevents the principal from inducing certain optional disclosures.

$s = \tilde{H}$, and $x^{P*} = x_k^P$ and $x^{A*} = x_k^A$ following $s = \tilde{L}$. Furthermore, the principal always has the lowest level of power possible given the items that are transparent.

The restriction to natural equilibria means that, if the target always communicates and the agent generates the signal, then the principal and agent select their respective optimal policies based on the signal when each has authority. Also, if a natural-message signal equilibrium exists, it is unique up to disclosures of items of information that do not increase the principal's power.

Theorem 2 has two implications for relationship between voluntary disclosure and power. First, because the agent maximizes his probability of selecting the final policy, he will never disclose an item that decreases his power unless he is required to. Second, if disclosure of an item does not reduce his power, he may disclose it in a natural message-signal equilibrium. Since the three players' payoffs are do not change due to the release of such an item, neither do their incentives to defect. This intuition implies the following corollary of Theorem 2:

Corollary 3. *(a) A natural message-signal equilibrium cannot be sustained in which the agent voluntarily discloses an item when doing so would increase the principal's power. (b) If a natural message-signal equilibrium exists, then there exists such an equilibrium in which he voluntarily discloses any item(s) when doing so does not increase the her power.*

This corollary indicates within this model, the agent definitely withholds information only when releasing it would increase the principal's power. Although part (b) allows agencies to keep information from public view even when disclosing it has no implications for power, withholding in these cases is inconsequential since the policies and power levels are the same. Thus, this result suggests that a key reason for agencies' withholding information relating to policy deliberations as well as information they receive from regulated parties, when it

matters, is that other participants in the regulatory policymaking process might be able to exercise more power with the release of information.

3.2 Existence of Natural Message-Signal Equilibria

With the types of natural message-signal equilibria that can exist and the disclosure patterns of the agent established, the next question is when a natural message-signal equilibrium can be sustained. Although in practice, the APA's notice-and-comment requirement makes the agent's proposal transparent, whether he must disclose it turns out to be unimportant. Instead, the key question is whether each of the target's message and agent's signal is transparent. Because these items come from different sources, they will often be separable. Two exemptions in the FOIA approximately track the distinction between these types of records: Exemption 4 , which consists of "trade secrets and commercial or financial information obtained from a person and privileged or confidential," and Exemption 5, which includes "intra-agency memorandums or letters which would not be available by law to a party other than an agency in litigation with the agency."[4] Thus, it is reasonable to consider four transparency modes. The extremes are considered first, followed by intermediate modes.

The first mode, in which only the proposal may be transparent, is arguably the default one. In general, agencies do not have to place all relevant information in a docket unless mandated by law (see Kerwin and Furlong 2011, 65). The agent in the model may take the opportunity to withhold information; however, the principal will be always able to determine what signal he has, even if he does not disclose the low-cost signal.

Proposition 4. *When disclosure of the message and signal is optional for the agent, there*

[4] 5 U.S.C. §§552(b)(4)-(5) (2006).

always exists a natural message-signal equilibrium, but the principal always has her baseline power. The principal's payoff from this equilibrium increases the preference divergence decreases or her baseline power increases.

This proposition contains a formal statement of the first key result mentioned in the introduction, since the principal has the same knowledge about the target's type in a message-signal equilibrium. Similar results are found in theoretical work which the principal can elicit voluntary disclosure of information from the high type based on skepticism when she lacks credible information that the target's type is high (see, e.g., Milgrom 1981, Okuno-Fujiwara, Postlewaite, and Suzumura 1990). It extends these results into a setting in which a mediating agent is deciding whether to process the sender's (i.e., target's) information and his ability to determine the type is imperfect. A limiting aspect of Proposition 4 is that, under the Refinement, the principal cannot induce the agent to disclose either the target's message or the low-cost signal when each increases her likelihood of selecting the final policy. Thus, she is never able to benefit from the empowering effect of these items of information.

The message-signal equilibrium is supported in three specific ways: First, the regulated party would rather transmit a message than be perceived as a low type if it does not convey a message because even when the low-cost signal appears, it is partially pooling and yields a level of regulation less costly to it than x_l^A or x_l^P. Second, the agent with the high-cost signal can disclose to induce the best policy he can reasonably expect from the principal, x_i^P. He can distinguish himself from the agent with no signal or a low-cost signal as necessary. Finally, an agent who has not interpreted the target's message can never distinguish himself from an agent with the low-cost signal, which means that the principal can prevent him from defecting from this equilibrium by not scrutinizing the target's message. More generally, the principal's ability to select policies less favorable to the target and the agent induces messaging, signal

generation, and disclosures sufficient for her to determine the agent's signal. In a reversal of Madison's aphorism, it is power that gives knowledge.

The equilibrium in Proposition 4 yields the principal

$$\tau_i(\pi EU_i^P(x_i^P) + (1-\pi)EU_i^P(x_i^A)) + \tau_k(\pi EU_k^P(x_k^P) + (1-\pi)EU_k^P(x_k^A)). \tag{2}$$

This expected utility can exceed her default payoff, $EU^P(x_j^P)$. Also, she can achieve her second-best payoff, $\tau_i EU_i^P(x_i^A) + \tau_k EU_k^P(x_k^A)$, if she has complete power. This fact foreshadows an important implication for institutional design, that a direct increase in power could better serve public interest groups than transparency in settings like rulemaking.

At the other extreme is transparency of both the message and signal. With mandatory disclosure, the principal can automatically increase her power whenever the message or low-cost signal is created since she will see these items. However, the agent and target are worse off when the principal increases her power, which means that they may have an incentive not to generate information in the first place. A natural message-signal equilibrium will not always exist, and the next result indicates when it does:

Proposition 5. *When the agent's message and signal are transparent, a natural message-signal equilibrium exists if and only if the following are satisfied respectively for the agent and low-cost target:*

$$\begin{aligned}(\pi + \Delta\pi_m)EU_j^A(x_j^P) + (1-\pi-\Delta\pi_m)EU_j^A(x_j^A) \\ \leq \tau_i((\pi+\Delta\pi_m)EU_i^A(x_i^P) + (1-\pi-\Delta\pi_m)EU_i^A(x_i^A)) \\ + \tau_k(\pi+\Delta\pi_m+\Delta\pi_{\tilde{L}})EU_k^A(x_k^P) + (1-\pi-\Delta\pi_m-\Delta\pi_{\tilde{L}})EU_k^A(x_k^A)),\end{aligned} \tag{3}$$

$$\textit{and } \pi c(x_l^P) + (1-\pi)c(x_l^A) \geq (1-\alpha)((\pi + \Delta\pi_m)c(x_i^P) + (1-\pi-\Delta\pi_m)c(x_i^A))$$
$$+ \alpha(\pi + \Delta\pi_m + \Delta\pi_{\tilde{L}})c(x_k^P) + (1-\pi-\Delta\pi_m-\Delta\pi_{\tilde{L}})c(x_k^A)). \quad (4)$$

If this equilibrium exists, then the principal's payoff is the highest among message-signal equilibria in any transparency mode. If only Inequality (3) *fails, then no natural equilibrium in which the agent generates a signal can be sustained. If only Inequality* (4) *fails, the low-cost type will choose not to send a message with some positive probability in any natural equilibrium.*

Proposition 5 implies that transparency will have one of two main effects. First, there may be a shift from message-signal equilibrium at the principal's baseline power to one in which the principal takes advantage of power increases. In that case her payoff is

$$\tau_i((\pi + \Delta\pi_m)EU_i^P(x_i^P) + (1-\pi-\Delta\pi_m)EU_i^P(x_i^A))$$
$$+ \tau_k(\pi + \Delta\pi_m + \Delta\pi_{\tilde{L}})EU_k^P(x_k^P) + (1-\pi-\Delta\pi_m-\Delta\pi_{\tilde{L}})EU_k^P(x_k^A)),$$

which exceeds her payoff in Expression (2) by

$$\tau_i\Delta\pi_m(EU_i^P(x_i^P) - EU_i^P(x_i^A)) + \tau_k(\Delta\pi_m + \Delta\pi_{\tilde{L}})(EU_k^P(x_k^P) - EU_k^P(x_k^A)) \geq 0.$$

This inequality holds strictly when the disclosure either item of information strictly increases her power.

The other possibility, however, is that the message-signal equilibrium is unable to hold, which occurs when either inequality in the proposition fails. Inequality (3) is the individual rationality constraint for the agent to prefer generating a signal based on a message he has

received. The agent's constraint exists largely because, with a transparency requirement, he is able to show that he has not generated a signal because no signal implies that the agent has no additional knowledge about the target's type. It did not exist under optional disclosure because in that setting the principal could confuse him for an agent withholding the low-cost signal. If the equilibrium fails because the agent would defect, the most likely outcome is that both target types communicate with the agent and the principal chooses uninformed. Then principal receives $(\pi + \Delta\pi_m)EU_j^P(x_j^P) + (1 - \pi - \Delta\pi_m)EU_j^P(x_j^A)$, which is less than her default payoff of $EU_j^P(x_j^P)$ unless $\pi + \Delta\pi_m = 1$. Thus, whereas the principal can benefit from the agent with just optional disclosure, she cannot expect to benefit, and her utility will quite possibly decrease if the agent does not generate a signal.[5]

Meanwhile, Inequality (4) is the individual rationality constraint that determines whether the low-cost target keeps sending a message.[6] This constraint exists because the target might prefer being identified as a low-cost target if it can more often receive the policy selection of the agent, who is more favorable to it. If the equilibrium fails because of this constraint, possibilities for equilibria are that the low-cost target chooses not to message with some positive probability while the high-cost target continues to message, and that neither type messages. The latter kind of equilibrium would clearly harm the principal, just like one in which the agent opts not to specialize. The former kind, however, could benefit the principal if the agent generates a signal, because then these players are acting on better information.

[5]There is a remote possibility that a natural equilibrium might be sustainable in which the high-cost target always messages, the low-cost target mixes between messaging and not messaging, and the agent does not generate a signal. However, the low-cost target would be willing to pool with the high-type if the agent would generate a signal, which means that it receives costlier policies on average when he does not generate a signal. This is unusual, because scrutiny should reduce the high-cost target's costs while raising those for the low-cost target. Even in this case, however, the benefits would arise not from content of any disclosed information, but from the inferences the agent and principal are able to make based on whether the target sent a message.

[6]The high-cost target also has an individual rationality constraint, but it is not binding.

However, gains compared to no transparency are not guaranteed and are limited by the fact that $x_l^A < x_k^P$ is necessary for the low type to want to defect.[7] In this alternative equilibrium, then, the agent facing the low-cost type is selecting quite a lower level of regulation than what the principal would select just based on a low-cost signal (when both types message).

Also, with a full transparency requirement, the message-signal equilibrium with the principal's baseline power is no longer available. Overall, increased transparency carries potential benefits and costs. The resulting equilibrium is better if a message-signal equilibrium still obtains, probably worse if only the agent would defect from this equilibrium, and better or worse when the low-cost type would defect from it. Importantly, if the principal benefits from transparency, it is *not* because she becomes better informed through the disclosure of the message or low-cost signal. In a natural message-signal equilibrium, the only function of information disclosure is to increase her power. In an equilibrium in which the low-cost type only sometimes messages, the she benefits as she infers from the target's *lack* of a message that it is a low-cost type. Thus, Proposition 5 is an example of the second main result in the introduction.

The intermediate transparency modes operate similarly to compete transparency, albeit in a lesser way. First, if the signal but not the message is transparent, the existence of a message-signal equilibrium depends on incentives for the agent and low-cost target:

Proposition 6. *When the signal is transparent but not the message, a natural message-signal equilibrium exists if and only if the following inequalities hold respectively for the agent and*

[7]Otherwise, defecting would cause the target always to receive regulations that are more stringent than any policy it would have received after messaging.

low-cost target:

$$\pi EU_j^A(x_j^P) + (1-\pi)EU_j^A(x_j^A) \leq \tau_i(\pi EU_i^A(x_i^P) + (1-\pi)EU_i^A(x_i^A)) + \tau_k(\pi + \Delta\pi_{\tilde{L}})EU_k^A(x_k^P) + (1-\pi-\Delta\pi_{\tilde{L}})EU_k^A(x_k^A)), \quad (5)$$

$$\textit{and } \pi c(x_l^P) + (1-\pi)c(x_l^A) \geq (1-\alpha)(\pi c(x_i^P) + (1-\pi)c(x_i^A)) + \alpha(\pi + \Delta\pi_{\tilde{L}})c(x_k^P) + (1-\pi-\Delta\pi_{\tilde{L}})c(x_k^A)). \quad (6)$$

If this equilibrium exists, then the principal's payoff is weakly less (more) than any natural message-signal equilibrium in which both the signal and message are (not) transparent. If only Inequality (5) *fails, then no natural equilibrium in which the agent generates a signal can be sustained. If only Inequality* (6) *fails, the low-cost type will choose not to send a message with some positive probability in any natural equilibrium.*

The possibility that the agent might not want to generate a signal when it would have to disclose it and lose power to the principal is roughly consistent with the notion that, with transparency of the agency's information, "officials will not engage in as probing and self-critical forms of deliberation because they know that outsiders . . . perhaps will try to use against them later what they say when simply 'thinking aloud'" (Coglianese 2009, 536). Inequality (6), which clearly could hold, suggests that mandating disclosure of agency's information could also discourage the regulated from communicating with the agency. This possibility makes sense because levels of regulation higher than the agent's optimum for the low-cost signal are also costlier for the target, but it does not appear to explored in prior studies. The logic for the remainder of the proposition is similar to that for analogous part of Proposition 5.

Meanwhile, if she chooses to make only the target's message transparent, the intuitive

result is that she only has to worry about the possibility that the regulated party would stop communicating with the agent.

Proposition 7. *When the message is transparent but not the signal, a natural message-signal equilibrium exists if and only if the following holds for the low-cost target:*

$$\pi c(x_l^P) + (1-\pi)c(x_l^A) \geq (1-\alpha)((\pi + \Delta\pi_m)c(x_i^P) + (1-\pi-\Delta\pi_m)c(x_i^A)) \\ + \alpha(\pi + \Delta\pi_m)c(x_k^P) + (1-\pi-\Delta\pi_m)c(x_k^A)). \tag{7}$$

If this equilibrium exists, then the principal's payoff is weakly less (more) than any natural message-signal equilibrium in which both the signal and message are (not) transparent. If Inequality (7) *fails, the low-cost type will choose not to send a message with some positive probability in any natural equilibrium.*

Inequality (7) expresses the intuitive notion that requiring an agency to disclose information provided voluntarily by regulated parties could cause these parties to withhold their information. The case *Critical Mass Energy Project v. Nuclear Regulatory Commission* embodies this logic;[8] however, it does not include the caveat that the principal could potentially benefit if the target communicates less often with the agency. In terms of the model, she might benefit from an equilibrium in which the low-cost target partially or completely separates from the high-cost type.

One final note for the two intermediate transparency forms is that they affect the agent and target's incentive to generate information differently. Suppose disclosure of the message and of the low-cost signal empower the principal equally. Message transparency is clearly less risky with the agent, who will generate a signal if the message is transparent, but not

[8]975 F.2d 871 (D.C. Cir. 1992).

necessarily if the low-cost signal is transparent. The comparison is the reverse for the target. It is more likely to withhold its message when it is transparent than if the low-cost signal is because the power increase in the former case applies to both signals. Thus, there is a general principal that mandating disclosure of each of these items will make its source less likely to create it. Compared to complete transparency, these intermediate modes allow the principal to surrender power from one kind of information in exchange for a reduced risk that the signal will not be generated and that the message-signal equilibrium will disappear

3.3 Equilibria with No Empowerment

One final way to suggest that transparency is more about power than about information is to consider the equilibrium results when the principal gains no power from either the disclosure of the target's message or a low-cost signal. Applying the other results from this section yields the following:

Proposition 8. *Suppose* $\Delta\pi_m = \Delta\pi_{\tilde{L}} = 0$. *If the signal is not transparent, natural message-signal equilibria exist, including one in which the agent discloses all of his items. If the signal is transparent, whether a natural message-signal equilibrium depends on whether Inequality* (3) *holds. If it does not, the agent and principal learn nothing about the target's type in any natural equilibria, in which case the principal receives no more than her default payoff.*

Remark. It is possible for signal transparency to eliminate a natural signaling equilibria, but not likely since the agent prefers to choose policies according to the signal when he has authority and probably prefers that the principal do the same when she has authority.

This last proposition formally states the third result from the introduction and implies that it is difficult to account for any withholding of information when its disclosure does not

increase the principal's power. Propositions 4 and 8 subtly different: when the message and signal are not transparent, the former provides only that the agent will provide information sufficient for the principal to have as much knowledge about the policy question as he does, while the latter states that he has no reason ever to withhold from her any information that leads to that knowledge.

4 Applicability of the Results

The results in the previous section are particular to the model, which has a clearly defined scope. Thus, it is worth considering what other aspects of regulatory politics could be incorporated without affecting the essence of these results. Thus, this section will check the robustness of two principles: (1) without transparency, the principal can know what the agent does about the regulated party's costs, and (2) without transparency and without any power increases from information disclosure, the agent has no reason to withhold any information he has about the target's costs.

4.1 Features that Maintain Both Principles

Since the second principle imposes an extra condition compared to the first, it applies whenever the first does. Thus, it is sufficient to consider whether the first principle still holds for each feature. The first feature that maintains both principles is uncertainty about the agent's preference divergence. This addition to the model captures the idea that public interest groups might want to discover the extent to which regulated parties have captured an agency. However, the principal only needs to know what the signal is to determine what policy to select when she has authority. She will be able to learn what the agent knows if

he discloses the high-cost signal when he has it, because then she can infer that the signal points to low costs if she does not receive one. Regardless of his type, the agent will disclose the high-cost signal because he faces no consequences for his policy selection.

As a practical matter, it is probably the case that agency policymakers, most of whom have civil service protections, are not likely to face direct sanctions for proposing a rule unduly favors regulated parties. Even if such discipline were theoretically possible, there would still be difficulties in determining whether capture had occurred (see Carpenter 2013). However, one can incorporate a punishment based on what the agent proposes and still achieve results that comport with both principles. An agent who has been captured might want to conceal the content of his proposed policy. However, will still want to provide the high-cost signal to induce the principal to select her optimal policy based on the signal, rather than what she would select if she thought he had a low-cost signal. Thus, a proposal-based punishment is a second element that would not affect the essence of the results.

One more feature that will not affect either principle is if the principal can credibly convey to the agent her own policy-relevant information that is independent of the target's message and the agent's signal. It is reasonable to suppose that entities other than the agency and regulated interests have some expertise, even if not the same amount (see Kerwin and Furlong 2011, 167–68). Then the agent would accept the information before the proposal, combine it with his signal, and propose policy accordingly. He would continue to disclose the high-cost signal, and the principal would infer from anything but a high-cost signal that he had the low-cost signal. Like the other two changes, the one leaves the principal with no informational reason to desire transparency. Admittedly, unilateral information provision does not capture more interactive types of communication (cf. Coglianese, Kilmartin, and Mendelson 2009, 932–33). If the principal's ability to contribute depends observing the

target's message or the agent's signal, then she might desire transparency of these items, even though in a narrow sense she would not be any less informed than the agent about the target's cost.

4.2 Features that Maintain the Second Principle

There are other amendments to the game that would only uphold the second principle. First, the high-cost signal might also confer power to the principal. For example, an email exchange might clearly indicate that agency official believes that costs are high but also contain unflattering remarks that generate negative attention for the agency and allow concerned citizens to exert more influence over the rulemaking process. Then the agent would not necessary disclose the high-cost signal when doing so would reduce his power. However, if no disclosures increase power, then the equilibrium results are the same as in the original model, and the agent has no reason to withhold any item of information.

Though the notion of transparency naturally suggests that concerned citizens will interpret disclosed information in a predictable way, such information may not have a clear meaning to public audiences (cf. Fenster 2006, 924–27). Thus, another assumption of the game that could be challenged is that the agent can credibly communicate the signal at all; instead, the principal might read both signals the same way. Then the agent with the low-cost signal might try to mimic the agent with the high-cost signal, in which case only he would know the signal. Just as the agent and principal can induce the target to communicate with a belief that it has low costs if it does not, she can induce the agent to communicate with a belief that he has the low-cost signal if he does not. In that case, the agent is still not withholding any information.

4.3 Reasons for Nondisclosure in the Absence of Power Increases

Finally, there are ways to change the model so that the agent would want to withhold particular items of information, even if doing so would not transfer power to the principal. The most general method is to ascribe some other type of cost to the disclosure of particular items of information. Financial costs could be a significant reason for withholding information; for example, the costs of implementing the FOIA have been much greater than anticipated (Wichmann 1998, 1220). However, agencies disclose much information in their dockets, and, in some cases, they may be revealing everything that is relevant to the policy question. Another other type of cost is psychological: Coglianese (2009) notes that transparency could "inhibit other, desirable behavior—such as internal dissent or asking the proverbial dumb question—that might be embarrassing but is still necessary for good decision making" (536). These types of costs yield alternative explanations agencies' desire information; however, they have in common with the empowerment theory that they are not based on an agent's desire to withhold information about the policy from the principal.

Another element that the model does not include is agent competence. This is a key feature in models like Prat (2005), in which transparency does grant the principal additional knowledge, albeit about the agent's capabilities. The game would have to be expanded substantially to incorporate this feature; for example, it might be necessary for the principal to learn the true costs of the industry and use that information to draw inferences about how capable the agent is. The model does not consider this characteristic of agents for two reasons: first, because the principal might not ever determine the regulated party's true cost level, and second, because public interest groups seem to be more concerned about agency officials' bias rather than their abilities.

Finally, the model is designed to match the rulemaking process, and its conclusions may

not be appropriate for settings that are very different from it. For example, in policy areas where government officials can make decisions secretly, such as defense and national security, citizens might not even be aware that a change has taken place apart from transparency requirements. In rulemaking, however, the APA "ensure[s] that agencies cannot secretly conspire against elected officials by presenting them with a fait accompli" (McCubbins, Noll, and Weingast 1987, 258). Although rulemaking is the core type of policymaking envisioned by the model, it could also be applied to similar decision-making formats, such as agency guidance.

Overall, while there are aspects of regulatory policymaking that could challenge the general results that the principal can learn what the agent knows about the policy question without transparency and that the agent has no reason to withhold information when there are no implications for power, this exploration of additional features suggests that there is a fairly broad range of circumstances in which the general logic of the model can operate. The potential for transparency to improve policy outcomes by increasing knowledge through release documents would appear to be limited once it is recognized that it is possible to make inferences about policy apart from released documents and that agencies might not have an incentive to withhold information.

5 Policy Implications

It is plausible that power considerations actually influence agency decisions as to whether to generate and disseminate information. Withholding information is a form of secrecy, and Stiglitz (2002) contends that "making decision in secret . . . is much easier than making them in full public view" (34). There is also survey evidence that arguably supports this theory in

the form of agency officials who reported that they stopped communicating with "affected interests" when information must be docketed in part because of fear that it could be used in a legal challenge (West 2004, 70). In addition to surveying agency officials, there are some ways to determine whether the logic of the model is operating in a given regulatory arena, and there are possibilities for institutional design whenever this logic proves important.

5.1 Empirical Implications

There are many kinds of evidence that would indicate that an agency is not withholding documents and information in a way that reduces concerned citizens' knowledge policy questions. However, because an agency might withhold information because of direct costs rather than from the implicit costs of lost power, more specific documentary evidence is necessary to identify power as the main reason. Nonetheless, there are at least three signs that would indicate that some aspect of the model is operating.

First, an agency may voluntarily release a large amount of relevant information relating to a proposed rule. If the information is adequate for the outside participants in the rulemaking process to ascertain their ideal policy, then it would show that, even without transparency, the agency provides enough documentation for these participants to have good knowledge about the policy. Following Proposition 4, the record could be adequate in the sense that no evidence or weak evidence for for some form of regulation implies that stronger evidence does not exist. In addition, the docket may include explicit information that supports a stringent level of regulation and information it received from regulated parties, even though the agency might be thought to be biased toward regulated parties. In that case, such a disclosure pattern would suggest that the agency does not expect outside participants to be able to use the information to increase their influence over the policy.

Second, the empowering effect of disclosure could be shown if settings in which an agency is required to keep a comprehensive docket engender more political or judicial intervention than comparable settings in which the agency can choose what information to select, and the inferences that can be drawn from the two settings about what regulation should be promulgated are approximately the same. A greater frequency of intervention distinguishes the empowerment effect of disclosure from any of its direct costs since merely embarrassing or expensive disclosures may impose costs, but they would not yield more policy changes. Admittedly, it would not be easy to find settings for comparison because issue areas that are more politically salient may be more likely to have transparency requirements. The key challenge would be determining what the frequency of these interventions would be apart from the transparency requirements.

Third, concerned citizens may be able to obtain information from the agency that it has neither voluntarily released nor disclosed because of some reporting requirement through a FOIA request or perhaps through some form of political pressure. If the information revealed does not really add to the requestors' understanding of what regulation is preferable or obviously embarrass agency officials, and if the agency did not resist disclosure out of monetary concerns, then it is reasonable to believe that the agency withheld the document because it believed that that document would lend support to some kind of challenge against the regulation. Here, concerns about what the baseline level of power is are relatively small since the same rulemaking is under consideration in both cases.

5.2 Institutional Design Implications

When the logic of the model is operating, there are two major implications for the value of transparency. First, instead of requiring disclosure of information, public interest groups can

make deductions based on the information available. They can claim that weak support for lenient regulation implies that a stricter policy is warranted. Elected officials to whom these groups might appeal for intervention should consider these claims. When these groups have standing to challenge regulation, then a court should be willing to make similar inferences based on an incomplete record. These techniques map onto skeptical beliefs from the lack of a high-cost signal in the model. If the model applies, then formally requiring disclosure may increase the power of public interest groups, but they may encounter resistance in compelling the release of documents, and there may be suspicious that an agency has not provided all information. Instead of requiring disclosure for more power, inducing disclosure with less power may be an attractive alternative.

A more important ramification of the model is that a higher baseline level of power combined with skeptical beliefs appears to be a superior alternative to relying on information disclosure to increase power. When certain information disclosures increase power in the model, individual rationality constraints arise. However, if the principal's power baseline were equal to her increased level of power to begin with, she could still induce enough information disclosure to infer what the agent knows. Thus, public interest groups would should prefer more formal power over greater transparency.

However, these groups have relatively little formal power. Even when they have standing to challenge regulation, their challenge may not be very beneficial because success typically yields not the regulation that they would prefer, but no regulation. The Administrative Procedures Act appears to have been designed to protect the status quo of New Deal Regulation (McNollgast 1999). For advocates of stricter regulation, maintaining the status quo means less progress (in their view) on various issues. When a proposed regulation would create more benefits than current law but not as much as pubic interest groups would like,

they have mixed motives about challenging the regulation. Regulated parties, in contrast, are unconflicted about filing a lawsuit if they can because doing so will delay and possibly invalidate the regulation.

It is possible to increase public interest advocates' power in court by attaching so-called hammer provisions to legislation, which set a default policy that applies if the agency does not promulgate a regulation by a specified deadline (see Kerwin and Furlong 2011, 226). Then the agency and regulated parties would be placed in a position of having to produce information to support regulation less stringent than the hammer's default. The information would not be compelled by a transparency law, but induced by the threat of adverse regulation. An example in which this dynamic worked is the Hazardous and Solid Waste Amendments (HSWA) of 1984, which prohibited land disposal of certain untreated hazardous wastes if EPA did not promulgate standards by various deadlines (Corwin 1992, 539). It caused the regulated industry to provide more data more quickly than was typical in that policy arena (id., 540). A hammer need not be as severe as an absolute prohibition of a substance. However, inserting defaults that are substantially more stringent that current regulation places advocates of stricter regulation in a stronger position. In addition to inducing disclosure of information, it may also force regulated parties to produce information that is more comprehensible and not overly voluminous, mitigating what Wagner (2010) calls "filter failure."

6 Conclusion

Although transparency in the form of mandatory information disclosure is designed to improve people's knowledge about regulatory policy questions, the model presented in this

paper suggests that there is a non-trivial set of cases in which it cannot be expected to have this effect. If policy-relevant information that would be transmitted with a transparency rule has a clear meaning, then the agency could always disclose it voluntarily. Although it may not disclose all of its information, the model suggests that citizens may be able to infer what the agency knows about what regulation would be optimal from information that is missing from as well as present in the record. The intuition is that, with the ability to select policy that is worse for the agency and regulated parties some of the time, outside participants in the rulemaking process can induce the agency to produce information that supports less stringent regulation.

When an agency discloses information sufficient for citizens to determine what regulation they would prefer, the impacts of transparency will not arise from what they learn from the content of released documents. Instead, they are likely to stem from the empowering effect of disclosures. Empowerment can be beneficial, but it also presents the risk that information will not be generated in the first place. In addition, the benefits could be achieved without the risk by increasing the principal's baseline power instead. Thus, for groups seeking more stringent regulation, more formal power to overturn an agency's proposal and substitute it with their own would seem to yield better results than transparency.

Starting from Madison, discussions about transparency have mentioned its effect on citizens' knowledge and on their power. The model presented in this paper contributes to the understanding of the benefits and costs of mandatory disclosure by more explicitly distinguishing these two effects. At least in rulemaking and similar settings, transparency seems to be more about increasing power than about increasing policy-related knowledge. Some empowering effect appears to be necessary, as it is difficult to account for information withholding when releasing documents has no direct impact on the agency. The relative

importance of the two effects may differ in other policymaking settings, but results from this model suggest that it is generally important to separate these effects to the extent possible when assessing the value of information transparency.

A Equilibrium Refinements

The reason for having equilibrium refinements is to prevent the principal from having arbitrary beliefs for disclosures off the equilibrium path. Clearly, the principal cannot credibly threaten to set policy above x_l^P, because even if she knows that the target's costs are low, she will not want to have regulation more stringent than this level. However, there are still implausible equilibria if the principal's beliefs are not restricted further. For instance, there may be equilibria in which the agent discloses everything out of fear that the principal will select x_l^P, even though the disclosure of a high-cost signal would indicate that she should not select x_l^P. As noted in the main text, the standard equilibrium refinements assume a finite action space and just two players, so it appears to be necessary to customize refinements for this model. Two refinements are offered to suggest that the choice of refinement is not arbitrary.

The first refinement is an adaptation of the Intuitive Criterion of Cho and Kreps (1987). Some new notation is needed: First, it is useful to define the partial strategy profile $\sigma^{-P} \equiv (\sigma^H, \sigma^L, \sigma^A)$ and $\theta \equiv (\sigma^{-P}, s)$ as a partial strategy profile combined with whatever signal the agent received, or colloquially, a "strategy-signal profile." To reflect the idea that defection by just one player or type at a time is considered, the relation $\sigma^{-P\prime} \approx \sigma^{-P}$ is defined to apply when exactly two of $\sigma^{H\prime} = \sigma^H$, $\sigma^{L\prime} = \sigma^L$, and $\sigma^{A\prime} = \sigma^A$ hold. Then for any disclosure off the equilibrium path $\mathring{d}$, $\Theta(\mathring{d}) \equiv \{\theta : \sigma^{-P} \approx \sigma^{-P*} \wedge \theta \Rightarrow \mathring{d}\}$, where σ^{-P*} is the equilibrium partial

strategy profile. Informally, this set consists of strategy-signal profiles that could yield the disclosure such that only one player or type is changing strategy. Finally, $q \in \{H, L, A\}$ will denote the player or type with $\sigma^q \neq \sigma^{q*}$ and have expected utility EU^q, defined from the point at which q defects. For consistency, the target's utility is the negative of its costs. Now the refinement can be stated:

Refinement A.1. For any $\mathring{d}$ off the equilibrium path, $\Pr(\theta) > 0$ only if $\theta \in \Theta(\mathring{d})$ and

$$EU^q(\sigma^*) < \max_{\substack{\beta_L^P \in [\min_{\theta \in \Theta(\mathring{d})} \beta_L^P(\theta)\}, \\ \max_{\theta \in \Theta(\mathring{d})} \beta_L^P(\theta)]}} EU^q(\theta, \sigma^P(\beta_L^P)), \tag{A.1}$$

If no $\theta \in \Theta(\mathring{d})$ satisfies Inequality (A.1), the principal may freely set $\beta_L^P(\mathring{d})$.

Refinement A.1, like the Intuitive Criterion, qualitatively dictates that the principal should place zero probability on strategy-signal profiles for which the player who is defecting could not gain from the deviation if the principal chose of her best responses based on the set of profiles with a single defector that could have produced the disclosure (cf. McCarty and Meirowitz 2007, 243).

The second refinement requires one more function to be defined: $\mathring{\beta}_L^P(\mathring{d}_x)$ will be the value of β_L^P that satisfies $\mathring{d}_x = \arg\max_x b(x) - (1+a)(\beta_L^P l + (1-\beta_L^P)h)c(x)$ if $\mathring{d}_x \in [x_h^A, x_l^A]$. Also, $\mathring{\beta}_L^P(\mathring{d}_x) \equiv 1$ when $\mathring{d}_x > x_l^A$, and $\mathring{\beta}_L^P \equiv 0$ when $\mathring{d}_x < x_h^A$ or $\mathring{d}_x = \emptyset$.

Refinement A.2. For any $\mathring{d}$ off the equilibrium path, $\Pr(\theta) > 0$ only if $\theta \in \Theta(\mathring{d})$ and

$$EU^q(\sigma^*) < \max_{\substack{\beta_L^P \geq \max\{\mathring{\beta}_L^P, \\ \min_{\theta \in \Theta(\mathring{d})} \beta_L^P(\theta)\}}} EU^q(\theta, \sigma^P(\beta_L^P)) \tag{A.2}$$

If no $\theta \in \Theta(\mathring{d})$ satisfies Inequality (A.2), the principal may freely set $\beta_L^P(\mathring{d})$.

This refinement is not based on any other standard refinement but captures the notion that the principal should not have a lower posterior probability that the regulated party's costs are low than the agent's proposal indicates. It implies that, if the agent wants the principal to select a lower policy, he should disclose a lower policy proposal or not disclose his proposal. In addition to corroborating the first refinement, this refinement is somewhat easier to apply and arguably provides more "intuitive support" for eliminating implausible equilibria.

Though the refinements operate differently, it makes no difference which refinement is applied for the purposes of the derived results. Thus, either there is a message-signal equilibrium satisfies both refinements or any such equilibrium fails both of them.

B Proofs of Numbered Results

Proof of Lemma 1 The first step is to show that $EU^q(r) \equiv \max_x b(x) - (1+a)rc(x)$ is convex with respect to r. The functional form assumptions on $b(\cdot)$ and $c(\cdot)$ allow the use of the envelope theorem to determine that $\frac{\partial}{\partial r}EU^q(r) = (1+a)c(x^*)$, where x^* maximizes $EU^q(r)$ for a given r. Then $\frac{\partial}{\partial r^2}EU^q(r) = -(1+a)c(x^*)\frac{\partial x^*}{\partial r}$. This expression is positive since $\frac{\partial x^*}{\partial r} = -\frac{(1+a)c'(x^*)}{(1+a)rc''(x)-b''(x)} < 0$. This convexity and the fact that $EU^q_t(x^q_t) = EU^q(t)$ for $t \in \{h,i,j,k,l\}$ implies that $EU^q_j(x^q_j) = EU^q(\tau_i i + \tau_k k) < \tau_i EU^q(i) + \tau_k EU^q(k) = \tau_i EU^q_i(x^q_i)) + \tau_k EU^q_k(x^q_k) = (\alpha\tau_h + (1-\alpha)\tau_l)EU^q\big(\frac{\alpha\tau_h h + (1-\alpha)\tau_l l}{\alpha\tau_h + (1-\alpha)\tau_l}\big) + (\alpha\tau_l + (1-\alpha)\tau_l)EU^q\big(\frac{\alpha\tau_l l + (1-\alpha)\tau_h h}{\alpha\tau_l + (1-\alpha)\tau_h}\big) < \alpha\tau_h EU^q(h) + (1-\alpha)\tau_l EU^q(l) + \alpha\tau_l EU^q(l) + (1-\alpha)\tau_h EU^q(h) = \tau_h EU^q_h(x^q_h)) + \tau_l EU^q_l(x^q_l)$. ■

For the remaining results, it will be useful to notate A_s as the agent with signal s for $s \in \{H, L\}$, A_m as the agent with m but no signal, and $A_\emptyset$ as the agent with no message.

Proof of Theorem 2 First, $A_{\tilde{H}}$ will always achieve $x^A = x_i^A$, $x^P = x_i^P$ and, when $\Delta\pi_m > 0$, $d_m = \delta$ only when m is transparent in a natural equilibrium. If $\sigma^H = \sigma^L = m$ and $\sigma_n^A(m) = n$, weak consistency requires $x^P \geq x_i^P$ on the equilibrium path. Then the following inequality holds:

$$(\pi + \mathbf{1}_m\Delta\pi_m)EU_i^A(x^P) + (1 - \pi - \mathbf{1}_m\Delta\pi_m)EU_i^A(x^A)$$
$$\leq (\pi + \mathbf{1}_m^{tr}\Delta\pi_m)EU_i^A(x_i^P) + (1 - \pi - \mathbf{1}_m^{tr}\Delta\pi_m)EU_i^A(x_i^A), \forall x^P \geq x_i^P, \quad \text{(B.1)}$$

where $\mathbf{1}_m = 1$ (0) when $d_m = \delta$ ($\emptyset$) in actuality, and $\mathbf{1}_m^{tr} = 1$ (0) when m is (not) transparent. Inequality (B.1) holds whenever $x^P \geq x_i^P$ because, on the right-hand side, A selects his best policy when he has authority, P selects the value of x that exceeds x_i^A by the least when she has authority, and A has maximum power to select policy rather than P (since $\mathbf{1}_m - \mathbf{1}_m^{tr} \geq 0$ and $\Delta\pi_m \geq 0$). This inequality holds strictly if $x^A \neq x_i^A$, $x^P > x_i^P$, or $(\mathbf{1}_m - \mathbf{1}_m^{tr})\Delta\pi_m > 0$, the last of which occurs when $A_{\tilde{H}}$ voluntarily discloses m and thereby increases P's power. If any of these three hold, then, under either refinement, $A_{\tilde{H}}$ could make $\mathring{d} = (\emptyset, \tilde{H}, x_i^A)$ and expect the P to believe $\beta_L^P = \frac{(1-\alpha)\tau_l}{\alpha\tau_h + (1-\alpha)\tau_l}$, in which case Inequalities (A.1) and (A.2) are satisfied because Inequality (B.1) holds strictly. Because only $A_{\tilde{H}}$ can produce this disclosure, P must believe $\beta_L^P = \frac{(1-\alpha)\tau_l}{\alpha\tau_h + (1-\alpha)\tau_l}$. Then Inequality (B.1) implies that $A_{\tilde{H}}$ would choose not to defect if and only if $x^A = x_i^A$, $x^P = x_i^P$, and $(\mathbf{1}_m - \mathbf{1}_m^{tr})\Delta\pi_m = 0$, so that he has maximum power given disclosure constraints.

Because of what follows $s = \tilde{H}$, $A_{\tilde{L}}$ will have $x^A = x_k^A$, $x^P = x_k^P$, and maximum power given the transparency constraints. $A_{\tilde{H}}$ can distinguish himself from $A_{\tilde{L}}$ by disclosing $\tilde{H}$ if needed. Then, when $\sigma^H = \sigma^L = m$ and $\sigma_n^A(m) = n$, weak consistency requires $x^P = x_k^P$ on

the equilibrium path after $s = \tilde{L}$. Analogous to $A_{\tilde{H}}$, the following inequality holds for $A_{\tilde{L}}$:

$$\begin{aligned}&(\pi + \mathbf{1}_m \Delta\pi_m + \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}}) EU_k^A(x^P) + (1 - \pi - \mathbf{1}_m \Delta\pi_m - \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}}) EU_k^A(x^A) \\ &\leq (\pi + \mathbf{1}_m^{tr} \Delta\pi_m + \mathbf{1}_{\tilde{L}}^{tr} \Delta\pi_{\tilde{L}}) EU_k^A(x_k^P) + (1 - \pi - \mathbf{1}_m^{tr} \Delta\pi_m - \mathbf{1}_{\tilde{L}}^{tr} \Delta\pi_{\tilde{L}}) EU_k^A(x_k^A), \forall x^P \geq x_k^P,\end{aligned} \tag{B.2}$$

where $\mathbf{1}_{\tilde{L}} = 1$ (0) when $d_s(\tilde{L}) = \delta$ ($\emptyset$) in actuality, and $\mathbf{1}_{\tilde{L}}^{tr} = 1$ (0) when s is (not) transparent. Inequality (B.2) holds because, on the right-hand side, A selects his best policy when he has authority and A has maximum power to select policy (since $\mathbf{1}_m - \mathbf{1}_m^{tr}$, $\Delta\pi_m$, $\mathbf{1}_{\tilde{L}} - \mathbf{1}_{\tilde{L}}^{tr}$, and $\Delta\pi_{\tilde{L}}$ are all weakly positive). This inequality holds strictly if $x^A \neq x_k^A$, $(\mathbf{1}_m - \mathbf{1}_m^{tr})\Delta\pi_m > 0$, or $(\mathbf{1}_{\tilde{L}} - \mathbf{1}_{\tilde{L}}^{tr})\Delta\pi_{\tilde{L}} > 0$. The last two occur respectively when $A_{\tilde{L}}$ voluntarily discloses m or $\tilde{L}$ and thereby increases P's power.

Now consider $\mathring{d}$ is such that $\mathring{d}_m = m$ if and only if m is transparent, $\mathring{d}_s = \tilde{L}$ if and only if s is transparent, and $\mathring{d}_x = x_k^A$. This disclosure can occur off the equilibrium path if at least one of the three conditions is met for strict satisfaction of Inequality (B.2). Assume that $\mathring{d}$ is *not* A's disclosure when R defects by not messaging. Then P cannot have β_L^P satisfying either refinement that would prevent A from defecting from a proposed message-signal equilibrium.

To begin with, Inequalities (A.1) and (A.2) are satisfied for since $A_{\tilde{L}}$ with this disclosure. The reason is that he strictly benefits by strict satisfaction of Inequality (B.2) if $\beta_L^P = \frac{\alpha\tau_l}{\alpha\tau_l + (1-\alpha)\tau_h}$, which is in the range of permissible beliefs under either refinement. Then $\Pr(\theta)$ can be strictly positive for $\theta \in \Theta(\mathring{d})$ involving $A_{\tilde{L}}$. If s is transparent, then P must assign all the probability to elements of $\Theta(\mathring{d})$ involving $A_{\tilde{L}}$. Then with $\beta_L^P = \frac{\alpha\tau_l}{\alpha\tau_l + (1-\alpha)\tau_h}$, the only permissible belief from $\mathring{d}$, $A_{\tilde{L}}$ will defect, and the proposed equilibrium fails both refinements. If s is not transparent, P may be able to assign strictly positive probabilities to $A_{\tilde{H}}$ or A_m.

From the first part of the proof, $A_{\tilde{H}}$ is receiving the right-hand side of Inequality (B.1), while his payoff from this defection would be a form of the left-hand side. He would expect P to act as though $\beta_L^P \geq \frac{(1-\alpha)\tau_l}{\alpha\tau_h+(1-\alpha)\tau_l}$, for $x^P \geq x_i^P$, so that Inequality (B.1) holds. Then Inequalities (A.1) and (A.2) are not satisfied for $A_{\tilde{H}}$, and P must set $\Pr(\theta) = 0$ for any $\theta \in \Theta(\mathring{d})$ involving $A_{\tilde{H}}$.

The only remaining $\theta \in \Theta(\mathring{d})$ involve A_m. If, by each refinement, P must set $\Pr(\theta) = 0$ for $\theta \in \Theta(\mathring{d})$ involving A_m, then $\beta_L^P = \frac{\alpha\tau_l}{\alpha\tau_l+(1-\alpha)\tau_h}$ is the only permissible belief after $\mathring{d}$), $A_{\tilde{L}}$ will defect, and the proposed equilibrium fails both refinements. Otherwise, it survives only if some $\beta_L^P \in \left[\tau_l, \frac{\alpha\tau_l}{\alpha\tau_l+(1-\alpha)\tau_h}\right]$ prevents both $A_{\tilde{L}}$ and A_m from defecting. If, for Refinement A.2, P can set $\Pr(\theta) > 0$ for some $\theta \in \Theta(\mathring{d})$ involving A_m, it must be possible for $\beta_L^P = \frac{\alpha\tau_l}{\alpha\tau_l+(1-\alpha)\tau_h}$, since it yields the lowest x^P, and $x^P > x_j^P$. Then Inequality (A.1) is also satisfied since $\beta_L^P = \frac{\alpha\tau_l}{\alpha\tau_l+(1-\alpha)\tau_h}$ is a permissible expectation for A_m. Then he is willing to defect for any $\beta_L^P \in \left[\tau_l, \frac{\alpha\tau_l}{\alpha\tau_l+(1-\alpha)\tau_h}\right]$, and the proposed equilibrium does not survive either refinement.

If, instead, $\Pr(\theta) > 0$ for some $\theta \in \Theta(\mathring{d})$ involving A_m only under Refinement A.1, it is only clear that he will defect for $\beta_L^P = \tau_l$ since $\beta_L^P \geq \tau_l$ implies $x^P \geq x_j^P$, the smallest value exceeding x_j^A. Then the utility of A_m from this defection decreases with x^P. Since A_m is not willing to defect when $x^P = x_k^P$, there exists $\hat{x}^P \in (x_j^P, x_k^P)$ such that he is indifferent about deviating, which means that $\tau_i[(\pi+\mathbf{1}_m^{tr}\Delta\pi_m)EU_i^A(x_i^P)+(1-\pi-\mathbf{1}_m^{tr}\Delta\pi_m)EU_i^A(x_i^A)]+\tau_k[(\pi+\mathbf{1}_m\Delta\pi_m+\mathbf{1}_{\tilde{L}}\Delta\pi_{\tilde{L}})EU_k^A(x_k^P)+(1-\pi-\mathbf{1}_m\Delta\pi_m-\mathbf{1}_{\tilde{L}}\Delta\pi_{\tilde{L}})EU_k^A(x^A)] = (\pi+\mathbf{1}_m^{tr}\Delta\pi_m)EU_j^A(\hat{x}^P)+(1-\pi-\mathbf{1}_m^{tr}\Delta\pi_m)EU_j^A(x_k^A)$. Because Inequality (B.1) holds, it must be that

$$(\pi + \mathbf{1}_m\Delta\pi_m + \mathbf{1}_{\tilde{L}}\Delta\pi_{\tilde{L}})EU_k^A(x_k^P) + (1 - \pi - \mathbf{1}_m\Delta\pi_m - \mathbf{1}_{\tilde{L}}\Delta\pi_{\tilde{L}})EU_k^A(x^A) < (\pi + \mathbf{1}_m^{tr}\Delta\pi_m)EU_k^A(\hat{x}^P) + (1 - \pi - \mathbf{1}_m^{tr}\Delta\pi_m)EU_k^A(x_k^A). \quad \text{(B.3)}$$

Since $EU_k^A(x)$ is concave with respect to x, Inequalities (B.2) and (B.3) imply that Inequality (A.1) is satisfied for all $x^P \in [\hat{x}^P, x_k^P]$ for $A_{\tilde{L}}$. Thus, at least one of $A_{\tilde{L}}$ and A_m would deviate for all $\beta_L^P \in \left[\tau_l, \frac{\alpha\tau_l}{\alpha\tau_l+(1-\alpha)\tau_h}\right]$. Then the message-signal equilibrium does not survive Refinement A.1.

Thus, if $x^A \neq x_k^A$, $(\mathbf{1}_m - \mathbf{1}_m^{tr})\Delta\pi_m > 0$, or $(\mathbf{1}_{\tilde{L}} - \mathbf{1}_{\tilde{L}}^{tr})\Delta\pi_{\tilde{L}} > 0$, a proposed message-signal equilibrium cannot satisfy either refinement if $\mathring{d}$ defined above is *not* the disclosure of $A_\emptyset$ off the equilibrium path. Otherwise, strict satisfaction of Inequality (B.2) implies that this inequality would still be strictly satisfied for some $x^{A\prime}$ slightly less than x_k^A. Then for $\mathring{d}'$, which differs from $\mathring{d}$ only in that $\mathring{d}' = x^{A\prime}$ instead of x_k^A and in that it cannot come from $A_\emptyset$ (since $\mathring{d}$ is assumed to come from $A_\emptyset$), P cannot have β_L^P satisfying either refinement that would prevent A from defecting from a proposed message-signal equilibrium. This fact can be shown by repeating the proof about P's beliefs for $\mathring{d}$, but replacing $\mathring{d}$ with $\mathring{d}'$ and x_k^A with $x^{A\prime}$, *mutatis mutandis*. Therefore a natural message-signal equilibrium cannot be sustained unless $x^A = x_k^A$, $x^P = x_k^P$, $(\mathbf{1}_m - \mathbf{1}_m^{tr})\Delta\pi_m = 0$, and $(\mathbf{1}_{\tilde{L}} - \mathbf{1}_{\tilde{L}}^{tr})\Delta\pi_{\tilde{L}} = 0$ after $s = \tilde{L}$. ■

The following Lemma will be involved in proving Propositions 4–8:

Lemma B.1. *In a proposed message-signal equilibrium, H's expected cost divided by h is less than L's divided by L, so H's individual rationality constraint is never binding.*

Proof. The comparison between H's and L's costs, each divided by t, is:

$$\begin{aligned}
&\alpha((\pi + \mathbf{1}_m \Delta\pi_m)c(x_i^P) + (1 - \pi - \mathbf{1}_m \Delta\pi_m)c(x_i^A)) \\
&\qquad + (1-\alpha)(\pi + \mathbf{1}_m \Delta\pi_m + \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}})c(x_k^P) + (1 - \pi - \mathbf{1}_m \Delta\pi_m - \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}})c(x_k^A)) \\
&< (1-\alpha)((\pi + \mathbf{1}_m \Delta\pi_m)c(x_i^P) + (1 - \pi - \mathbf{1}_m \Delta\pi_m)c(x_i^A)) \\
&\qquad + \alpha(\pi + \mathbf{1}_m \Delta\pi_m + \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}})c(x_k^P) + (1 - \pi - \mathbf{1}_m \Delta\pi_m - \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}})c(x_k^A)), \qquad \text{(B.4)}
\end{aligned}$$

where $\mathbf{1}_m$ and $\mathbf{1}_{\tilde{L}}$ respectively denote the indicator functions for when $\mathring{d}_m = m$ and $\mathring{d}_{\tilde{L}} = \tilde{L}$. The reason is that $1 - \alpha < \frac{1}{2} < 1$ while $(\pi + \mathbf{1}_m \Delta\pi_m + \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}})c(x_k^P) + (1 - \pi - \mathbf{1}_m \Delta\pi_m - \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}})c(x_k^A)$ exceeds $(\pi + \mathbf{1}_m \Delta\pi_m)c(x_i^P) + (1 - \pi - \mathbf{1}_m \Delta\pi_m)c(x_i^A)$ by $(\pi + \mathbf{1}_m \Delta\pi_m)(c(x_k^P) - c(x_i^P)) + \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}}(c(x_k^P) - c(x_i^A)) + (1 - \pi - \mathbf{1}_m \Delta\pi_m - \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}})c(x_k^A)(c(x_k^A) - c(x_i^A)) > 0$. The rest of the lemma follows from the fact that H and L will receive the same policies from a deviation since R has not messaging as its only method of defection. ■

Proof of Proposition 4 The following message-signal equilibrium in which P always selects policy with probability π will be shown always to exist and satisfy Refinement A.1 and A.2: $\sigma^H = \sigma^L = m$, $\sigma_n^A(m) = n$; $x^A(\emptyset, \emptyset) = x_l^A$, $x^A(m, \tilde{H}) = x_i^A$, $x^A(m, \tilde{L}) = x_k^A$, $\sigma_d^A(m, \tilde{H}, x_i^A) = (\emptyset, \delta, \delta)$, $\sigma_d^A = (\emptyset, \emptyset, \delta)$ otherwise; and

$$\sigma^P(\mathring{d}_m, \mathring{d}_s, \mathring{d}_x) = \begin{cases} x_i^P & \text{if } \mathring{d}_s = \tilde{H}, \\ x_l^P & \text{if } \mathring{d}_s = (\emptyset, \emptyset, x_l^A), \\ x_k^P & \text{otherwise,} \end{cases}$$

except that, under certain conditions listed below, P may select a different $\sigma^P(\emptyset, \emptyset, x_l^A)$. Note that, since this equilibrium always has $d_x = \delta$, this proof applies whether or not x^A is

transparent. A's and P's beliefs, β_L^A and β_L^P, follow from their strategies for policy selection, so showing sequential rationality implies weak consistency.

Sequential rationality is shown via backward induction. At stage 5 the two disclosures that P will see are $(\emptyset, \tilde{H}, x_i^A)$ when $s = \tilde{H}$ and $(\emptyset, \emptyset, x_j^A)$ when $s = \tilde{L}$. P's given strategy implies that she is selecting optimally based on these signals. At stage 4 A obtains his best payoff from his strategy with each signal for the following reasons: (1) he obtains his ideal policy for the signal (x_i^A or x_k^A) when he has authority, along with the value of x among those that P might choose that exceeds A's ideal policy by the least when she has authority (which is x_k^P for $A_{\tilde{L}}$ since he cannot produce $\tilde{H}$), and (3) A has maximum power. At stage 3 A prefers to process m rather than receive his maximum payoff from not doing so:

$$\tau_i(\pi EU_i^A(x_i^P) + (1-\pi)EU_i^A(x_i^A)) + \tau_k(\pi EU_k^A(x_k^P) + (1-\pi)EU_k^A(x_k^A)) \\ > \pi EU_j^A(x^P) + (1-\pi)EU_j^A(x^A), \forall x^P \geq x_k^P, \quad \text{(B.5)}$$

with $x^A = x_j^A$ and $x^P = x_k^P$ in this case. This inequality holds since $\tau_i EU_i^A(x_i^A)) + \tau_k EU_k^A(x_k^A) > EU_j^A(x_j^A)$ by Lemma 1, and since $x^P \geq x_k^P > x_i^P > x_i^A$ implies for all $x^P \geq x_k^P$ that $\tau_i EU_i^A(x_i^P) + \tau_k EU_k^A(x_k^P) > \tau_i EU_i^A(x^P) + \tau_k EU_k^A(x^P) = EU_j^A(x^P)$. Finally, at stage 2, each type prefers messaging over not messaging: since $x_k^P > x_i^P$ and $x_l^A > x_k^A > x_i^A$, L has $l(\pi(\alpha c(x_k^P) + (1-\alpha)c(x_i^P)) + (1-\pi)(\alpha c(x_k^A) + (1-\alpha)c(x_i^A) < l\pi c(x_k^P) + (1-\pi)c(x_l^A)$, and Lemma B.1 implies that H will not defect.

Robustness of this equilibrium to the refinements: If R defects, $\mathring{d} = (\emptyset, \emptyset, x_l^A)$ results. The goal is to find beliefs for P that satisfy each refinement and that prevent all defections. Under Refinement A.2, $\beta_L^P = 1$ for Inequality (A.2) and is not satisfied for any defector, who would receive worse policies (higher, and for A, further from his optimum) in every situation.

This refinement, then, allows $\beta_L^P = 1$ and x_l^P for P's strategy as initially stated.

Showing that beliefs exist that prevent all defections under Refinement A.1 is more complex. Since H, L, and A can all defect, any $\beta_L^P \in [0,1]$ can be used to satisfy Inequality (A.1). If, for L, this inequality is satisfied with $\beta_L^P = 0$, the most favorable belief, P can set $\beta_L^P = 1$ and, as above, can prevent all defections. If not, L would not defect, and neither would H by Lemma B.1. Then the appropriate β_L^P depends on whether $A_{\tilde{H}}$, A_m, or $A_{\tilde{L}}$ can satisfy Inequality (A.1) when P selects the policy he most prefers between x_h^P and x_l^P. If none of these can, then $\Pr(\theta) = 0$, for all $\theta \in \Theta(\mathring{d})$, P can set $\beta_L^P = 1$, and no defections will occur. If $A_{\tilde{H}}$ alone satisfies Inequality (A.1), then $\beta_L^P = \frac{(1-\alpha)\tau_l}{\alpha\tau_h + (1-\alpha)\tau_l}$, leading to x_i^P. He receives the right-hand side of Inequality (B.1) in the proposed equilibrium, while the left-hand side encompasses his payoff from defection. Since $x^P = x_i^P$, this inequality holds, and he will not defect. Thus, though P cannot set $\sigma^P(\emptyset, \emptyset, x_l^A) = x_l^P$, she can set $\sigma^P(\emptyset, \emptyset, x_l^A) = x_i^P$ to satisfy Refinement A.1, per the exception above.

If A_m, alone or along with $A_{\tilde{H}}$, but not $A_{\tilde{L}}$, satisfies Inequality (A.1), P can set $\beta_L^P = \tau_l$, which implies $x^P = x_j^P$. Again, with defection switching the payoff of $A_{\tilde{H}}$ from the right-hand side to the left-hand side of Inequality (B.1) and $x^P = x_j^P > x_i^P$, this inequality holds, and $A_{\tilde{H}}$ will not defect. This belief will also prevent A_m from defecting. Otherwise, if $\tau_i(\pi EU_i^A(x_i^P) + (1-\pi)EU_i^A(x_i^A)) + \tau_k(\pi EU_k^A(x_k^P) + (1-\pi)EU_k^A(_k^A)) > \pi EU_j^A(x_j^P) + (1-\pi)EU_j^A(x_l^A)$, satisfaction of Inequality (B.1) implies $\pi EU_k^A(x_k^P) + (1-\pi)EU_k^A(x_k^A) > \pi EU_j^A(x_j^P) + (1-\pi)EU_j^A(x_l^A)$, which would imply that $A_{\tilde{L}}$ also satisfies Inequality (A.1), which contradicts the assumption. Thus, P cannot set $\sigma^P(\emptyset, \emptyset, x_l^A) = x_l^P$, but she can set $\sigma^P(\emptyset, \emptyset, x_l^A) = x_j^P$ to satisfy Refinement A.1, per the exception above.

Finally, if $A_{\tilde{L}}$ satisfies Inequality (A.1), P can set $\beta_L^P = \frac{\alpha\tau_l}{\alpha\tau_l + (1-\alpha)\tau_h}$, for x_k^P. In defecting, $A_{\tilde{L}}$ would see his payoff go from the right-hand side of Inequality (B.2) to a form of the

left-hand side. With $x^P = x_k^P$, this inequality is satisfied, and he would not defect. As in the previous two cases, defection would switch the payoff for $A_{\tilde{H}}$ from the right-hand side of Inequality (B.1) to the left-hand side, and $x^P = x_k^P > x_i^P$ implies satisfaction of this inequality and no defection for him. Finally, since $x^P = x_k^P$, Inequality (B.5) applies and A_m will not defect. Thus, P cannot set $\sigma^P(\emptyset, \emptyset, x_l^A) = x_l^P$, but she can set $\sigma^P(\emptyset, \emptyset, x_l^A) = x_k^P$ to satisfy Refinement A.1, per the exception above.

Any disclosure not on the equilibrium path other than $\mathring{d} = (\emptyset, \emptyset, x_l^A)$ is caused by a defection by A. If the deviation involves $\mathring{d}_s = \tilde{H}$ ($\tilde{L}$), $A_{\tilde{H}}$ ($A_{\tilde{L}}$) should expect P to act as if $\beta_L^P \geq \frac{(1-\alpha)\tau_l}{\alpha\tau_h+(1-\alpha)\tau_l}$ ($\frac{\alpha\tau_l}{\alpha\tau_l+(1-\alpha)\tau_h}$), so that $\mathbb{E}(r) \geq i$ (k), and to select $x^P \geq x_i^P$ (x_k^P). Defecting would cause his payoff to change from the right-hand side of Inequality (B.1) [(B.2)] to a form of the left-hand side. Since $x^P \geq x_i^P$ (x_k^P), this inequality holds, and Inequalities (A.1) and (A.2) do not hold. Then $\Pr(\theta) = 0$ for any θ involving that defection, and P's equilibrium strategy when $\mathring{d}_s = \tilde{H}$ ($\tilde{L}$) is supportable.

If the deviation involves $\mathring{d}_s = \emptyset$, P must set $\Pr(\theta) = 0$ for any θ involving $A_{\tilde{H}}$. Since a defection by R has been ruled out, he would expect P to act as though $\beta_L^P \geq \frac{(1-\alpha)\tau_l}{\alpha\tau_h+(1-\alpha)\tau_l}$, which was just shown to lead to the fact that Inequalities (A.1) and (A.2) do not hold. Thus, P can assign strictly positive probability only to strategy-signal profiles involving $A_{\tilde{L}}$ or A_m. However, if P can believe that $\Pr(\theta) > 0$ for some θ involving A_m, she can believe that $\Pr(\theta) > 0$ for some θ involving $A_{\tilde{L}}$ (which exists since $A_{\tilde{L}}$ can withhold s, propose, and disclose the same thing as A_m). If $\Pr(\theta) > 0$ for A_m, the refinements imply minimally that

$$\tau_i((\pi + \mathbf{1}_m^{tr}\Delta\pi_m)EU_i^A(x_i^P) + (1 - \pi - \mathbf{1}_m^{tr}\Delta\pi_m)EU_i^A(x_i^A)) + \tau_k\big[(\pi + \mathbf{1}_m^{tr})EU_k^A(x_k^P) + (1 - \pi - \mathbf{1}_m^{tr}\Delta\pi_m)EU_k^A(x_k^A)\big] < (\pi + \mathbf{1}_m\Delta\pi_m)EU_j^A(x^P) + (1 - \pi - \mathbf{1}_m\Delta\pi_m)EU_j^A(x^A) \tag{B.6}$$

for some x^A and $x^P \geq x_i^P$. Because Inequality (B.1) holds for $A_{\tilde{H}}$ (whose equiibrium payoff is on the right-hand side), it must be the case that

$$(\pi + \mathbf{1}_m^{tr}\Delta\pi_m)EU_k^A(x_k^P) + (1 - \pi - \mathbf{1}_m^{tr}\Delta\pi_m)EU_k^A(x_k^A) \\ < (\pi + \mathbf{1}_m\Delta\pi_m)EU_k^A(x^P) + (1 - \pi - \mathbf{1}_m\Delta\pi_m)EU_k^A(x^A), \quad \text{(B.7)}$$

which means that $A_{\tilde{L}}$ would also defect if P would respond with the same x^P. Then Inequalities (A.1) and (A.2) hold for $A_{\tilde{L}}$ as well as A_m, in which case P assign all the probability to strategy-signal profiles involving $A_{\tilde{L}}$. Doing so supports her equilibrium strategy x_k^P. Therefore, the given equilibrium always exists and satisfies the refinements. In addition, Theorem 2 implies that P always has her minimum level of power, which is the baseline π.

For the second statement, P's payoff from the separating equilibrium is

$$\tau_i(\pi EU_i^P(x_i^P) + (1-\pi)EU_i^P(x_i^A)) + \tau_k(\pi EU_k^P(x_k^P) + (1-\pi)EU_k^P(x_k^A)). \quad \text{(B.8)}$$

The derivative with respect to a is $(1-\pi)(\tau_i EU_i^P(x_i^A)\frac{\partial x_i^A}{\partial a} + \tau_k EU_k^P(x_k^A)\frac{\partial x_k^A}{\partial a})$. This expression is negative because, for any r, the optimal x^* satisfies $b'(x^*) = (1+a)rc'(x^*)$, and $\frac{\partial x^*}{\partial a} = -\frac{rc'(x^*)}{(1+a)rc''(x) - b''(x)} < 0$. The derivative of P's payoff with respect to π is $\tau_i(EU_i^P(x_i^P) - EU_i^P(x_i^A)) + \tau_k(EU_k^P(x_k^P) - EU_k^P(x_k^A)) > 0$ whenever $a > 0$, in which case $x_i^A \neq x_i^P$ and $x_k^A \neq x_k^P$. ■

Proof of Proposition 5 Preliminarily, it is worth noting that transparency of x^A is irrelevant because the principal's policy choice depends only on her knowledge about m and s, both of which are assumed to be transparent. For any message-signal equilibrium with these items transparent, $\sigma^H = \sigma^L = m$, and $\sigma_n^A(m) = n$. Weak consistency requires

beliefs leading to $\sigma^P(m, \tilde{H}, \cdot) = x_i^P$, $\sigma^P(m, \tilde{L}, \cdot) = x_k^P$, which are sequentially rational; while weak consistency and sequential rationality imply $x^A(m, \tilde{H}) = x_i^A$ and $x^A(m, \tilde{L}) = x_k^A$. Now strategies and beliefs off the equilibrium path are defined to sustain the equilibrium. At stage 4, there cannot be defections in the form of withholding m or s, and for each signal, A will not deviate because he is selecting his optimal policy and cannot induce P to select a different policy choosing x^A differently. At stage 3, preventing defection is best served by setting $\beta_L^P = 1$ (although this will prove not to satisfy either refinement). Finally, at stage 2, the strategies for A and P that will maximize R's costs from not messaging are $x^A(\emptyset, \emptyset) = x_l^A$ and $\sigma^P(\emptyset, \emptyset, \cdot) = x_l^P$. R's defection payoff divided by t is the left-hand side of Inequality 4, while the right-hand side is L's equilibrium payoff under full transparency, divided by L. Thus, there exists a proposed equilibrium in which L chooses not to defect only if Inequality 4 holds. By Lemma B.1 H would not defect unless L would also defect.

Robustness of this equilibrium to the refinements: Apart from the refinements, the equilibrium exists if and only if neither L nor A_m defects. The refinements restrict β_L^P for certain disclosures off the equilibrium path. For $\mathring{d} = (\emptyset, \emptyset, \cdot)$, only R can defect by not messaging. Lemma B.1 implies that, if P can assign $\Pr(\theta) > 0$ to some θ with H defecting, she can also assign $\Pr(\theta) > 0$ to some θ with L defecting. Then $\beta_L^P = 1$ survives both refinements, as do the strategies that this belief entails. Therefore, a necessary and sufficient condition for R not to defect in a natural equilibrium is satisfaction of Inequality (4).

For $\mathring{d} = (m, \emptyset, \cdot)$, the refinements together imply that A_m can expect P to act as though $\beta_L^P \geq \tau_l$. Satisfaction of Inequality (3) implies that, under either refinement, A_m would not defect even with the most favorable belief $\beta_L^P = \tau_l$ and his best policy when he has authority, x_j^A. Then the equilibrium is sustainable with any belief $\beta_L^P \geq \tau_l$, which is allowed by both refinements. If this inequality is reversed, then Inequalities (A.1) and (A.2) are satisfied

with $\beta_L^P = \tau_l$ when $x^A = x_j^A$. Since only A_m can defect in this manner, satisfaction of these inequalities implies that P has $\beta_L^P = \tau_l$ (i.e., assigns probability only to strategy-signal profiles involving A_m), in which case A would defect by setting $\sigma_n^A(m) = \emptyset$ and $x^A = x_j^A$, and the equilibrium does not satisfy either refinement. Thus, Inequality (3) determines whether A would defect by not generating a signal in any proposed natural equilibrium.

Finally, for $\mathring{d} = (m, \tilde{H}, \cdot)$ $[(m, \tilde{L}, \cdot)]$, $A_{\tilde{H}}$ $(A_{\tilde{L}})$ expects $\beta_L^P \geq \frac{(1-\alpha)\tau_l}{\alpha\tau_h+(1-\alpha)\tau_l}$ $(\frac{\alpha\tau_l}{\alpha\tau_l+(1-\alpha)\tau_h})$ under the refinements. Because $x^P \geq x_i^P$ (x_k^P), Inequality (B.1) [(B.2)] holds, with his equilibrium payoff on the right-hand side and defection payoff on the left-hand side. Then Inequalities (A.1) and (A.2) cannot be satisfied. Thus, P can respectively assign any $\beta_L^P \geq \frac{(1-\alpha)\tau_l}{\alpha\tau_h+(1-\alpha)\tau_l}$ $(\frac{\alpha\tau_l}{\alpha\tau_l+(1-\alpha)\tau_h})$ to prevent $A_{\tilde{H}}$ $(A_{\tilde{L}})$ from defecting. Overall, analysis of all the off-equilibrium path disclosures shows no further necessary conditions for the equilibrium to satisfy the refinements. Thus, satisfaction of Inequalities (3) and (4) are sufficient, as well as necessary, for a natural message-signal equilibrium.

For the first statement after Inequality (4), Theorem 2 implies that P and A choose their respective ideal policies after each signal. P's payoff increases with her power as she substitutes her ideal policy for the agent's over the difference in power, and her power is at a maximum when m and s are transparent.

If Inequality (3) fails while Inequality (4) holds, the then A does not generate a signal in any natural equilibrium. First, there cannot be an equilibrium in which L messages with probability less than one while H messages. If A does not defect, then satisfaction of Inequality (4) implies that L would defect by always messaging. The reason is that the policies after each signal are more favorable to L while chosen with the same corresponding probabilities than if L had adopted a pure strategy of messaging. The only remaining way in which A could generate a signal is if L messages with strictly positive probability while H

does not message. However, L would defect by pooling with H for more favorable policy in all situations. If Inequality (4) fails while Inequality (3) holds, then the only other potential equilibrium in which L messages with probability 1 has H not messaging, which is a case of the scenario just mentioned in which L would defect by not messaging. ■

Proof of Proposition 6 The following message-signal equilibrium will be shown to exist and satisfy the refinements if and only if Inequalities (5) and (6) are satisfied:

$\sigma^H = \sigma^L = m$, $\sigma_n^A(m) = n$; $x^A(\emptyset, \emptyset) = x_l^A$, $x^A(m, \tilde{H}) = x_i^A$, $x^A(m, \tilde{L}) = x_k^A$, $\sigma_d^A(m, \cdot, x^A) = (\emptyset, \delta, \delta)$, $\sigma_d^A((\emptyset, \emptyset, x^A)) = (\emptyset, \emptyset, \delta)$; and

$$\sigma^P(\mathring{d}_m, \mathring{d}_s, \mathring{d}_x) = \begin{cases} x_i^P & \text{if } \mathring{d}_s = \tilde{H}, \\ x_k^P & \text{if } \mathring{d}_s = \tilde{L}, \\ x_l^P & \text{if } \mathring{d}_s = \emptyset, \end{cases}$$

except that, under certain conditions listed below, P may set $\sigma^P(\emptyset, \emptyset, x_l^A) = x_j^P$. Note that, since this equilibrium always has $d_x = \delta$, this proof applies whether or not x^A is transparent. A's and P's beliefs, β_L^A and β_L^P, follow from their strategies for policy selection, so showing sequential rationality implies weak consistency. Checking for deviations, P does not defect at stage 5, as she selects her optimal policy on the equilibrium path: x_i^P after $s = \tilde{H}$ and x_k^P after $s = \tilde{L}$. Given P's policy selections and a message, A at stage 4 is optimizing. For each signal A selects his optimal policy, while the principal's policy is already determined by s, and he does not change his disclosure because has the most power possible given transparency of s ($1 - \pi - \Delta\pi$ or $1 - \pi - \Delta\pi_{\tilde{L}}$). At stage 3, preventing defection is best served by setting $\beta_L^P = 1$ (although this will prove not to satisfy either refinement). Finally, at stage 2, a necessary and sufficient condition for L not to defect is embodied in Inequality

(6), and Lemma B.1 implies that H's individual rationality constraint does not bind.

Robustness of this equilibrium to the refinements: Because s is transparent, lack of a signal implies that R defected by not messaging or A defected by not generating a signal. In the former case, the resulting disclosure is $\mathring{d} = (\emptyset, \emptyset, x_L^A)$. Lemma B.1 implies that, under either refinement, if P can set $\Pr(\theta) > 0$ for some θ involving deviation by H, she can also set $\Pr(\theta) > 0$ for some θ involving deviation by L. Thus, P can always assign $\Pr(\theta) = 0$ for every θ involving deviation by H.

The remaining deviation is if A sets $\sigma_n^A(m) = \emptyset$, proposes x_L^A, and discloses it. If A would be willing to defect when $\beta_L^P = 1$, then $\pi EU_j^A(x_l^P) + (1-\pi)EU_j^A(x_l^A) > \tau_i(\pi EU_i^A(x_i^P) + (1-\pi)EU_i^A(x_i^A))\tau_k(\pi + \Delta\pi_{\tilde{L}})EU_k^A(x_k^P) + (1 - \pi - \Delta\pi_{\tilde{L}})EU_k^A(x_k^A))$, in which case Inequality (5) does not hold (since $\pi EU_j^A(x_j^P) + (1-\pi)EU_j^A(x_j^A) > \pi EU_j^A(x_l^P) + (1-\pi)EU_j^A(x_l^A)$, and the equilibrium will not survive either refinement for some other reason. If A is not willing to defect when $\beta_L^P = 1$, then under Refinement A.2, P can assign zero probability to all defections: by A because disclosing x_L^A implies $\beta_L^P = 1$ in Inequality (A.2)), and by H and L, since for the equilibrium to exist, Inequality (6) must be satisfied. Then she can sustain her belief $\beta_l^P = 1$.

Under Refinement A.1, she is unable to sustain $\beta_l^P = 1$ if and only if (1) A satisfies Inequality A.1 while L does not. This is the exception for P's strategy given at the start of the proof. If she cannot sustain $\beta_l^P = 1$, then $\beta_l^P = \tau_l$, leading to $x^P = x_j^P$ since only A with $s = \emptyset$ could defect. Possibly, A would not defect given these beliefs. If he would not, then setting $\sigma^P(\emptyset, \emptyset, x_L^A) = x_j^P$ according to the exception above will prevent any defection. If A would defect, however, and $\pi EU_j^A(x_j^P) + (1-\pi)EU_j^A(x_l^A) > \tau_i(\pi EU_i^A(x_i^P) + (1-\pi)EU_i^A(x_i^A))\tau_k(\pi + \Delta\pi_{\tilde{L}})EU_k^A(x_k^P) + (1 - \pi - \Delta\pi_{\tilde{L}})EU_k^A(x_k^A))$, then Inequality (5) again does not hold (since $\pi EU_j^A(x_j^P) + (1-\pi)EU_j^A(x_j^A) > \pi EU_j^A(x_j^P) + (1-\pi)EU_j^A(x_l^A)$, and the

equilibrium will not survive Refinement A.1 for some other reason.

Overall, either P can sustain a belief on $\mathring{d} = (\emptyset, \emptyset, x_l^A)$ that prevents defection by A and R or the belief she can sustain dissuades only R from defecting. In the latter case, Inequality (5) fails. Then A can defect by setting with $\mathring{d} = (\emptyset, \emptyset, x_j^A)$ $\sigma_n^A(m) = \emptyset$, proposing x_j^A and disclosing it, for $\mathring{d} = (\emptyset, \emptyset, x_j^A)$. Under either refinement, A can expect P to set $\beta_L^P = \tau_L$, and would want to deviate. Then P must set $\beta_L^P = \tau_l$. Since Inequality (5) does not hold, A will deviate in this way given this belief and the equilibrium does not survive either refinement. In the former case, independent failure of Inequality (5) implies again that A will deviate with $\mathring{d} = (\emptyset, \emptyset, x_j^A)$, and again the equilibrium does not survive. If, however, Inequality (5) holds, then A would not defect by setting $\sigma_n^A(m) = \emptyset$ and $x^A = x_j^A$, even with the most favorable belief under either refinement, $\beta_L^P = \tau_l$. Then the equilibrium can be sustained with any belief $\beta_L^P \geq \tau_l$, which is allowed by both refinements. Thus, Inequality (5) determines whether A would make this defection in any proposed natural equilibrium.

Finally, $A_{\tilde{H}}$ ($A_{\tilde{L}}$) would not pursue defections after either signal. The part of the proof of Proposition 5 involving $\mathring{d} = (m, \tilde{H}, \cdot)$ [$(m, \tilde{L}, \cdot)$] can be applied. (In Inequality (B.2) $\mathbf{1}_{\tilde{L}}^{tr} = \mathbf{1}_{\tilde{L}} = 1$.)

Considering beliefs for off-equilibrium path disclosures yielded no additional conditions for this equilibrium. Therefore, Inequalities (5) and (6) are not just necessary, but sufficient for this message-signal equilibrium to exist and satisfy each refinement. If either of these inequalities does not hold, no other natural message-signal equilibrium can be sustained. Theorem 2 limits the possibilities for such an equilibrium to those that yield L the same payoff, the right-hand side of Inequality (7). Meanwhile the left-hand side is the highest cost that L can incur if it defects. Thus, if L would defect from the given proposed equilbrium, it would defect from any other proposed signal-separating equilibrium that could

exist according to Theorem 2. Theorem 2 also limits the possibilities for a natural message-signal equilibrium to those that yield A the same payoff, the right-hand side of Inequality (5). Failure of this inequality is sufficient for A to cause any equilibrium not to survive the refinements. Even if $\mathring{d} = (\emptyset, \emptyset, x_j^A)$ were somehow the disclosure that occurred after a defection by R, A could select and disclose some $x^{A\prime}$ slightly less than x_j^A, but still satisfying Inequality (5), with $x^{A\prime}$ substituting for x_j^A on the left-hand side. The proof starting from the introduction of $\mathring{d} = (\emptyset, \emptyset, x_j^A)$ can be redone, with $x^{A\prime}$ replacing x_j^A, *mutatis mutandis*.

For the first statement after Inequality (6), Theorem 2 implies that P and A choose their respective ideal policies after each signal. P's payoff increases with her power as she substitutes her ideal policy for the agent's over the difference in power. Disclosure of s weakly increases her power after $s = \tilde{L}$. By Theorem 2, she can take advantage of this power increase, but not any from disclosure of m. Thus, after each signal, her power lies in the interval $[\pi, \pi + \Delta\pi_m + \mathbf{1}_{s=\tilde{L}}\Delta\pi_{\tilde{L}}]$, where $\mathbf{1}_{s=\tilde{L}} = 1$ (0) when $s = \tilde{L}$ ($\tilde{H}$). The remaining statements can be shown applying the analogous part of the Proof of Proposition 5 and substituting these inequalities respectively for Inequalities (3) and (4). ■

Proof of Proposition 7 The following message-signal equilibrium will be shown to exist and satisfy the refinements if and only if Inequality (7) is satisfied: $\sigma^H = \sigma^L = m$, $\sigma_n^A(m) = n$; $x^A(\emptyset, \emptyset) = x_l^A$, $x^A(m, \tilde{H}) = x_i^A$, $x^A(m, \tilde{L}) = x_k^A$, $\sigma_d^A(m, \tilde{H}, x^A) = (\delta, \delta, \delta)$, $\sigma_d^A(m, \cdot, x^A) = (\delta, \emptyset, \delta)$ for $s \in \{\emptyset, \tilde{L}\}$, $\sigma_d^A(\emptyset, \emptyset, x^A) = (\emptyset, \emptyset, \delta)$; and

$$\sigma^P(\mathring{d}_m, \mathring{d}_s, \mathring{d}_x) = \begin{cases} x_i^P & \text{if } \mathring{d}_s = \tilde{H}, \\ x_k^P & \text{if } \mathring{d}_s \neq \tilde{H} \text{ and } \mathring{d}_m = m, \\ x_l^P & \text{if } \mathring{d}_m = \emptyset. \end{cases}$$

Note that, since this equilibrium always has $d_x = \delta$, this proof applies whether or not x^A is transparent. A's and P's beliefs, β_L^A and β_L^P, follow from their strategies for policy selection, so showing sequential rationality implies weak consistency. Checking for deviations, P does not defect at stage 5, as she selects her optimal policy on the equilibrium path: x_i^P after $\tilde{H}$ and x_k^P after $\mathring{d} = (m, \emptyset, \cdot)$ since A has produced $\tilde{L}$. Given P's policy selections and a message, A at stage 4 is optimizing. For each signal A selects his optimal policy and receives the most favorable possible policy from the principal (x_k^P for $A_{\tilde{L}}$, who cannot produce $\tilde{H}$), and he not does change his disclosure because has the most power possible given transparency of m (i.e., $1-\pi-\Delta\pi_m$). At stage 3 A prefers generating a signal to receiving his best payoff from not doing so: $\tau_i((\pi+\Delta\pi_m)EU_i^A(x_i^P) + (1-\pi-\Delta\pi_m)EU_i^A(x_i^A)) + \tau_k((\pi+\Delta\pi_m)EU_k^A(x_k^P) + (1-\pi-\Delta\pi_m)EU_k^A(x_k^A)) > (\pi+\Delta\pi_m)EU_j^A(x_k^P) + (1-\pi-\Delta\pi_m)EU_j^A(x_j^A)$ for the same reasons that Inequality (B.5) holds. Finally, at stage 2, a necessary and sufficient condition for L not to defect is embodied in Inequality (7), and Lemma B.1 implies that H's individual rationality constraint does not bind.

Robustness of this equilibrium to the refinements: Because m is transparent, it is clear whether A or R has defected. If $\mathring{d}_m = \emptyset$ off the equilibrium path, R must have deviated by not messaging. For the same reasons as in the proof of Proposition 5 when $\mathring{d} = (\emptyset, \emptyset, \cdot)$, satisfaction of Inequality (7) is necessary and sufficient for R not to defect. If $\mathring{d}_m = m$ off the equilibrium path, A has defected, and the argument used to support P's beliefs and strategies after disclosures off the equilibrium path other than $\mathring{d} = (\emptyset, \emptyset, x_l^A)$ in the proof of Proposition 4 can be used to support P's off-equilibrium path beliefs and strategies for the equilibrium proposed here. (For the relevant inequalities, $\mathbf{1}_m^{tr} = 1$, so $\mathbf{1}_m = 1$.) Subjecting P's off-equilibrium beliefs to the refinements yields no additional conditions for the proposed equilibrium, so Inequality (7) is a sufficient condition for it.

If Inequality (7) does not hold, no other natural message-signal equilibrium can be sustained. Theorem 2 limits the possibilities for such an equilibrium to those that yield L the same payoff, the right-hand side of Inequality (7). Meanwhile the left-hand side is the highest cost that L can incur if it defects. Thus, if L would defect from the given proposed equilbrium, it would defect from any other proposed signal-separating equilibrium that could exist according to Theorem 2.

For the first statement after Inequality (7), Theorem 2 implies that P and A choose their respective ideal policies after each signal. P's payoff increases with her power as she substitutes her ideal policy for the agent's over the difference in power. Disclosure of m weakly increases her power. By Theorem 2, she can take advantage of this power increase, but not any from disclosure of $\tilde{L}$. Thus, after each signal, her power lies in the interval $[\pi, \pi + \Delta\pi_m + \mathbf{1}_{s=\tilde{L}}\Delta\pi_{\tilde{L}}]$, where $\mathbf{1}_{s=\tilde{L}} = 1$ (0) when $s = \tilde{L}$ ($\tilde{H}$). For the last statement of the proposition, the only other potential equilibrium in which L messages with probability 1 is one in which H does not message. However, sequential rationality and weak consistency imply H would receive x_h^A and x_h^P from A and P respectively with P having minimal power. L would receive costlier policies, with more power for the principal and would deviate. ■

Proof of Proposition 8 When $\Delta\pi_m = \Delta\pi_{\tilde{L}} = 0$, the low-cost target's individual rationality constraint becomes

$$\pi c(x_l^P) + (1-\pi)c(x_l^A) > \pi((1-\alpha)c(x_i^P) + \alpha c(x_k^P)) + (1-\pi)((1-\alpha)c(x_i^A) + \alpha c(x_k^A)), \quad \text{(B.9)}$$

which holds since $x_l^P > x_k^P > x_i^P$ and $x_l^A > x_k^A > x_i^A$. Then Inequalities (4), (6), and (7) are all satisfied, and L would not defect in any transparency mode. By Lemma B.1, H would not defect, either. Thus, only Inequality (3) or (5), which are the same when $\Delta\pi_m = \Delta\pi_{\tilde{L}} = 0$,

can fail. Then Propositions 4 and 7 imply that there always exist natural message-signal equilibria, and Corollary 3 implies that there exists one in which A discloses everything. Propositions 5 and 6, imply Inequality (3) determines whether a natural message-signal equilibrium exists, and that A will not generate a signal in any natural equilibrium if this inequality does not hold. Because $\Delta\pi_m = \Delta\pi_{\tilde{L}} = 0$, there cannot be any equilibria in which L adopts a different strategy from H, since the policies after H's action would dominate those after the other action. Thus, H and L are pooling, and with no signal generated, $\beta_L^P = \beta_L^A = \tau_l$, the prior belief. Sequential rationality implies an equilibrium in which P receives $\pi EU_j^P(x_j^P) + (1-\pi)EU_j^P(x_j^A) \leq EU_j^P(x_j^P)$. ■

References

Banks, Jeffrey S., and Joel Sobel. 1987. "Equilibrium Selection in Signaling Games." *Econometrica* 55 (May): 647–61.

Carpenter, Daniel. 2013. "Detecting and Measuring Capture." In *Preventing Regulatory Capture: Special Interest Influence, and How to Limit It*, ed. Daniel Carpenter and David Moss. New York: Cambridge University Press. Forthcoming.

Cho, In-Koo, and David M Kreps. 1987. "Signaling Games and Stable Equilibria." *The Quarterly Journal of Economics* 102 (May): 179–221.

Coglianese, Cary. 2009. "The Transparency President? The Obama Administration and Open Government." *Governance* 22 (October): 529–44.

Coglianese, Cary, Heather Kilmartin, and Evan Mendelson. 2009. "Transparency and Public Participation in the Federal Rulemaking Process: Recommendations for the New Administration." *George Washington Law Review* 77 (June): 924–72.

Corwin, Erik H. 1992. "Congressional Limits on Agency Discretion: A Case Study of the Hazardous and Solid Waste Amendments of 1984." *Harvard Journal on Legislation* 29 (Summer): 517–60.

Fenster, Mark. 2006. "The Opacity of Transparency." *Iowa Law Review* 91 (March): 885–949.

Hicks, Josh. 2013. "Federal Openness Gets Mixed Reviews." *Washington Post* (March 13): A12.

Kerwin, Cornelius M., and Scott R. Furlong. 2011. *Rulemaking: How Government Agencies Write Law and Make Policy*. 4th ed. Washington: CQ Press.

Levine, Michael E., and Jennifer L. Forrence. 1990. "Regulatory Capture, Public Interest, and the Public Agenda: Toward a Synthesis." *Journal of Law, Economics, and Organization* 6 (April): 167–98.

Madison, James. 1999. "To William T. Barry." In *James Madison: Writings*, ed. Jack N. Rakove. New York: Library of America.

McCarty, Nolan, and Adam Meirowitz. 2007. *Political Game Theory: An Introduction*. New York: Cambridge University Press.

McCubbins, Mathew D., and Thomas Schwartz. 1984. "Congressional Oversight Overlooked: Police Patrols versus Fire Alarms." *American Journal of Political Science* 28 (February): 165–79.

McCubbins, Matthew D., Roger G. Noll, and Barry R. Weingast. 1987. "Administrative Procedures as Instruments of Political Control." *Journal of Law, Economics, and Organization* 3 (Autumn): 243–77.

McNollgast. 1999. "The Political Origins of the Administrative Procedure Act." *Journal of Law, Economics, and Organization* 15 (March): 180–217.

Milgrom, Paul R. 1981. "Good News and Bad News: Representation Theorems and Applications." *Bell Journal of Economics* 12 (Autumn): 380–91.

Moffitt, Susan L. 2010. "Promoting Agency Reputation Through Public Advice: Advisory Committee Use in the FDA." *Journal of Politics* 72 (July): 880–93.

Obama, Barack. 2009. Memorandum for the Heads of Executive Departments and Agencies: Freedom of Information Act. January 21. 74 Fed. Reg. 4683.

Okuno-Fujiwara, Masahiro, Andrew Postlewaite, and Kotaro Suzumura. 1990. "Strategic Information Revelation." *Review of Economic Studies* 57 (January): 25–47.

Potters, Jan, and Frans Van Winden. 1992. "Lobbying and Asymmetric Information." *Public Choice* 74 (3): 269–92.

Prat, Andrea. 2005. "The Wrong Kind of Transparency." *American Economic Review* 95 (June): 862–77.

Sloof, Randolph. 1998. *Game-theoretic Models of the Political Influence of Interest Groups*. Boston: Kluwer Academic Publishers.

Stiglitz, Joseph. 2002. "Transparency in Government." In *The Right to Tell: The Role of Mass Media in Economic Development*, ed. Roumeen Islam. Washington: World Bank.

Ting, Michael M. 2011. "Organizational Capacity." *Journal of Law, Economics, and Organization* 27 (August): 245–71.

Wagner, Wendy E. 2010. "Administrative Law, Filter Failure, and Information Capture." *Duke Law Journal* 59 (April): 1321–1432.

West, William F. 2004. "Formal Procedures, Informal Processes, Accountability, and Responsiveness in Bureaucratic Policy Making: An Institutional Policy Analysis." *Public Administration Review* 64 (January-February): 66–80.

Wichmann, Charles J. 1998. "Ridding FOIA of Those 'Unanticipated Consequences': Repaving a Necessary Road to Freedom." *Duke Law Journal* 47 (April): 1213–56.