HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Public Policy

have examined a dissertation entitled

"Hierarchical Game-theoretic Models of Transparency in the Administrative State"

presented by Laurence Tai

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Professor Daniel Carpenter

Signature _____

Typed name: Professor Kenneth Shepsle

Signature _____

Typed name: Professor Matthew Stephenson

Signature _____

Typed name: Professor Richard Zeckhauser

Date: April 22, 2013

Hierarchical Game-theoretic Models of Transparency in the Administrative State

A dissertation presented
by
Laurence Tai
to
The Department of Public Policy

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Public Policy

Harvard University
Cambridge, Massachusetts

April 2013

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Dissertation Advisor: Professor Daniel Carpenter Laurence Tai

Hierarchical Game-theoretic Models of Transparency in the Administrative State

Abstract

This dissertation develops three game-theoretic models in each of its three chapters to explore the strategic implications of transparency in the administrative state. Each model contains a similar set of three players: a political principal, an agent representing an agency or a bureaucrat, and an interested third party. The models consider the utility of transparency as a tool for mitigating regulatory capture, in which the third party influences the agent to serve its interest rather than the principal's.

Chapter 1, "Transparency and Media Scrutiny in the Regulatory Process," models transparency as the volume of records that the media receives from the agent, which raises the likelihood of news alleging low costs to the interest group after the agent's proposal of lax regulation. Such reports cost these two players and may deter the group from capturing the agent. Among other things, the model describes costs due to distorted policy proposals and loss of information when greater transparency causes inaccurate reports to increase along with accurate ones.

In Chapter 2, "Transparency and Power in Rulemaking," transparency is a requirement for the agent to disclose an item of information, such as his message from the regulated party or his signal about the cost of regulation. The agent can always disclose this information, but doing so may increase the principal's power to set regulation higher than he or the regulated party desires. A key result is that transparency is not necessary for the principal to know as much as the agent does but may discourage the generation of the message or signal.

Chapter 3, "A Reverse Rationale for Reliance on Regulators," suggests that an agent can benefit a principal not by gathering information from an outsider that she cannot access, but by preventing her from obtaining or acting on this information. The agent benefits the principal when he induces additional effort in the outside party's information generation because he is more adversarial toward that party than she is. Mandatory disclosure of the agent's information is harmful because it effectively allows the outsider to communicate directly with the principal and provide lower quality information.

Table of Contents

Acknowledgments

I would like to thank the many people who made this dissertation possible:

First, beyond reviewing various drafts of the dissertation chapters, my committee members have provided invaluable guidance on the general direction for my research. Dan Carpenter, the chair, has supported my interest in government bureaucracies and has helped extend the scope of inquiry in my research beyond transparency to larger questions about the administrative state. Ken Shepsle gave the initial encouragement and suggestions for exploring the field of political economy when I was a newcomer. Matthew Stephenson was my instructor for multiple classes at Harvard Law School, and I have benefited from his insight in formal modeling and especially from his substantive expertise on the major questions of institutional design. Richard Zeckhauser has been a great mentor since I took his course on analytic frameworks in policy. I appreciate his very detailed and diverse comments and his challenges for me to broaden the applicability of my research.

Several other professors have been very helpful throughout my graduate education. I learned a great deal of what I know about game theory from John Patty. He also closely supervised my early work and was my examiner in formal political theory, and he continues to provide research and career advice. Oliver Hart taught and served as my examiner for contracts and organizations. He has reviewed much of my work and kept it connected to the

relevant economics literature. In addition to providing feedback, Steven Shavell has helped situate my research in the field of law and economics both personally and institutionally. Specifically, I have been a Terence M. Considine Fellow in Law and Economics at the Law School for the last four years and have received support for all three chapters from the Considine Family Foundation and the School's John M. Olin Center for Law, Economics, and Business, of which Steven is the director.

Besides the individuals listed above, the following people have provided suggestions for one or more chapters: Jim Alt (Chapter 3), Louis Kaplow (Chapter 1), Yuki Takagi (Chapters 2 and 3), Craig Volden (Chapter 2), and Galina Zudenkova (Chapter 3), as well as participants in the Harvard seminars in Contracts and Organizational Economics, (Chapter 3), Law, Economics, and Organizations (Chapter 1), and Political Economy (all chapters), and participants at the APSA Annual Meeting in 2010 (Chapter 2) and at the MPSA Annual Conferences in 2010 (Chapter 2), 2012 (Chapter 1), and 2013 (Chapter 3). Louisa Van Baalen and Nicole Tateosian provided excellent administrative support. I also acknowledge financial support from the Harvard Kennedy School and the Harvard GSAS.

Finally, I could not have completed this dissertation without family and friends. My parents and my younger brother have always offered their love and support. Also, my housemates and other members of Antioch Baptist Church provided much-needed prayer and encouragement. It was a privilege to see many of them at my defense. I thank Pastor Paul and Becky Jdsn for founding Antioch, where I could learn the truth of the Word of God and the meaning of the family of God. I am grateful to Thomas Chen, my spiritual leader throughout grad school. Most of all, I thank my personal Lord and Savior, Jesus Christ, who has been faithful, even when I have not (2 Timothy 2:13). I want to go wherever He leads and for my work, inside and outside of research, to be used for His Kingdom.

Executive Summary

This dissertation explores the benefits and costs of transparency in the form of mandatory information disclosure by the government. It derives much of its motivation from two themes: First, information transparency has been an important element of the Obama's Administration approach to governing, which includes encouraging agencies to proactively release documents and other information in their possession. The recent salience of this issue implies that policy changes in this area are plausible in practice. Second, academics and practitioners have had a long-standing concern about regulatory capture, a phenomenon in which agencies in the administrative state serve the parties they are charged with regulating rather than the general public interest. The potential for this kind of influence suggests that there exists significant scope for transparency to improve outcomes from policymaking in the administrative state.

The key method in analyzing the effects of transparency is game theory, which, applied in political science, is known as formal or positive political theory. This method is useful for determining how different actors might respond to changes in their institutional environment. Rules for what kinds of information agencies must release constitute an institutional design feature. Also, the assumption that these players strategically strive to maximize their payoff is appropriate because it provides consistency in discussing their behavior. In the case of

transparency, the same agencies that currently resist granting certain kinds of information are likely to react to additional disclosure obligations in ways more complex than simply giving the information demanded of them. Positive political theory highlights not only the likelihood of complications in implementing institutional change, but the nature of these complications in a broader policymaking context. Even when the results of a particular model depend on its assumptions, this method nonetheless elucidates the link between the assumptions and the conclusions and thereby provides more solid grounding for hypotheses about the effects of an institutional change.

This dissertation consists of three self-contained chapters, each of which presents a different model involving transparency in the administrative state. These models are hierarchical, in that they contain three active players: a principal, and agent, and an interested third party. While many games dealing with administrative agencies and their information assume that the agency gathers information on its own, these models reflect a common setting in which a regulated party has a key item of information that the agency needs in deciding what policy to promulgate or propose to the principal. Having two players that will strategically respond to increases in transparency increases the scope for difficulties in the policymaking process that may leave a principal worse off. Together, the three chapters yield a substantial set of cases in which transparency increases should either be avoided or be more narrowly tailored to particular situations.

Chapter 1, "Transparency and Media Scrutiny in the Regulatory Process," focuses on the media as a channel through which transparency might plausibly yield better policies. Here, greater transparency means that agencies must release more documents. A greater volume of documents, in turn, provides media outlets with more material upon which they can base a report that an agency's decision reflects regulatory capture. The principal has the final say

on regulation but needs the agent to learn about the cost of regulation to an affected interest group. She prefers a level of regulation corresponding to an interest group's costs: regulation should be stringent when its costs are low but lax when its costs are high. Thus, regulatory capture occurs when the interest group induces an agent to aim for lenient regulation even though the agent's information suggests that the group's costs are low. An upright agent cannot be captured, but a venal agent is susceptible to the group's influence. The principal does not know what type of agent is gathering information from the group, but she would like to prevent the venal agent from improperly proposing a low level of regulation when the costs are low.

A media report is unpleasant to both the agent and to the interest group and lowers these players' payoffs. Since a report occurs only when the agent proposes a lenient regulation, he can avoid it by proposing stricter regulation. Media reports are assumed to be more likely when transparency increases because the additional government documents released make it easier to write a story detailing how the agency has been captured. As a result, increasing transparency has the potential to deter the group with low costs from influencing the venal agent through a greater incidence and cost of adverse reports. When there are only accurate media reports, which correctly indicate that the agent had information indicating low costs for the group, mandating the release of more documents can only be beneficial.

However, the model considers the potential for media misreporting, which means that the agent's information actually implied high group costs. If incorrect reports rise along with correct ones when transparency increases, then policy outcomes can be worse in two ways. First, due to fear of a report, the upright agent might not be willing to propose low levels of regulation even when his information points to high costs. Second, the low-cost group may become more willing to engage in capture. These two effects, as well as possibly the

direct effect of a less informative media signal, mean that the principal will have less precise information upon which to make her final decision. Though it may not be surprising that inaccurate reports make transparency less useful, the model nonetheless highlights an issue that does not appear to have received much attention and provides more details on the ways in which transparency can yield worse policy outcomes.

Chapter 2, "Transparency and Power in Rulemaking," considers a similar scenario in which the principal's optimal level of regulation depends on whether the costs for the target of regulation are high or low. As in Chapter 1, she needs an agent to learn about the target's costs. Unlike in Chapter 1, where the agent is one of two types, there is a single agent who is captured in the sense that he prefers a lower policy than the principal for any cost level. However, he, like the principal, wishes to match the level of regulation to the group's cost. Another change in this model is that the principal and agent each have some likelihood of making the final policy selection. This likelihood represents power in the model.

Information and transparency also work differently. First, the target of regulation decides whether to communicate with the agent. If and only if it does, the agent can generate a signal about these costs and chooses whether to do so. Transparency in this model entails requiring the agent to disclose any message from the target or signal. The agent can always release each of these items of information that he has, but doing so may increase the principal's power to select the final policy. This feature of the model can represent the idea that interest groups aligned with the principal have an easier time overturning an agency's decision if they have the agency's information in their possession and can respond to it.

Without transparency, the principal can know as much about the target's cost as the agent does. This result relies on the agent's ability to credibly disclose the content of the signal, i.e., whether it points to high or low costs. Since an agent with the high-cost signal

can costlessly display it, the principal can assume that the costs are low if the agent does not disclose any signal. Since there is no transparency requirement, the agent cannot credibly convey that he lacks a signal, and the principal cannot distinguish between a low signal and no signal. Meanwhile, transparency of an item forces its disclosure and can increase the principal's power, which can benefit her. However, it also allows the agent to show that he lacks that item. In addition, the target and the agent may prefer not to produce information in the first place so that the principal does not increase her power to select policies less favorable to them. This result is typically, although not always, harmful for the principal. As a whole, the model suggests that, when an agent can disclose information credibly, the benefits of transparency come not from increased knowledge about policy, but from increased power, which in turn might discourage information generation.

Chapter 3, "A Reverse Rationale for Reliance on Regulators," does not have transparency as its central topic, but it still has implications for what kind of information agencies should disclose. The model in this chapter features a decision between two policies. Unlike in the other two models, the agent does not directly learn anything about which policy would be appropriate; instead, an outside party, or researcher, generates all the information in the game. The first item is a signal as to which policy is better. The signal becomes more accurate with more effort, which constitutes the second item. Effort, however, is costly to the outsider, so one of this party's considerations is minimizing the effort it must exert. All three players have a policy that they presumptively prefer in the absence of a signal but are willing to select the other policy if there is a signal pointing to it supported by enough effort. Once generated, both items can be credibly relayed from one player to another.

If the principal directly faces the researcher, it has no reason to withhold either item of information from her. However, it will typically exert only enough effort for the principal

to be willing to follow either signal, so her benefits from the outsider's research are limited. By delegating the decision or limiting the researcher's communications to the right kind of agent, however, the principal can induce additional effort. The right kind of agent is one who is more opposed to the outsider's presumptive preference than she is, although not so opposed that he discourages research altogether. Such an agent can credibly threaten always to enact or induce the policy that is the opposite of its presumptive preference apart from enough effort supporting a contrary signal because his preferences are strongly opposed to the outsider's. This result implies a reverse rationale for having an agent: instead of gathering information that the principal cannot access, he prevents her from obtaining information that she is perfectly capable of understanding.

Because this institutional arrangement is beneficial, the outsider might attempt to capture the agent to make him less adversarial. With enough influence, it might be able to induce policies that are sometimes different from what the principal would prefer given the information it has generated. When the principal has decision-making authority but the researcher can communicate only with the agent, requiring the agent to disclose whatever information it has would prevent this effect of capture. However, it would also effectively allow the researcher to directly convey information to the principal, in which case it can once again produce a signal with effort that barely satisfies her. As a result, transparency allows the researcher to achieve the benefits of capture without actually engaging in such influence. Meanwhile, if the principal has successfully delegated decision-making authority to the agent, then transparency cannot make any difference, as she cannot act on any information that comes to light. To prevent or mitigate regulatory capture, it is better to keep the agent highly adversarial than to require him to disclose his information.

Even in the context of agency policymaking and potential regulatory capture, the three

models in this dissertation do not exhaust the range of settings for assessing the benefits and costs of mandatory disclosure. However, they collectively suggest that the effects of transparency on policy outcomes can be expected to be mixed. More importantly, they show that, in analyzing the results of transparency, it is necessary to consider not only the immediate effect of additional disclosure of existing information, but also the less direct effects of transparency on other facets of regulatory policymaking, such as agencies' policy proposals, outside party's willingness to generate high-quality information, and both kinds of stakeholders' willingness to create information in the first pace. Although it is possible to conjecture these kinds of consequences, the use of game theory provides an internally consistent logic for understanding the conditions under which one can expect various results from greater transparency.

Chapter 1

Transparency and Media Scrutiny in the Regulatory Process

Contents

1.1 Introduction

Regulatory agencies are supposed to act according to some notion of the public interest as they administer the nation's laws; however, scholars generally believe that agents are susceptible to capture by special interests (Levine and Forrence 1990). Specifically, the information that agencies gain from parties that they regulate, but which they can withhold from the public, is hypothesized to enable this kind of influence (Dal Bó 2006). Given these two premises, it seems to follow logically that greater transparency, in the form of making more of regulators' information publicly accessible, may mitigate the potential for favoritism (Coglianese, Kilmartin, and Mendelson 2009). Transparency in this and other forms has become a major theme in discussions about good governance and accountability in the regulatory state (Hood 2006, Lodge 2004). Political leaders have created initiatives for executive branch transparency in both the United States and the European Union (Coglianese 2009, Cini 2008). The rhetoric in leaders' announcements promotes the concept with largely unqualified praise (see Obama 2009, Kallas 2005). Some scholarship appears similarly to support greater transparency with limited exceptions (Rose-Ackerman 1999, 164–65; Stiglitz 2002).

A growing body of work has challenged the intuitive appeal of increasing the degree to which government information is made public. Possible problems with transparency are that too much of it deprives regulators of the space they need for private discussions (Heald 2006,

68–69; Coglianese 2009, 536); that greater information disclosure, such as under the Freedom of Information Act (FOIA), carries significant administrative costs (Wichmann 1998); that regulators will resist compliance requirements (Roberts 2006); and that they may release information, but in a way citizens are unable to understand (Weil et al. 2006, O'Neill 2006). For the most part, these challenges seem to suggest a need to carefully design transparency laws so that meaningful information disclosure actually occurs in a way that is cost-effective, while not being overinclusive.

Greater difficulties arise from the potential of increased information disclosure to induce undesirable policy distortions. For example, parties that would have provided information to an agency under secrecy may not if they expect that the agency must release it to the public (Coglianese, Zeckhauser, and Parson 2004). In a seminal work, transparency can induce undesirable conformance in behavior among agents (Prat 2005). To be effective, greater transparency should induce different policy outcomes, but its logic may undercut if those different outcomes are worse.

Whether supportive or skeptical about transparency in regulatory policymaking, less clear from these studies as a whole is how making more information accessible to the public will result in changed policies. One method is electoral accountability (Stiglitz 2002, Besley 2006). However, many transparency measures are directed at administrative agencies, whose officials cannot be voted out of office. An alternative explanation is that the negative attention that might arise from information revealed through transparency measures factors into agents' and agencies' utility, in which case they might be incentivized to pursue policies in the public interest. The idea that bureaucrats might fear reports based on the actions they take can be found in Leaver (2009). This explanation is consistent with the idea that agencies have a reputation that they might seek to preserve (Carpenter 2010), and consonant with the idea

that an agent's utility may depend on others' perception of his or her confidence (Prat 2005).

One institution upon which citizens might rely to report on the government is the media (Rose-Ackerman 1999, 165–67; Stiglitz 2002). However, there is an additional complication associated with relying on the media to bring about the benefits of more information disclosure: the media may not always report on the agency's policymaking correctly or objectively. The media tend to report on most agencies infrequently but tend to portray them in a negative light when they do so, causing many bureaucrats to have a fearful attitude toward the media (Lee 1999). There is anecdotal evidence that bureaucrats strive to avoid adverse publicity (Nownes 2006, 72) and, more generally, bureaucrats are thought to have a mentality of blame-avoidance (Hood 2007). In practice, agencies need to devote significant attention and resources to public relations (Graber 2003). This function is a specialized one performed by public information officers who speak for them (Morgan 1986). In general, media investigative reports sometimes make factual errors (Greenwald and Bernt 2000). Thus, even agents who are not susceptible to capture and have nothing to hide may have reason to be concerned about negative reporting on their policy decisions. The media may report on more than just actual capture.

A recent example that raises the question of whether media report accurately when they imply that an agency is not acting in the public interest is Bloomberg News' struggle with the Federal Reserve to obtain documents relating to emergency loans that the central bank made in 2008 during the financial crisis. The news company asserted its right to the records with the Freedom of Information Act (FOIA), but the central bank denied Bloomberg's request (Appelbaum 2011). After a protracted lawsuit ending in a denial of certiorari by the Supreme Court, the agency released the relevant documents in March 2011 (*id.*). Bloomberg proceeded to publish various articles based on the information, including

one highlighting, among other things, how large banks benefited from the Federal Reserve's lending (Ivry, Keoun, and Kuntz 2011). The Federal Reserve produced a memo to respond to the points that the Bloomberg reports raised (Bernanke 2011), and Bloomberg responded in turn ("Bloomberg News Responds" 2011). This episode allows two alternative conclusions of the effects of transparency: either FOIA enabled Bloomberg to hold the Federal Reserve to account or the law merely enabled the news outlet to appear more convincing in its misreporting. Records are perhaps inherently susceptible to alternative interpretations, and one does not have to claim that "government information has no meaning" (Fenster 2006, 924) to acknowledge that investigators must process the records since they are too voluminous to speak for themselves before making their claims.

The possibility of media misreporting suggests that increasing transparency can carry two additional costs that appear not to have been accounted for in the literature on transparency. First, compared to a world with only accurate media reports, a world with media errors implies that political leaders will gain less information about the optimal policy when a report occurs. Second, if media reports occur frequently enough or are costly enough to agents, they may propose different policies in a way that provides less information to political principals about the optimal policy. The second is similar to the distortion in actions that can occur when actions are observable and a principal is trying to measure the agent's competence (Prat 2005), but here the principal is seeking information simply to decide what policy would be optimal. This paper develops a model capturing these two intuitions in a regulatory setting to suggest how media reports, one common avenue through transparency might operate, can ambiguously impact social welfare.

1.2 The Model

1.2.1 Players and Policies

There are three strategic players in the game: a political principal (P, she), an agent (A, he), and an interest group (G, it). The policy in question is a regulation $r \in \mathbb{R}_+$. Mechanically, the political principal can be understood as a unitary leader who perfectly represents her electorate's preferences, or at least the part of the public that is interested in the issue at hand. The relevant public has well-defined preferences over the regulation, and, depending on the circumstances, may prefer a higher or lower level of it. The group, however, always prefers as low a value of regulation as possible. For a concrete example, one can consider a regulation as to how much to reduce emissions of some pollutant (so greater r corresponds to less pollution), where the group is some industry that will have to bear costs under the regulation and thus would like the smallest r that it can secure from the policymaking process.

This model need not be restricted to regulations per se or to corporate interests. Instead, one could imagine a permitting decision related to land use in a municipality, with a preservationist group wanting a parcel of land to stay as close to its original condition as possible. Then the level of policy would reflect the degree to which the land is allowed to be developed, and the preservationist group would incur some net cost taking into account how much they value the natural features of the land compared to the lost opportunity of future jobs and property tax collections resulting from development. However, because business interests, rather than other kinds of interests, are generally thought to be in a position to capture regulators (Ayres and Braithwaite 1991, Laffont and Tirole 1991), the leading example will be about an industry group aiming for as little regulation as possible.

Whatever type of group is envisioned, it may prefer lower regulation with high or low intensity, and the intensity is relevant to the principal's preferred policy. To continue with the example, the regulation may be very costly or only somewhat costly for any level of r. The level of cost would be relevant to the principal because very strict regulation when its cost would be very high would lead to higher prices and/or lost jobs in the industry that the principal and the public do not want. Thus, the group has two types, high (H) and low (L), reflecting the intensity with which it wants the lowest level of regulation possible. At the beginning of the game, only the group knows what type it is. The probability of each group type i is p_i, and this distribution is common knowledge.

The preferences for the principal and for the group are motivated by the following policy payoff functions: First, the group incurs a cost $\gamma^i c(r)$, where $c(\cdot)$ is a twice continuously differentiable function with $c(0) = c'(0) = 0$, $c''(0) > 0$, and where γ^i is a scalar parameter, with $\gamma^H > \gamma^L > 0$. In contrast, the principal and the public balance the benefits and costs of the regulation according to $b_P(r) - \gamma_P^i c_P(r)$, where $\gamma_P^H > \gamma_P^L > 0$, $c_P(\cdot)$ has the same properties as $c(\cdot)$, and $b_P(\cdot)$ is a twice continuously differential function with $b_P'(0) > b_P(0) = 0$, $b_P''(0) < 0$, and $\lim_{r\to\infty} b_P'(r) = 0$.[1] Since the principal does not know the group's type, its preferred policy depends on its beliefs about how costly it is to the group. It is worth noting that with different cost functions, it is immaterial whether the public actually incurs the group's costs, or whether the higher costs for the group correspond perfectly to higher costs for the public.

The agent, who works in the executive branch for the public involved in this game (federal state, or local), is the initial policy proposer, and, like the group, he can be one of two types.

[1]The group may also have modest benefits, but these can be assumed to be negligible and subsumed in the costs.

First, there is an upright agent (U), who, given the same information as the public or political principal, prefers the same level of regulation. Thus, his policy preferences are motivated by the function $\alpha(b_P(r) - \gamma_P^i c_P(r))$, where $\alpha > 0$ is a scalar parameter. Second, there is a venal agent (V), subject to capture, who does not care about policy but seeks to extract rents from the group in exchange for policy more favorable to the group. The mechanism and payoffs related to the venal agent's rent-seeking activities will be described below. The probabilities of the agent types j, p_j, are common knowledge. Besides the agent himself, the group knows the agent's type. Mechanically, this scenario can be thought of as occurring when a representative of the group tests, perhaps through conversation, whether the agent is susceptible to influence before he proposes a policy.[2] In contrast, the public and the principal can never observe the group's type directly.

In describing the model in terms of the agent, susceptibility to capture is treated as an individual characteristic. Even if an individual bureaucrat does not directly have concerns about negative media reports, this concern may be induced by leaders within the agency who can discipline him or her or otherwise respond negatively after adverse publicity. Such a formulation suggests that some agents are fully honest, while others are subject to influence. However, nothing is lost in the model if the player is a whole agency that can or cannot be captured, with some probability of being in each state.

There is a fourth, non-strategic player in this game: the media. There are studies of how media outlets consciously choose what to cover (e.g. Hamilton 2004), but the stylized fact that they cover most agencies only once in a while for their real or apparent failings

[2]Discovering the agent's type at this point will turn out to be as effective for the group as knowing it from the start. Having the group know the agent's type prevents the high-cost group from distinguishing itself in front of the upright agent by offering a different kind of benefit to the venal agent from the low-cost group.

seems consistent enough that the media can be modeled as a producer of reports, depending on the agency's policy and the information it has access to. This kind of reporting differs from "squawking" by regulated industry in Leaver (2009). The reporting in this model is more neutral because it does not directly serve the purpose of firms. The inferences that the principal can draw from the reports are also more straightforward because the probability of a report is determined automatically, rather than based on some calculation by a media outlet. Transparency has the potential to affect the kinds of reports that emanate from the media. In addition to regular media outlets, the media may be thought of as any watchdog group that produces reports that sometimes successfully direct negative attention to an agency.

The two players involved in making policy, however, are the agent and the principal. First, the agent proposes a level of regulation r_A. Then the principal selects the final policy, r_P. In this costless decision-making structure, the principal is given the maximum amount of formal authority possible. This level of power is perhaps greater than political principals have in some actual policymaking settings, given the relatively small number of regulations that principals end up overturning, but it is illustrative of the case in which she can act fully upon the information she receives.

1.2.2 Types of Information

Transparency in the context of regulatory capture is about making information publicly available during the policymaking process so that other actors can potentially influence the final outcome. The two important types of information in the game are the information about the agent's policy decision and the information that the agent gains about the group's type.

The agent's policy proposal, r_A, is verifiably observed by all players. Not only does this simplify the question about the benefits and costs of transparency, but it also reflects empirically how U.S. federal agencies actually operate. The Administrative Procedures Act requires agencies to announce their regulations before they take effect in the Federal Register, making it very difficult for them to hide the content of their policy (Gersen and O'Connell 2009, 1161–62). It is illegal to deliberately choose not to enforce a regulation, and the explanatory materials that accompany a policy announcement are usually detailed enough to make the meaning of the policy clear. It might be possible for the agency or one of its staffers to hide the true intention of a policy in documents that would be disclosed under greater transparency, but doing so requires extremely careful crafting such that the content seems unambiguous but is actually vague enough to allow an agent to implement a different policy. In any case, the nature of the second type of information provides a much easier path for an agent aiming to obfuscate.

The second type of information surrounds the group's type. The agent receives a random signal, $S \in \{H, L\}$, pointing to the group's type. This gain in information can be motivated by the group's communications with the agent, for instance, by submitting evidence that it believes indicates the high costs of the regulation.[3] The signal is properly understood as the agent's impression of the group's claims and works as follows: the signal is $s = L$ only if the group is actually of the low-cost type, and then only with probability $q < 1$. With the remaining probability $p_H + p_L(1 - q)$, the signal is $s = H$.[4] The interpretation of this kind

[3]The possibility that the group might not communicate with the agent will be considered in an extension of the model.

[4]The reason for this asymmetric signal is that, to correspond with what one expects to occur after a negative media report (described in the next paragraph), a media report should always indicate a greater likelihood of the group's costs being low. The equilibrium solutions are more complicated, but the qualitative result that increasing transparency is not always beneficial remains when the agent can misread high-cost type as a low-cost group.

of signal is that the agent may be able to determine conclusively that the interest group can easily afford the regulation's implementation costs, but it cannot definitively find that the industry cannot afford the costs.

Like the group's type, the agent's signal is private information. Importantly, he cannot credibly communicate the signal to the principal or public. Instead, they learn imperfectly about the agent's signal through the media (or other reporter). This imperfection is represented by the media's producing a media report, M, that the agent's signal was L ($m = 1$) with some positive probability if and only if the agent selects a policy below some threshold $\tilde{r}$. Otherwise, it produces no media report ($m = 0$). If the proposal is below the threshold, then the probability of the presence or absence of negative media coverage is p_m^s, with $p_1^L \geq p_1^H$. This ordering can be rationalized by noting that, if the agent signal is low, either the media is more likely to interpret the agent's information as indicating a low-cost group or it can more easily and convincingly create a negative report against the agency.

The effect of the media report is to impose a cost on the agent $k_A > 0$ (assumed to be the same for both types of agents). This cost can represent the embarrassment an agent faces from being named in a news piece, the disapproval from a leader who sees the agency mentioned in the news, or if the player is a full agency, it can represent the extra resources that it must devote to damage control.[5] The group also incurs a similar kind of cost $k_G > 0$ from the negative publicity. The public clearly cannot compensate for embarrassment and disapproval costs, and it is difficult to imagine that it would pay to support additional

[5]The need for damage control can be reconciled with a fully rational political principal. One can suppose that, even though the average (or median) voter treats the media report simply as information, there are other voters who express their outrage, which produces costs for the agent and group. In the alternative construction of the media as a watchdog organization, such an organization may be able to bring shame to the agent or agency. Spontaneous reactions like disappointment and outrage, along with bureaucrats' perception of them, cannot easily be suppressed. Even the narrower policymaking public might not be able to refrain from reacting in these ways and imposing costs on regulators.

public relations expenditures. Also, as a simplifying assumption, there are no positive media reports, to follow the general conclusion that such reports are extremely rare compared to negative ones (Lee 2008). The costs and the threshold for avoiding a media report are common knowledge, and they make the agent's proposal less like cheap talk.

1.2.3 The Role of Transparency

Greater transparency means more information disclosure. Transparency is represented as a real variable $t \in \mathbb{R}_+$, which could represent the number of documents or number of categories of documents that the agency must release to the public. There may be a maximum value for t. Since information about the agency's signal is communicated through the media, transparency has an impact when it changes values of m_S. With $\mathbf{p}_1 \equiv (p_1^L, p_1^H)$, the probability of a negative media report can be expressed as $\mathbf{p}_1(t)$, where $p_1(\bar{t}) \geq p_1(\underline{t})$ when $\bar{t} > \underline{t}$, with the constraint that $p_1^L \geq p_1^H$ for any value of t. Note that this leaves the possibility that the media might perfectly report instances in which a low policy was based on a low cost signal without reporting any instances in which it was based on a high cost signal, even as transparency increases.

On the other hand, p_1^H may increase with t if the media is not quite objective or is susceptible to incorrectly reporting on capture. It may be digging to find anything that it can report to the public, even though the agent's conscience about his role in the policy process may be perfectly clear. Alternatively, the media may simply be mistaken (at least from the agent's view) about the interpretation of the documents it receives. In the leading example, the agent may be fully convicted, based on his reading of the evidence, that regulation incurs high costs, but the media may nonetheless declare that he has unduly favored the industry in his policymaking. The possibility of media misreporting begins to suggest that there may

be some costs to heightened transparency.

Thus, it is the imperfect chain of information transmission from the agent to the public, mediated by news reports, that is the focus of this model and at least some questions about the instrumental value of transparency. Notably, it departs from other models of capture and information transmission (e.g., Tirole 1986, Laffont and Tirole 1991), by denying the group and the agent the ability to credibly signal its type or information. The high-cost group has no independent way of definitively indicating to the agent (or the media or the public) that its costs are high, although its communications with the agent will always hold up at least as well under scrutiny as the low-cost type's. Similarly, the agent has no way of conveying to the media or the public what its signal was. In these ways, the model, while portraying information in signals, still reflects the notions that information consists of documents that need to be interpreted rather than just signals.[6]

1.2.4 Influence (Capture)

The final elements of the game relate to the undue influence that greater transparency is designed to prevent and the public's response. Influence will take the form of a transfer payment that takes some form that is legal. In the U.S., at least, outright bribery of bureaucrats is rare. Instead, industry influence of regulators tends to take more subtle forms, like the implicit promise of employment within the industry after the regulator leaves his or her agency (Quirk 1981). Although other acts are punishable by ethics rules, there remain many legal channels through which firms can implicitly compensate agency officials for favorable policy. Furthermore, since the group identifies the agent's type before the agent makes his

[6]Fenster (2006) observes, "[T]he subset of government texts that are ultimately disclosed does not appear to the public as raw information that is ready, in its capacity as the carrier of the stuff of government and politics, to enable democracy and produce the consequences anticipated by transparency advocates" (927).

proposal, it can choose to attempt to influence only the venal agent. Although contractual agreements to induce policy changes are illegal, the model treats the transfer as if it were enforceable, following the idea of a "quasi-contract" in Laffont and Tirole (1991).

Despite the implicit nature of the bargain, capture is treated as if the group makes a take-it-or-leave offer to the venal agent.[7] It offers a benefit d, which can be made contingent on any observable features of the game. Thus, it can offer different values d for different policies that result in the end, and it can offer some benefit to the venal agent merely for proposing a policy, regardless of any media reports or policy changes. When the quasi-contract executes, the group loses d from its utility and the agent receives d toward his utility. Since he does not care about policy, the venal agent's utility is simply any transfer it receives from the firm minus any media penalty (k_A) it receives.

Because the venal agent is not doing anything illicit, the public does not have any recourse to legal sanctions. Its only defense against undue agency influence is to have the principal change the policy from what the agent has proposed to something else.

1.2.5 Summary

The various components in this model are organized as follows:

1. Nature selects types for the group and the agent.

2. The agent receives signal about the group's type.

3. The group presents an offer to the venal type of transfer payments based on observable

[7]Reversing the bargaining power in favor of the venal agent would allow him to screen the agents, in which case the possibility of capture would appear to be beneficial. The idea that a group with high costs might pay compensation to demonstrate its high costs is intriguing but is beyond the scope of this paper and conventional regulatory policy. In any case, the goal is to have transparency, rather than the venal agent, to improve upon regulatory outcomes, possibly through screening.

variables.

4. The agent publicly proposes a policy.

5. The media reviews any information about the agency's signal that it has and produces a negative report with probability conditioned on the agency's actual signal and the policy that the agent has selected.

6. The public, through the political principal, decides whether to change the policy from what the agent has proposed. Then policy payoffs are realized.

The goal is to determine the impact of increased transparency, which means an increase in p_1^L, p_1^H, or both. Since this is a signaling game, there will be multiple equilibria and thus a question of equilibrium selection. Increased transparency is always beneficial only if, for any $\mathbf{p}_1(t)$, the public's payoff increases with t. Otherwise, it is not clear whether an increase in transparency is beneficial unless the value of t is calibrated to maximize the public's payoff. Such ambiguity would imply that transparency policies need to be tailored to different agencies and possibly to different decisions. Whether or not such specificity is feasible, it contrasts with the simpler call in President Obama's (2009) memorandum on FOIA for agencies to "adopt a presumption in favor of disclosure" (p. 4683) and does not directly specify cases in which some intermediate level might be more appropriate.

1.3 Equilibrium Results

The equilibrium concept for this game is perfect Bayesian equilibrium: players have the correct beliefs about player types on the path of play, and their strategies are optimal given their beliefs on and off the equilibrium path.

1.3.1 Basic Properties

Bayesian Updating Both the upright agent and the political principal engage in Bayesian updating in equilibrium; at the end of the game, each of these players has a posterior probability λ that the group has lower costs for any level of regulation. For the agent, the L signal implies $\lambda = 1$ since only the low-type generates that signal. Meanwhile, the H signal implies a posterior probability of

$$\lambda_{U_H} \equiv \lambda_{V_H} \equiv \frac{p_L(1-q)}{p_H + p_L(1-q)}. \tag{1.1}$$

Then the principal updates her probability of the low type based on the proposal and, if the proposal is below the threshold $\tilde{r}$, whether there is a media report. Just after the proposal, her value of λ is determined by the circumstances under which the agent would propose that value. For the upright agent, the relevant scenario is simply whether he saw s=H or s=L. We can denote these situations as U_H and U_L. Because the venal agent is susceptible to influence from the group, the relevant scenarios for him involve both the group's type and the agent's signal. These situations, which constitute the sample space Ω, can be denoted as V_s^i, with $V_H \equiv V_H^H \cup V_H^L$, $V^L \equiv V_H^L \cup V_L^L$, $V_L \equiv V_L^L$, and $V \equiv V_H^H \cup V^L$. We can further define $A_s \equiv U_s \cup V_s$. Since an agent in a particular setting may choose to mix among different strategies, it is also useful to place a fraction in front of any of these scenarios to denote the probability with which the agent proposes a particular level of regulation. Then the posterior probability after a proposal can be presented with a subscript for the set of events under which the proposal occurs. For example, if the agent proposes the same policy under settings U_H, V_H^H, V_H^L, and some fraction of V_L^L, θV_L^L, the public and principal's revised probability

of the low type becomes

$$\lambda_{A_H \cup \theta V_L^L} = \frac{p_L(1-q) + \theta p_V p_L q}{p_H + p_L(1-q) + \theta p_V p_L q}. \tag{1.2}$$

Other posterior probabilities can be calculated placing the sum of the probability masses associated with a proposal in the denominator and sum of those masses containing p_L in the numerator.

If the proposal falls below $\tilde{r}$, then at the media report stage, principal can further update her probability. Then symbols $\bar{\lambda}$ and $\underline{\lambda}$ can be used to represent, respectively, the updated probabilities with and without a media report. Continuing with the example including all the agents other than U_L and $(1-\theta)V_L^L$, the probabilities after the media reporting stage are

$$\bar{\lambda}_{A_H \cup \theta V_L^L} = \frac{p_L(1-q)p_1^H + \theta p_V p_L q p_1^L}{(p_H + p_L(1-q))p_1^H + \theta p_V p_L q p_1^L} \tag{1.3}$$

$$\text{and } \underline{\lambda}_{A_H \cup \theta V_L^L} = \frac{p_L(1-q)p_0^H + \theta p_V p_L q p_0^L}{(p_H + p_L(1-q))p_0^H + \theta p_V p_L q p_0^L}. \tag{1.4}$$

The Principal's Decision Rule Based on this updating, the political principal's decision rule can be derived. Her overall utility is

$$f(r, \lambda) \equiv b_P(r) - (\lambda \gamma_P^L + (1-\lambda)\gamma_P^H) c_P(r). \tag{1.5}$$

The assumptions on $b_P(\cdot)$ and $c_P(\cdot)$ guarantee a uniquely optimal policy, $\tilde{r}(\lambda)$, for any posterior probability.[8] Since she can set policy freely after the agent's proposal and any

[8]The quantity $\tilde{r}(\lambda)$ satisfies $b_P'(\tilde{r}(\lambda)) = (\lambda \gamma_P^L + (1-\lambda)\gamma_P^H) c_P'(\tilde{r}(\lambda))$.

media report, her decision rule becomes

$$r_P^* = \tilde{r}(\lambda). \tag{1.6}$$

This decision rule is not only optimal for the principal, but it also matches what one expects from a media report: the final policy is weakly higher with a media report than without one if the proposal was below the threshold. This fact is implied by the following lemma:

Lemma 1.1. *The principal's choice of regulation increases with her posterior probability on the low-cost type. When the proposal is below the media threshold,* $\bar{\lambda} \geq \underline{\lambda}$.

Proof. Proofs of all numbered results in this chapter are in Appendix B.1. ■

Mechanism of Regulatory Capture The group is able to identify the venal agent after the first stage and influence him. Since it presents an offer to the agent, it has all the bargaining power. The agent's costs come from adverse media reports, so the group only has to compensate for the costs associated with negative reports. Thus, it need not offer anything to have the agent propose a policy $\tilde{r}$ or above. For policies below this threshold, the likelihood of bad publicity depends on the signal the agent (if any) received. Because news is intrinsically publicly observable information, the group can ensure that the venal agent receives no surplus by paying $d = k_A$ only in the event of a media report. If $p_1^L > p_1^H$, it is cheaper for the low-cost group to influence the agent under V_H^L than under V_L^L, and it can choose to influence only the venal agent with the high signal. The mechanism is to offer $d = p_1^H k_A$ to the venal agent for proposing a policy or one of set of policies below $\tilde{r}$, regardless of whether a media report occurs. Then the venal agent under V_H^L receives zero

in expectation, while under V_L^L, he receives $\left(p_1^H - p_1^L\right) k_A < 0$.[9]

Equilibrium Viability and Selection With various types and a continuous policy space, an infinite number of perfect Bayesian equilibria are possible. These can be reduced into a smaller set of equivalence classes based on the payoffs to the three strategic players. Then the equilibria in an equivalence class share the following characteristics: (a) the same pooling among the agent settings described above on proposals, and (b) the same side of the media reporting threshold for each setting. There will often be more than one possible equivalence class, so one more refinement is that only equilibria in which U_L (always) proposes a policy above the media threshold will be considered. While this restriction is not necessary to prove most of the propositions that follow, it is sensible because agreement between the upright agent and the principal implies that U_L should propose relatively high policies.[10]

Among these equilibria, that one that will be selected is the one that yields the public its highest expected payoff. This criterion is useful because many of the equilibria that follow given a set of parameters can be ranked, and because it allows for the derivations of comparative statics on the media report probabilities. Focusing on the public's best equilibrium provides a starting point for alternative selection criteria, which are discussed in the following section. The analysis will also consider whether there is any equilibrium selection criterion under which more transparency is preferred for any function $\mathbf{p}_1(t)$.

The Default Equilibrium The agent can always avoid a media report by proposing a policy that is at least the threshold and thus always engage in cheap talk through its

[9]This option makes unnecessary the need to consider how V_H and V_L should different when they are indifferent among a large number of policies because it is less arbitrary when V_L can be induced not to choose a policy to which V_H is amenable.

[10]The level of the threshold does not matter for the results in the basic model, but, for realism, one may imagine that $\tilde{r} \in (\tilde{r}(0), \tilde{r}(1))$, or even that $\tilde{r} \in (\tilde{r}(\lambda_{U_H}), 1)$.

proposals. One type of equilibria in which the agent always proposes above $\tilde{r}$ are as follows: The venal agent and upright agent with the high-cost signal propose one level of regulation that avoids a media report, and the upright agent with the low-cost signal proposes a different regulation, also at least $\tilde{r}$. Then there are no transfer payments to the venal agents, and the public chooses the final regulation according to its decision rule. The first proposition effectively makes this equilibrium the default for the public and the principal:

Proposition 1.2. *As long as $p_1^L < 1$, the default equilibrium can always be sustained. The default equilibrium yields a higher payoff to the public and the principal than any fully pooling one.*

While, in many signaling games with interests opposed, only a fully pooling equilibrium obtains, this game provides the public with a better default payoff because the upright agent is able to help by obtain information about the group. Thus, it is always possible to achieve an equilibrium that yields the public at least $\Pr(\Omega \setminus U_L)\hat{f}(\lambda_{\Omega \setminus U_L}) + \Pr(U_L)\hat{f}(1) > \hat{f}(p_L)$, where $\hat{f}(\lambda) \equiv f(\tilde{r}(\lambda), \lambda)$.

1.3.2 Equilibria with Only True Positives

Equilibria with No Media Reports A setting without media reports provides a baseline from which to consider the effects of increasing transparency. The transparency variable t can be scaled arbitrarily. Suppose there is a value t such that $p_1^L = p_1^H = 0$. This setting would represent a world in which FOIA does not exist and government agencies can operate without disclosing any relevant documents until they announce their proposed policies. Then the threshold for media reporting is irrelevant, and all agency proposals become cheap talk as costly signaling becomes impossible. In this case, the low-cost group can always induce

the venal agent to imitate the proposals from U_H and/or V_H^H. As a result, the best the principal can do is to have U_L separated from the other agent scenarios, which is what she is always able to do:

Proposition 1.3. *When $p_1^L = 0$, there exists no equilibrium that yields the public a greater payoff than the default equilibrium.*

Thus, the no transparency case leaves the public with only the default payoff, suggesting that some media scrutiny would be beneficial. Even though there may be drawbacks to having too much transparency, neither critics of over-transparent government nor this model promotes the opposite extreme of having no transparency.

Equilibria with Media Reports Only after the Low-cost Signal With additional transparency, the media can potentially have the information necessary to create a report on an agency policymaking decision. If, with a higher value of t, p_1^L increases but there remains no risk of a media report after a high signal, that means that every media report indicates perfectly that the agent observed $S = L$, and thus that the group is of the low-cost type. Lack of a media report does not point to the agent's having seen $S = H$ unless $p_1^L = 1$, but it does point to a greater likelihood that the proposal is supported by a high-type signal, provided that any agent seeing the high-type signal has proposed that policy.

The public benefits because, while agents having seen the high signal can still freely choose to propose policies below $\tilde{r}$, the group will need to pay the venal agents who have seen the low-cost signal to induce them to propose under the threshold. Additionally, the group incurs k_G if a news report occurs. Thus, the media report serves two related functions: First, its presence or absence provides information to the political principal that allows her optimize the policy selection further. Even if all the agents besides U_L still pool on the

same proposal, the public benefits from media reporting if that proposal is below the media threshold. Then there is a chance of a negative publicity that allows the principal to adjust the final policy upward to $r_P = \tilde{r}(1)$.

Second, the low-cost group facing the venal agent with $s = L$ must consider whether it is worth risking a media report to induce a proposal below the threshold. If there is no media report, such a proposal yields $r_P = \hat{r}(\underline{\lambda}_x)$, where $\underline{\lambda}_x$ is the principal's lowest value of $\underline{\lambda}$ for a proposal less than $\tilde{r}$. With a media report, the low-cost group receives $r_P = \hat{r}(1)$ because the principal is aware that only the low-cost group produces a media report. Furthermore, it loses k_G after a media report and in expectation, pays $p_1^L k_A$ to compensate the venal agent in the event of a report. On the other hand, the low-cost group can induce a proposal at least the threshold for a safe $r_P = \hat{r}(\lambda_y)$, where λ_y is the principal's lowest value of λ for any $r_A \geq \tilde{r}$. Therefore, to decide whether to propose at least $\tilde{r}$, both at least and less than $\tilde{r}$, or only below the threshold, the low-cost group's incentive compatibility test facing V_L^L is

$$p_0^L \gamma^L \hat{c}(\underline{\lambda}_x) + p_1^L \left(\gamma^L \hat{c}(1) + k_A + k_G\right) \gtreqless \gamma^L \hat{c} \lambda_y, \tag{1.7}$$

where $\hat{c}(\lambda) \equiv c(\hat{r}(\lambda))$.

The cost to the low-cost group on the right-hand side of the constraint is independent of the probabilities of media reports, but the cost on the left-hand side does depend on p_1^L. Intuitively, increasing p_1^L should make it less attractive to propose below the threshold. The additional likelihood of receiving a media report and the weakly higher policy that goes with it[11] should outweigh the fact that the policy without a media report is lower. A general set of circumstances in which increasing p_1^L makes proposing below $\tilde{r}$ more costly can be

[11]The policy is strictly higher when $p_1^H = 0$.

articulated:

Lemma 1.4. *Suppose some fraction η_U of U_H and some fraction η_V each of V_H^H and V_H^L (at least one of them strictly positive) are fully pooled with some $\theta > 0$ of V_L^L proposals below the threshold in equilibrium. Then the cost to the low-cost group of inducing a proposal below $\tilde{r}$ increases with p_1^L whenever $1 > p_1^L \geq p_1^H$ and $\hat{c}(\lambda)$ is convex with respect to λ.*

Convexity of $\hat{c}(\lambda)$ with respect to λ is a fairly weak condition, as explained in the proof to Lemma 1.4. Based on this lemma, one natural possibility for equilibria in this setting is to have one proposal below $\tilde{r}$ after A_H and some fraction $\theta \in [0, 1]$ of V_L^L and a second proposal of at least $\tilde{r}$ after the remaining $1-\theta$ of V_L^L and all of U_L. For the first proposal, the absence of a media report implies $\lambda = \underline{\lambda}_{A_H \cup \theta V_L^L}$, while a media report and the second proposal imply $\lambda = 1$. Thus, the low-cost group facing V_L^L applies the incentive compatibility test in Equation (1.7) with $\underline{\lambda}_x = \underline{\lambda}_{A_H \cup \theta V_L^L}$ and $\lambda_y = 1$. Because $\underline{\lambda}_x$ increases with θ and all the other quantities stay constant with θ, the V_L^L incentive compatibility constraint is satisfied for exactly one value of θ. With the proper beliefs, there is no deviation in the other situations. While there are other equilibria, it turns out that these natural ones are the ones that yield the principal the highest possible payoff. Thus, equilibria with $p_1^L > p_1^H = 0$ can be formally characterized as follows:

Proposition 1.5. *Suppose $p_1^H = 0$ for any t. When $p_1^L > 0$, equilibria exist with some $r_A < \tilde{r}$ after A_H and fraction $\theta \in [0, 1]$ of V_L^L, and a second $r_A \geq \tilde{r}$ that appears after U_L and the other $1 - \theta$ of V_L^L.*

(a) For given values of p_1^L, k_A, k_G, and γ^L, one sustainable equilibrium is among of the following three mutually exclusive types of equilibria:

(i) If $p_1^L(k_A + k_G) \leq p_0^L \gamma^L \left(\hat{c}(1) - \hat{c}\left(\underline{\lambda}_{A_H \cup V_L^L}\right)\right)$, V_L^L pools fully with A_H.

(ii) *If* $p_1^L(k_A + k_G) = p_0^L \gamma^L \left(\hat{c}(1) - \hat{c}\left(\underline{\lambda}_{A_H \cup \theta V_L^L}\right)\right)$, *for some* $\theta \in (0,1)$, *then* V_L^L *pools with* A_H *with probability* θ *and with* U_L *with probability* $1 - \theta$.

(iii) *If* $p_1^L(k_A + k_G) \geq p_0^L \gamma^L \left(\hat{c}(1) - \hat{c}(\lambda_{A_H})\right)$, V_L^L *pools fully with* U_L.

(b) *Each of the equilibria in (a) yields the principal more than the default payoff. The lower the fraction of* V_L^L *pooled with* A_H *in the equilibrium, the higher her expected utility.*

(c) *For each of the incentive compatibility scenarios in (a), the equilibrium that can obtain yields the principal her highest possible payoff.*

(d) *Suppose that the principal achieves her highest payoff given* p_1^L, k_A, k_G, *and* γ^L. *If raising the level of transparency sufficiently high allows* $p_1^L = 1$, *then the principal prefers to increase* t *so as high as possible.*

(e) *Even if* $p_1^L = 1$ *is not achievable, the principal strictly prefers that* t *increase as much as possible until* V_L^L *proposals are all at least* $\tilde{r}$, *provided that* $\hat{c}(\lambda)$ *is convex.*

Since, as noted above, it is relatively easy for $\hat{c}(\lambda)$ to be convex, it is fair to say that increased transparency is beneficial to the principal and the public when media reports never misidentify the signal. If $p_1^L = 1$ is achievable, then the media report is a perfect indicator of the signal, and the public can achieve the same payoff as it would achieve if there were only upright agents. However, this is not the highest payoff that the principal can achieve. The upright agent can never know for certain whether a high-cost signal definitely indicates high costs, and the principal cannot infer the group's type from his proposals. In contrast, the principal may be able to distinguish the two group types completely from the venal agent's proposals. However, she cannot do so if media reports occur only after the low-cost signal.

1.3.3 Equilibria with False Positives

If there are media reports after high-cost signals as well as low-cost signals, the optimal level of transparency becomes more complicated. To begin with, it is not the case that increasing p_1^H always makes the principal worse off. Instead, false positives can benefit the principal if they discourage the low-cost group from compensating the venal agent with that signal for a negative media report. In fact, if $k_A + k_G$ is large enough and p_1^L and p_1^H are at the right levels, the principal may be able to achieve her highest payoff:

Proposition 1.6. *An equilibrium in which different policies follow from V_H^H, U_H, and the other agent scenarios exists only when p_1^H is greater than zero and $k_A + k_G$ is sufficiently high. It yields the principal her highest possible payoff in the game, but it may not exist for any values of p_1^L and p_1^H.*

Qualitatively, this equilibrium requires three incentive compatibilities: (1) the low-cost group must find it too costly to compensate the venal agent for a media report and incur its own costs for negative publicity, (2) the high-cost group must find it not too costly to compensate the venal agent, and (3) the upright agent with $s = H$ must be prefer to incur the costs of a media report rather than propose at least $\tilde{r}$. The first condition requires a sufficiently high p_1^H, while the third requires a sufficiently low p_1^H, and there may not be a value that satisfies both, even if $k_A + k_G$ is high enough to allow for screening the low-cost group from the high-cost group when they face the venal agent. Still, if this equilibrium can exist, then the greater payoff from this equilibrium constitutes and improvement from when $p_1^H = 0$.

As the maximum payoff possible, the principal's expected utility from the equilibrium in

Proposition 1.6 is also greater than it could be if there were only upright agents in the game:

$$\begin{aligned} & p_U(p_H + p_L(1-q))\hat{f}(\lambda_{U_H}) + p_V p_H \hat{f}(0) + p_V p_L(1-q)\hat{f}(1) + p_L q \hat{f}(1) \\ & \quad > p_U(p_H + p_L(1-q))\hat{f}(\lambda_{U_H}) + p_V p_H f(\lambda_{U_H}, 0) + p_V p_L(1-q) f(\lambda_{U_H}, 1) + p_L q \hat{f}(1) \\ & \quad = (p_H + p_L(1-q))\hat{f}(\lambda_{U_H}) + p_L q \hat{f}(1) \quad (1.8) \end{aligned}$$

Thus, this payoff requires venal agents, and it requires the venal agent to accept compensation from the high-cost agent. If even this compensation is successfully deterred, the principal may end up not being able to achieve more than her default payoff.

Proposition 1.7. *For sufficiently large k_A and k_G, there always exists some value of p_1^H such that the principal can receive no more than her default equilibrium payoff.*

The situation described in Proposition 1.7 requires occurs when the high-cost and low-cost group types are both deterred from inducing the venal agent propose below the media threshold and the upright agent with the high-cost signal is discouraged from proposing less than $\tilde{r}$. The scenarios in Propositions 1.6 and 1.7 entail high media costs. Since a group may have more at stake than what they would need to compensate a venal agent, these situations seem to necessitate substantial direct costs to the group from negative publicity[12]. For smaller media costs, it is more likely that the two benefits from increasing p_1^L while keeping $p_1^H = 0$ will be reversed. First, the information that the principal gains from the media report becomes less valuable, since, with $p_1^H > 0$, it is possible that $s = H$ preceding negative publicity.[13] Second, in certain circumstances, inducing a proposal below the media

[12]However, if, following Laffont and Tirole (1991), one supposes that a transfer d to the agent costs more than d to the group, the cost of compensation to the venal agent may also become high enough to deter it.

[13]If $p_H^1 = p_L^1$, the principal and public gain no information based on the presence or absence of a media

threshold will become at least weakly more attractive for the low-cost group facing V_L^L as p_1^H increases:

Lemma 1.8. *Suppose some fraction η_U of U_H and some fraction η_V each of V_H^H and V_H^L (at least one of them strictly positive) are fully pooled with some $\theta > 0$ of V_L^L proposals below the threshold in equilibrium. Then the cost to the low-cost group of inducing a proposal below $\tilde{r}$ decreases with p_1^H whenever $p_1^L > p_1^H$ and $\hat{c}(\lambda)$ is convex with respect to λ.*

Lemma 1.8 is the converse of Lemma 1.4. In this case the reductions in costs due to$\bar{\lambda}$ decreasing for the low-cost group in the event of a media report outweigh the increase in cost from $\underline{\lambda}$ rising when there is no media report. If U_H is not discouraged below $\tilde{r}$ when $p_1^H = 0$, then many of the equilibria analogous to those in Proposition 1.5 are worse:

Proposition 1.9. *Suppose it remains incentive-compatible for U_H to propose below the media threshold, $p_1^H(k_A + k_G) < \min\{\gamma^L(\hat{c}(1) - \hat{c}(\lambda_{A_H})), \gamma^H(\hat{c}(p_L) - \hat{c}(\lambda_{A_H})\}$, and $0 < p_1^H < p_1^L$.*

(a) For given values of p_1^L, p_1^H, k_A, k_G, and γ^L, one sustainable equilibrium is among of the following three mutually exclusive types of equilibria:

(i) If $p_1^L(k_A + k_G) \leq \gamma^L(p_0^L\left(\hat{c}(1) - \hat{c}\left(\underline{\lambda}_{A_H \cup V_L^L}\right)\right) + p_1^L(\hat{c}(1) - \gamma^L\hat{c}(\bar{\lambda}_{A_H \cup V_L^L})))$, then V_L^L pools fully with A_H.

(ii) If $p_1^L(k_A + k_G) = \gamma^L(p_0^L\left(\hat{c}(1) - \hat{c}\left(\underline{\lambda}_{A_H \cup \theta V_L^L}\right)\right) + p_1^L(\hat{c}(1) - \hat{c}(\bar{\lambda}_{A_H \cup \theta V_L^L})))$ for some $\theta \in (0,1)$, then V_L^L pools with A_H with probability θ and with U_L with probability $1 - \theta$.

(iii) If $p_1^L(k_A + k_G) \geq \gamma^L(\hat{c}(1) - \hat{c}(\lambda_{A_H}))$, V_L^L pools fully with U_L.

---

report. However, the possibility of a media report still allows for the possibility of screening the low-cost group from the high-cost group.

(b) Holding p_1^L constant, equilibrium (i) and equilibrium (ii) (for a given θ) in (a) yield a lower payoff for the principal for any given $p_1^H > 0$ compared to when $p_1^H = 0$, while the payoff for type (iii) is the same as for type (iii) in Proposition 1.5(b). Holding p_1^L and p_1^H constant, the lower the fraction of V_L^L pooled with A_H in the equilibrium, the higher her expected utility.

(c) The equilibria in (a) achieve the principal's highest payoff.

(d) If $p_1^L(k_A + k_G) < \gamma^L\left(\hat{c}(1) - \hat{c}(\lambda_{A_H})\right)$, convexity of $\hat{c}(\lambda)$ implies that the principal's maximum payoff in (a) increases with p_1^L while p_1^H is held constant and decreases with p_1^H when p_1^L is held constant.

(e) If $p_1^L(k_A + k_G) \leq \gamma^L(\hat{c}(1) - \hat{c}(\lambda_{A_H \cup V_L^L}))$ and $\hat{c}(\lambda)$ is convex, then as p_1^H approaches p_1^L, the principal's payoff approaches the default payoff.

Proposition 1.9 focuses on conditions under which media reports after the high-cost signal do not serve to screen the agents with the high-cost signal from each other. In that case, increasing p_1^H only results in information loss, which, under weak conditions, makes the principal worse off. Part (e) indicates that, if p_1^L does not sufficiently serve to induce some V_L^L proposals to at least $\tilde{r}$, that the principal's expected utility falls all the way to the default payoff. This can occur even when V_L^L proposals were partially or fully separated from A_H proposals with $p_1^H = 0$. The reason is that $p_1^L(k_A + k_G) \geq p_0^L \gamma^L\left(\hat{c}(1) - \hat{c}(\lambda_{A_H})\right)$ from Proposition 1.5(a)(iii) does not imply that $p_1^L(k_A + k_G) > \gamma^L\left(\hat{c}(1) - \hat{c}(\lambda_{A_H \cup V_L^L})\right)$. Thus, if increasing transparency causes an increase in both p_1^L and p_1^H, the overall effect is ambiguous.

The remaining set of circumstances to consider is U_H's proposing at least the threshold, which is possible since the upright agent incurs k_A for a media report when $p_1^H > 0$. The up-

right agent could then conceivably make a different proposal after each signal.[14] It would be appealing if these proposals were distinct from the venal agents' proposals below $\tilde{r}$. However, it is not incentive compatible for the low-cost agent facing V_L^L to continue proposing below the threshold when it can pool with U_H instead. Instead, if U_H sets $r_A \geq \tilde{r}$, his proposals must be pooled with venal agent proposals, with proposals from U_L, or with both. Because U_H proposals are never by themselves, cannot produce equilibria as good for the principal as the one in Proposition 1.6. On the other hand, the low-cost agent has more of an incentive to induce V_L^L to pool with U_H than with U_L. Equilibria with the upright proposing different value of $r_A \geq \tilde{r}$ for each signal that can be more formally characterized as follows:

Proposition 1.10. *Consider equilibria in which U_H proposes $r_A \geq \tilde{r}$, but always separately from U_L.*

(a) U_H proposals cannot be separated from venal agent proposals. If $\hat{c}(\lambda)$ is convex, then a fraction of V_L^L proposals exceeding p_U must be pooled with U_H, so that $\lambda > \lambda_{A_H \cup V_L^L}$ for proposals involving U_H.

(b) If $\hat{c}(\lambda)$ is convex and some fraction of V_L^L proposals are originally below $\tilde{r}$, the fraction of V_L^L proposals pooled with U_H proposals will increase if p_1^L increases and decrease if p_1^H increases (provided that $p_1^H < p_1^L$).

(c) For a given p_1^L and p_1^H, this type of equilibrium may or may not exist.

(d) Suppose the best equilibrium payoff for the principal in which U_H proposes below $\tilde{r}$ (if any) involves some V_L^L proposals below the threshold. Then there will be fewer venal

[14] U_H pooling with U_L below the threshold results in an additional loss of information for the principal, and the resulting equilibrium would almost certainly not be the best one available to her.

agent proposals below $\tilde{r}$ in any equilibrium in which U_H proposes $r_A \geq \tilde{r}$ and separately from U_L.

(e) Suppose the best equilibrium payoff for the principal in which U_H proposes below $\tilde{r}$ (if any) is achievable with V_L^L proposals all at least $\tilde{r}$ and some venal agent proposals below $\tilde{r}$ that are all distinct from U_H proposals. Then there will be weakly fewer venal agent proposals below $\tilde{r}$ in any equilibrium in which U_H proposes $r_A \geq \tilde{r}$ and separately from U_L.

Propositions 1.6, 1.7, 1.9, and 1.10 show that the principal gains from increases in p_1^H only when the resulting best equilibrium is one that splits from each other some of the proposals resulting from the high-cost signal, i.e., one in which V_H^H, V_H^L, and U_H proposals are not all on the same side of the threshold. Essentially, p_1^H needs to be high enough to screen apart these agent scenarios, but it also needs to be not so high as to always deter the agent from proposing below the threshold. Otherwise, the false positives only make the principal worse off via loss of information about the agent's signal. More generally, Propositions 1.7 and 1.9 show that there exist functions $\mathbf{p}_1(t)$ under which more transparency does not make the principal better off. Overall, there exists a non-trivial set of cases in which increasing transparency need not improve outcomes for the public. In particular, because the upright agent with the high-cost signal can be discouraged from proposing below the threshold, the principal and public can be worse off even if the increase in transparency is informative (i.e., the ratio p_1^L/p_1^H increases) but p_1^H also increases.

1.3.4 Alternative Equilibrium Selection Criteria

Given that the public receives its highest payoff, the preceding discussion characterizes circumstances under which its expected payoff increases. However, the game does not intrinsically require that the public receive its greatest payoff. The question is whether any other reasonable criterion would make increasing transparency unambiguously beneficial for any $\mathbf{p}_1(t)$.

Criteria Based on Other Players' Payoffs Instead of the public's payoff, one might suppose that some other player's payoff is maximized, like the high-cost group or the upright agent. For the high-cost group, the rationale is that the high-cost group will decide whether it prefers to distinguish itself from the low-cost group or finds it too costly to do so, and that it has the political savvy to minimize its cost. By Proposition 1.2, the default equilibrium, which yields the high-cost group $\gamma^H \hat{c}\left(\lambda_{A_H \cup V_L^L}\right)$, is always achievable. If the high-cost group achieves its highest payoff, then some of the highly informative equilibria that occur with relatively large values of p_1^H might no longer be selected. For example, an equilibrium of the type described in Proposition 1.6 might be sustainable, with $\gamma^H \hat{c}(0) + p_1^H(k_A + k_G) \leq \gamma^H \hat{c}(1)$. Even when this condition holds, it is possible that its overall payoff, $p_1^H k_G + p_U \gamma^H \hat{c}\left(\lambda_{U_H}\right) + p_V(\gamma^H \hat{c}(0) + p_1^H k_A)$, is not as good for it as the default payoff since $\gamma^H \hat{c}\left(\lambda_{A_H \cup V_L^L}\right) < \gamma^H \hat{c}(1)$. If $\hat{c}(\cdot)$ is convex, it is also possible that the high-cost group's cost from an equilibrium in Proposition 1.9 will exceed the default equilibrium cost, since $p_0^H \gamma^H \hat{c}\left(\bar{\lambda}_{(\cdot)}\right) + p_1^H\left(\gamma^H \hat{c}\left(\underline{\lambda}_{(\cdot)}\right) + k_G + p_V k_A\right) > \gamma^H \hat{c}\left(\lambda_{A_H \cup V_L^L}\right)$ is consistent with $p_0^H \gamma^H \hat{c}\left(\bar{\lambda}_{(\cdot)}\right) + p_1^H\left(\gamma^H \hat{c}\left(\underline{\lambda}_{(\cdot)}\right) + k_G + k_A\right) < \gamma^H \hat{c}(1)$. Since there are values of p_1^H small enough for the high-cost group to propose below the threshold, there are also paths $\mathbf{p}_1(t)$ for which increasing transparency would not increase the public's payoff. With this criterion,

even paths in which p_1^H sometimes increases, but less quickly than p_1^L, could yield this effect.

A similar challenge arises the upright agent is assumed to achieve his greatest payoff. His incentive compatibility constraint in an equilibrium will often entail avoiding $\alpha f(1, \lambda_{U_H})$, but the default equilibrium yields $\alpha f\left(\lambda_{A_H \cup V_L^L}, \lambda_{U_H}\right) > \alpha f\left(1, \lambda_{U_H}\right)$. A third possibility is to eliminate equilibria in which both the upright agent and the high-cost group do worse than in the default equilibrium, since the public cannot expect both the high-cost group and the upright agent to incur large costs solely for its benefit. However, this criterion still leaves functions $\mathbf{p}_1(t)$ where the public's expected utility does not increase with t, because increasing p_1^H slightly yields losses as described in Proposition 1.9, and increasing it more may yield the default equilibrium even when better equilibria for the public could be sustained. Overall, it seems difficult to construct a criterion in which the public improves its payoff with p_1^H as with p_1^L based on comparable payoffs for one or more players.

Criteria Based on Proposal Content Another method of selecting equilibria is to have the proposals correspond to the posterior probabilities to some extent. In the current form of the game, the principal has full authority to adjust policy, so the only features of an agent proposal are whether it is below the threshold and which agent scenarios produce that proposal. The value of the threshold has been entirely irrelevant due to the principal's power. However, it might be supposed that the agent's proposal has some connection to the final policy in the equilibrium that actually obtains. For example, the upright agent would propose $\hat{r}(1)$ after the low-cost signal if $\tilde{r} \leq \hat{r}(1)$, even though any proposal of at least the threshold distinct from all the others would also result in the same policy.

More generally, the value of the threshold might influence equilibrium selection. For example, if $\tilde{r} > \hat{r}(\lambda_{U_H})$, one might expect the upright agent to propose $r_A = \hat{r}(\lambda_{U_H})$ as

long as at least one equilibrium with that proposal can be sustained. However, this still leaves the possibility of losses according to Propositions 1.7 and 1.9. On the other hand, if $\tilde{r} \leq \hat{r}(\lambda_{U_H})$, the upright agent might be expected to propose $r_A = \hat{r}(\lambda_{U_H})$, in which case the default equilibrium would always obtain since the venal agent proposals would pool with the upright agent's.

Overall, it appears difficult to construct a selection criterion in which the public's expected payoff increases both with p_1^L and p_1^H that is not identical to deliberately selecting equilibria to yield this outcome. In particular, with some pairs of these probabilities yielding only the default payoff at most (e.g., $p_1^L = p_1^H = \epsilon > 0$ for sufficiently small ϵ), other pairs of these probabilities would need to be adjusted downward toward the default equilibrium or even worse. Thus, the challenge of calibrating the level of transparency generally remains.

1.4 Extensions

The principal's role in this game modeled on regulation has been idealized and simplified in a few ways. First, it was assumed that the group would automatically communicate, but this need not necessarily be the case. Second, the principal might have additional sanctions available for certain mechanisms of regulatory capture, or she might opt to criminalize some forms of influence. Third, the principal might not be able to intervene as often, leaving the courts to limit regulatory capture. With the basic equilibrium results above providing a framework, these extensions can be considered separately. The ambiguities in transparency's impact will remain.

1.4.1 Optional Communication for the Group

The standard game assumes that the group will generate a signal about its cost level and leaves as the main strategic questions how the group will influence the venal agents and whether the upright agent will propose a policy below the threshold. Suppose, in the first stage, that the firm has the option of not providing any information to the agent from which he can derive a signal. Since this is a signaling game, there will exist an equilibrium in which the group does not communicate with the agent, and the principal always assigns $r_P = \hat{r}(p_L)$. If, as in the standard game, the group discovers the agent's type before deciding what kind of transfers to offer, then it will generally be the case that $r_P = \hat{r}(p_L)$ in all agent situations in any equilibrium in which the group does not communicate.[15] Such equilibria yield the public $\hat{f}(p_L)$, which is less than its default expected utility by Proposition 1.2. They also yield a lower payoff to both the high-cost group and the upright agent, which means that this equilibrium would not occur under any of the criteria listed above.

However, a fully pooling equilibrium may be appropriate, depending on the equilibrium that is selected when the group provides information. If it is expected, according to any rule that does not leave the high-cost group as well off as possible, that the high-cost group will end up paying more than $\gamma^H \hat{c}(p_L)$ if the group communicates, then the high-cost group might try to have the pooling equilibrium occur. For example, suppose $\hat{r}(p_L) \geq \tilde{r}$. If the principal sees $r_A = \hat{r}(p_L)$ after the agent sees no signal from the group, the principal and the upright agent could realize that $\lambda = p_L$ because the high-cost group and low-cost group both

[15]The group might be able to determine the agent's type by talking, but not formally submitting (as much) evidence pointing to its high costs. An alternative equilibrium may occur if a media report can occur even with no communication and a proposal $r_A < \tilde{r}$, and if the low-cost group finds it too costly to compensate the venal agent for that risk. However, this probability, which might be denoted as $p_1^{\emptyset}$, would most likely be less than p_1^L, since it is easier to generate a media report with substantive information than without. Thus, it is unlikely that the high-cost group would separate itself from the low-cost group.

prefer a pooling equilibrium.[16] In contrast, if the high-cost group pays less than $\gamma^H \hat{c}(p_L)$ in a communicating equilibrium, it would not try to induce $r_A = \hat{r}(p_L)$. Then the low-cost group would not deny information and try to claim that the principal and agent should believe $\lambda = p_L$, because then its claim that the high-cost group would also want to be in a pooling equilibrium would not be credible.

The plausibility of a fully pooling equilibrium applies even though the informative equilibrium it replaces is perfectly sustainable with proper beliefs. In general, the high-cost group will compare its equilibrium payoff to $\gamma^H \hat{c}(1)$ to decide whether to deviate, but a comparison to $\gamma^H \hat{c}(p_L)$ implies a stricter test for equilibrium selection, akin to the intuitive criterion. Like other results of the model, the possibility that the final policy might always be $r_P = \hat{r}(p_L)$ does not arise if $p_1^H = 0$.

Also, as long as the low-cost group can sometimes generate a high-cost signal, there cannot be equilibria in which only the high-cost group communicates with the upright agent, who then distinguishes the two groups in his proposals. Then the low-cost group would receive policy $\hat{r}(1)$, whereas it would get better policy if it communicated, since then it would sometimes receive the same policy as the high-cost group, which is lower. A similar logic applies if the low-cost group communicates but the high-cost group does not. Meanwhile, communication is harmless with a venal agent, since the group can always opt not to induce a proposal below the threshold and thereby avoid all media reports. Thus, the main function of this extension appears to be to allow for a pooling equilibrium, which would only be worse for the principal than the equilibria that occur when the group is assumed to communicate with the agent.

[16]The venal agent who has not seen any information does not care about the group's type and only cares about being compensated if it proposes a policy that can trigger a media report.

1.4.2 Punishing Influence

In the standard model, the principal's power is limited to adjusting policy to her posterior belief. However, it might be thought that the principal has additional powers to punish influence. The availability of additional sanctions is not obvious. Among the forms of influence listed in Laffont and Tirole (1991, 1090–91), bribery is already a crime. Some other activities are also illegal for government officials because of ethics regulations applying only to them. Furthermore, activities that constitute capture are difficult to detect since there are frequently other explanations for behavior that seems unduly to favor industry (Carpenter 2013). Still, one might suppose that an ethical code is partially definable, and that, media reports might help the principal discipline ethics violations by agents. Then greater transparency could be helpful because in the increase in the probability of a media report.

If the punishment automatically occurs after the media report, then this is like k_A increasing. If there are false positives, this means that upright agents will be sanctioned, along with venal agents who actually committed the actions worthy of punishment. If the principal is unconcerned with incorrect judgments (a doubtful assumption), then the net effect could be positive, leading to an equilibrium like the type in Proposition 1.6. On the other hand, if not carefully calibrated, the punishment could yield the default equilibrium, which would be worse than many of the equilibria in which influence occurs. Unless the form of influence is in monetary bribes, calibrating punishment to effective level of the transfer is very difficult.

The difficulty can possibly be mitigated, but not necessarily adequately resolved, by investigating further into whether undue influence has occurred, given that there has been a media report, even if investigation is costless (given a media report) and principal is perfect at distinguishing between upright and venal agents after the investigation. If upright

agents incur a cost merely due to the fact of an investigation, such as time lost to other administrative pursuits, pain and suffering, these costs are not likely to be compensated. The venal agent likely incurs investigational costs, as well. The levels of these costs, unlike the level of punishment, are definitely not under the control of the principal. If the investigational costs are sufficient or nearly sufficient to deter both types of agents from proposing below the media threshold, even carefully calibrated punishments will not deter influence in a way that increases the principal's payoff. Overall, if the effectiveness of punishment and investigation depend on greater transparency, then having the options of increasing punishments and investigating suspected wrongdoing does not eliminate the essential ambiguities involved in increasing transparency.

1.4.3 Less Principal Power and Judicial Review

The baseline model assumes that the principal has the full power to determine the final policy, regardless of what the agent proposes. In reality, the principal's power to adjust policy may depend on media reports. Specifically, media reports might be necessary to alert elected officials to the salience of a particular rulemaking process, given the large number of issues a principal must deal with. Suppose that the principal can only change the policy when an agent or agency receives negative publicity, and that, without a media report, the agency's proposal becomes the final policy. Then the level of the media threshold, which is immaterial in the standard model (unless equilibrium selection follows proposal content), becomes important. If the principal can only change the policy after a media report, the agent can guarantee a particular policy by selecting a policy that of at least $\tilde{r}$.

Greater transparency is most helpful if the threshold is above zero and the agent may freely choose any nonnegative level of regulation. Then the group will consider inducing

$r_A = 0$ from the venal agent, in which case increases in p_1^H would generally increase the principal's payoff, since even the venal agent with the low-cost signal should be selecting $\hat{r}(\lambda_{V_H})$ instead of 0. Whether the agent is freely able to select any policy in an environment in which increased transparency would lead to more frequent media reports is an open question, but two of the cases in which the agency's proposal would be effectively zero do not fit the model very neatly. First, it may be that $r_A = 0$ for a venal agent that does not contemplate initiating a rulemaking in a new area. In this case, though, the agent has not solicited any information specific to the issue upon which a media report might be based. With many possible regulations given the information that an agency has, it would be difficult to make a case that an agent or agency has been captured by not pursuing a particular avenue of regulation. Second, $r_A = 0$ for a rulemaking that has started, but in which a captured agency is engaged in delay. However, delay, by definition, means that the media will not report for a while on the lack of progress, which again limits the benefits to the principal of greater transparency until later in the rulemaking process. Furthermore, with a variety of other institutions that can cause delay in regulations (Yackee and Yackee 2010), media reporting would be more difficult and might have other targets. In terms of the model, these cases would represent situations in which the media threshold is zero, and transparency would make no difference.

The scenario in which greater transparency is most likely to be relevant and in which the agency's proposal might be effectively zero is when the agency has completed the rulemaking and has decided not to issue a rule. However, because the agency has compiled a record in these cases, judicial review is available to compel agency action just as in cases in which the regulation is alleged to be too lax (Lubbers 2006, 541–56). Judicial review operates independently of media reports. Under judicial review of agency rulemakings, agencies are

allowed a good deal of discretion, but they can be overruled for "arbitrary and capricious" rulemaking. Here, the media threshold would be above zero. Applying judicial review to extend the baseline model, a weak form of judicial review would imply that the agency would have to set the level of regulation to at least $\hat{r}(0)$, since no matter how high the costs are, the agency could not justify a lower policy than this. A stronger form would require the agency to set the level of regulation to at least $\hat{r}\left(\lambda_{A_H}\right)$, since, no matter what the evidence states, the agency basing its decision on the signal could not justify any lower level of regulation.

Suppose the media threshold is greater than the minimum allowable policy for the respective forms of judicial review. In the weaker version, increases in p_1^L would increase the principal's payoff as before, but increasing p_1^H could may be better or worse for the principal because (1) the policy after V_L^L, $\hat{r}\left(\bar{\lambda}_{V_L}\right)$, would be lower than before, and (2) for the venal agent with the high-cost signal $\hat{r}\left(\bar{\lambda}_{V_L}\right)$ might be farther from $\hat{r}\left(\lambda_{V_H}\right)$ than $\hat{r}(0)$. In the second form, an increase in p_1^H would result in a lower payoff for the principal because more of the agents with the high-cost signal would be moved away from $\hat{r}(\lambda_{A_H})$. In both cases, increases in p_1^H result in informational loss and the possibility of screening agents from each other. The result, in the weaker form of judicial review, might be an equilibrium like the one in Proposition 1.6. At the other extreme, if all agents are deterred from proposing below the threshold, the principal will end up with less than the default equilibrium unless $\tilde{r} = \hat{r}\left(\lambda_{A_H \cup V_L^L}\right)$.[17]

Overall, these three extensions add realism to the policymaking setting. However, they do not remove the ambiguities involved in increasing transparency in the most likely scenarios.

[17] If $\tilde{r} < \hat{r}\left(\lambda_{A_H \cup V_L^L}\right)$, then U_H selects $\hat{r}\left(\lambda_{A_H \cup V_L^L}\right)$ while the group induces $\tilde{r}$ from the venal agent, which is further away from $\hat{r}(p_L)$ than $\hat{r}\left(\lambda_{A_H \cup V_L^L}\right)$ in the default equilibrium. If $\tilde{r} > \hat{r}\left(\lambda_{A_H \cup V_L^L}\right)$, then all agents except U_L always select $\tilde{r}$, yielding $f\left(\tilde{r}, \lambda_{A_H \cup V_L^L}\right) < \hat{f}\left(\lambda_{A_H \cup V_L^L}\right)$ from those agents.

1.5 Policy Implications

The discussion of the impacts of greater transparency is motivated significantly by recent efforts in both the U.S. and the European Union to increase transparency generally. The model presented in this paper points to some suggestions for how to think about and design transparency policies. This is true even when there is no mechanical cost to making documents available and when the political principal has full control over the final policy.

1.5.1 Accounting for Policy Losses

The political principal in the model is perfectly rational and uses Bayesian updating in determining the level of regulation. However, there remains plenty of room for executive agencies or agents and an interest group to incur costs due to public opprobrium. If an agent cannot directly communicate the substance of his information via the documents that are released, there is a risk that they will be misinterpreted. Thus, the potential gains from transparency will be limited to the extent that the media makes errors in interpreting which policies should follow from the documents that it would have access to. If, between a lower level of transparency and a higher level of transparency, the increase in false positives is sufficiently high compared to increase in cases in which a lenient policy is correctly seen as not supported by the evidence, increasing transparency can result in worse policy outcomes for the principal due to the loss of information at the media reporting stage. Therefore, scholars and practitioners weighing the benefits and costs of transparency need to consider the possibility that released documents will be misinterpreted among the costs.

Another distinct possibility from increased transparency is that agents and agencies with documents that actually support a lower level of regulation may propose higher levels of

regulation to avoid media scrutiny. In the baseline case, in which the principal can readjust policy from whatever is proposed, policy losses result from the fact that proposals are less distinct from one another. In the extension in which the agency's proposal can be binding, these proposals can result in higher levels of regulation than the principal desires. Undesirable policy distortions might also take the form of interest groups not providing information to agencies. Thus, an increase in transparency can have two chilling effects: one on agency's willingness to propose low levels of regulation and one on groups' willingness to provide information in the first place. These also need to be considered when weighing proposals about increasing transparency. Even if the possibility for losses of information and policy distortions will not appear in politicians' official rhetoric, they should at least appear in more private discussions about transparency policy.

1.5.2 General Transparency Policies

The policy implications for a general transparency policy, one that broadly encourages greater transparency in every agency, depend on how much more likely media reports will become and how costly those reports will be for agents and the group. If the probability or costs of media reports with greater transparency are thought to be relatively low, then if false positives do not increase much compared to true reports that more stringent regulation is justified given the evidence, then increasing transparency is most likely to be beneficial. However, if there are enough false positives and agents are strongly averse to negative publicity, then some intermediate level of transparency is more likely to be warranted. Similarly, if it is thought, that, in general, that, increasing transparency from its current level would produce a high proportion of additional false positives, the intermediate transparency may also be advisable.

1.5.3 Tailored Transparency Policies

In general, however, different levels of transparency may produce better policy outcomes from some agencies and worse in others, making it difficult to optimize transparency across agencies. In this case, the model suggests an approach tailored to different agencies. Rather than assign one level of transparency, increasing transparency as much as possible across all agencies, a more tailored approach would increase transparency only to the extent that it is beneficial in each agency. Following the logic above, agencies for which greater transparency leads to mostly accurate media reports in the event of lax regulation not supported by the evidence should be pressed to release more documents, whereas a more moderate level of transparency would be better for agencies for which greater transparency would produce a large number of false positives or deter agents who have evidence that low regulation is proper from proposing such regulation.

Predicting what would happen to an agency requires empirical analysis. One dimension, how large costs from media reporting are should be measurable, based on surveys of agency perceptions of media reports. It should also be quite possible to determine how often agencies appear in the news. The most difficult challenge would be determining how often the reports are accurate. While the media will change its story, most likely the agent who was the subject of a report and the media will disagree about who is correct. Here, a researcher would need some independent criterion for discerning whether a report is a true or false positive so as to determine the amount of information gained from the media reporting stage.[18] In contrast to reporting accuracy, it should be easier to determine the extent to which agents might be deterred from proposing their preferred policies by media reports based on insider accounts.

[18]A researcher might also make errors in determining whether the media or the agent was correct in a particular case. This possibility provides further support for the idea that the media might report incorrectly.

If one supposes that most gains in predicting the impacts of transparency accurately arise from the first efforts in research, then, while perfect calibration may be impossible, it should be possible to be more discerning among agencies. Since the government already exempts some information from disclosure for different agencies and counts this exemption as one of the reasons not to release a document under FOIA, the idea of at least partial tailoring of transparency to different agencies may not be so foreign.

1.5.4 Agency Resistance to Transparency

As noted above, agencies have been observed to respond to transparency initiatives by resisting information disclosure requirements rather than by changing its proposals (Roberts 2006). The model has implications for this kind of resistance, as well. First, the possibility of incorrect reporting by the media means that agencies are not resisting transparency merely because they have "something to hide." To be more precise, an agent whose evidence properly supports lower regulation has documents to hide that perhaps should remain hidden if they are likely to be misinterpreted. Second, because greater transparency can lead to worse results for a political principal, the solution to resistance to transparency is not necessarily to redouble efforts to increase transparency, even if these efforts are costless. Even if a political principal is successful in forcing agencies to disclose more documents, and her payoff may not necessarily be better. Finally, the possibility that different agencies will resist to varying degrees means that political principal may be able to effectuate varying levels of transparency by devoting different amounts of effort to overcoming this resistance.

1.6 Conclusion

In contrast to statements by leaders, transparency non-governmental organizations, and some scholars about the values of transparency, the model presented in this paper suggests that the potential benefits to transparency need to be considered with more nuance. Greater transparency can result in loss of information about the meaning of an agent's evidence as well as undesired changes in policy proposals. Since the U.S. and many other industrialized societies already have substantial amounts of transparency in the disclosure of documents, it is not obvious that more transparency would improve policy outcomes (cf. Coglianese 2009, 538). Contrary to President Obama's assertion that "the Government should not keep information confidential merely because public officials might be embarrassed by disclosure, because errors and failures might be revealed, or because of speculative or abstract fears" (2009, 4683), fears arising from adverse media reports complicate the use of transparency and suggest that it should applied in a tailored, rather than general fashion.

Chapter 2

Transparency and Power in Rulemaking

Contents

> "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).

2.1 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.

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

[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.

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 2.2 sets up a formal model designed to match the rulemaking process, and Section 2.3 describes the results that follow from it. Section 2.4 the importance of power in understanding transparency beyond the confines of the model. Section 2.5 discusses policy implications for the model, and Section 2.6 concludes.

2.2 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.

2.2.1 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.

2.2.2 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 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

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

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.

2.2.3 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).

2.2.4 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 2.1. *For $q \in \{P, A\}$, the following inequality holds:*

$$EU_j^q(x_j^q) < \tau_i EU_i^q(x_i^q) + \tau_k EU_k^q(x_k^q) < \tau_h EU_h^q(x_h^q)) + \tau_l EU_l^q(x_l^q). \tag{2.1}$$

Proof. Proofs of all numbered results for this chapter except Corollary 2.3 are in Appendix B.2. ■

For the principal, the lowest utility in Inequality 2.1 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.

2.3 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.

2.3.1 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.

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

Theorem 2.2. *In any natural message-signal equilibrium,* $x^{P*} = x_i^P$ *and* $x^{A*} = x_i^A$ *following* $s = \tilde{H}$*, 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 2.2 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 2.2:

Corollary 2.3. *(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.

2.3.2 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.

[4] 5 U.S.C. §§552(b)(4)-(5) (2006).

Proposition 2.4. *When disclosure of the message and signal is optional for the agent, there 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 2.4 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 2.4 yields the principal

$$\tau_i(\pi EU_i^P(x_i^P) + (1-\pi)EU_i^P(x_i^A)) + \tau_k(\pi EU_k^P(x_k^P) + (1-\pi)EU_k^P(x_k^A)). \tag{2.2}$$

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 2.5. *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:*

$$\begin{aligned}(\pi + \Delta\pi_m)EU_j^A(x_j^P) + (1-\pi-\Delta\pi_m)EU_j^A(x_j^A) \\ \leq \tau_i((\pi+\Delta\pi_m)EU_i^A(x_i^P) + (1-\pi-\Delta\pi_m)EU_i^A(x_i^A)) \\ + \tau_k(\pi+\Delta\pi_m+\Delta\pi_{\tilde{L}})EU_k^A(x_k^P) + (1-\pi-\Delta\pi_m-\Delta\pi_{\tilde{L}})EU_k^A(x_k^A)),\end{aligned} \tag{2.3}$$

and $\pi c(x_l^P) + (1-\pi)c(x_l^A) \geq (1-\alpha)((\pi + \Delta\pi_m)c(x_i^P) + (1-\pi-\Delta\pi_m)c(x_i^A))$

$$+ \alpha(\pi + \Delta\pi_m + \Delta\pi_{\tilde{L}})c(x_k^P) + (1-\pi-\Delta\pi_m-\Delta\pi_{\tilde{L}})c(x_k^A)). \quad (2.4)$$

If this equilibrium exists, then the principal's payoff is the highest among message-signal equilibria in any transparency mode. If only Inequality (2.3) *fails, then no natural equilibrium in which the agent generates a signal can be sustained. If only Inequality* (2.4) *fails, the low-cost type will choose not to send a message with some positive probability in any natural equilibrium.*

Proposition 2.5 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 (2.2) 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 (2.3) 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 (2.4) is the low-cost target's individual rationality constraint for 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. However, gains compared

[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.

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 2.5 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 2.6. *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*

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

and low-cost target:

$$\pi EU_j^A(x_j^P) + (1-\pi)EU_j^A(x_j^A) \leq \tau_i(\pi EU_i^A(x_i^P) + (1-\pi)EU_i^A(x_i^A)) + \tau_k(\pi + \Delta\pi_{\tilde{L}})EU_k^A(x_k^P) + (1-\pi-\Delta\pi_{\tilde{L}})EU_k^A(x_k^A)), \quad (2.5)$$

$$\textit{and } \pi c(x_l^P) + (1-\pi)c(x_l^A) \geq (1-\alpha)(\pi c(x_i^P) + (1-\pi)c(x_i^A)) + \alpha(\pi + \Delta\pi_{\tilde{L}})c(x_k^P) + (1-\pi-\Delta\pi_{\tilde{L}})c(x_k^A)). \quad (2.6)$$

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 (2.5) *fails, then no natural equilibrium in which the agent generates a signal can be sustained. If only Inequality* (2.6) *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 (2.6), 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 2.5.

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 2.7. *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:*

$$\pi c(x_l^P) + (1-\pi)c(x_l^A) \geq (1-\alpha)((\pi + \Delta\pi_m)c(x_i^P) + (1-\pi-\Delta\pi_m)c(x_i^A)) \\ + \alpha(\pi + \Delta\pi_m)c(x_k^P) + (1-\pi-\Delta\pi_m)c(x_k^A)). \tag{2.7}$$

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 (2.7) *fails, the low-cost type will choose not to send a message with some positive probability in any natural equilibrium.*

Inequality (2.7) 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 Reguatory Comm'n* (1992) embodies this logic; 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 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

2.3.3 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 2.8. *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* (2.3) *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 2.4 and 2.8 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.

2.4 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.

2.4.1 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.

2.4.2 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.

2.4.3 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.

2.5 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.

2.5.1 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 2.4, 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.

2.5.2 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."

2.6 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.

Chapter 3

A Reverse Rationale for Reliance on Regulators

Contents

3.1 Introduction

Scholarship dating at least from Weber's (1922/1978) and Landis' (1938) seminal works on bureaucracy has identified expertise as a fundamental rationale for agencies in the modern administrative state. A recent review of models in this area observes that "[t]he bureaucratic agent in these models typically possesses (or may come to possess) some information that the leader would like to extract to make a decision" (Gailmard and Patty 2012, 354). That elected officials rely on agencies because the latter have better access to information than the former is a foundational principle of bureaucracy studies that does not appear to have been seriously challenged. Instead, because bureaucrats are unelected, a large body of literature focuses on how to address the risk that an agency might use its informational advantage to pursue policies that differ from what political leaders would prefer if they had the same information as the agency (e.g., Epstein and O'Halloran 1999, Huber and Shipan 2002).

More recent work has problematized bureaucratic agents' ability and willingness to secure the specialized information that justifies their role. First, if information is costly to acquire, a political leader may have to incentivize an agent to gather higher quality information, and this task may conflict with her desire for agency decision-making that conforms to her policy preferences (e.g., Gailmard and Patty 2013a). Second, agencies may need to gather information from outside parties like firms to effectively implement policy and may have

difficulties doing so (e.g., McCarty 2013). However, works focusing on these adjustments to the theory of bureaucratic expertise seem implicitly to affirm that agencies serve the purpose of gather information that leaders cannot themselves obtain.

This paper considers a reverse rationale for reliance on agencies, that they may benefit a leader by preventing her from receiving or acting on information that she is perfectly capable of obtaining. One half of this logic, that a leader is capable of gaining information, requires only that she can understand information that others have produced. For example, even if she cannot conduct scientific experiments, she may be able to assess the quality of these experiments and interpret the results, either directly or through trusted staff. Steps to obtain passive expertise in practice have occurred in both Congress and the White House: Congress engages in oversight through the Government Accountability Office and the Congressional Research Service (Beermann 2006, 127–30), which may allow its members to rely on in-house expertise. Meanwhile, during the George W. Bush Administration, the Office of Information and Regulatory Affairs added scientific experts to its personnel (Graham, Noe, and Branch 2006).

The other half of this logic, that an agency would helpfully prevent the leader from receiving information, implies minimally that some outside party is generating the information. Much information for agency policymaking does actually comes from firms. In the area of industrial regulation, firms are likely to have private knowledge about their manufacturing processes and the costs of potential regulations (Coglianese, Zeckhauser, and Parson 2004). Other examples are tests of new chemicals under the Toxic Substances and Control Act and project proposals for government contracts. Thus, there exists a variety of cases in which regulated parties, rather than agencies, act as the primary researchers for decisions involving the executive branch.

Although a typical problem is outside parties' reluctance to provide their information (see, e.g., Coglianese, Zeckhauser, and Parson 2004, McCarty 2013), they may, in other circumstances, be too willing to provide information. This scenario arises when the leader and agent can observe their information. This situation can harm a leader if they produce lower quality information as a result. The area of drug approvals serves as a useful illustration: Congress generally allows Food and Drug Administration (FDA) to investigate pharmaceutical firms' clinical trials, rather than itself examining the results. This mode of operation may be desirable for Congress even if it is perfectly capable of reviewing the data and has no time constraints. The reason is that, if Congress were to directly review a firm's application, the company could create evidence just barely favorable enough to warrant approval. FDA might require stronger evidence, perhaps because, as an unelected bureaucracy, it is less swayed by constituencies interested in new treatments for diseases. Congress might be able to induce the drug manufacturer to produce higher quality information by forcing it to communicate only with FDA or by delegating the final decision to the agency.

Therefore, instead of securing information from outsiders that the leader is unable to access or process, an agent may have the opposite function of preventing leaders from acting on or receiving outsiders' information. Generalizing from the above scenario, this paper presents a model in which the leader can do better if the outside party cannot directly convey information to her and motivate her to act but must instead work through an agent. Her payoff can improve if the agent initially disfavors the third party's preferred policy and requires more accurate information than the leader does to favor that policy. More accurate information, in turn, allows the selection of policy that is more likely to be correct.

This rationale for agencies has several implications for understanding regulatory capture: First, institutional arrangements to have agencies make decisions or filter information may

mitigate this phenomenon rather than enhance it. Second, it relates to work on how capture of agencies may (not) be inferred from their behavior (e.g., Carpenter 2004). In particular, seemingly discomforting disclosure patterns, such as (1) release only of information that favors interest groups and (2) release of no information at all, are not necessarily signs of capture but instead may further public interest. Finally, to the extent that interest groups might attempt to influence agencies' policy preferences, the model suggests that mandatory disclosure is less promising than ethics rules for agents in combating capture. The difference in the desirability of different measures against capture is especially important since the Obama Administration has initiated policies relating both to transparency and ethics (Coglianese 2009, Thurber 2011).

The rest of this paper proceeds as follows: Section 3.2 presents a simple model of agency in which policy may be based on an interested outside party's research. Section 3.3 compares equilibrium outcomes in which the leader can receive the third party's information and select the policy, those in which the agent selects the policy, and those in which the leader retains decision-making authority but allows the third party to communicate only with the agent. Section 3.4 discusses the implications for capture from the baseline model and an extension. Section 3.5 suggests that the FDA drug approval process is a plausible example of the reverse rationale operating in practice. Section 3.6 concludes.

3.2 The Model

The game features a political leader (L, she), whose preferences, until Section 3.4, are identical to those of some "public," represented as a passive principal (P); an agent or agency (A, he), who proposes policy; and an outsider or researcher (R, it), which generates new

information about policy. The goal is to structure information channels and decision-making authority to benefit the public.

3.2.1 Basic Elements

There are two policies, $x \in X = \{0, 1\}$, from which to choose, and the leader may either make the final selection or delegate that decision to the agent. The state of the world, $w \in \{0, 1\}$, is unknown, but the probability of each state w is $q_w \in (0, 1)$. The outsider can conduct research to reduce the initial uncertainty by expending effort, $e \in E = \mathbb{R}_+$. If its effort is zero, it can be said to be doing no research. Lack of research may be understood as literally no effort or as an idealization of doing some minimum amount of work to produce a report, but not any real work that contributes to the players' understanding of the situation. Any effort, even zero, generates a signal, $s \in S = \{0, 1\}$. $\Pr(s = w|e) \equiv g(e)$ is an accuracy function, which is a continuous, increasing, and concave function, with $g(0) = 1/2$ and $\lim_{e\to\infty} g(e) = 1$. The cost of research effort, $c(e)$, is continuous, increasing, and convex, with $c(0) = c'(0) = 0$. Together, the accuracy and cost functions will be referred to as the outsider's research technology, which is the same as that in Prendergast (2007).

For each player $i \in \{L, P, A, R\}$, the benefits when $x \neq w$ are normalized to 0 in each state, so that net benefits to player i in state w when $x = w$ are denoted by b^i_w. For now, the leader has the same preferences as the principal, so whatever benefits the leader benefits the public, and the principal will not mentioned again until Section 3.4. The player's preferences are restricted as follows: First, the leader and agent strictly prefer $x = w$ in each state, i.e., $b^i_w > 0, i \in \{L, A\}, w \in \{0, 1\}$, which means that they agree on which policy is better in both states of the world. However, because the benefits for matching the state can differ for each state and each player, there remains scope for disagreement about what policy to pursue in

the face of uncertainty about the state. The outsider has $b_1^R > 0$, but b_0^R may be positive or negative, which means that it may prefer policy 1 in both states. Interest groups might be thought to have preferences of this sort. However, so that the researcher may end up exerting effort in equilibrium, its preferences are bounded so that $q_0 b_0^R + q_1 b_1^R > 0$. Additionally, to avoid borderline cases, it will be assumed that $q_0 b_0^i \neq q_1 b_1^i$ for each player i. Finally, without loss of generality, the researcher will have $q_1 b_1^R > q_0 b_0^R$.

Of the above elements, the realized effort level and signal begin as the third party's private information, but the other players may be able to learn them. The others are common knowledge, including the prior distribution on the states of the world, the players' preferences and the outsider's research technology. In the end, one of the players makes a final decision. If it is the agent, then $x^A \in X^A = \{0, 1\}$ represents his decision. Otherwise, x^A represents his cheap-talk proposal to the leader, followed by her policy selection, $x^L \in \{0, 1\}$.

3.2.2 Communication, Authority, and Game Stages

The main focus of information transmission will be on the outsider's effort level and signal. These items are observable, which means that they can be truthfully conveyed or withheld, but not faked. However, these items of information will not be verifiable in the sense that a court could review effort, so contracts cannot be based on them. In making information observable, the model follows Ting (2008) and Gailmard and Patty (2013b). In particular, the model will assume that the leader and agent are equally capable of comprehending any information they receive, although the discussion will also consider the implications of a less capable leader.

Although communications between any two players will generally be unregulated, the rules for decision-making reduce the scope of communications that need to be analyzed. The

game will take one of three forms: First, the leader can place the decision under her direct *administration*, so that she can communicate with the outsider and select the final policy, just as Congress can seek input from various interest groups in the legislative process. Second, she can commit to *delegation*, so that agent instead of the leader makes the final policy choice. Finally, she can engage in *oversight*, in which she retains decision-making authority but cannot communicate with the outsider. Instead, she can only receive the researcher's information if it transmits it to the agent and he relays it to her. The second two modes of the game apply to scenarios in which the leader lacks the time or ability to actively monitor the player exerting effort (see Tirole 1986, Aghion and Tirole 1997). With a large volume of regulatory policy under consideration each year, it is quite plausible that the leader will end up not directly observing many activities of outsiders (McCubbins, Noll, and Weingast 1987).

The three important directions for communicating information are from the researcher to the agent, from the researcher to the leader, and from the agent to the leader. In particular, since the outsider will always prefer to provide both items of information to the decision-maker, communications to the researcher are nugatory. Also, because researcher can communicate freely with the agent, information transmission from the leader to the agent can be ignored. Then, using ϵ_j^i and σ_j^i to denote player i's decision of whether to transmit to player j, respectively, the outsider's effort level and signal, when player i has the information and is allowed to relay the information, the decisions that can appear in the game are ϵ_A^R, σ_A^R, ϵ_L^R, σ_L^R, ϵ_L^A, and σ_L^A. For each of these variables, δ (ν) can be used to represent (non)disclosure.

In addition, the timing of disclosures can be substantially limited without loss of generality. When the leader has authority, the researcher's disclosures to the agent are effective

only before his proposal. Meanwhile, disclosure of the researcher's information by the agent can only affect the outcome before the leader's policy decision. Thus, agent can make his disclosures simultaneously with his proposal. Furthermore, disclosures by the outsider to the leader, if allowed, will take place at the same time as the agent's proposal and disclosure decisions for notational convenience, since the agent's strategy will not affect that part of its communication strategy and vice versa. Similar timing restrictions can be applied when the agent has authority, except that the agent's proposal and leader's decision are substituted with the agent's decision.

With these restrictions on the nature and timing of information disclosure, the stages of the game can be succinctly stated as follows:

1) Nature chooses the state of the world $w \in \{0, 1\}$.

2) The researcher chooses the level of research effort e and receives a random signal s about the state of the world, whose accuracy increases with e.

3) The researcher decides whether to convey each of e and s to the agent.

4) The agent makes or proposes policy, $x^A \in \{0, 1\}$, and decides whether to relay each item of information he received from the researcher to the leader. Except under oversight, the researcher decides which of e and s to disclose to the leader.

5) Under administration or oversight, the leader makes the final policy decision, $x^L \in \{0, 1\}$.

3.2.3 Strategies and Beliefs

To notate pure strategies, it helps to distinguish between intended transmission and actual reception of information. The variables ϵ_j^i and σ_j^i defined above indicate whether player

i would transmit information to player j given the opportunity. Reception of an item of information will be represented either by the true value in the case of transmission or by $\emptyset$ in the case of no transmission. Then, for the agent and leader, the sets $\mathring{E}_j \equiv \emptyset \cup \mathbb{R}_+$ and $\mathring{S}_j \equiv \emptyset \cup \{0,1\}$ will represent the possibilities for what player j has learned about e and s, respectively, and $\mathring{e}_j$ and $\mathring{s}_j$, will respectively represent elements of these sets. Although the leader may receive information from either other player, the identity of the transmitter will turn out to be irrelevant, so additional notation for the sender can be omitted.

Now the parties' strategies can be expressed as follows: The researcher selects effort level e. After receiving the signal s, it decides what to communicate to the agent, and, when possible, to the leader. Its strategy is unaffected by the agent's, so its communications can be represented as ordered pairs $(\epsilon_i^R, \sigma_i^R) : E \times S \to \{\delta, \nu\}^2$, $i \in \{L, A\}$. Overall, the researcher's strategy can be denoted by $\Sigma^R \equiv (e; (\epsilon_A^R, \sigma_A^R); (\epsilon_L^R, \sigma_L^R))$. The agent's strategy consists of her proposal or decision and intentions to disclose and can be written as $\Sigma^A \equiv (x^A; (\epsilon_L^A, \sigma_L^A)) : \mathring{E}_A \times \mathring{S}_A \to \{0,1\} \times \{\delta, \nu\}^2$. Finally, the leader's strategy is just her policy choice when she has one, $\Sigma^L \equiv x^L : X^A \times \mathring{E}_L \times \mathring{S}_L \to \{0,1\}$. Notation for pure strategies is sufficient since mixed strategies do not play an important role except in borderline cases.

The fundamental set of beliefs center on the state of the world, which the players update as they receive information. The outsider's beliefs derive from its effort and the signal, which it always observes, so its posterior probabilities will not be notated. For the other active players $i \in \{L, A\}$, let β_1^i map the information s/he receives to a posterior probability (belief) that $w = 1$: i.e., $\beta_1^A : \mathring{E}_A \times \mathring{S}_A \to [0,1]$, and $\beta_1^L : X^A \times \mathring{E}_L \times \mathring{S}_L \to [0,1]$. Since full strategy-belief profiles are very extensive, only the most essential parts will be highlighted in the main text, and the propositions will describe only equilibrium path strategies.

3.2.4 Additional Terminology

In categorizing various scenarios, it will be useful to introduce some more terms. First, the limitations on the players' preferences described above imply that each strictly prefers one of the policies initially. It can be said that the player has a *bias* toward that policy, denoted by $\tilde{x}^i \equiv \arg\max_w q_w b_w^i$, and against the other policy, $1 - \tilde{x}^i$. Thus, the outsider is always biased toward policy 1, which is appropriate since interest groups are likely to consistently favor one side of an issue. For example, pharmaceutical companies generally want approval for their drugs, at least in the absence of contradictory information. In relation to the outsider, each of the other active players will be termed *advocative* if s/he also has a bias toward policy 1 and *adversarial* if s/he is biased toward policy 0. Players can differ not only in terms of the policy toward which they are biased, but also in the strength of their biases. The degree of a player's bias toward policy x can be measured as $B_x^i \equiv 2q_x^i b_x^i/(q_0 b_0^i + q_1 b_1^i) - 1$, so that the quantity is negative if the player is biased against that policy.

Intuitively, a player will always want policy to follow a signal that follows that player's bias, while that player will want policy to match a contradictory signal only if it is supported by enough effort. A formal condition can be stated:

Lemma 3.1. *After research, a player prefers to have* $1 - \tilde{x}^i$ *enacted if and only if* $s = 1 - \tilde{x}^i$ *and* $e \geq g^{-1}(q_{\tilde{x}^i}^i b_{\tilde{x}^i}^i/(q_0 b_0^i + q_1 b_1^i)) \equiv e^i$. *Otherwise, that player strictly prefers* $\tilde{x}^i$.

Proof. Proofs of all numbered results except Corollaries 3.4 and 3.13 are in Appendix B.3. ■

The quantity e^i can be called a player's *standard of proof*, which increases with the bias toward the policy toward the policy that player initially prefers. This quantity is the minimum effort level at which the leader and agent prefer to have the policy follow the signal rather than to always match his or her presumptive preference. Following Bayes' rule, the agent or

leader's expected payoff when the policy follows the signal is

$$EU_f^i(e) \equiv g(e)(q_0 b_0^i + q_1 b_1^i), i \in \{L, A\}, \tag{3.1}$$

which is increasing in e. Then the standard of proof e^i is the minimum level of effort that satisfies $EU_f^i(e) \geq q_{\tilde{x}^i} b_{\tilde{x}^i}^i$.

A few more terms for the researcher are worth defining. First is its *signal-constrained optimum*, the amount that it would devote to its research knowing that the signal would be followed. Bayes' Rule implies that this expected payoff is

$$EU_f^R(e) \equiv g(e)(q_0 b_0^R + q_1 b_1^R) - c(e). \tag{3.2}$$

Then the signal-constrained optimum, denoted by $\hat{e}$, satisfies the first-order condition

$$g'(\hat{e})(q_0 b_0^R + q_1 b_1^R) = c'(\hat{e}). \tag{3.3}$$

Related to this effort level is how the payoff in Equation (3.2) compares to $q_1 b_1^R$, its payoff if it selects policy 1, toward which it is biased, with no research effort. It is *(un)motivated* (i.e., to do research) if $EU_f^R(\hat{e}) > (<) q_1 b_1^R$.[1] Finally, the most effort that the researcher is willing to expend and have the signal be followed, rather than have policy x always be chosen, is its *discouragement point* for that policy. This point is defined as $\bar{e}_x \equiv \max\{e : EU_f^R(e) \geq q_x b_x^R\}$, with $\bar{e}_1$ existing only for a motivated researcher. The functional form assumptions on $g(\cdot)$ and $c(\cdot)$ imply that $\bar{e}_0 > \hat{e}$, and, when $\bar{e}_1$ exists, that $\bar{e}_1 \in (\hat{e}, \bar{e}_0)$.

[1]The borderline case of $EU_f^R(\hat{e}) = q_1 b_1^R$ does not add any insight and is omitted.

3.3 Model Results

The solution concept for this game is perfect Bayesian equilibrium (PBE) in pure strategies. Understanding why the leader would want to delegate to an agent or cut off her communications with the outsider requires an analysis of the equilibria under each of the three game forms and comparison of her payoff given various parameters for the three players. Since the researcher generates the information, a useful benchmark is how the leader would fare if it had authority to set the policy. Depending on its preferences and research technology, it would either maximize its payoff under Equation (3.2) or summarily choose policy 1:

Proposition 3.2. *If allowed to select the policy, an unmotivated researcher would set* $e = 0$ *and* $x = 1$*, while a motivated researcher would set* $e = \hat{e}$ *and* $x = s$*.*

This result, can be used to represents total lack of regulation. In the drug approval setting, it implies that a firm might market a drug without doing any research on it. Though it may be difficult to imagine a setting in which people would dare to sell drugs without doing any research, Congress estimated in 1906, the year when it first legislated federal controls on drugs, that there were 50,000 so-called "patent medicines" in the drug industry (Carpenter 2010, 77-78). This proposition indicates that there is much scope for improvement. For example, an adversarial leader facing an unmotivated researcher would do better at least by always selecting policy 0.

3.3.1 Equilibria under Administration

In the administration game form, the leader has final policymaking authority and can scrutinize whatever information that the researcher offers. It turns out that, in equilibrium, researcher has no problem disclosing its information to the leader. The next proposition

states the equilibria under administration in general terms:

Proposition 3.3. *Under administration the unique PBE with respect to effort and policy choice is* $e^* = \max\{\hat{e}, e^L\}$ *and* $x^{L*} = s$ *when* $e^L < \bar{e}_{\tilde{x}^L}$ *and* $e^* = 0$ *and* $x^{L*} = \tilde{x}^L$ *when* $e^L \not\leq \bar{e}_{\tilde{x}^L}$*, except that when* $e^L = \bar{e}_{\tilde{x}^L}$*, both equilibria can obtain. The researcher can induce the first equilibrium with* $\epsilon_L^R = \sigma_L^R = \delta$ *when* $s = 1 - \tilde{x}^L$ *and the second with* $\epsilon_L^R = \delta$*.*

Remark. The condition that $e^L \not\leq \bar{e}_{\tilde{x}^L}$ can result from $e^L > \bar{e}_{\tilde{x}^L}$ or an unmotivated outsider's lack of a value of $\bar{e}_{\tilde{x}^L}$ when $\tilde{x}^L = 1$. Also, though it may communicate with the agent, that player is unnecessary because it can directly convey information to the leader.

The specific equilibrium results depend on the leader's bias. For an advocative leader, whether the researcher is motivated also matters. With an unmotivated researcher, effort is zero and policy 1 always obtains, the same result as if the researcher were acting on its own. With a motivated researcher, how much effort she induces depends on her standard of proof. When $e^L \leq \hat{e}$, her standard of proof is not binding, and the outsider expends effort at its signal-constrained optimum for policy matching the signal, as if it had decision-making authority. When $e^L \in (\hat{e}, \bar{e}_1]$, her standard of proof is binding and induces the outsider to exert additional effort to meet the standard of proof, after which policy matches the signal. Finally, when $e^L > \bar{e}_1$, the researcher is unwilling to incur the cost needed to meet the standard of proof and instead sets effort at zero for policy 1. When the leader is adversarial, then only her standard of proof is relevant. The results are the same as those for an advocative leader facing a motivated agent, except that the upper bound for effort that she can extract is the discouragement point for policy 0.

Compared to allowing the researcher to decide policy, taking control of decision-making only helps the leader, primarily because she can always summarily select the policy toward

which she is biased to earn a reservation payoff of $q_{\tilde{x}^L} b^L_{\tilde{x}^L}$. Both adversarial and advocative leaders can induce motivated researchers to expend additional research to meet their standard of proof. Adversarial leaders particularly benefit because they can also stimulate unmotivated researchers to exert effort in the first place, and because they can incentivize motivated researchers to increase their effort up to a higher limit, since $\bar{e}_0 > \bar{e}_1$.

However, given that the leader has authority to choose the policy, waiting for the outsider's research often does not benefit her compared to summarily selecting the policy for which she has a bias as if the researcher were not present. Whenever her standard of proof is binding and does not discourage effort altogether, her payoff is the same as if she committed to selecting $\tilde{x}^L$ from the beginning. Meanwhile, when a motivated researcher's effort is its signal-constrained optimum, the result is the same as if it were by itself. The only case in which the leader exceeds the reservation payoff of $q_{\tilde{x}^L} b^L_{\tilde{x}^L}$ from research that the outsider would not have conducted by itself is when she is adversarial, it is unmotivated, and its signal-constrained optimum exceeds her standard of proof.

These difficulties result from the fact that, when the standard of proof binds, the researcher can expend just enough research to make her weakly prefer to select policy according to the signal, in which case it effectively denies the leader any surplus over her reservation value. The effort and signal are observable, and the outsider is willing to disclose both of these items, so these limits on the leader's payoff do not result from any informational advantage that the researcher retains. Because she can receive both items of information, she cannot commit to summarily select the policy toward which she is biased for any effort level above her standard of proof. Thus, she would prefer to prevent low-quality information from reaching her, but she cannot when she has decision-making authority and the "freedom" to communicate with the third party.

3.3.2 Equilibria under Delegation

One potential solution to the problem that the leader faces in having to receive and respond to information from the outsider is to prevent herself from doing the latter by irrevocably granting an agent the authority to set policy. Under delegation the agent assumes the leader's role, receiving information and making a decision identically. Substituting the agent for the leader in Proposition 3.3 yields an analogous equilibrium result:

Corollary 3.4. *Under administration the unique PBE with respect to effort and policy choice is* $e^* = \max\{\hat{e}, e^A\}$ *and* $x^{A*} = s$ *when* $e^A < \bar{e}_{\tilde{x}^A}$ *and* $e^* = 0$ *and* $x^{A*} = \tilde{x}^A$ *when* $e^A \not\leq \bar{e}_{\tilde{x}^A}$*, except that when* $e^A = \bar{e}_{\tilde{x}^A}$*, both equilibria can obtain. The researcher can induce the first equilibrium with* $\epsilon_A^R = \sigma_A^R = \delta$ *when* $s = 1 - \tilde{x}^A$ *and the second with* $\epsilon_A^R = \delta$*.*

Whether the leader's payoff is higher or lower delegating to an agent depends on their preferences and whether the agent is motivated. Ignoring borderline cases, one can formally state when delegation is better for her as follows:

Proposition 3.5. *The leader's payoff is higher from delegation than from administration: (a) when* $\max\{e^L, \hat{e}\} < e^A \leq \bar{e}_{\tilde{x}^A}$ *and (b) when* $e^A \leq \hat{e}$ *and* $e^L < \hat{e}$*, with* $\tilde{x}^A = 0$*,* $\tilde{x}^L = 1$*, and an unmotivated researcher.*

The generally necessary condition $\max\{e^L, \hat{e}\} < e^A < \bar{e}_{\tilde{x}^A}$ corresponds to two intuitive principles. First, the agent's standard of proof must not exceed the discouragement point corresponding to his bias. Otherwise, the outsider will do no research, and the agent will summarily select the policy for which he has a bias, which the leader cannot strictly prefer to making this kind of choice herself. Second, among the leader's and agent's standards of proof and the outsider's signal constrained optimum, the agent's standard of proof must be the highest. If the leader's standard of proof is the highest, then even if the agent induces

research, the effort level will less than needed to satisfy her, in which case she would better off summarily selecting the policy according to her bias. If the signal-constrained optimum is the highest, either the agent or the leader will, for the most part, induce effort at $\hat{e}$, making administration and delegation equally good for the leader. In contrast, when the agent's standard of proof is the highest and does not exceed the relevant discouragement point, he can induce extra effort in a way that satisfies the leader and increases her expected payoff in the form of policy that is more likely to be correct.

For Proposition 3.5(b), an unmotivated researcher would induce policy 1 from an advocative leader by exerting and disclosing zero effort, but it must meet an adversarial agent's standard of proof. The policy can at best (for the outsider) match the signal, so it maximizes at its signal-constrained optimum, which meets both players' standards of proof.

Overall, when committing authority to an agent with the right preferences benefits the leader, it does so usually by forcing the researcher to satisfy the agent's standard of proof that exceeds hers. This mechanism differs from the logic that agencies gather information that political leaders cannot (Bendor and Meirowitz 2004). Here, delegation prevents the leader from acting on information that she might receive from the researcher. As in other agency models, whether the leader can commit to delegate is a significant issue; however, unlike in canonical models, this difficulty does not persist past the agent's policy choice, since the leader would not want to reverse the agent's decision ex post (cf. Callander 2008).

In situations outside those in Proposition 3.5, delegation does not improve the leader's payoff and will often reduce it. If she can choose whether to delegate, she can avoid cases in which she would do worse than under administration. However, if an agent must be chosen for many decisions, then she may have to trade off cases in which she gains against those in which she loses. Surrendering authority to an agent is a rather blunt way of avoiding

the challenge of a researcher too willing to provide low-quality information. The remaining form of the game provides another way for the leader to increase her policy payoff over administration.

3.3.3 Equilibria under Oversight

Oversight is an intermediate form of control: she can still make the final decision, but she cannot obtain information directly from the researcher. Instead, it can only communicate with the agent, who then decides what, if anything, to convey to the leader along with his policy proposal. This game form helps distinguish the effect of withholding information from the effect of committing authority. Equilibria are no longer necessarily unique in this setting, even in terms of effort and policy choices. However, it is always possible to identify the PBE that yields the leader her highest payoff.

To begin with, whenever delegation yields the leader a higher payoff than administration, there exists a functionally equivalent oversight PBE in which the agent never discloses the researcher's information and the leader always ratifies the agent's proposal:

Proposition 3.6. *Under oversight, when* $e^A \leq \bar{e}_{\tilde{x}^A}$ *and* $e^L \leq \max\{\hat{e}, e^A\}$*, there exists a PBE that maximizes the leader's equilibrium payoff with* $e^* = \max\{\hat{e}, e^A\}$*,* $\epsilon_A^{R*} = \sigma_A^{R*} = \delta$ *when* $s = 1 - \tilde{x}^A$*,* $\epsilon_L^{A*} = \sigma_L^{A*} = \nu$*, and* $x^{A*} = x^{L*} = s$*.*

In an overlapping set but not identical of circumstances, there exists another PBE that maximizes the leader's payoff, one in which an adversarial agent proposes policy 1 when discloses the researcher's effort level and signal when the former meets his standard of proof and the latter points to policy 1, but discloses nothing and proposes policy 0 otherwise. As in the other equilibrium, the leader always accedes to the agent's proposal, although in fact

a proposal is not necessary, since the agent's intentions can be inferred from his disclosure choices.

Proposition 3.7. *Suppose the game form is one of oversight,* $\tilde{x}^A = 0$*, and* $e^A \leq \bar{e}_0$*. In addition, if* $\tilde{x}^L = 1$ *and* $e^L \leq \max\{e^A, \hat{e}\}$*, or if* $\tilde{x}^L = 0$ *and* $e^L \leq \bar{e}_0$*, there exists a PBE that maximizes the leader's equilibrium payoff with* $e^* = \max\{\hat{e}, e^A, e^L\}$*,* $x^{A*} = x^{L*} = s$*,* $\epsilon_A^{R*} = \sigma_A^{R*} = \epsilon_L^{A*} = \sigma_L^{A*} = \delta$ *when* $s = 1$*, and* $\epsilon_L^{A*} = \sigma_L^{A*} = \nu$ *when* $s = 0$*.*

Compared to delegation, oversight does not increase the leader's payoff above that under administration in many more cases. However, oversight also does not result in a lower utility for the leader than administration in many situations in which delegation would. Intuitively, the leader preserves her payoff under administration with the decision-making authority that she retains under oversight, often by summarily selecting the policy toward which she is biased. The only scenario in which oversight underperforms administration is when an adversarial agent discourages research in the former game form, whereas an adversarial leader in the latter induces research that yields her a surplus above her reservation payoff from always selecting policy 0, $q_0 b_0^L$. In these scenarios, however, delegation yields an equally low payoff.

Overall, when all parameter values are considered, one can find that oversight achieves all the benefits of delegation compared to administration with few of delegation's costs.

Theorem 3.8. *Assume that, under oversight, a PBE that maximizes the leader's equilibrium payoff obtains. Then the three game forms can be ranked in terms of her utility as follows:*

(a) Whenever delegation outperforms administration, oversight does so equally.

(b) Administration outperforms delegation in these cases: (i) $\tilde{x}^A = \tilde{x}^L$ *and* $\max\{\hat{e}, e^A\} < \min\{e^L, \bar{e}_{\tilde{x}^L}\}$*, (ii)* $\tilde{x}^A = \tilde{x}^L$*, the outsider is motivated, and* $e^L < \hat{e} < \bar{e}_{\tilde{x}^L} < e^A$*, and (iii)*

$\tilde{x}^A \neq \tilde{x}^L$ *and either* $e^A \not\leq \bar{e}_{\tilde{x}^A}$ *or* $\max\{e^A, \hat{e}\} < e^L$. *Oversight yields her as much as administration, but not more, except possibly when* $\max\{e^L, \hat{e}\} \leq e^A$ *and* $e^L \leq \bar{e}_{\tilde{x}^L}$ *in case (iii).*

(c) Administration equally outperforms delegation and oversight when $\tilde{x}^A = \tilde{x}^L = 0$, *the outsider is unmotivated, and* $e^L < \hat{e} < \bar{e}_0 < e^A$.

(d) Three game forms yield the same payoff in the remaining cases: (i) $\max\{e^A, e^L\} \leq \hat{e}$, *apart from when* $\tilde{x}^L = 0$ *and* $\tilde{x}^A = 1$ *with an unmotivated outsider, and (ii)* $\tilde{x}^L = \tilde{x}^A$, $\hat{e} < e^L$, *and* $e^A \not\leq \bar{e}_{\tilde{x}^A}$.

3.3.4 Optimal Choice of Game Form and Agent

Oversight and delegation each have the potential to benefit the leader under in various settings. For institutional design purposes, however, the most useful equilibria for her in these two modes are those that can apply regardless of whether the leader is adversarial and regardless of whether the researcher is motivated or not. The motivation for this criterion is that the leader cannot control her preferences or those of the outsider, but she may be able to influence the agent preferences that apply through the choice of agent (see Bertelli and Feldmann 2007). Then the most readily helpful equilibria are those involving an adversarial agent in Propositions 3.5–3.7.

In general, an adversarial agent with a greater standard of proof is better up to a point, since he induces additional effort. However, a standard of proof that is too high can discourage the outsider from research altogether. This intuition underlies the next result:

Proposition 3.9. *Suppose* $\bar{e}_0$ *is fixed and the leader can select among a set of adversarial agents with* $e^A \leq \bar{e}_0$. *Also, suppose she does not know* $\tilde{x}^L$ *or* e^L *when she selects the game form*

and agent but will know after the agent's proposal. If oversight is available, she maximizes her utility with that game form and the agent with the highest e^A. If not, she maximizes her utility with delegation or administration and the same agent.

Not counting the potential equilibria in Theorem 3.8(b)(iii), if the leader can select an agent with any preferences, her best agent is one who requires the maximum amount of effort that does not discourage research (cf. Bueno de Mesquita and Stephenson 2007, 614–15). When the leader cannot select an agent for each policy decision, calibrating his standard of proof to match the researcher's discouragement point for policy 0 is likely infeasible. Nonetheless, since the leader's payoff increases with effort when policy matches the signal, the usefulness of an agent who can enforce a higher standard of proof remains important. Since an adversarial agent with a higher standard of proof also has a greater bias toward policy 0 than the leader, his and the researcher's biases toward that policy will sometimes lie on opposite sides of hers.

It is worth noting that the "right" agent benefits the leader solely by virtue of his preferences, rather than because of his expertise. Other models that rely just on the agent's preferences have differing results about what kinds of agents are beneficial. First, Proposition 3.9 contrasts considerably with models in which the best agent preferences lie in between the leader's and the researcher's so that he can elicit more precise messages about its private information in a delegated cheap-talk setting (Dessein 2002, Gailmard and Patty 2013a). When the agent merely proposes rather than sets a policy, however, the leader benefits instead from a well-chosen agent with preferences relative to hers on the opposite side of the third party's (Ivanov 2010, Ambrus, Azevedo, and Kamada 2013). In these mediated cheap-talk models, the third party is willing to transmit more detailed messages because, in response to the agent's incentive to propose policies further away from the other two players'

preferred ones, the leader will select policies closer to her and its preferred ones. Here, in contrast, the result lies in inducing additional effort from the third party by making it fearful of policy that is, in expectation, more adverse to its interests.

Other models highlighting the benefits of a more adversarial agent can be found in Bertelli and Feldmann (2007), in which his extreme preferences offset those of an interest group in policy bargaining; and Rogoff (1985), in which a conservative central banker with extra concern about reducing inflation beneficially does so given wage-setters' attempts to anticipate his response to economic shocks. Of these models, Rogoff's model most closely approximates the reverse rationale of having an agent to prevent the leader from receiving or using the same information,[2] since a conservative central banker does better than an equally expert leader with preferences matching social welfare. The current game points to a larger set of policymaking settings in which the reverse rationale may apply.

Overall, the model suggests that incorporating an adversarial agent with a high standard of proof and giving him the exclusive authority to make policy or the sole ability to communicate with an outside group helps the leader avoid the problem of an outsider's providing information of just barely sufficient to satisfy her. The next section considers ways in which the researcher might try to frustrate this institutional arrangement.

3.4 Extension of the Model to Regulatory Capture

The assumption, maintained until now, that the leader has preferences identical to the "public," the true principal, is relaxed with the possibility of regulatory capture. Although capture has various definitions (see Levine and Forrence 1990, Dal Bó 2006), the term here

[2]In particular, cheap-talk models involve only messages from the third party about its private information, because by assumption, it is not able to credibly disclose the information that is generated.

can be understood as steps by the researcher to influence the leader or agent such that the principal's payoff decreases. By Proposition 3.2, a motivated researcher would like to research at its signal-constrained optimum and have policy follow the signal, while an unmotivated researcher would like to avoid expending any effort and induce summary selection of policy 1. Thus, it has a reason to attempt regulatory capture as well as a maximum degree to which it is willing to do so.

The outsider can attempt to influence either player. As Carpenter (2013) observes, one can distinguish statutory capture, which occurs apart from any agency action, from agency capture, in which an outside frustrates legislative intent through its influence on the agency. In this section, statutory capture will be represented by attempts to influence the leader, whereas agency capture will be modeled as steps to influence the agent. For each other active player, it has two techniques, τ, for capture: First, it can engage in *bias-shifting* (β), in which it causes a player's policy preferences, b_w^i, to change so that he or she has a different bias with respect to the two policies. Second, following Laffont and Tirole (1991), it can effectuate a *quasi-contract* (κ), in which a player is compensated for taking a different action than his or her policy preferences would dictate.

The outsider's cost for bias-shifting directed at player i can be denoted as a function $c_\beta^i(B_0^i - \check{B}_0^i)$, where B_0^i is that player's natural bias toward policy 0 and $\check{B}_0^i$ is that player's final bias when captured. Meanwhile, the cost of quasi-contractual compensation can be represented as $c_\kappa^i(V^i - \check{V}^i)$, where V^i is the player's policy payoff in an equilibrium without capture[3] and $\check{V}^i$ is that player's payoff from policy set according to the quasi-contract. It is convenient to further define $\Delta B_0^I = B_0^i - \check{B}_0^i$ and $\Delta V^i = V^i - \check{V}^i$. Since the goal is merely to understand how the different mechanisms operate for each player, rather than to define

[3]In the case of oversight, the relevant equilibrium is the one yielding the principal the highest payoff.

the researcher's optimal combination of capture strategies, it is sufficient to specify that, $\forall \tau \in \{\beta, \kappa\}, \forall i \in \{L, A\}$, c_τ^i is a strictly increasing function of its argument, to indicate roughly that more capture is more difficult for the researcher.

3.4.1 Attempts at Statutory Capture

For statutory capture, the two methods of influencing the leader have different effects because she can select delegation or oversight. First, if the leader is using oversight and delegation in a particular case and it will yield more than her administration payoff, bias-shifting requires a fixed cost for her to be willing to return to administration. Proposition 3.9 implies that small values of ΔB_0^i do not help the outsider:

Proposition 3.10. *Suppose $e^L \leq e^A \leq \bar{e}_0$ and $\tilde{x}^A = 0$ and the principal can select among game forms. Bias-shifting of the leader does not affect the research effort or policy selection as long as $e^L \leq e^A$ continues to hold.*

Thus, if the leader can rule out quasi-contracts, she can mitigate capture with a strongly adversarial agent, even though it is subject to bias-shifting that might come from political pressure. To benefit from bias-shifting, the researcher would need to make ΔB_0^L large enough for leader to prefer administration and summarily selection of policy 1.

For a quasi-contract, small amounts of compensation to the leader, (i.e., values of ΔV^L near zero) can cause her to select an agent with less bias toward policy 0. As compensation to the principal increases, the researcher can induce her to take actions that are correspondingly more favorable to it.

Proposition 3.11. *Suppose that $e^L < \bar{e}_0$ and the leader can choose the game form and an adversarial agent with any $e^A \bar{e}_0$. For quasi-contracts with the leader, a non-empty interval*

$[0, \Delta_1 V^L)$ *exists in which the researcher can only induce her to select an adversarial agent with a lower standard of proof. For some* $\Delta_2 V^L \geq \Delta_1 V^L$*, it can induce her to adopt her strategy under administration. If this policy outcome differs from what it would select acting alone, there exists* $\Delta_3 V^L > \Delta_2 V^L$ *such that the policy outcomes of Proposition 3.2 obtain.*

Therefore, a key contrast between statutory bias-shifting and quasi-contracts is that the former operates in an all-or-nothing fashion, while the latter can achieve more graduated results, starting from minimal costs. If the two methods are combined, then they may substitute for each other. For example, if an unmotivated researcher agrees with the leader in principle to have an agent with a lower standard of proof, then it needs a lower level of bias-shifting to induce the leader to always select policy 1.

3.4.2 Attempts at Agency Capture

Agency capture becomes relevant when the leader chooses delegation or oversight. Unlike for the leader, bias-shifting and quasi-contracts for the agent are essentially equivalent methods in the following sense:

Proposition 3.12. *Starting from any adversarial agent with* $e^A \in (\max\{\hat{e}, e^L\}, \bar{e}_0]$ *in oversight or delegation, the researcher can effect any standard of proof under capture,* $\check{e}^A < e^A$*, with* $\check{e}^A \in (\max\{\hat{e}, e^L\}, \bar{e}_0]$*, while remaining adversarial, with some level of bias-shifting* ΔB_0^A *or amount of compensation* ΔV^A*.*

The result that bias-shifting can have the same impact as a quasi-contract is consistent with the notion that so-called cultural capture, by which an interest group influences agency officials' preferences through human contact and can thereby sway regulation, as well as the argument that focusing on interest-based capture is incomplete (Kwak 2013). However, the

exchangeability of cultural capture and interest-based capture does not extend to statutory capture because the leader has delegation and oversight. She can escape bias-shifting through the use of another player, but, by assumption, the agent cannot.[4]

For completeness' sake, it is worth observing that the degree to which the outsider would want to influence the agent depends on what game form the leader selects and whether she can change the game form based on whether capture is occurring. If the leader delegates and cannot revoke the agent's authority when the outsider exerts its influence, then the outsider may want to influence the agent so that his standard of proof falls below the principal's. Otherwise, it can only recreate the policy outcome that would obtain under administration.

3.4.3 Inferences about Capture

With the mechanisms of capture clarified, it becomes possible to determine how the outsider's influence can be inferred from actions taken by the players. While it might be possible to observe capture directly, such as with a recording of a conversation about a quid pro quo, it is realistically likely that a player consciously subject to capture would act so that such evidence cannot be discovered. Thus, for the remainder of the discussion, actions from which the public can detect capture will be limited to what the leader can observe in the baseline model: her choice of game form and agent, her or his policy decision (or, under oversight, the agent's proposal), and any of the researcher's information that she receives.

In the context of this model, statutory capture is likely to be quite difficult to detect. Choosing an agent with a low standard of proof would show that she was subject to capture by quasi-contract. However, it may not be clear which available agent has the greatest bias

[4]Even if an agency could employ yet another party to avoid directly facing the interest group, doing so would only remove the problem one step, as the group could seek to capture that party.

toward policy 0 that benefits the principal. Furthermore, if there is ex ante uncertainty about the researcher's preferences or research technology, the leader would be right to select an agent with a somewhat lower standard of proof to prevent him from discouraging the outsider's research.

A clearer sign of statutory capture would the leader's decision not to use an agent at all, but to make the policy decision herself via administration. This evidence is also not unambiguous since she might not have any agent available who would benefit her under delegation or oversight. However, it can be argued that, in relative terms, opting for administration is stronger evidence than choosing an agent different from what the principal would prefer. In the case of Congress, this intuition contrasts with iron-triangle style arguments that agencies exist to benefit the interest groups that they regulate, and even that Congress creates this arrangement. Though delegation to or oversight of an agency may represent an attempt to avoid "making a hard decision," this avoidance can be socially beneficial when the problem is that outside groups will only submit information that barely satisfies legislators. Thus, employing an agent is not only not a indication of congressional capture, but it can also be a means to mitigate capture in the form of statutory bias-shifting.

Meanwhile, agency capture could be inferred if the agent transmits information about an effort level below his standard of proof, assuming that the latter is known. If the researcher has engaged in capture, she might try to withhold the outsider's effort level from the leader. However, the converse is not necessarily true, as Proposition 3.6 indicates that, given the right parameters, she (and thus the principal) can benefit when the agent adopts a policy of nondisclosure. Applied to Congress, this result responds to claims that it has neglected its oversight responsibility in a way that is different from arguments that oversight is actually robust (Aberbach 1990) or that it has fire alarms as an alternative (McCubbins and Schwartz

1984). Instead, lack of oversight facilitates better policy-making by inducing more policy research effort from outside research groups.

A similar argument can be made about the oversight equilibrium in Proposition 3.7, in which an adversarial agent only discloses the researcher's information when the effort is high enough and the signal contradicts her bias toward policy 0. Although the agent seems to be reporting only the researcher's successes in this equilibrium, the effort that he reveals exonerates him of capture. Moreover, the leader would not want the agent to report all research results, because then the researcher would effectively be able to transmit low-quality information to her, which would result in a lower payoff for the leader and the principal. Like the leader's decision to use an agent, these disclosure patterns are not only not necessarily signs of capture, but they also can facilitate her attempt to avert the effects of statutory bias-shifting.

3.4.4 Measures Against Capture

The model also has implications for what kinds of measures are likely to be effective in combating capture. In theory, three methods can be considered: (1) complete transparency of the researcher's information whenever the agent has them, (2) transparency of its information only when the agent proposes policy 1, and (3) efforts to keep the agent's bias for policy 0 relatively high. In practice, the first measure roughly corresponds to President Obama's Open Government Initiative (see Coglianese 2009), while the third corresponds to his attempts to tighten ethics rules for executive branch officials (see Thurber 2011).

The discussion about detecting agency capture makes clear that the first measure, transparency of both items of the research information, will generally be useless or even counterproductive. The nature of delegation and oversight implies the following result:

Corollary 3.13. *Disclosure of the researcher's effort level and signal to the leader does not affect policy outcomes under delegation but causes the administration policy outcomes to obtain under oversight.*

In particular, complete transparency means that the researcher can effectively communicate directly with the leader as under administration. Under oversight, this measure, if intended to mitigate capture, will ironically allow the researcher to achieve what it would want from capture without having to engage this type of activity. The Obama Administration's exhortation to agencies to release documents more proactively under the Freedom of Information Act (Coglianese 2009, 533) may work for prior policymaking decisions that were not transparent, but it may have unintended consequences for future decisions to the extent that it "successfully" induces more disclosure.

Corollary 3.13 also adds some nuance to one of the key results in Ting (2008), which states that having an employee report to the principal when the manager would reject a project regardless of its quality can only benefit the principal. That result roughly corresponds to the case in Theorem 3.8(c), in which the agent's standard of proof is so high that it discourages the outsider from researching at all. However, part (a) of Theorem 3.8 highlights cases in which a moderately high standard of proof for an adversarial agent can benefit the principal, but only when the agent can withhold either the researcher's information. Ting (2008) does not include an analogous result because it considers only two quality levels. The oversight equilibrium results in the present model suggest that, if multiple quality levels are possible in a whistleblowing setting and the manager accepts projects at fewer quality levels than the employee and the principal, then the principal's desire to have the employee report project quality might be less absolute. Discouraging whistleblowing could encourage the employee to exert more effort so that the quality is high enough for the manager to approve, whereas

encouraging it might incentivize the employee to put in less effort and indicate a quality level that is satisfactory for the politician but not the manager, resulting in more approvals of lower quality projects.

The next option for combating is a conditional form of transparency, in which the agent reports the researcher's information only when he proposes policy 1. Since method would effectively detect capture, an important empirical question whether it can be implemented. One challenge that may arise is defining "policy 1," although this identification is simple for some categories of policymaking, like drug approvals. Also, in the face of uncertainty about the agent's preferences, it may be unclear upon observing a fairly low level of effort whether the agent has been captured or is merely acting according to a weaker bias toward policy 0 that he naturally has. In addition, since there are two oversight equilibria that yield the same policy outcomes under the conditions in Theorem 3.8(a), changing from the one in which the agent discloses nothing to one in which the agent transmits information with a proposal of policy 1 could be a challenge. To the extent that equilibrium selection represents culture (see Kreps 1990), it may not be easy for an agent accustomed to the former equilibrium to transition to the latter one.

The third possibility is preventing the adversarial agent from lowering his standard of proof. In the model, this entails increasing the cost of bias-shifting and quasi-contractual compensation. The Office of Government Ethics (OGE) proscribes for executive branch officials various kinds of behavior linked to influence by interest groups, such as gifts exceeding a nominal value and employment in a related industry after too short a period of time. These measures are designed in part to prevent officials from biasing their policymaking toward interest groups, including in an unconscious way. If it is difficult to stop capture at the policymaking stage, it is arguably helpful for OGE to prevent bureaucrats from becoming less

adversarial in the first place through interactions outside of any decision-making processes.

The main challenge is in enforcement. Unambiguously illicit activity, like bribery, requires substantial resources to punish and deter, although the same might be said for information nondisclosure if an agency can claim that it lacks information. However, much of the behavior that ethics regulations target is legal except for government officials. Thus, if they are aware of what actions are improper, they are likely to abstain from them and truthfully certify that they have done so on reporting forms. Furthermore, some restricted activities, like post employment lobbying, cannot be hidden. If maintaining agency officials' preference is feasible, then the Obama Administration's ethics reforms measures for executive branch officials (see Thurber 2011) are likely to be more effective than unconditionally applied transparency measures, as the Obama Administration's seem to be (see Coglianese 2009).

3.5 Application to FDA Pharmaceutical Regulation

A general implication of the model is that, rather than administer a policy program herself, a principal can do better if she employs the right kind of agent either to make the final policy decisions or to be the sole collector of information. The reason is not so that the agent can obtain knowledge that the principal cannot, but something of the reverse: so that he can prevent the principal from receiving information that she is perfectly capable of understanding. A policy area that arguably implicates many of the features of this model is the FDA drug approval process. In terms of the game, either the agency or its employees serve as adversarial agents facing the outsiders, pharmaceutical companies. This application heavily on the account in Carpenter (2010), so unless otherwise noted, page numbers in this section are citations to this source.

3.5.1 The FDA as Adversarial Gatekeeper

When an agent benefits the principal, it is because he stimulates more research effort from the outsider. The outsider is not formally required to engage in any level of research, but it cannot have policy 1 enacted if it does not submit research that satisfies the him. Carpenter observes this kind of dynamic operating in drug regulation when he observes, "the Administration's gatekeeping power enacts a system of incentives that induces the production of far more information (and higher quality information) from drug companies and medical researchers than would otherwise have occurred " (751). In particular, the FDA's gatekeeping power "stems from its ability to veto product entry" (16).

In addition to showing the value of gatekeeping, the model also suggests that this power encourages the most research if the agent is highly adversarial toward the outsider. In the case of the FDA, its reviewers induce large amounts of research arguably because they would prefer that the drug not be marketed in the absence of sufficient evidence supporting the drug. This notion is consonant with the idea that "the agency would have to negate an appreciable fraction of new drug applications. If approval became so happily predictable as to become perceivably deserved, the incentives for drug companies to conduct exhaustive, careful, and clinical trials would vanish" (493). Although rejecting some applications, regardless of personal preferences, might be a viable strategy in a repeated game context, such a strategy is at least easier to pursue if FDA reviewers actually value safety over drug innovation a priori.

There is some evidence that these reviewers are adversarial. To begin with, the FDA was one of the two main forces that maneuvered for amendments to the 1906 Pure Food and Drugs Act (80), and the act that passed gave the agency its current gatekeeping authority. In general, the FDA seems to have been more consistently adversarial compared to

the general public and the most vocal interest groups. It has had to withstand criticisms of a so-called "drug lag," according to which it was allegedly taking too long to approve new medicines (374); campaigns by patient advocacy groups to make new cancer and AIDS drugs available (410–11, 429), and calls to use external reviewers for new drug applications (NDAs) (458). The highly adverse media response to the agency's initial reaction of the drug Activase in 1987 (3–4) supports the general intuition that the people, directly or through their elected representatives, might more readily approve a drug if it could directly access the relevant information and make the final decision. Although there have also been congressional hearings questioning whether the FDA should have allowed particular drugs, they do not establish that Congress or the relevant committees are generally more adversarial than the agency's policymakers: first, there have been hearings expressing concern about slow approvals and lack of innovation (337), and second, as described in more detail below, even the first type of hearings may reflect institutional design concerns rather than committee members' underlying policy preferences.

It is harder to show directly that an adversarial stance is necessary for the FDA's gatekeeping authority to be effective, since there does not appear to be a period in which the FDA consistently approved drugs with too little evidence. However, there is evidence in other settings suggesting that, in general, gatekeeping power alone is insufficient to induce probing research. In related area of medical devices, Harris (2008) has reported in the *New York Times* that "disputes tend to pit agency managers, who often lean toward approving drugs or devices when the data are equivocal, against agency scientists, who want more certain trial results before allowing the products to be sold" (A15). A different agency, the now-defunct Minerals Management Service (MMS), had the authority to reject oil and gas lease applications based on safety an environment concerns, but it appears to have approved

applications even when its scientists concluded that these were significant issues (see Urbina 2010, May 14). More generally, the MMS "faced criticism . . . for generally favoring the oil industry over public and environmental safety concerns" (Neill and Morris 2012, 636). A more adversarial agency could conceivably have induced more research as to whether prospective lessees could adequately and cost-effectively address potential hazards.

3.5.2 Oversight and Delegation

The second element of the model that appears to operate in the FDA's pharmaceutical regulation is in the idea that a leader can benefit when the agent prevents her from receiving information or from acting on it. One can view the game form as one of oversight or delegation, depending on which actors are assigned the key roles in the game. There is modest support for the idea that the public benefits from congressional oversight if Congress is the leader and the FDA as a whole is the agent. Meanwhile, Carpenter's account provides rather strong evidence of delegation premised on the reverse rationale if an FDA manager is the leader and scientists lower in the hierarchy play the role of agent.

Though it is intuitive to view the game involving Congress and the agency as one of delegation, it is also possible to interpret it as one of oversight. Delegation assumes an act of commitment, and in theory, at least, Congress can reassert its authority through new legislation (see Callander 2008, 124). Members of Congress can also attempt to influence FDA informally. Carpenter reports "numerous cases in which legislator applied pressure behind the scenes and lobbied for the approval of a particular drug" (337). In the related area of the FDA's monitoring activities (inspections and analyses of product samples), Shipan (2004) finds that the agency can sometimes be responsive to congressional committees. In addition, the leader always always accedes to FDA's proposal in the equilibria in Propositions

3.6 and 3.7, so lack of oversight in the model's definition does not follow from rare decisions by members of Congress to reverse FDA drug approval decisions. Overall, it seems plausible that the oversight game could be in effect.

Perhaps the most difficult aspect of the reverse rationale to establish for Congress is that some of its members can sufficiently understand the clinical trials to determine whether a new drug is safe and effective. Admittedly, it would rare for a member of Congress to have the skill to generate the studies that come from clinical trials. However, inability to create information need not imply that they cannot comprehend information. Even if they cannot personally dissect a study, they may be able to rely on trusted staff or outside scientists for their opinions about the studies. Empirically, there is mild support for the proposition that members of the relevant committees would feel confident in drawing their own conclusions about a drug's safety and efficacy. First, a few oversight hearings in the past have focused on particular drugs on the market (338–39). Second, the agency's "technical reputation" has come under attack in the past by AIDS activists (456), and it has more recently "suffered as top scientists have fled the agency or have complained publicly about being overruled or ignored" (748). Thus, even if their level of scientific knowledge is not as high as those of FDA officials, the gap in expertise may be small enough for some congresspersons to believe that they can interpret experimental data, either directly or through surrogates they trust more than FDA reviewers.

If legislators are exercising oversight in pharmaceutical regulation and some of them feel they have sufficient expertise to understand the evidence supporting a drug application, then the information filtering is clearly occurring and supports their continuing oversight, even though the primarily motive for this filtering is not to withhold information from Congress. Currently, the FDA discloses documents related to a NDA if the agency approves

the medicine, but it does not release any materials related to the application if it rejects the prospective drug (Lurie and Zieve 2006, 89). This pattern of disclosure decisions corresponds to the equilibrium in Proposition 3.7. McGarity and Shapiro (1980) indicates that a primary reason that the FDA has cited for withholding this information is that it constitutes trade secrets (868–69). This work argues that the information should be disclosed, with an embargo on its use to support future applications to "ensure adequate research incentives" (884). However, the model suggests that the current withholding incentivizes research in its own way. Specifically, if data from all NDAs were released, firms might be able to expend less effort in research and rely on political pressure to have their drugs approved, anyway.

As for the working relationship between agency scientists and managers, Carpenter's account provides strong support for the notion that delegation prevents less adversarial players from making decisions based on information that they are capable of understanding. Through rulemaking, the FDA formally delegated authority as far down as the "directors and deputy division directors of the various drug review devisions" (484). Informally, true authority may lie in entry-level medical officers (see 483). Although entry-level officers may have more specialized expertise (id.), it is less plausible that their immediate supervisors and some higher officers lack sufficient expertise for an informed review. Instead, commitment of authority to entry-level officers might be motivated by the belief that they are the most adversarial agents within the FDA. Support for this notion comes from an activist who asserted that this kind of delegation insulates them not only from sponsoring firms, but also their overseers (490). Analogously, formal delegation to relatively junior directors might also be rationalized by the idea that they are more adversarial than more] senior officers, even if they are not as adversarial as entry-level reviewers.

3.5.3 Mitigating Capture

One important principle from the results on capture is that the use of an agent can protect against this phenomenon when bias-shifting is the key method since small amounts of bias-shifting directed at the leader will likely yield nothing for the outsider. If quasi-contracts with the leader are impossible, pharmaceutical firms should direct their efforts at influence toward the agent. If the FDA as a whole is perceived as the agent, this idea implies that they should target the agency rather than Congress. Unfortunately, it is difficult to compare the degree to which these companies have tried to exert their influence on each institution. However, the fact that committees have held hearings questioning the approval of particular drugs may reflect cases in which the sponsoring firm tried to influence the review process at the FDA rather than Congress. Such hearings would be consistent with a recognition that the agency should be adversarial, even though the committee members themselves might have difficulty rejecting a product with the same information.

Within the FDA, the idea that entry-level medical reviewers are more adversarial than directors was discussed as if these people's preferences were not susceptible to influence. However, formal delegation to division directors and informal delegation to initial reviewers could reflect an awareness by officers that they are susceptible to influence and should thus grant authority to less senior employees. Then, up to a point, the review process remains intact even if these officers are somewhat swayed, provided that the medical officers are not.

Because agency capture is a risk, a second principle is that ethics rules to keep the agent adversarial are better than additional disclosures in mitigating capture of the agent. This distinction is relevant to current policy discussions at the FDA. Specifically, the FDA has considered disclosing more information from NDAs, including those that are ultimately rejected (Asamoah and Sharfstein 2010). Although the standard tradeoff is between current

knowledge about potential treatments and future innovation, the model suggests that releasing information about failed NDAs might empower firms and patient advocates who disagree with the rejection. They might be able to appeal to legislators or more senior agency officials to intervene with arguments that supposed risks are not as great as reviewer concluded, especially compared to the benefits

Instead, the more important challenge is keeping medical officers or the agency as a whole more adversarial. For the agency, the delegations already mentioned can be portrayed as one method of maintaining a high standard of proof. For individual officers, ethics rules could be helpful. Five FDA employees were convicted for accepting bribes to approve generic drugs in 1989 (Gibbons 1991), so officials are clearly subject to capture. Though bribery has always been illegal, ethics rules could prevent officers from being influenced to that degree or to a lesser extent. Pharmaceuticals constitute an area in which various ethical issues arise, so the model suggests that the ethics of government officials' policymaking should be as thoroughly examined as the ethics of other actors' decisions.

3.6 Conclusion

Based on information that is observable (albeit not contractible) and whose quality depends on an outside party's effort, the model presented in this paper offers what appears to be a new logic for a leader's use of an agent to prevent her from obtaining information that she could understand. It differs not only from the idea that an agent exists to apply his expertise and gather information that she cannot comprehend, but also from the notion that he exists to elicit information from a regulated party by virtue of policy preferences that are closer to that party's preferences than those of the leader. This rationale clearly contrasts with the

expertise purpose, and it also differs from the standard information elicitation reason since the leader will want an agent who is more opposed to the outsider's preferences than she.

The "reverse" nature of this rationale continues into the analysis of capture. This model presents a plausible situation in which the purpose of an agent is not to facilitate capture by allowing interest groups to obtain favorable policy away from public scrutiny, but to reduce the incidence of regulatory capture by forcing interest groups to face agencies rather than political leaders. If leaders are subject to pressure that shifts their bias but can avoid quasi-contracts, then can and prefer to pass authority, or at least information-gathering, to an agent who can credibly threaten unfavorable policy can elicit higher quality information because he is naturally set against the interested party. Though the agent himself can be captured, incomplete disclosures that might seem to evince capture not only do not necessarily indicate influence by interest groups but instead may be essential for the leader to benefit from oversight.

The FDA's drug approval process arguably provides a concrete example in which this reverse rationale operates. The FDA and its scientific reviewers can generally be expected to be more adversarial than the general public and thereby induce more research from drug sponsors than if, in theory, each drug approval were decided according to popular will. There are even some hints, in the agency's delegations to junior-level employees and in congressional investigations of approved drugs, that actors might be aware of this dynamic. Since FDA officials do not have a monopoly on the relevant scientific knowledge, it is plausible that the current system usefully denies other actors either certain types information or the ability to act on that information. Even if specialized knowledge is one rationale for FDA regulation, the model at least indicates the desirability of relying on a regulator simply because she is adversarial and requires a high standard of proof—independently of any expertise advantage

she may have over other decision-makers.

The overall logic of the model is that, when an outsider directly faces a leader who can understand it, it can generate information that just barely satisfies her so that she will base her policy on that evidence. Unable to commit to demand higher quality evidence, she can benefit by relying on an agent who is not satisfied except by much higher-quality evidence. Thus, the agent functions as a natural commitment device. Though committing to delegation or oversight may be a challenge, the example of FDA drug approvals suggests that it is possible and can yield public benefits. Therefore, more exploration of the presence and potential usefulness of this alternative informational rationale for agencies in the administrative state is warranted.

Appendix A

Equilibrium Refinements for Chapter 2

The reason for having equilibrium refinements in Chapter 2 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.

Appendix B

Proofs of Numbered Results

Contents

B.1 Proofs of Results for Chapter 1

Proof of Lemma 1.1 The first statement follows from the first-order condition $b'_P(\hat{r}(\lambda)) = (\lambda\gamma_P^L + (1-\lambda)\gamma_P^H)c'_P(\hat{r}(\lambda))$. Differentiating with respect to λ yields

$$\frac{\partial \hat{r}}{\partial \lambda} = \frac{\left(\gamma_P^H - \gamma_P^L\right) c'_P\left(\hat{r}(\lambda)\right)}{\left(\lambda\gamma_P^L + (1-\lambda)\gamma_P^H\right) c''_P\left(\hat{r}(\lambda)\right) - b''_P\left(\hat{r}(\lambda)\right)} > 0. \tag{B.1}$$

For the second statement, given any fractions (or pdf values) θ_1, θ_2, θ_3, θ_4, and θ_5, respectively of U_H, U_L, V_H^H, V_H^L, and V_L^L,

$$\begin{aligned}
\bar{\lambda} &= \frac{(\theta_1 p_U + \theta_4 p_V) p_L (1-q) p_1^H + (\theta_2 p_U + \theta_5 p_V) p_L q p_1^L}{(\theta_1 p_U + \theta_3 p_V) p_H p_1^H + (\theta_1 p_U + \theta_4 p_V) p_L (1-q) p_1^H + (\theta_2 p_U + \theta_5 p_V) p_L q p_1^L} \\
&\geq \frac{(\theta_1 p_U + \theta_4 p_V) p_L (1-q) p_1^H + (\theta_2 p_U + \theta_5 p_V) p_L q p_1^H}{(\theta_1 p_U + \theta_3 p_V) p_H p_1^H + (\theta_1 p_U + \theta_4 p_V) p_L (1-q) p_1^H + (\theta_2 p_U + \theta_5 p_V) p_L q p_1^H} \\
&= \frac{(\theta_1 p_U + \theta_4 p_V) p_L (1-q) p_0^H + (\theta_2 p_U + \theta_5 p_V) p_L q p_0^H}{(\theta_1 p_U + \theta_3 p_V) p_H p_0^H + (\theta_1 p_U + \theta_4 p_V) p_L (1-q) p_0^H + (\theta_2 p_U + \theta_5 p_V) p_L q p_0^H} \\
&\geq \frac{(\theta_1 p_U + \theta_4 p_V) p_L (1-q) p_0^H + (\theta_2 p_U + \theta_5 p_V) p_L q p_0^L}{(\theta_1 p_U + \theta_3 p_V) p_H p_0^H + (\theta_1 p_U + \theta_4 p_V) p_L (1-q) p_0^H + (\theta_2 p_U + \theta_5 p_V) p_L q p_0^L} = \underline{\lambda}
\end{aligned}$$

because $\frac{(\theta_1 p_U + \theta_4 p_V) p_L (1-q) p_1^H}{(\theta_1 p_U + \theta_3 p_V) p_H p_0^H + (\theta_1 p_U + \theta_4 p_V) p_L (1-q) p_1^H)} < 1$. ■

Proof of Proposition 1.2 The default equilibrium includes one proposal, $r_A^1 \geq \tilde{r}$, under the circumstances $\Omega \backslash U_L$ and a second proposal, $r_A^2 \geq \tilde{r}$ ($r_A^2 \neq r_A^1$), under U_L. The principal's belief after r_A^1 is $\lambda = \lambda_{\Omega \backslash U_L} \equiv \lambda_A^1$, leading to $r_P^1 = \hat{r}(\lambda_A^1)$, and her belief after r_A^2 is $\lambda = 1$, leading to $r_P^2 = \hat{r}(1)$. By setting $p_1^L < 1$, the agent cannot propose below the threshold and use lack of a media report as proof that the signal was $s = H$. Then, the principal can believe that $\lambda = 1$ and select $r_P = \hat{r}(1)$ for proposals off the equilibrium path. The group ends up paying zero for the equilibrium proposals that meet the threshold to avoid a media report. Going back one more step, the agent's beliefs follow the signal: $\lambda = \lambda_{U_H}$ for any agent who has seen the high signal and $\lambda = 1$ for any agent who has seen the low signal. The structure of the game assures that the upright agent does not update his posterior probability, while the group's decision to always induce the same policy from the venal agent assures that the venal agent's probability is not updated, either.

The description in the previous paragraph implies the weak consistency necessary for a perfect Bayesian equilibrium. For sequential rationality, we start with the group's decision

about transfers. The group cannot influence the upright agent's decisionmaking, but it could consider inducing different policies from the venal agent. However, the group will not do so because the following incentive compatibility conditions are satisfied for various scenarios involving the venal agent, where $\mathbf{1}_{\{r<\tilde{r}\}}$ is the indicator function as to whether the group induces a policy below the threshold:

$$\begin{aligned}
\gamma^H \hat{c}(\lambda_A^1) &\leq \gamma^H \hat{c}(1) + \mathbf{1}_{\{r<\tilde{r}\}} p_1^H (k_A + k_G) \text{ for } V_H^H, \\
\gamma^L \hat{c}(\lambda_A^1) &\leq \gamma^H \hat{c}(1) + \mathbf{1}_{\{r<\tilde{r}\}} p_1^H (k_A + k_G) \text{ for } V_H^L, \text{ and} \\
\gamma^L \hat{c}(\lambda_A^1) &\leq \gamma^H \hat{c}(1) + \mathbf{1}_{\{r<\tilde{r}\}} p_1^L (k_A + k_G) \text{ for } V_L^L.
\end{aligned}$$

These conditions are satisfied because $c(\cdot)$ is an increasing function. The group ensures that venal agent is receiving zero in expectation at all times, so the venal agent's strategy is incentive compatible. The upright agent's conditions, respectively after the high and low signals, are as follows:

$$\begin{aligned}
\alpha f(\hat{r}(\lambda_A^1), \lambda_{U_H}) &\geq \alpha f(\hat{r}(1), \lambda_{U_H}) - \mathbf{1}_{\{r<\tilde{r}\}} p_1^H k_A \\
\text{and } \alpha \hat{f}(1) &\geq \alpha \hat{f}(1) - \mathbf{1}_{\{r<\tilde{r}\}} p_1^H k_A
\end{aligned}$$

The condition for U_H is satisfied because $\hat{r}(1) > \hat{r}(\lambda_A^1) > \hat{r}(\lambda_{U_H})$, while the condition for U_L is automatically satisfied due to U_L's receiving his optimum payoff. Finally, the public's condition is incentive compatible by construction.

A fully pooling equilibrium may or may not exist, but even if it does exist, the public's payoff from it is less than the payoff from this default equilibrium: $\hat{f}(p_L) = \Pr(\Omega \setminus U_L) f(\hat{r}(p_L), \lambda_A^1) + \Pr(U_L) f(\hat{r}(p_L), 1) < \Pr(\Omega \setminus U_L) \hat{f}(\lambda_A^1) + \Pr(U_L) \hat{f}(1)$. ■

The following lemma will be used in the proofs of Propositions 1.3, 1.5, and 1.7:

Lemma B.1. *When U_H, V_H^H, and V_H^L are known not to incur any media costs, then, in equilibrium, all proposals of at least $\tilde{r}$ yield $\lambda \geq \lambda_{U_H}$ and any proposals below $\tilde{r}$ must yield $\underline{\lambda} \geq \lambda_{U_H}$.*

Proof. If a proposal does not involve V_H^H, the principal must always believe $\lambda \geq \lambda_{U_H}$ and $\underline{\lambda} \geq \lambda_{U_H}$ (if applicable), since all distinct agent scenarios other than V_H^H have a posterior probability of at least λ_{U_H}. Meanwhile, if a proposal involving V_H^H implies $\lambda < \lambda_{U_H}$, then every proposal by V_H^H must yield $\lambda < \lambda_{U_H}$ for $r_A \geq \tilde{r}$ or $\underline{\lambda} < \lambda_{U_H}$ below the threshold. V_H^H cannot be fully pooled with V_H^L, because then the posterior probability (λ or $\underline{\lambda}$) is at least λ_{U_H}, regardless of how much additional pooling occurs with U_H, U_L, or V_L^L, none of which could pull the posterior probability below λ_{U_H}. Then there is at least one proposal which involves V_H^L but not V_H^H, which it has already been shown must have the principal believing $\lambda \geq \lambda_{U_H}$or $\underline{\lambda} \geq \lambda_{U_H}$. Then V_H^L would deviate from any equilibrium of this form by choosing one of the proposals that follows from V_H^H. Thus, there is no proposal in equilibrium that yields $\lambda < \lambda_{U_H}$, with or without a media report. ■

Proof of Proposition 1.3 Lemma B.1 implies that U_H, along with the venal agent, can (be induced to) achieve the lowest policy possible. Then all proposals from any venal agent or U_H must yield the same policy. If U_L pools at all with any of the other agent scenarios, they and the fraction of U_L pooled must receive the same policy which is less than $r_P = \hat{r}(1)$, in which case U_L must fully pool for $\hat{r}(p_L)$, or else U_L would deviate by separating completely (rather than partially pooling) for $r_P = \hat{r}(1)$. The default when U_L fully pools is less than the default payoff: $\hat{f}(p_L) = \Pr(\Omega \setminus U_L) f(p_L, \lambda_{\Omega \setminus U_L}) + \Pr(U_L) f(p_L, 1) < \Pr(\Omega \setminus U_L) \hat{f}(\lambda_{\Omega \setminus U_L}) + \Pr(U_L) \hat{f}(1)$. If U_L is by itself, then non-U_L proposals are such that they all lead to the same

policy, which must be $r_P = \hat{r}(\lambda_{\Omega \setminus U_L})$ for weak consistency. Meanwhile, U_L receives $r_P = \hat{r}(1)$, also to satisfy weak consistency. The principal's payoff is $\Pr(\Omega \setminus U_L)\hat{f}(\lambda_{\Omega \setminus U_L}) + \Pr(U_L)\hat{f}(1)$, the default payoff. ■

Proof of Lemma 1.4 The pooling with some A_H proposals is equivalent to having $\eta = \eta_U p_U + \eta_V p_V$ of A_H proposals pooled with $\theta > 0$ of V_L^L proposals. Bayes rule implies that the posterior probabilities after the media reporting stage for proposals below the threshold can be expressed as

$$\bar{\lambda}_{\eta A_H \cup \theta V_L^L} = \frac{\eta p_L(1-q)p_1^H + \theta p_V p_L q p_1^L}{\eta(p_H + p_L(1-q))p_1^H + \theta p_V p_L q p_1^L} \tag{B.2}$$

$$\text{and } \underline{\lambda}_{\eta A_H \cup \theta V_L^L} = \frac{\eta p_L(1-q)p_0^H + \theta p_V p_L q p_0^L}{\eta(p_H + p_L(1-q))p_0^H + \theta p_V p_L q p_0^L}. \tag{B.3}$$

For convenience the subscript $\eta A_H \cup \theta V_L^L$ will be suppressed for the remainder of this proof. Then $C \equiv p_0^L \gamma^L \hat{c}(\underline{\lambda}) + p_1^L(\gamma^L \hat{c}(\underline{\lambda}) + k_A + k_G)$ is the low-cost group's cost of proposing below the threshold, and the goal is to show that $\frac{\partial C}{\partial p_1^L} > 0, \forall p_1^L < 1$. Differentiating yields $\frac{\partial C}{\partial p_1^L} = k_A + k_G + \gamma^L(\hat{c}(\bar{\lambda}) - \hat{c}(\underline{\lambda}) + \hat{c}'(\bar{\lambda})p_1^L \frac{\partial \bar{\lambda}}{\partial p_1^L} + \hat{c}'(\underline{\lambda})\left(1 - p_1^L\right)\frac{\partial \underline{\lambda}}{\partial p_1^L})$. Differentiating Equations (B.2) and (B.3) with respect to p_1^L yields respectively

$$p_1^L \frac{\partial \bar{\lambda}}{\partial p_1^L} = (1 - \bar{\lambda}) \frac{\theta p_V p_L q p_1^L}{\eta(p_H + p_L(1-q))p_1^H + \theta p_V p_L q p_1^L} \tag{B.4}$$

$$\text{and } p_0^L \frac{\partial \underline{\lambda}}{\partial p_1^L} = -(1 - \underline{\lambda}) \frac{\theta p_V p_L q p_0^L}{\eta(p_H + p_L(1-q))p_0^H + \theta p_V p_L q p_0^L}. \tag{B.5}$$

Substituting in these expressions and rearranging yields

$$\frac{\partial C}{\partial p_1^L} = k_A + k_G + \gamma^L \Bigg[\hat{c}\big(\bar{\lambda}\big) - \hat{c}(\underline{\lambda}) - \hat{c}'(\underline{\lambda})\big(\bar{\lambda} - \underline{\lambda}\big) \frac{\theta p_V p_L q p_0^L}{\eta(p_H + p_L(1-q))p_0^H + \theta p_V p_L q p_0^L}$$

$$+\big(1-\bar{\lambda}\big)\left(\hat{c}'(\bar{\lambda})\frac{\theta p_V p_L q p_1^L}{\eta(p_H+p_L(1-q))p_1^H+\theta p_V p_L q p_1^L}-\hat{c}'(\underline{\lambda})\frac{\theta p_V p_L q p_0^L}{\eta(p_H+p_L(1-q))p_0^H+\theta p_V p_L q p_0^L}\right)\Bigg].$$

Note that $p_H^1 \leq p_L^1$ implies $\frac{\theta p_V p_L q p_1^L}{\eta(p_H+p_L(1-q))p_1^H+\theta p_V p_L q p_1^L} \geq \frac{\theta p_V p_L q p_0^L}{\eta(p_H+p_L(1-q))p_0^H+\theta p_V p_L q p_0^L}$, while Lemma 1.1 and convexity of $\hat{c}(\lambda)$ implies $\hat{c}'(\bar{\lambda}) > \hat{c}'(\underline{\lambda}) > 0$, so that $\hat{c}'(\bar{\lambda})\frac{\theta p_V p_L q p_1^L}{\eta(p_H+p_L(1-q))p_1^H+\theta p_V p_L q p_1^L} > \hat{c}'(\underline{\lambda})\frac{\theta p_V p_L q p_0^L}{\eta(p_H+p_L(1-q))p_0^H+\theta p_V p_L q p_0^L}$. Since k_A, k_G, and γ are all positive, $\frac{\partial C}{\partial p_1^L} > 0$ if $\hat{c}(\bar{\lambda}) - \hat{c}(\underline{\lambda}) > \hat{c}'(\underline{\lambda})(\bar{\lambda}-\underline{\lambda})\frac{\theta p_V p_L q p_0^L}{\eta(p_H+p_L(1-q))p_0^H+\theta p_V p_L q p_0^L}$. This inequality holds: convexity of $\hat{c}(\lambda)$ implies that $\hat{c}(\bar{\lambda}) - \hat{c}(\underline{\lambda}) > \hat{c}'(\underline{\lambda})(\bar{\lambda}-\underline{\lambda})$, and $\frac{\theta p_V p_L q p_0^L}{\eta(p_H+p_L(1-q))p_0^H+\theta p_V p_L q p_0^L} < 1$ since $\theta p_V p_L q p_0^L, \eta(p_H + p_L(1-q))p_0^H > 0$. ■

Remark. Convexity of $\hat{c}(\lambda)$ with respect to λ is a fairly weak condition, since $c(r)$ is already assumed to be strictly convex with respect to r. From Lemma 1.1, $\frac{\partial \hat{r}}{\partial \lambda} > 0$. If the third derivatives of $b_P(r)$ and $c_P(r)$ exist, then

$$\frac{\partial^2 \hat{r}}{\partial \lambda^2} = \frac{\partial \hat{r}}{\partial \alpha}\left(\frac{\partial \hat{r}}{\partial \alpha} + \frac{\big((\lambda\gamma_P^L + (1-\lambda)\gamma_P^H)c_P'''(\hat{r}) - b_P'''(\hat{r})\big)\frac{\partial \hat{r}}{\partial \alpha} + (\gamma_P^L - \gamma_P^H)c_P''(\hat{r})}{(\lambda\gamma_P^L + (1-\lambda)\gamma_P^H)\,c_P''(\hat{r}) - b_P''(\hat{r})}\right). \tag{B.6}$$

As long as $(\lambda\gamma_P^L + (1-\lambda)\gamma_P^H)c_P'''(\hat{r}) - b_P'''(\hat{r})$ is not negative and too large in magnitude compared to $(\lambda\gamma_P^L + (1-\lambda)\gamma_P^H)c_P''(\hat{r}) - b_P''(\hat{r})$, $\frac{\partial^2 \hat{r})}{\partial \lambda^2} \geq 0$, so that $\hat{c}(\lambda)$ is convex with respect to λ. Since $c(r)$ is convex with respect to r, $\hat{c}(\lambda)$ can be convex even if $\frac{\partial^2 \hat{r}}{\partial \lambda^2}$ is somewhat negative. The third derivatives of the public's benefit and cost functions may not exist, but Equation (B.6) suggests that the conditions under which $\hat{c}(\lambda)$ is convex are broader than the conditions under which it is not. □

Proof of Proposition 1.5 (a) For the equilibrium in (i), the inequality is just a rearrangement of the incentive compatibility condition in (1.7), with beliefs that $\lambda = 1$ for any proposal off the equilibrium path. Pointing in the direction indicated, it properly implies

that risk of a media report is worth the chance of a lower level of regulation so that V_L^L is fully pooled with A_H. The other incentive compatibility constraints are satisfied: the group facing the venal agent with $s = H$ receives the lowest policy available in equilibrium with probability 1, the upright agent with $s = L$ receives its optimal polity of $\hat{r}(1)$ with no media cost, and the upright agent with $s = H$ prefers $\hat{r}\left(\underline{\lambda}_{A_H \cup V_L^L}\right)$ to $\hat{r}(1)$ and can achieve it without any media cost. Weak consistency for the players is satisfied by construction, and the principal's decision rule described in the text implies that her strategy is sequentially rational. The proofs of the sustainability for the equilibria in (ii) and (iii) are analogous, except that in (iii), the test points in the other direction because V_L^L is pooled with U_L rather than A_H, and that in (ii), the test is for equality because V_L^L is pursuing a mixed strategy. Since $\underline{\lambda}_{A_H \cup \theta V_L^L}$ increases with θ (by inspection), there is no conflict between these tests, and exactly one of the three kinds of equilbria is possible.

(b) The payoff to the public from the equilibria can also be expressed as $\left(\Pr\left(A_H\right) + \theta p_0^L \Pr\left(V_L^L\right)\right)\hat{f}\left(\underline{\lambda}_{A_H \cup \theta V_L^L}\right) + \left(\Pr\left(U_L\right) + \left(1 - \theta p_0^L\right)\Pr\left(V_L^L\right)\right)\hat{f}(1)$, for some $\theta \in [0,1]$, since Equilibrium (i) represents $\theta = 1$, while Equilibrium (iii) represents $\theta = 0$. This expected utility exceeds the default payoff:

$$\begin{aligned}
&\left(\Pr A_H + \theta p_0^L \Pr\left(V_L^L\right)\right)\hat{f}\left(\underline{\lambda}_{A_H \cup \theta V_L^L}\right) + \left(\Pr\left(U_L\right) + \left(1 - \theta p_0^L\right)\Pr\left(V_L^L\right)\right)\hat{f}(1) \\
&\quad > \left(\Pr A_H + \theta p_0^L \Pr\left(V_L^L\right)\right) f\left(\lambda_{\Omega \setminus U_L}, \underline{\lambda}_{A_H \cup \theta V_L^L}\right) + \left(1 - \theta p_0^L\right)\Pr\left(V_L^L\right) f\left(\lambda_{\Omega \setminus U_L}, 1\right) \\
&\qquad\qquad + \Pr\left(U_L\right)\hat{f}(1) = \Pr\left(\Omega \setminus U_L\right)\hat{f}\left(\lambda_{\Omega \setminus U_L}\right) + \Pr\left(U_L\right)\hat{f}(1).
\end{aligned}$$

Also, given any two equilibria with fractions of V_L^L $\underline{\theta} < \bar{\theta}$, the payoff from $\underline{\theta}$ is higher:

$$\left(\Pr A_H + \underline{\theta} p_0^L \Pr\left(V_L^L\right)\right)\hat{f}\left(\underline{\lambda}_{A_H \cup \underline{\theta} V_L^L}\right) + \left(\Pr\left(U_L\right) + \left(1 - \underline{\theta} p_0^L\right)\Pr\left(V_L^L\right)\right)\hat{f}(1)$$

$$> \left(\Pr A_H + \underline{\theta} p_0^L \Pr\left(V_L^L\right)\right) f\left(\underline{\lambda}_{A_H \cup \bar{\theta} V_L^L}, \underline{\lambda}_{A_H \cup \underline{\theta} V_L^L}\right) + \left(\bar{\theta} - \underline{\theta}\right) p_0^L \Pr\left(V_L^L\right) f\left(\underline{\lambda}_{A_H \cup \bar{\theta} V_L^L}, 1\right)$$
$$+ \left(\Pr\left(U_L\right) + \left(1 - \bar{\theta} p_0^L\right) \Pr\left(V_L^L\right)\right) \hat{f}(1)$$
$$= \left(\Pr A_H + \bar{\theta} p_0^L \Pr\left(V_L^L\right)\right) \hat{f}\left(\underline{\lambda}_{A_H \cup \bar{\theta} V_L^L}\right) + \left(\Pr\left(U_L\right) + \left(1 - \bar{\theta} p_0^L\right) \Pr\left(V_L^L\right)\right) \hat{f}(1).$$

(c) To identify the highest achievable payoff for the public, the first step is to identify the necessary conditions for various types of equilibria. This proof holds even when U_L is not restricted from proposing below the threshold. The proof follows by a series of claims:

Claim 1.5(c)-1. Proposals from the agent scenarios U_H, V_H^H, and V_H^L will all yield the lowest equilibrium policy from the public.

Proof. This claim follows from Lemma B.1, since U_H, V_H^H, and V_H^L can all (be induced to) deviate to a lower policy if there is one, the public always chooses a single policy after the media report stage, and these agent types never receive a media report. □

In any equilibrium, it must be the case that V_L^L always proposes above the media threshold, mixes above and below it, or always proposes below it. The types of equilibria in these categories whose payoffs to the public exceed the default are limited and share certain characteristics:

Claim 1.5(c)-2. Any equilibrium in which proposals after V_L^L are always at least the threshold and in which the public exceeds the default payoff entails proposals after A_H that are always below the threshold and lead to $r_P = \hat{r}(\lambda_{A_H})$ and U_L's always proposing at least the threshold and receiving $r_P = \hat{r}(1)$ along with V_L^L. The equilibrium requires $p_1^L(k_A + k_G) \geq p_0^L \gamma^L \left(\hat{c}(1) - \hat{c}\left(\underline{\lambda}_{A_H}\right)\right)$ for V_L^L.

Proof. If A_H always has proposals meeting or exceeding the threshold, one of three things happens: (1) U_L always proposes $r_A \geq \tilde{r}$, which the proof of Proposition 1.3 implies can't

yield the principal above her default payoff; (2) U_L always proposes below the threshold, in which case the remaining agent types pool together, and the resulting equilibrium (even if sustainable) yields the same as the default payoff; or (3) U_L randomizes between proposing on each side of the threshold. In the third case, U_L must always pool when it proposes at least the threshold. If he ever separates, weak consistency implies that he always separates to achieve his maximum payoff, in which case he would never propose below $\tilde{r}$. If U_L pools fully, every proposal $r_A \geq \tilde{r}$ must yield the same level of regulation and yield the same policy, because the equilibrium is only incentive-compatible only if the other agent types have proposals leading to the same λ. All the proposals that are at least $\tilde{r}$ must lead to the same policy to prevent deviations by A_H agents and U_H. The expected payoff to the public is less than the default equilibrium payoff: $\theta \Pr(U_L)\hat{f}(1) + \Pr(\Omega \setminus \theta U_L)\hat{f}(\Omega \setminus \theta U_L) = \theta \Pr(U_L)\hat{f}(1) + (1-\theta)\Pr(U_L)f(\hat{r}(\Omega \setminus \theta U_L), 1) + \Pr(\Omega \setminus U_L)f(\hat{r}(\Omega \setminus \theta U_L), \Omega \setminus U_L) < \Pr(U_L)\hat{f}(1) + \Pr(\Omega \setminus U_L)\hat{f}\left(\lambda_{\Omega \setminus U_L}\right)$, where θ is the frequency with which U_L chooses to propose below $\tilde{r}$.

Next to be considered is A_H randomizing above and below the threshold. Then the same value of λ must follow from each proposal involving part of A_H. With A_H and any fraction θ of U_L, $\lambda < 1$, so V_L^L will pool with A_H at at least the threshold, which means that the upright agent with the low-cost signal must pool with A_H below the threshold and may also pool on the other side. Since it pools with A_H below the threshold, it cannot propose any separate policy below $\tilde{r}$, or else it would benefit by deviating to it. For $r_A \geq \tilde{r}$, it also cannot propose any separate policy from A_H or else it would benefit by always making that proposal. Thus, the only possibly incentive-compatible behavior entails that all agents, except U_L after a media report, receive the same policy. (For A_H, $\underline{\lambda}$ for proposing below the threshold must equal λ for proposing at least $\tilde{r}$.) The resulting payoff is also less than the default payoff for

any fraction θ of U_L that pools with A_H: $\theta p_1^L \Pr(U_L)\hat{f}(1) + \left(1 - \theta p_1^L \Pr(U_L)\right) \hat{f}\left(\lambda_{\Omega \setminus \theta p_1^L U_L}\right) = \theta p_1^L \Pr(U_L)\hat{f}(1) + \left(1 - \theta p_1^L\right) \Pr(U_L) f\left(\hat{r}\left(\lambda_{\Omega \setminus \theta p_1^L U_L}\right), 1\right) + \Pr(\Omega \setminus U_L) f\left(\hat{r}\left(\lambda_{\Omega \setminus \theta U_L}\right), \Omega \setminus U_L\right) < \Pr(U_L)\hat{f}(1) + \Pr(\Omega \setminus U_L)\hat{f}\left(\lambda_{\Omega \setminus U_L}\right)$.

Thus, with V_L^L proposing at least the threshold, a higher payoff than the default accrues only if A_H always proposes below $\tilde{r}$. Weak consistency implies $r_P = \hat{r}(1)$ for whatever V_L^L proposes, in which case U_L finds it optimal to pool with V_L^L so that they both receive $r_P = \hat{r}(1)$. Claim 1.5(c)-1 implies the A_H agents all get $r_P = \hat{r}((\lambda_{A_H}))$. The principal's expected utility, $\Pr A_H \hat{f}((\lambda_{A_H})) + \Pr(A_L)\hat{f}(1)$, is the same as the payoff for Equilibrium (iii). The equilibrium policies also imply that this inequality is required to hold to be incentive compatible for V_L^L. □

Claim 1.5(c)-3. Any equilibrium in which proposals after V_L^L are below the threshold with probability θ and at least the threshold otherwise, and in which the public exceeds the default payoff, entails proposals after A_H that are always below the threshold and lead to $r_P = \hat{r}\left(\underline{\lambda}_{A_H \cup \theta V_L^L}\right)$ and U_L's proposing at least the threshold for $r_P = \hat{r}(1)$. After V_L^L, the policy is $r_P = \hat{r}\left(\underline{\lambda}_{A_H \cup \theta V_L^L}\right)$ with probability p_0^L after proposing below $\tilde{r}$ and $r_P = \hat{r}(1)$ otherwise. The equilibrium requires $p_1^L(k_A + k_G) = p_0^L \gamma^L \left(\hat{c}(1) - \hat{c}\left(\underline{\lambda}_{A_H \cup \theta V_L^L}\right)\right)$ for V_L^L.

Proof. The low-cost group facing V_L^L is mixing and so must achieve the same cost on each side of the threshold, and the expected payoffs on the two sides must be equal: $p_0^L \gamma^L \hat{c}(\underline{\lambda}_x) + p_1^L(\gamma^L \hat{c}(1) + k_A + k_G) = \gamma^L \hat{c}(\lambda_y)$. By inspection $\underline{\lambda}_x < \lambda_y$ for the expected media cost to justify proposing below $\tilde{r}$. Also, λ_y must be the lowest value of λ for proposals of at least $\tilde{r}$ (including off the equilibrium path). No agent in one of the situations U_H, V_H^H, and V_H^L will ever propose at least $\tilde{r}$. If he did, he would propose whatever yields $r_P = \hat{r}(\lambda_y)$, but then he (or the group influencing him) would prefer to deviate by proposing below the threshold to yield $r_P = \hat{r}(\underline{\lambda}_x)$. Instead, all proposals after A_H fall below the threshold and yield the

same policy. By weak consistency, $\lambda_y = 1$, making it optimal for U_L to pool with V_L^L. Claim 1.5(c)-1 and the ability of the low-cost group with agent setting V_L^L to select any proposal below the threshold for the same cost imply that $\lambda = \lambda_{A_H \cup \theta V_L^L}$ for the proposals below the threshold. Meanwhile, weak consistency implies $\lambda = 1$ for U_L and V_L^L proposals that are at least $\tilde{r}$. The resulting polices after the media reporting stage follow from these probabilities. The resulting payoff to the public is $\Pr\left(A_H \cup \theta p_0^L V_L^L\right) \hat{f}\left(\underline{\lambda}_{A_H \cup \theta V_L^L}\right) + \Pr\left(A_L \setminus \theta p_0^L V_L^L\right) \hat{f}(1)$, the same as the payoff for Equilibrium (ii). Substituting the posterior probabilities for the equilibrium into the incentive compatibility condition for V_L^L implies that this equality is also required for the value of θ. □

Claim 1.5(c)-4. Any equilibrium in which proposals after V_L^L are all below the threshold entail that agents in A_H scenarios fully pool with V_L^L to induce the same policy in the event of no media report. Among these equilibria, those in which U_L always proposes at least the threshold yield the highest payoff for the public. These equilbria require $p_1^L(k_A + k_G) \leq p_0^L \gamma^L \left(\hat{c}(1) - \hat{c}\left(\underline{\lambda}_{A_H \setminus V_L^L}\right)\right)$.

Proof. This time, the incentive compatibility condition for V_L^L is $p_0^L \gamma^L \hat{c}(\underline{\lambda}_x) + p_1^L(\gamma^L \hat{c}(1) + k_A + k_G) \leq \gamma^L \hat{c}(\lambda_y)$, where $\hat{c}(\lambda_y)$ is the minimum the low-cost group would pay if it induced a policy meeting the threshold for V_L^L. The same steps as in the proof of the previous claim can be applied to establish that V_L^L proposals and A_H proposals are all below the threshold. Because there is no cost to any agent in these scenarios for changing among proposals below $\tilde{r}$, all the proposals involving these agent scenarios must lead to the same $\underline{\lambda}$ when there is no media report.

Three types of strategies are possible for U_L: First, if U_L ever proposes at least $\tilde{r}$, weak consistency implies that he always does to get $\hat{r}(1)$ all the time. Then weak consistency and the non-U_L agents' ability to pick any policy below $\tilde{r}$ for the same cost imply $\lambda = \lambda_{A_H \cup V_L^L}$ for

any proposal below the threshold and $\lambda = 1$ for any proposal at least the threshold. Then the incentive compatibility condition for V_L^L is $p_0^L \gamma^L \hat{c}\left(\underline{\lambda}_{A_H \cup V_L^L}\right) + p_1^L(\gamma^L \hat{c}(1) + k_A + k_G) \leq \gamma^L \hat{c}(1)$, which is equivalent to the condition in the claim. The principal's expected utility is $\Pr\left(A_H \cup p_0^L V_L^L\right) \hat{f}\left(\underline{\lambda}_{A_H \cup V_L^L}\right) + \Pr\left(U_L \cup p_1^L V_L^L\right) \hat{f}(1)$, the same as that for Equilibrium (i).

Second, if U_L always proposes less than $\tilde{r}$ and ever proposes a separate policy from the agent in other scenarios, then weak consistency again implies that U_L always does so in equilibrium to get $\lambda = 1$ all the time. Again, weak consistency and the non-U_L agents' ability to pick any policy below $\tilde{r}$ for the same cost imply $\lambda = \lambda_{A_H \cup V_L^L}$ for those agent settings. With the principal choosing the same policy after each agent scenario, her expected utility must be the same.

The only remaining possibility is that U_L always proposes below $\tilde{r}$ and fully pools with the other agents. Then Lemma B.1 implies that, below $\tilde{r}$, all agent scenarios other than U_L must have the lowest value of $\underline{\lambda}$. If U_L has a different value of $\underline{\lambda}$, its proposal would be different from the other agents', which contradicts full pooling by U_L. Then all proposals yield the same $\underline{\lambda}$. This payoff is less than the payoff of the first two equilibria: $\Pr\left(A_H \cup p_0^L A_L^L\right)\hat{f}(\underline{\lambda}_\Omega) + p_1^L \Pr(A_L)\hat{f}(1) = \Pr\left(A_H \cup p_0^L V_L^L\right) f\left(\hat{r}(\underline{\lambda}_\Omega), \underline{\lambda}_{A_H \cup V_L^L}\right) + p_0^L \Pr(U_L) f(\hat{r}(\underline{\lambda}_\Omega), 1) + p_1^L \Pr(A_L)\hat{f}(1) < \Pr\left(A_H \cup p_0^L V_L^L\right)\hat{f}\left((\underline{\lambda}_{A_H \cup V_L^L}\right) + \Pr\left(U_L \cup p_1^L V_L^L\right)\hat{f}(1)$. □

Part (b) indicates that the less pooling by V_L^L with A_H, the better. The second and third types of equilibria from Claim 1.5(c)-4 do not yield payoffs better than any of the equilibria in part (a), so they can be ignored. The possibilities for different types of equilibria have been exhausted based on the type of strategy that the low-cost group pursues facing V_L^L. The best equilibria corresponding to these strategies correspond 1-to-1 with the equilibria described for the incentive compatibility conditions in part (a), so those equilibria yield the greatest payoff conditional on the relevant incentive compatibility constraint for V_L^L.

(d) From part (b), given any value of p_1^L, the greatest expected utility for the principal occurs when $\theta = 0$. Substituting this value into the expression for the payoff yields $(\Pr A_H)\hat{f}(\lambda_{A_H}) + \left(\Pr(U_L) + \Pr\left(V_L^L\right)\right)\hat{f}(1)$, which does not depend on the value of p_1^L. Meanwhile, as p_1^L approaches 1, the left-hand side of $p_1^L(k_A + k_G) \geq p_0^L\gamma^L(\hat{c}(1) - \hat{c}(\lambda_{A_H}))$ approaches $k_A + k_G$ while the right-hand side approaches zero, so the incentive compatibility condition for Equilibrium (iii) is automatically satisfied. Thus, the principal can achieve this maximum expected utility when $p_1^H = 0$ if it can increase t such that $p_1^L = 1$.

(e) Based on the equilibria in part (a), the low-cost group cares about $p_0^L\gamma^L\hat{c}\left(\underline{\lambda}_{A_H \cup \theta V_L^L}\right) + p_1^L(\gamma^L\hat{c}(1) + k_A + k_G) - \gamma^L\hat{c}(1)$. Applying Lemma 1.4 with $\eta_U = \eta_V = 1$ and $p_1^H = 0$ implies that any equilibrium that previously existed with $\theta \in (0, 1)$ is no longer incentive compatible for the low-cost group facing V_L^L. Rebalancing requires decreasing θ until equality is restored or until $\theta = 0$ for equilibrium (iii). For $\theta = 0$, the positive derivative causes the low-cost group to favor V_L^L proposals that avoid a media report even more. For $\theta = 1$, the positive derivative implies that increasing transparency either breaks the equilibrium, requiring $\theta < 1$ or simply reduces the benefit to the low-cost group of inducing V_L^L proposals below $\tilde{r}$. Thus, an increase in transparency means that the incentive compatibility condition that is satisfied will be for a weakly lower value of θ or for the same value of θ. Part (b) states that, for a given value of p_1^L, the payoff increases as θ decreases. However, p_1^L increases, so the principal's payoff is even higher. For any $\underline{p}_0^L < \bar{p}_0^L$, the comparison with the subscript $A_H \cup \theta V_L^L$ for posterior probabilities suppressed is

$$\begin{aligned}
&\left(\Pr A_H + \theta\underline{p}_0^L \Pr\left(V_L^L\right)\right)\hat{f}\left(\underline{\lambda}\left(\underline{p}_0^L\right)\right) + \left(\Pr(U_L) + \left(1 - \theta\bar{p}_0^L\right)\Pr\left(V_L^L\right)\right)\hat{f}(1) \\
&\qquad > \left(\Pr A_H + \theta\underline{p}_0^L \Pr\left(V_L^L\right)\right) f\left(\underline{\lambda}\left(\bar{p}_0^L\right), \underline{\lambda}\left(\underline{p}_0^L\right)\right) \\
&\quad + \theta\left(\bar{p}_0^L - \underline{p}_0^L\right)\Pr\left(V_L^L\right) f\left(\underline{\lambda}\left(\bar{p}_0^L\right), 1\right) + \left(\Pr(U_L) + \left(1 - \theta\bar{p}_0^L\right)\Pr\left(V_L^L\right)\right)\hat{f}(1)
\end{aligned}$$

$$= \left(\Pr A_H + \theta \bar{p}_0^L \Pr\left(V_L^L\right)\right) \hat{f}\left(\underline{\lambda}\left(\bar{p}_0^L\right)\right) + \left(\Pr\left(U_L\right) + \left(1 - \theta \bar{p}_0^L\right) \Pr\left(V_L^L\right)\right) \hat{f}(1).$$

Thus, the principal's payoff is weakly increasing with t whenever $p_1^H = 0$ and strictly increasing when $\theta > 0 = p_0^H$ originally. ■

Proof of Proposition 1.6 ' V_H^H and U_H must propose below the media threshold to prevent pooling by V_L^L and V_H^L, since proposals of at least $\tilde{r}$ are cheap talk. If V_H^H and U_H separate from the other agent scenarios, then V_L^L and V_H^L will be assigned $r_P = \hat{r}(1)$, in which case the low-cost group is better off not inducing a proposal less than $\tilde{r}$ and thus cannot end up proposing below $\tilde{r}$. Then, since U_L has an option to pool with V_L^L or V_H^L to achieve $\hat{r}(1)$, he also will not propose below $\tilde{r}$. Thus, proposals that can trigger a media report only come from V_H^H and U_H. Since V_H^H would achieve $r_P = \hat{r}(0)$, the high-cost group and U_H prefer different policies and do not need to be screened from each other. If this equilibrium policy works, the payoff to the principal is as high as possible: She assigns $r_P = \hat{r}(0)$ to V_H^H and $r_P = \hat{r}(1)$ to V_L^L, V_H^L, and U_L, which are optimal with those agents. She also assigns $r_P = \hat{r}(\lambda_{U_H})$ to U_H, which is as good as possible because the upright agent has no other information than $s = H$ and is effectively transmitting that message perfectly through its proposal.

However, this equilibrium may not always exist. Based on the policies chosen, the low-cost group facing the venal agent would deviate by aiming for the V_H^H proposal if it is willing to pay media costs in expectation, while the upright agent with $s = H$, which generally prefers $r_P = \hat{r}(\lambda_{U_H})$, would deviate only to avoid a media report. Thus, there are three

binding incentive compatibility constraints:

$$\gamma^H \hat{c}(0) + p_1^H(k_A + k_G) \leq \gamma^H \hat{c}(1) \text{ for } V_H^H, \tag{B.7}$$

$$\gamma^L \hat{c}(0) + p_1^H(k_A + k_G) \geq \gamma^L \hat{c}(1) \text{ for } V_H^L, \text{ and} \tag{B.8}$$

$$\alpha \hat{f}(\lambda_{U_H}) - p_1^H k_A \geq \alpha f(1, \lambda_{U_H}) \text{ for } U_H. \tag{B.9}$$

The incentive compatibility condition for V_L^L is automatically satisfied because $p_1^H \leq p_1^L$, while U_L's is satisfied because he receives his optimal utility $\alpha \hat{f}(1)$. Combining the three conditions together yields $\frac{\gamma^L}{k_A+k_G}(\hat{c}(1) - \hat{c}(0)) \leq p_1^H \leq \min\left\{\frac{\gamma^H}{k_A+k_G}(\hat{c}(1) - \hat{c}(0)), \frac{\alpha}{k_A}(\hat{f}(\lambda_{U_H}) - f(1, \lambda_{U_H}))\right\}$. Fixing γ^L, letting α be arbitrarily small makes it possible for $\frac{\alpha}{k_A}(\hat{f}(\lambda_{U_H}) - f(1, \lambda_{U_H})) < \frac{\gamma^L}{k_A+k_G}(\hat{c}(1) - \hat{c}(0))$, or Inequality (B.7) may fail if $k_A + k_G < \gamma^L(\hat{c}(1) - \hat{c}(0))$ so that the equilibrium cannot be sustained. ■

Proof of Proposition 1.7 V_H and U_H can be deterred from proposing below the media threshold if $\alpha \hat{f}(\lambda_{U_H}) - p_1^H k_A < \alpha f(1, \lambda_{U_H})$ and $\gamma^H \hat{c}(0) + p_1^H(k_A + k_G) > \gamma^H \hat{c}(\lambda_{U_H \cup A_L})$, or if $\gamma^H \hat{c}(0) + p_1^H(k_A + k_G) > \gamma^H \hat{c}(1)$ and $\alpha \hat{f}(\lambda_{U_H}) - p_1^H k_A < \alpha f(\lambda_{V \cup U_L}, \lambda_{U_H})$. In the first case, U_H will not propose below $\tilde{r}$ even if it receives its ideal policy, in which case the high-cost group facing the venal agent will not receive more than $r_P = \hat{r}(\lambda_{U_H \cup A_L})$ if it deviates from proposing below $\tilde{r}$. The conditions for the low-cost group facing the venal agent to deviate from any equilibrium in which it (sometimes) proposes below $\tilde{r}$ are not binding: $\gamma^H \hat{c}(0) + p_1^s(k_A + k_G) > \gamma^H \hat{c}\left(\lambda_{U_H \cup U_L \cup V_{-s}^L}\right)$ is satisfied for each signal because $\gamma^L < \gamma^H$, $p_1^L \geq p_1^H$, and $\lambda_{U_H \cup U_L \cup V_{-s}^L} < \lambda_{U_H \cup A_L}$. In the second case, the condition for V_H^H implies that proposals from V_H^L and V_L^L will also be below the threshold because $\gamma^L < \gamma^H$ and $p_1^L \geq p_1^H$. Thus, the worst policy U_H can receive is $r_P = \hat{r}(\lambda_{V \cup U_L})$, because if U_H deviates by proposing

at least $\tilde{r}$, pooling among venal agents implies that he would seek out the lowest policy. (The lowest value for the lowest policy among V and U_L is $\hat{r}(p_L)$, so U_H will not seek $r_P < \hat{r}(U_H)$.) After both cases, at least the venal agent and U_H are proposing at least $\tilde{r}$.

Lemma B.1 and the fact that V_L^L proposals are also at least $\tilde{r}$ imply that proposals by the venal agent and U_H must all lead to the same policy to prevent deviations among them. Then, if U_L partially pools with any other agent scenario with a proposal, all proposals that meet the threshold must lead to the same policy. U_L might also propose below $\tilde{r}$, but the payoff to the principal can be expressed in the form $(1-\theta \Pr(U_L))\hat{f}\left(\lambda_{\Omega\setminus\theta U_L}\right)+\theta \Pr(U_L)\hat{f}(1)$ for some $\theta \in [0,1]$. This payoff does not exceed the default payoff:

$$
\begin{aligned}
&(1-\theta \Pr(U_L))\hat{f}\left(\lambda_{\Omega\setminus\theta U_L}\right)+\theta \Pr(U_L)\hat{f}(1) \\
&= (1-\Pr(U_L))\Pr(U_H \cup V) f\left(\lambda_{\Omega\setminus\theta U_L}, \lambda_{\Omega\setminus U_L}\right) + (1-\theta)\Pr(U_L) f\left(\lambda_{\Omega\setminus\theta U_L}, 1\right) + \theta \Pr(U_L)\hat{f}(1) \\
&\leq (1-\Pr(U_L))\hat{f}\left(\lambda_{\Omega\setminus U_L}\right) + \Pr(U_L)\hat{f}(1).
\end{aligned}
$$

The first case requires (1) $p_1^H > \max\left\{\frac{\alpha}{k_A}(\hat{f}(\lambda_{U_H}) - f(1,\lambda_{U_H})), \frac{\gamma^H}{k_A+k_G}(\hat{c}(\lambda_{U_H\cup A_L}) - \hat{c}(0))\right\}$, (2) $k_A \geq \alpha(\hat{f}(\lambda_{U_H}) - f(1,\lambda_{U_H}))$, and (3) $k_A + k_G \geq \gamma^H(\hat{c}(\lambda_{U_H\cup A_L}) - \hat{c}(0))$. The second case requires (1) $p_1^H > \max\left\{\frac{\alpha}{k_A}(\hat{f}(\lambda_{V\cup U_L}) - f(1,\lambda_{U_H})), \frac{\gamma^H}{k_A+k_G}(\hat{c}(1) - \hat{c}(0))\right\}$, (2) $k_A \geq \alpha(\hat{f}(\lambda_{U_H}) - f(\lambda_{V\cup U_L},\lambda_{U_H}))$, and (3) $k_A + k_G \geq \gamma^H(\hat{c}(1) - \hat{c}(0))$. ■

Proof of Lemma 1.8 The pooling with some A_H proposals is equivalent to having $\eta = \eta_U p_U + \eta_V p_V$ of A_H proposals pooled with $\theta > 0$ of V_L^L proposals. The posterior probabilities after the media reporting stage for proposals below the threshold can be expressed as

$$
\bar{\lambda}_{\eta A_H \cup \theta V_L^L} = \frac{\eta p_L(1-q)p_1^H + \theta p_V p_L q p_1^L}{\eta(p_H + p_L(1-q))p_1^H + \theta p_V p_L q p_1^L} \tag{B.10}
$$

$$\text{and } \underline{\lambda}_{\eta A_H \cup \theta V_L^L} = \frac{\eta p_L(1-q)p_0^H + \theta p_V p_L q p_0^L}{\eta(p_H + p_L(1-q))p_0^H + \theta p_V p_L q p_0^L}. \tag{B.11}$$

For convenience the subscript $\eta A_H \cup \theta V_L^L$ will be suppressed for the rest of the proof. The low-cost group's cost of proposing below the threshold is $C \equiv p_0^L \gamma^L \hat{c}(\underline{\lambda}) + p_1^L(\gamma^L \hat{c}(\bar{\lambda}) + k_A + k_G)$, so that $\frac{\partial C}{\partial p_1^H} = \gamma^L(p_0^L \hat{c}'(\underline{\lambda})\frac{\partial \underline{\lambda}}{\partial p_1^H} + p_1^L \hat{c}'(\bar{\lambda})\frac{\partial \bar{\lambda}}{\partial p_1^H})$. Differentiating Equations (B.10) and (B.11) yields

$$p_1^L \frac{\partial \bar{\lambda}}{\partial p_1^H} = -\eta\theta p_V p_H p_L q \left(\frac{p_1^L}{\eta(p_H + p_L(1-q))p_1^H + \theta p_V p_L q p_1^L}\right)^2 \tag{B.12}$$

$$\text{and } p_0^L \frac{\partial \underline{\lambda}}{\partial p_1^H} = \eta\theta p_V p_H p_L q \left(\frac{p_0^L}{\eta(p_H + p_L(1-q))p_0^H + \theta p_V p_L q p_0^L}\right)^2. \tag{B.13}$$

Substitution yields

$$\frac{\partial C}{\partial p_1^H} = \gamma^L \eta\theta p_V p_L p_H q \left[\hat{c}'(\underline{\lambda})\left(\frac{p_0^L}{\eta(p_H + p_L(1-q))p_0^H + \theta p_V p_L q p_0^L}\right)^2 - \hat{c}'(\bar{\lambda})\left(\frac{p_1^L}{\eta(p_H + p_L(1-q))p_1^H + \theta p_V p_L q p_1^L}\right)^2\right].$$

Convexity implies $\hat{c}'(\underline{\lambda}) \leq \hat{c}'(\bar{\lambda})$, while $p_1^H < p_1^L$ implies

$$\frac{p_1^L}{(p_H + p_L(1-q))p_1^H + \theta p_V p_L q p_1^L} > \frac{p_0^L}{(p_H + p_L(1-q))p_0^H + \theta p_V p_L q p_0^L}.$$

Then $\frac{\partial C}{\partial p_1^H} < 0$, and the cost to the low-cost group strictly decreases with p_1^H when $p_1^H < p_1^L < 1$. For $p_L^1 = 1$, the cost of proposing below the threshold is $\gamma^L \hat{c}(\bar{\lambda}) + k_A + k_G$. This cost decreases because $\bar{\lambda}$ decreases with p_1^H, as shown by Equation (B.10). ■

Proof of Proposition 1.9 (a) For the equilibrium in (i), the incentive compatibility condition for V_L^L is equivalent to $p_0^L \gamma^L \hat{c}\left(\underline{\lambda}_{A_H \cup V_L^L}\right) + p_1^L\left(\gamma^L \hat{c}\left(\bar{\lambda}_{A_H \cup V_L^L}\right) + k_A + k_G\right) \leq \gamma^L \hat{c}(1)$, with beliefs that $\lambda = 1$ for any proposal off the equilibrium path and implies that V_L^L prefers to propose below the threshold. The other incentive compatibility constraints are satisfied: U_L's because it receives its preferred policy at no cost; U_H's by assumption; V_H^L's because replacing $\frac{p_1^L}{p_0^L}$ with $\frac{p_1^H}{p_0^H}$ in the constraint reduces the left-hand side so that V_H^L will not (be induced to) defect; and V_H^H's because compared to the constraint for V_H^L, the V_H^H more strongly keeps those proposals below $\tilde{r}$ because $\gamma^H > \gamma^L$ on the right-hand side. Weak consistency for the players is satisfied by construction, and the principal's decision rule described in the text implies that her strategy is sequentially rational. The proofs of the sustainability for the equilibria in (ii) is analogous, except that the test is for equality because V_L^L is pursuing a mixed strategy. In (iii), the test points in the opposite direction of (i) because V_L^L is pooled with U_L rather than A_H. Meanwhile, $p_1^H(k_A + k_G) < \gamma^L(\hat{c}(1) - \hat{c}(\lambda_{A_H}))$ means that the group will continue to induce the venal agent with $s = H$ to propose below the threshold. Then the upright agent follows this equilibrium just as it would follow equilibrium (i). Since $\underline{\lambda}_{A_H \cup \theta V_L^L}$ increases with θ, there is no conflict between these tests, and exactly one of the three kinds of equilibria is possible.

With post-media stage probabilities listed explicitly as a function of p_1^H, the payoffs from the equilibria in part (a) following a proposal below $\tilde{r}$ can be expressed as $\left(p_0^H \Pr A_H + \theta p_0^L \Pr\left(V_L^L\right)\right)\hat{f}\left(\underline{\lambda}_{A_H \cup \theta V_L^L}\left(p_1^H\right)\right) + \left(p_1^H \Pr A_H + \theta p_1^L \Pr\left(V_L^L\right)\right)\hat{f}\left(\bar{\lambda}_{A_H \cup \theta V_L^L}\left(p_1^H\right)\right)$, for some $\theta \in [0,1]$, since Equilibrium (i) represents $\theta = 1$, while Equilibrium (iii) represents $\theta = 0$. Compared to equilibria in which $p_1^H = 0$, the payoffs from proposals above $\tilde{r}$ from U_L and $1-\theta$ of V_L^L are the same. However, when $\theta > 0$, the equilibrium payoff following the other proposals when $p_1^H > 0$ is less than when $p_1^H = 0$. Suppressing the subscript $A_H \cup \theta V_L^L$ for posterior

probabilities, the comparison is

$$\begin{aligned}
&\left(p_0^H \Pr A_H + \theta p_0^L \Pr\left(V_L^L\right)\right)\hat{f}\left(\underline{\lambda}\left(p_1^H\right)\right) + \left(p_1^H \Pr A_H + \theta p_1^L \Pr\left(V_L^L\right)\right)\hat{f}\left(\bar{\lambda}\left(p_1^H\right)\right) \\
&\quad = \left(p_0^H \Pr A_H + \theta p_0^L \Pr\left(V_L^L\right)\right) f\left(\underline{\lambda}\left(p_1^H\right), \underline{\lambda}(0)\right) \\
&\qquad + p_1^H \Pr A_H f\left(\bar{\lambda}\left(p_1^H\right), \underline{\lambda}(0)\right) + \theta p_1^L \Pr\left(V_L^L\right) f(\bar{\lambda}(p_1^H), 1) \\
&\quad < \left(\Pr A_H + \theta p_0^L \Pr\left(V_L^L\right)\right)\hat{f}\left(\underline{\lambda}(0)\right) + \theta p_1^L \Pr\left(V_L^L\right)\hat{f}(1). \qquad \text{(B.14)}
\end{aligned}$$

Since $\underline{\lambda}_{A_H}\left(p_1^H\right) = \bar{\lambda}_{A_H}\left(p_1^H\right) = \underline{\lambda}_{A_H}(0)$, substituting $\theta = 0$ into Inequality (B.14) changes it to equality.

Also, given any two equilibria with fractions of V_L^L $\underline{\theta} < \bar{\theta}$, the payoffs from proposals above $\tilde{r}$ from U_L and $1 - \bar{\theta}$ of V_L^L are the same. However, the payoff following the other proposals is higher when $\theta = \underline{\theta}$:

$$\begin{aligned}
&\left(p_0^H \Pr A_H + \underline{\theta} p_0^L \Pr\left(V_L^L\right)\right)\hat{f}\left(\underline{\lambda}_{A_H \cup \underline{\theta} V_L^L}\right) \\
&\qquad + \left(p_1^H \Pr A_H + \underline{\theta} p_1^L \Pr\left(V_L^L\right)\right)\hat{f}\left(\bar{\lambda}_{A_H \cup \underline{\theta} V_L^L}\right) + \left(\bar{\theta} - \underline{\theta}\right) \Pr\left(V_L^L\right)\hat{f}(1) \\
&\quad > \left(p_0^H \Pr A_H + \underline{\theta} p_0^L \Pr\left(V_L^L\right)\right) f\left(\underline{\lambda}_{A_H \cup \bar{\theta} V_L^L}, \underline{\lambda}_{A_H \cup \underline{\theta} V_L^L}\right) + \Big[\left(p_1^H \Pr A_H + \underline{\theta} p_1^L \Pr\left(V_L^L\right)\right) \\
&\times f\left(\bar{\lambda}_{A_H \cup \bar{\theta} V_L^L}, \bar{\lambda}_{A_H \cup \underline{\theta} V_L^L}\right)\Big] + \left(\bar{\theta} - \underline{\theta}\right)\left(p_0^L \Pr\left(V_L^L\right) f\left(\underline{\lambda}_{A_H \cup \bar{\theta} V_L^L}, 1\right) + p_1^L \Pr\left(V_L^L\right) f\left(\bar{\lambda}_{A_H \cup \bar{\theta} V_L^L}, 1\right)\right) \\
&\quad = \left(p_0^H \Pr A_H + \bar{\theta} p_0^L \Pr\left(V_L^L\right)\right)\hat{f}\left(\underline{\lambda}_{A_H \cup \bar{\theta} V_L^L}\right) + \left(p_1^H \Pr A_H + \bar{\theta} p_1^L \Pr\left(V_L^L\right)\right)\hat{f}\left(\bar{\lambda}_{A_H \cup \bar{\theta} V_L^L}\right).
\end{aligned}$$

(c) If V_L^L proposals are randomized on both sides of the threshold, the low-cost group facing V_L^L faces the incentive compatibility condition

$$p_0^L \gamma^L \hat{c}(\underline{\lambda}_x) + p_1^L\left(\gamma^L \hat{c}\left(\bar{\lambda}_x\right) + k_A + k_G\right) = \gamma^L \hat{c}(\lambda_y).$$

By inspection, $p_0^L \hat{c}(\underline{\lambda}_x) + p_1^L \hat{c}(\bar{\lambda}_x) < \hat{c}(\lambda_y)$ for the expected media cost to justify proposing below $\tilde{r}$. Also, λ_y must be the lowest value of λ for proposals of at least $\tilde{r}$ (including off the equilibrium path). This condition implies the group would induce the other venal agents to propose below the threshold. If not, they would choose $r_A \geq \tilde{r}$ for λ_y but then deviate to propose whatever induces $\underline{\lambda}_x$ and $\bar{\lambda}_x$ because $p_1^H < p_1^L$ and $\bar{\lambda}_x > \underline{\lambda}_x$. Meanwhile, U_L can achieve his highest utility by pooling with V_L^L and will propose at least $\tilde{r}$. The following claim will prove useful here and later in determining the possibilities for equilibria:

Claim 1.9(c)-1. When all of A_H and some or all of V_L^L propose below the threshold and U_L proposes separately, only one set $\left(\underline{\lambda}, \bar{\lambda}\right)$ can occur.

Proof. All proposals not coming from U_L must involve V_L^L. Otherwise, any proposal with initial $\lambda = \lambda_y$ that does not involve V_L^L comes only after the high-cost signal, in which case $\underline{\lambda} = \bar{\lambda} = \lambda_y$. For V_L^L not to defect from any of its proposals, say, one that yields λ_x before the media test, it must be that

$$p_0^L \hat{c}(\underline{\lambda}_x) + p_1^L \hat{c}\left(\bar{\lambda}_x\right) \leq \hat{c}(\lambda_y),$$

where λ_y is the lowest value of λ for a proposal without V_L^L. It must be that $\lambda_y > \underline{\lambda}_x$. (Only if $\lambda_x = 1$ can $\underline{\lambda}_x = \bar{\lambda}_x$, in which case the group would induce V_L^L to deviate, even to a policy that yields at least $\tilde{r}$.) Then V_H cannot be involved in proposals leading to a λ_y. If $\lambda_y \geq \bar{\lambda}_x$, the V_H agents would be induced to pool with V_L^L. If $\lambda_y \leq \bar{\lambda}_x$, $p_1^L\left(\hat{c}\left(\bar{\lambda}_x\right) - \hat{c}(\lambda_y)\right) \leq p_0^L(\hat{c}(\lambda_y) - \hat{c}(\underline{\lambda}_x))$ implies $p_1^H\left(\hat{c}\left(\bar{\lambda}_x\right) - \hat{c}(\lambda_y)\right) < p_0^H(\hat{c}(\lambda_y) - \hat{c}(\underline{\lambda}_x))$ (since $p_1^H < p_1^L$), and again these agents would be induced to deviate. That means that any proposal(s) without V_L^L consist(s) only of U_H, which has $\lambda = \underline{\lambda} = \bar{\lambda} = \lambda_{U_H}$. Let $\phi_m = p_L(1-q)p_m^H$, $\chi_m = (p_H + p_L(1-q))p_m^H$, and

$\psi_m = p_L q p_m^L$. This means that the posteriors after V_L^L that can be expressed as

$$\underline{\lambda}_{\theta_1 U_H \cup V_H \cup \theta_5 V_L^L} = \frac{(\theta_1 p_U + p_V)\phi_0 + \theta_5 p_V \psi_0}{(\theta_1 p_U + p_V)\chi_0 + \theta_5 p_V \psi_0} \text{ and } \bar{\lambda}_{\theta_1 U_H \cup V_H \cup \theta_5 V_L^L} = \frac{(\theta_1 p_U + p_V)\phi_1 + \theta_5 p_V \psi_1}{(\theta_1 p_U + p_V)\chi_1 + \theta_5 p_V \psi_1},$$

where θ_1 and θ_5 are the fractions of U_H and V_L^L involved in these proposals, relative to the fraction of V_H. Because $\bar{\lambda}_{\theta_1 U_H \cup V_H \cup \theta_5 V_L^L} > \lambda_{U_H}$, the venal agents would deviate unless the proposal yielded some $\underline{\lambda} \leq \lambda_{U_H}$. However, there would then need to be another proposal for which $\bar{\lambda} > \underline{\lambda} > \underline{\lambda}_{avg} > \lambda_{U_H}$, from which the venal agents would be induced to deviate. Thus, all proposals not coming from U_L must involve V_L^L.

If there are multiple values of $\left(\underline{\lambda}, \bar{\lambda}\right)$ it must be the case that, for any $\left(\underline{\lambda}_x, \bar{\lambda}_x\right)$ and $\left(\underline{\lambda}_y, \bar{\lambda}_y\right)$ with $\underline{\lambda}_x < \underline{\lambda}_y$, $\bar{\lambda}_y < \bar{\lambda}_x$ so that $p_0^L \hat{c}(\underline{\lambda}_x) + p_1^L \hat{c}\left(\bar{\lambda}_x\right) = p_0^L \hat{c}\left(\underline{\lambda}_y\right) + p_1^L \hat{c}\left(\bar{\lambda}_y\right)$. Because V_H has $p_1^L < p_1^H$, $p_1^L\left(\hat{c}\left(\bar{\lambda}_x\right) - \hat{c}\left(\bar{\lambda}_y\right)\right) = p_0^L\left(\hat{c}\left(\underline{\lambda}_y\right) - \hat{c}(\underline{\lambda}_x)\right)$ implies $p_1^H\left(\hat{c}\left(\bar{\lambda}_x\right) - \hat{c}\left(\bar{\lambda}_y\right)\right) < p_0^H\left(\hat{c}\left(\underline{\lambda}_y\right) - \hat{c}(\underline{\lambda}_x)\right)$. Then all proposals after V_H must yield $(\min(\underline{\lambda}), \max(\bar{\lambda}))$ for the principal. Since this is a single set of values, they can be expressed as $\min(\underline{\lambda}) = \underline{\lambda}_{\theta_1 U_H \cup V_H \cup \theta_5 V_L^L}$ and $\max\left(\bar{\lambda}\right) = \bar{\lambda}_{\theta_1 U_H \cup V_H \cup \theta_5 V_L^L}$ for some values of θ_1 and θ_5, because the final policy depends only on whether there was a media report. Meanwhile, the other values of $\left(\underline{\lambda}, \bar{\lambda}\right)$, which are from proposals that do not include V_H, can be expressed as

$$\left(\underline{\lambda}_{\dot{\theta}_1 U_H \cup \dot{\theta}_5 V_L^L}, \bar{\lambda}_{\dot{\theta}_1 U_H \cup \dot{\theta}_5 V_L^L}\right) = \left(\frac{\dot{\theta}_1 p_U \phi_0 + \dot{\theta}_5 p_V \psi_0}{\dot{\theta}_1 p_U \chi_0 + \dot{\theta}_5 p_V \psi_0}, \frac{\dot{\theta}_1 p_U \phi_1 + \dot{\theta}_5 p_V \psi_1}{\dot{\theta}_1 p_U \chi_1 + \dot{\theta}_5 p_V \psi_1}\right)$$

for some $\dot{\theta}_1$ and $\dot{\theta}_5$. Thus, all values of $\left(\underline{\lambda}, \bar{\lambda}\right)$ for proposals below $\tilde{r}$ are expressible as

$$\left(\underline{\lambda}, \bar{\lambda}\right) = \left(\frac{\zeta\phi_0 + \eta\psi_0}{\zeta\chi_0 + \eta\psi_0}\bar{\lambda}, \frac{\zeta\phi_1 + \eta\psi_1}{\zeta\chi_1 + \eta\psi_1}\right).$$

However, $\underline{\lambda}_x \equiv \frac{\zeta_x\phi_0 + \eta_x\psi_0}{\zeta_x\chi_0 + \eta_x\psi_0} < \underline{\lambda}_y \equiv \frac{\zeta_y\phi_0 + \eta_y\psi_0}{\zeta_y\chi_0 + \eta_y\psi_0}$ if and only if $(\zeta_x\phi_0 + \eta_x\psi_0)(\zeta_y\chi_0 + \eta_y\psi_0) < (\zeta_x\chi_0 +$

$\eta_x\psi_0)(\zeta_y\phi_0 + \eta_y\psi_0)$, or $\zeta_y\eta_x < \zeta_x\eta_y$, and the same condition implies$\bar{\lambda}_x < \bar{\lambda}_y$, analogously defined. This contradicts $\underline{\lambda}_x < \underline{\lambda}_y$ implying $\bar{\lambda}_y < \bar{\lambda}_x$ for multiple values of $\left(\underline{\lambda}, \bar{\lambda}\right)$. Therefore, a single $\left(\underline{\lambda}, \bar{\lambda}\right)$ obtains. □

Since a single $\left(\underline{\lambda}, \bar{\lambda}\right)$ occurs, the principal can optimally select policy solely based on whether there is a media report. The values are $\underline{\lambda}_{A_H \cup \theta V_L^L}$ and $\bar{\lambda}_{A_H \cup \theta V_L^L}$. With the other $(1-\theta)$ of V_L^L and U_L^L being assigned $r_P = \hat{r}(1)$, the resulting equilibrium has the same payoff and incentive compatibility condition for V_L^L equilibrium (ii) in part (a).

If V_L^L proposals are all below the threshold, and U_L proposes at least threshold, the incentive compatibility condition for V_L^L is

$$p_0^L \gamma^L \hat{c}(\underline{\lambda}_x) + p_1^L(\gamma^L \hat{c}(\bar{\lambda}_x) + k_A + k_G) \leq \gamma^L \hat{c}(\lambda_y).$$

Following the proof for V_L^L randomizing on both sides, the group must be incentivized to induce V_H agents to propose below the threshold, as well. With A_H and V_L^L proposals less than $\tilde{r}$, Claim 1.9(c)-1 implies a single set of $\left(\underline{\lambda}, \bar{\lambda}\right)$, i.e., $\left(\underline{\lambda}_{A_H \cup V_L^L}, \bar{\lambda}_{A_H \cup V_L^L}\right)$, which results in the same payoff as equilibrium (i) in part (a). (Here, the set of equilibria are restricted to those with U_L proposing at least $\tilde{r}$.)

If V_L^L and U_L proposals are all at least $\tilde{r}$, while U_H proposals are less than $\tilde{r}$, left to be determined are the fractions θ_1 of V_H^H proposals and θ_2 of V_H^L proposals that are below $\tilde{r}$. Proposals with fraction $\theta_1 < 1$ and $\theta_2 > 0$ are not incentive compatible. The two group types seek the lowest policies for proposals on both sides of the threshold, and with only one signal below the threshold, only one policy results from a particular proposal. Thus, on either side of the threshold, the group facing the venal agent with $s = H$ must pool so as to receive the same policy. The low-cost group tests $\gamma^L \hat{c}(\lambda_x) + p_1^H(k_A + k_G) \leq \gamma^L \hat{c}(\lambda_y)$. If this constraint is

satisfied, $\gamma^H > \gamma^L$ implies $\gamma^H \hat{c}(\lambda_x) + p_1^H(k_A + k_G) < \gamma^H \hat{c}(\lambda_y)$, which means $\theta_1 = 1$. Proposals with $\theta_1 = 1$ and $\theta_2 < 1$ are prevented by the restriction $k_A + k_G < \gamma^L(\hat{c}(1) - \hat{c}(\lambda_{A_H}))$. The low-cost group would deviate from any such equilibrium, since its worst payoff from deviating would be $\gamma^L \hat{c}\Big(\lambda_{U_H \cup V_H^H \cup \theta V_H^L}\Big) + p_1^H(k_A + k_G) < \gamma^L \hat{c}(\lambda_{A_H}) + (k_A + k_G) < \gamma^L \hat{c}(1)$. Finally, proposals with $\theta_2 = 0$ are precluded by the restriction $p_1^H(k_A + k_G) < \gamma^H(\hat{c}(p_L) - \hat{c}(\lambda_{A_H}))$. The high-cost group's worst payoff from deviating would be $\gamma^L \hat{c}\Big(\lambda_{U_H \cup \theta V_H^H}\Big) + p_1^H(k_A + k_G) \leq \gamma^L \hat{c}(\lambda_{A_H}) + p_1^H(k_A + k_G) < \gamma^L \hat{c}(p_L)$, the best payoff when U_L proposals are separated from the other proposals. Thus, only $\theta_1 = \theta_2$ is possible. Then only one value of λ, i.e., λ_{A_H}, is possible. If there were more than value of λ, the minimum value would some $\lambda_{min} < \lambda_{A_H}$. To have $\theta_1 = \theta_2$ would imply $\lambda_{min} = \lambda_{A_H}$, so $\lambda_{min} < \lambda_{A_H}$ implies $\theta_1 < \theta_2$. Then some percentage of V_H^L proposals would yield more than λ_{min}, in which case the low-cost group would deviate. With only one value of λ, the payoff must be the same as the equilibrium described in (a)(iii).

(d) If $p_1^L(k_A + k_G) < \gamma^L(\hat{c}(1) - \hat{c}(\lambda_{A_H}))$, then some $\theta > 0$ of V_L^L proposals appear below the media threshold. For p_1^L increasing, the low-cost group tests the difference $p_0^L \gamma^L \hat{c}\Big(\underline{\lambda}_{A_H \cup \theta V_L^L}\Big) + p_1^L\Big(\gamma^L \hat{c}\Big(\bar{\lambda}_{A_H \cup \theta V_L^L}\Big) + k_A + k_G\Big) - \gamma^L \hat{c}(1)$. Applying Lemma 1.4 with $\eta_U = \eta_V = 1$ reveals that the derivative of this expression is positive. Then any equilibrium that previously existed with $\theta \in (0, 1)$ is no longer incentive compatible for the low-cost group facing V_L^L. Rebalancing requires decreasing θ until equality is restored or until $\theta = 0$ for equilibrium (iii). For $\theta = 0$, the positive derivative causes the low-cost group to favor V_L^L proposals that avoid a media report even more. For $\theta = 1$, the positive derivative implies that increasing transparency either breaks the equilibrium, requiring $\theta < 1$ or merely reduces the benefit to the low-cost group of inducing V_L^L proposals below $\tilde{r}$. Thus, an increase in transparency means that the incentive compatibility condition that is satisfied

will be for a weakly lower value of θ or for the same value of θ. Part (b) states that, for given values of p_1^L and p_1^H, the payoff increases as θ decreases. However, p_1^L increases, so the principal's payoff is even higher. The payoffs from the proposals that are at least $\tilde{r}$ are the same. From proposals below the threshold, for any $\underline{p}_0^L < \bar{p}_0^L$, the payoff is higher with $\underline{p}_0^L$. Suppressing the subscript $A_H \cup \theta V_L^L$ for the posterior probabilities, the comparison is $\big(p_0^H \Pr A_H + \theta \underline{p}_0^L \Pr\big(V_L^L\big)\big)\hat{f}\big(\underline{\lambda}\big(\underline{p}_0^L\big)\big) + \big(p_1^H \Pr A_H + \theta\big(1-\underline{p}_0^L\big)\Pr\big(V_L^L\big)\big)\hat{f}\big(\bar{\lambda}\big(\underline{p}_0^L\big)\big) > \big(p_0^H \Pr A_H + \theta \underline{p}_0^L \Pr\big(V_L^L\big)\big) f\big(\underline{\lambda}\big(\bar{p}_0^L\big), \underline{\lambda}\big(\underline{p}_0^L\big)\big) + \theta\big(\bar{p}_0^L - \underline{p}_0^L\big)\Pr\big(V_L^L\big) f\big(\underline{\lambda}\big(\bar{p}_0^L\big), \bar{\lambda}\big(\underline{p}_0^L\big)\big) + \big(p_1^H \Pr A_H + \theta\big(1 - \bar{p}_0^L\big)\Pr\big(V_L^L\big)\big) f\big(\bar{\lambda}\big(\bar{p}_0^L\big), \bar{\lambda}\big(\underline{p}_0^L\big)\big) = \big(p_0^H \Pr A_H + \theta \bar{p}_0^L \Pr\big(V_L^L\big)\big)\hat{f}\big(\underline{\lambda}\big(\bar{p}_0^L\big)\big) + \big(p_1^H \Pr A_H + \theta\big(1 - \bar{p}_0^L\big)\Pr\big(V_L^L\big)\big)\hat{f}\big(\bar{\lambda}\big(\bar{p}_0^L\big)\big)$. (From Equations (1.3) and (1.4), one can derive that $\underline{\lambda}_{A_H \cup \theta V_L^L}$ increases while $\bar{\lambda}_{A_H \cup \theta V_L^L}$ decreases with p_0^L.) Thus, the principal's payoff is strictly increasing when $\theta > 0$ originally.

For p_1^H increasing, if $p_0^L \gamma^L(\hat{c}(1) - \hat{c}(\lambda_{A_H})) \leq p_1^L(k_A + k_G) < \gamma^L(\hat{c}(1) - \hat{c}(\lambda_{A_H}))$ when $p_1^H = 0$, increasing p_1^H from zero by any amount will cause a switch from $\theta = 0$ of V_L^L proposals below $\tilde{r}$ in equilibrium to some $\theta > 0$ in equilibrium. When it is not the case that $p_1^H = \theta = 0$, the low-cost group will again test the difference $p_0^L \gamma^L \hat{c}(\underline{\lambda}_{A_H \cup \theta V_L^L}) + p_1^L(\gamma^L \hat{c}(\bar{\lambda}_{A_H \cup \theta V_L^L}) + k_A + k_G) - \gamma^L \hat{c}(1)$. Applying Lemma 1.8 with $\eta_U = \eta_V = 1$ yields the fact that the derivative of the expression is negative. If the equilibrium has $\theta \in (0, 1)$, then the incentive compatibility constraint no longer holds; $p_0^L \gamma^L \hat{c}(\underline{\lambda}_{A_H \cup \theta V_L^L}) + p_1^L(\gamma^L \hat{c}(\bar{\lambda}_{A_H \cup \theta V_L^L}) + k_A + k_G) < \gamma^L \hat{c}(1)$, and rebalancing requires θ to increase until equality is restored or until $\theta = 1$. If $\theta = 1$, the negative derivative implies that V_L^L proposals remain fully below the threshold. Thus, in all cases, the equilibrium value of θ weakly increases. For the same p_1^H, part (b) indicates that a higher θ reduces the principal's payoff.

Increasing p_1^H also affects the value of the equilibrium at θ due to the reduction in information from the media signal. The proof is complete if increasing p_1^H decreases the value

of an equilibrium at θ, because then whether θ increases or stays the same, the principal's payoff is lower. Looking at Equations (1.3) and (1.4) reveals that $\underline{\lambda}_{A_H \cup \theta V_L^L}(p_1^H)$ increases while $\bar{\lambda}_{A_H \cup \theta V_L^L}(p_1^H)$ decreases with p_1^H. For any two values of p_1^H, $\underline{p}_1^H < \bar{p}_1^H$, the expected payoff is lower with $\bar{p}_1^H$ from the proposals below the threhsold. The comparison, with the subscript $A_H \cup \theta V_L^L$ omitted, is $\big(\big(1-\bar{p}_1^H\big)Pr(A_H) + \theta p_0^L \Pr\big(V_L^L\big)\big)\hat{f}\big(\underline{\lambda}\big(\bar{p}_1^H\big)\big) + \big(\bar{p}_1^H \Pr(A_H) + \theta p_1^L \Pr\big(V_L^L\big)\big)\hat{f}\big(\bar{\lambda}\big(\bar{p}_1^H\big)\big) = \big(\big(1-\bar{p}_1^H\big)\Pr(A_H) + \theta p_0^L \Pr\big(V_L^L\big)\big) f\big(\underline{\lambda}\big(\bar{p}_1^H\big), \underline{\lambda}\big(\underline{p}_1^H\big)\big) + \big(\bar{p}_1^H - \underline{p}_1^H\big)\Pr(A_H) f\big(\bar{\lambda}\big(\bar{p}_1^H\big), \underline{\lambda}\big(\underline{p}_1^H\big)\big) + \big(\underline{p}_1^H \Pr(A_H) + \theta p_1^L \Pr\big(V_L^L\big)\big) f\big(\bar{\lambda}\big(\bar{p}_1^H\big), \bar{\lambda}\big(\underline{p}_1^H\big)\big) < \big(\big(1 - \underline{p}_1^H\big)\Pr(A_H) + \theta p_0^L \Pr\big(V_L^L\big)\big)\hat{f}\big(\underline{\lambda}\big(\underline{p}_1^H\big)\big) + \big(\underline{p}_1^H \Pr(A_H) + \theta p_1^L \Pr\big(V_L^L\big)\big)\hat{f}\big(\bar{\lambda}\big(\underline{p}_1^H\big)\big)$. The payoff from the proposals that are at least $\tilde{r}$ is the same for either value of p_1^H, so for any $\theta \in (0,1]$, the payoff from the equilibrium decreases with p_1^H.

(e) As p_1^H approaches p_1^L, $\underline{\lambda}_{A_H \cup \theta V_L^L}$ and $\bar{\lambda}_{A_H \cup \theta V_L^L}$ both approach $\lambda_{A_H \cup \theta V_L^L}$. If $p_1^H = p_1^L$, the low-cost group proposes at least sometimes above the threshold if $p_0^L \gamma^L \hat{c}\Big(\lambda_{A_H \cup \theta V_L^L}\Big) + p_1^L\Big(\gamma^L \hat{c}\Big(\lambda_{A_H \cup \theta V_L^L}\Big) + k_A + k_G\Big) \geq \gamma^L \hat{c}(1)$. By inspection, $\lambda_{A_H \cup \theta V_L^L}$ increases with θ, so the condition $p_1^L(k_A + k_G) \geq \gamma^L\Big(\hat{c}(1) - \hat{c}\Big(\lambda_{A_H \cup V_L^L}\Big)\Big)$ implies that $p_1^L(k_A + k_G) < \gamma^L\Big(\hat{c}(1) - \hat{c}\Big(\lambda_{A_H \cup \theta V_L^L}\Big)\Big)$ for any $\theta < 1$, so $p_1^H = p_1^L$ in this situation implies that V_L^L proposals and A_H proposals all are below the threshold and yield $\lambda_{A_H \cup V_L^L}$. Then the principal's payoff is $\big(\Pr(A_H) + \Pr\big(V_L^L\big)\big)\hat{f}\Big(\lambda_{A_H \cup V_L^L}\Big) + \Pr(U_L)\hat{f}(1) = \Pr(\Omega \backslash U_L)\hat{f}\big(\lambda_{\Omega \backslash U_L}\big) + \Pr(U_L)\hat{f}(1)$. This part of the proposition then follows from the fact that $\underline{\lambda}_{A_H \cup \theta V_L^L}$ and $\bar{\lambda}_{A_H \cup \theta V_L^L}$ vary continuously with p_1^H. ■

Proof of Proposition 1.10 The proof begins by establishing a claim similar to Claim 1.9(c)-1.

Claim 1.10-1. In this form of equilibrium only one $\big(\underline{\lambda}, \bar{\lambda}\big)$ obtains below the threshold.

Proof. We can apply many of the steps from the proof of Claim 1.9(c)-1. A key step was that

all proposals not involving V_L^L there had to consist only of U_H. However, because there are no U_H proposals below $\tilde{r}$, all proposals under the threshold involve V_L^L. If there are multiple values of $\left(\underline{\lambda}, \bar{\lambda}\right)$, we can apply more steps from the proof of Claim 1.9(c)-1 to shows that all proposals after V_H must yield $(\min(\underline{\lambda}), \max(\bar{\lambda}))$ for the principal. All proposals must involve V_H; if they didn't, they would involve V_L^L only with $\lambda = 1$, which would lead to defection by V_L^L. □

This claim implicitly allows the direct use of particular values of $\underline{\lambda}$ and $\bar{\lambda}$ in any proposed equilibrium.

(a) If U_H were by himself, incentive compatibility for V_L^L would require $p_0^L \gamma^L \hat{c}(\underline{\lambda}_V) + p_1^L\left(\gamma^L \hat{c}\left(\bar{\lambda}_V\right) + k_A + k_G\right) \leq \gamma^L \hat{c}(\lambda_{U_H})$. However, $\underline{\lambda}_V \geq \lambda_{U_H}$, so V_L^L would defect from this equilibrium. Meanwhile, for any equilibrium in which some $\theta \in (0,1)$ of V_L^L proposals remain pooled with V_H, $\theta < p_V$ if $\hat{c}(\lambda)$ is convex. Then $p_0^L \gamma^L \hat{c}\left(\underline{\lambda}_{V_H \cup \theta V_L^L}\right) + p_1^L \gamma^L \hat{c}\left(\bar{\lambda}_{V_H \cup \theta V_L^L}\right) \leq \gamma^L \hat{c}\left(\lambda_{U_H \cup (1-\theta) V_L^L}\right) - k_A - k_G$ implies that $\lambda_{V_H \cup \theta V_L^L} < \lambda_{U_H \cup (1-\theta) V_L^L}$, since convexity means $p_0^L \hat{c}\left(\underline{\lambda}_{V_H \cup \theta V_L^L}\right) + p_1^L \hat{c}\left(\bar{\lambda}_{V_H \cup \theta V_L^L}\right) \geq \hat{c}\left(\lambda_{V_H \cup \theta V_L^L}\right)$. Solving the inequality $\lambda_{V_H \cup \theta V_L^L} < \lambda_{U_H \cup (1-\theta) V_L^L}$ for θ implies $1 - \theta > p_U$.

(b) Starting from an equilibrium in which fraction θ of V_L^L proposals are below the threshold, the incentive compatibility condition for V_L^L is

$$p_0^L \gamma^L \hat{c}\left(\underline{\lambda}_{V_H \cup \theta V_L^L}\right) + p_1^L \left(\gamma^L \hat{c}\left(\bar{\lambda}_{V_H \cup \theta V_L^L}\right) + k_A + k_G\right) = \gamma^L \hat{c}\left(\lambda_{U_H \cup (1-\theta) V_L^L}\right). \tag{B.15}$$

Lemma 1.4 implies that the left-hand side of Inequality (B.15) increases with p_1^L, while Lemma 1.8 implies that it decreases with p_1^H. Meanwhile, the right-hand side of (B.15) does not change with either of these parameters. Thus, to be incentive compatible for V_L^L, θ must decrease with p_1^L and increase with p_1^H.

(c) In the non-existent equilibrium for U_H, U_H's incentive compatibility constraint would be fully met:

$$p_0^H \alpha f(\underline{\lambda}_V, \lambda_{U_H}) + p_1^H (\alpha f(\bar{\lambda}_V, \lambda_{U_H}) - k_A) < \alpha \hat{f}(\lambda_{U_H}). \tag{B.16}$$

As more V_L^L proposals pool with U_H (i.e., θ decreases), $\underline{\lambda}_{V_H \cup \theta V_L^L}$ and $\bar{\lambda}_{V_H \cup \theta V_L^L}$ decrease toward λ_{U_H}, and $\lambda_{U_H \cup (1-\theta) V_L^L}$ increases away from λ_{U_H}. Thus, the left-hand side of Inequality (B.16) increases and the RHS decreases as θ approaches 0. For sufficiently large p_1^L, θ will be small enough that $\bar{\lambda}_{V_H \cup \theta V_L^L} < \lambda_{U_H \cup (1-\theta) V_L^L}$. For sufficiently small p_1^H, V_H proposals will stay below the threshold. Then $p_0^H f\left(\underline{\lambda}_{V_H \cup \theta V_L^L}, \lambda_{U_H}\right) + p_1^H f\left(\bar{\lambda}_{V_H \cup \theta V_L^L}, \lambda_{U_H}\right) > f\left(\lambda_{U_H \cup (1-\theta) V_L^L}, \lambda_{U_H}\right)$, and for α large enough, U_H can be made to deviate to propose below the threshold with V_H (and θ of V_L^L). Then for some smaller values of p_1^L, it will also be the case that $p_0^H f\left(\underline{\lambda}_{V_H \cup \theta V_L^L}, \lambda_{U_H}\right) + p_1^H f\left(\bar{\lambda}_{V_H \cup \theta V_L^L}, \lambda_{U_H}\right) > f\left(\lambda_{U_H \cup (1-\theta) V_L^L}, \lambda_{U_H}\right)$, even if $\bar{\lambda}_{V_H \cup \theta V_L^L} > \lambda_{U_H \cup (1-\theta) V_L^L}$, since $\lambda_{V_H \cup \theta V_L^L} < \lambda_{U_H \cup (1-\theta) V_L^L}$ for any equilibrium of the form described in this proposition. In this case, as well, a sufficiently large α makes it possible that U_H will defect below the threshold.

(d) Claim 1.10-1 implies that the best equilibrium payoff with U_H proposing $r_A < \tilde{r}$ when there are V_L^L proposals also below the threshold involve full pooling. In this case, the incentive compatibility condition is $p_0^L \gamma^L \hat{c}\left(\underline{\lambda}_{V_H \cup \theta V_L^L}\right) + p_1^L \left(\gamma^L \hat{c}\left(\bar{\lambda}_{V_H \cup \theta V_L^L}\right) + k_A + k_G\right) = \gamma^L \hat{c}(1)$ for $\theta \in (0, 1)$. Since $\lambda_{U_H \cup (1-\eta) V_L^L} < 1$, for any η, the incentive compatibility condition $p_0^L \gamma^L \hat{c}\left(\underline{\lambda}_{V_H \cup \eta V_L^L}\right) + p_1^L \left(\gamma^L \hat{c}\left(\bar{\lambda}_{V_H \cup \eta V_L^L}\right) + k_A + k_G\right) = \gamma^L \hat{c}\left(\lambda_{U_H \cup (1-\eta) V_L^L}\right)$ is satisfied with $\eta < \theta$, or the rebalancing is accomplished with $\eta = 0$. If the original equilibrium involves only U_L proposing at least the threshold, then part (a) of this proposition automatically implies that less of the V_L^L proposals will be below the threshold.

(e) For a venal agent proposing below $\tilde{r}$, incentive compatibility constraint is then of the

form

$$\gamma^i \hat{c}(\lambda_{V_H}) + p_1^H(k_A + k_G) \leq \gamma^i \hat{c}(1) \text{ or } \gamma^i \hat{c}(\lambda_{V_H}) + p_1^H(k_A + k_G) = \gamma^i \hat{c}(1) \tag{B.17}$$

for some $i \in \{H, L\}$. Either way, moving U_H proposals to at least $\tilde{r}$ means that the right-hand side of either constraint in (B.17) will decrease while the left-hand stays the same, meaning that there can be only more venal agents with U_H setting $r_A \geq \tilde{r}$. ■

B.2 Proofs of Results for Chapter 2

Proof of Lemma 2.1 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 $EU_t^q(x_t^q) = EU^q(t)$ for $t \in \{h, i, j, k, l\}$ implies that $EU_j^q(x_j^q) = EU^q(\tau_i i + \tau_k k) < \tau_i EU^q(i) + \tau_k EU^q(k) = \tau_i EU_i^q(x_i^q)) + \tau_k EU_k^q(x_k^q) = (\alpha\tau_h + (1-\alpha)\tau_l) EU^q\big(\frac{\alpha\tau_h h + (1-\alpha)\tau_l l}{\alpha\tau_h + (1-\alpha)\tau_l}\big) + (\alpha\tau_l + (1-\alpha)\tau_l) EU^q\big(\frac{\alpha\tau_l l + (1-\alpha)\tau_h h}{\alpha\tau_l + (1-\alpha)\tau_h}\big) < \alpha\tau_h EU^q(h) + (1-\alpha)\tau_l EU^q(l) + \alpha\tau_l EU^q(l) + (1-\alpha)\tau_h EU^q(h) = \tau_h EU_h^q(x_h^q)) + \tau_l EU_l^q(x_l^q)$. ■

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 2.2 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:

$$(\pi + \mathbf{1}_m \Delta\pi_m) EU_i^A(x^P) + (1 - \pi - \mathbf{1}_m \Delta\pi_m) EU_i^A(x^A)$$
$$\leq (\pi + \mathbf{1}_m^{tr} \Delta\pi_m) EU_i^A(x_i^P) + (1 - \pi - \mathbf{1}_m^{tr} \Delta\pi_m) EU_i^A(x_i^A), \forall x^P \geq x_i^P, \quad \text{(B.18)}$$

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.18) 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.18) 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.18) 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 $A_{\tilde{L}}$:

$$\begin{aligned}&(\pi + \mathbf{1}_m \Delta\pi_m + \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}}) EU_k^A(x^P) + (1 - \pi - \mathbf{1}_m \Delta\pi_m - \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}}) EU_k^A(x^A) \\ &\leq (\pi + \mathbf{1}_m^{tr} \Delta\pi_m + \mathbf{1}_{\tilde{L}}^{tr} \Delta\pi_{\tilde{L}}) EU_k^A(x_k^P) + (1 - \pi - \mathbf{1}_m^{tr} \Delta\pi_m - \mathbf{1}_{\tilde{L}}^{tr} \Delta\pi_{\tilde{L}}) EU_k^A(x_k^A), \forall x^P \geq x_k^P,\end{aligned} \tag{B.19}$$

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.19) 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.19). 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.19) 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.18), 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.18) 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.18) holds, it must be that

$$(\pi + \mathbf{1}_m\Delta\pi_m + \mathbf{1}_{\tilde{L}}\Delta\pi_{\tilde{L}})EU_k^A(x_k^P) + (1 - \pi - \mathbf{1}_m\Delta\pi_m - \mathbf{1}_{\tilde{L}}\Delta\pi_{\tilde{L}})EU_k^A(x^A) < (\pi + \mathbf{1}_m^{tr}\Delta\pi_m)EU_k^A(\hat{x}^P) + (1 - \pi - \mathbf{1}_m^{tr}\Delta\pi_m)EU_k^A(x_k^A). \quad \text{(B.20)}$$

Since $EU_k^A(x)$ is concave with respect to x, Inequalities (B.19) and (B.20) 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.19) 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 2.4–2.8:

Lemma B.2. *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:

$$\begin{aligned}
&\alpha((\pi + \mathbf{1}_m \Delta\pi_m)c(x_i^P) + (1 - \pi - \mathbf{1}_m \Delta\pi_m)c(x_i^A)) \\
&\qquad + (1-\alpha)(\pi + \mathbf{1}_m \Delta\pi_m + \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}})c(x_k^P) + (1 - \pi - \mathbf{1}_m \Delta\pi_m - \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}})c(x_k^A)) \\
&< (1-\alpha)((\pi + \mathbf{1}_m \Delta\pi_m)c(x_i^P) + (1 - \pi - \mathbf{1}_m \Delta\pi_m)c(x_i^A)) \\
&\qquad + \alpha(\pi + \mathbf{1}_m \Delta\pi_m + \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}})c(x_k^P) + (1 - \pi - \mathbf{1}_m \Delta\pi_m - \mathbf{1}_{\tilde{L}} \Delta\pi_{\tilde{L}})c(x_k^A)), \qquad \text{(B.21)}
\end{aligned}$$

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 2.4 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:

$$\tau_i(\pi EU_i^A(x_i^P) + (1-\pi)EU_i^A(x_i^A)) + \tau_k(\pi EU_k^A(x_k^P) + (1-\pi)EU_k^A(x_k^A)) \\ > \pi EU_j^A(x^P) + (1-\pi)EU_j^A(x^A), \forall x^P \geq x_k^P, \quad \text{(B.22)}$$

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 2.1, 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.2 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.2. 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.18) 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.18) 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.18) 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.19) 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.18) 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.22) 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.18) [(B.19)] 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

$$\tau_i((\pi + \mathbf{1}_m^{tr}\Delta\pi_m)EU_i^A(x_i^P) + (1 - \pi - \mathbf{1}_m^{tr}\Delta\pi_m)EU_i^A(x_i^A)) + \tau_k\big[(\pi + \mathbf{1}_m^{tr})EU_k^A(x_k^P) + (1 - \pi - \mathbf{1}_m^{tr}\Delta\pi_m)EU_k^A(x_k^A)\big] < (\pi + \mathbf{1}_m\Delta\pi_m)EU_j^A(x^P) + (1 - \pi - \mathbf{1}_m\Delta\pi_m)EU_j^A(x^A) \tag{B.23}$$

for some x^A and $x^P \geq x_i^P$. Because Inequality (B.18) holds for $A_{\tilde{H}}$ (whose equiibrium payoff is on the right-hand side), it must be the case that

$$(\pi + \mathbf{1}_m^{tr}\Delta\pi_m)EU_k^A(x_k^P) + (1 - \pi - \mathbf{1}_m^{tr}\Delta\pi_m)EU_k^A(x_k^A) \\ < (\pi + \mathbf{1}_m\Delta\pi_m)EU_k^A(x^P) + (1 - \pi - \mathbf{1}_m\Delta\pi_m)EU_k^A(x^A), \quad \text{(B.24)}$$

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 2.2 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

$$\tau_i(\pi EU_i^P(x_i^P) + (1-\pi)EU_i^P(x_i^A)) + \tau_k(\pi EU_k^P(x_k^P) + (1-\pi)EU_k^P(x_k^A)). \quad \text{(B.25)}$$

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 2.5 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 2.4, 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 2.4 holds. By Lemma B.2 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.2 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 (2.4).

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 (2.3) 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 (2.3) 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.18) [(B.19)] 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 (2.3) and (2.4) are sufficient, as well as necessary, for a natural message-signal equilibrium.

For the first statement after Inequality (2.4), Theorem 2.2 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 (2.3) fails while Inequality (2.4) 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 (2.4) 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 (2.4) fails while Inequality (2.3) 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 2.6 The following message-signal equilibrium will be shown to exist and satisfy the refinements if and only if Inequalities (2.5) and (2.6) 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

(2.6), and Lemma B.2 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.2 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 (2.5) 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 (2.6) 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 (2.5) 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 (2.5) 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 (2.5) 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 (2.5) implies again that A will deviate with $\mathring{d} = (\emptyset, \emptyset, x_j^A)$, and again the equilibrium does not survive. If, however, Inequality (2.5) 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 (2.5) 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 2.5 involving $\mathring{d} = (m, \tilde{H}, \cdot)$ $[(m, \tilde{L}, \cdot)]$ can be applied. (In Inequality (B.19) $\mathbf{1}_{\tilde{L}}^{tr} = \mathbf{1}_{\tilde{L}} = 1$.)

Considering beliefs for off-equilibrium path disclosures yielded no additional conditions for this equilibrium. Therefore, Inequalities (2.5) and (2.6) 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 2.2 limits the possibilities for such an equilibrium to those that yield L the same payoff, the right-hand side of Inequality (2.7). 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 2.2. Theorem 2.2 also limits the possibilities for a natural message-signal equilibrium to those that yield A the same payoff, the right-hand side of Inequality (2.5). 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 (2.5), 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 (2.6), Theorem 2.2 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 2.2, 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 2.5 and substituting these inequalities respectively for Inequalities (2.3) and (2.4). ■

Proof of Proposition 2.7 The following message-signal equilibrium will be shown to exist and satisfy the refinements if and only if Inequality (2.7) 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.22) holds. Finally, at stage 2, a necessary and sufficient condition for L not to defect is embodied in Inequality (2.7), and Lemma B.2 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 2.5 when $\mathring{d} = (\emptyset, \emptyset, \cdot)$, satisfaction of Inequality (2.7) 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 2.4 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 (2.7) is a sufficient condition for it.

If Inequality (2.7) does not hold, no other natural message-signal equilibrium can be sustained. Theorem 2.2 limits the possibilities for such an equilibrium to those that yield L the same payoff, the right-hand side of Inequality (2.7). 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 2.2.

For the first statement after Inequality (2.7), Theorem 2.2 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 2.2, 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 2.8 When $\Delta\pi_m = \Delta\pi_{\tilde{L}} = 0$, the low-cost target's individual rationality constraint becomes

$$\pi c(x_l^P) + (1-\pi)c(x_l^A) > \pi((1-\alpha)c(x_i^P) + \alpha c(x_k^P)) + (1-\pi)((1-\alpha)c(x_i^A) + \alpha c(x_k^A)), \quad \text{(B.26)}$$

which holds since $x_l^P > x_k^P > x_i^P$ and $x_l^A > x_k^A > x_i^A$. Then Inequalities (2.4), (2.6), and (2.7) are all satisfied, and L would not defect in any transparency mode. By Lemma B.2, H would not defect, either. Thus, only Inequality (2.3) or (2.5), which are the same when $\Delta\pi_m = \Delta\pi_{\tilde{L}} = 0$, can fail. Then Propositions 2.4 and 2.7 imply that there always exist natural message-signal equilibria, and Corollary 2.3 implies that there exists one in which A discloses everything. Propositions 2.5 and 2.6, imply Inequality (2.3) 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 $\pi EU_j^P(x_j^P) + (1-\pi)EU_j^P(x_j^A) \leq EU_j^P(x_j^P)$. ■

B.3 Proofs of Results for Chapter 3

Proof of Lemma 3.1 Using Bayes' rule, if $s = \tilde{x}^i$, then $x = s$ is always preferred to $x = 1 - s$:

$$\frac{g(e)q_{\tilde{x}^i}b^i_{\tilde{x}^i}}{g(e)q_{\tilde{x}^i} + (1-g(e))q_{1-\tilde{x}^i}} \geq \frac{(1-g(e))q_{1-\tilde{x}^i}b^i_{1-\tilde{x}^i}}{g(e)q_{\tilde{x}^i} + (1-g(e))q_{1-\tilde{x}^i}} \quad \text{(B.27)}$$

since $q_{\tilde{x}^i} b^i_{\tilde{x}^i} > q_{1-\tilde{x}^i} b^i_{1-\tilde{x}^i}$ by definition and $g(e) \geq \frac{1}{2}$. On the other hand, if $s = 1 - \tilde{x}^i$, Bayes' rule implies $x = s$ is preferred to $x = 1 - s$ when

$$\frac{g(e) q_{1-\tilde{x}^i} b^i_{1-\tilde{x}^i}}{g(e) q_{1-\tilde{x}^i} + (1 - g(e)) q_{\tilde{x}^i}} \geq \frac{(1 - g(e)) q_{\tilde{x}^i} b^i_{\tilde{x}^i}}{g(e) q_{1-\tilde{x}^i} + (1 - g(e)) q_{\tilde{x}^i}}. \tag{B.28}$$

Algebra yields $e \geq e^i$ as defined in the Lemma to satisfy the inequality in the Lemma. ■

The next two lemmas are each used to prove more than one of the numbered results in the text and build on Lemma 3.1:

Lemma B.3. *The strategy for a decision-maker $i \in \{L, A\}$ includes the following components:*

$$x^i = \begin{cases} \tilde{x}^i & \text{if } \mathring{e}_i < e^i \text{ or } \mathring{s}_i = \tilde{x}^i \\ 1 - \tilde{x}^i & \text{if } \mathring{e}_i > e^i \text{ and } \mathring{s}_i = 1 - \tilde{x}^i. \end{cases} \tag{B.29}$$

Proof. This result follows from Lemma 3.1 and the decision-maker's power to act on what information s/he observes. ■

Lemma B.4. *If the oversight game form applies and $x^{L*} = s$, $e^* \leq \max\{e^L, \hat{e}, e^A\}$. If an equilibrium exists in which $e^* = \max\{e^L, \hat{e}, e^A\}$ and $x^{L*} = s$, L cannot receive a higher payoff in equilibrium.*

Proof. Suppose $e > \max\{e^L, \hat{e}, e^A\}$. Since $e > e^L$, Lemma B.3 implies that, if A has the effort level and the signal, he can induce L to select $x^L = s, \forall s$, setting $\epsilon^A_L = \sigma^A_L = \delta, \forall s$. Because $e > e^A$, A prefers $x^L = s$ and will prefer to disclose information as described such that L selects $x^L = s$, provided that he can do so. Then, for any $e > \max\{e^L, \hat{e}, e^A\}$, R can receive $EU^R_f(e)$ with $\epsilon^R_A = \sigma^R_A = \delta$, so that, through A's disclosures, L selects $x^L = s$.

In any proposed equilibrium with $x^L = s$ and effort at some $\dot{e} > \max\{e^L, \hat{e}, e^A\}$, R would receive $EU^R_f(\dot{e})$. However, the above paragraph implies that R can increase his utility by

selecting some $\ddot{e} \in (\max\{e^L, \hat{e}, e^A\}, \dot{e})$ and setting $\epsilon_A^R = \sigma_A^R = \delta$. Then R would receive $EU_f^R(\ddot{e}) > EU_f^R(e^*)$. This inequality holds because the concavity of $g(\cdot)$ and convexity of $c(\cdot)$ imply that $EU_f^R(e)$ decreases with effort $e \geq \hat{e}$. Thus, R is not best-responding if $x^L = s$ and $e > \max\{e^L, \hat{e}, e^A\}$.

The first statement is thus established. The second statement follows from the first statement, which implies that any equilibrium with $e > \max\{e^L, \hat{e}, e^A\}$ would not have $x^{L*} = s$ and so would yield L a weakly lower payoff, and from the fact that any equilibrium with $e^* < \max\{e^L, \hat{e}, e^A\}$ also would yield a weakly lower payoff. ■

Proof of Proposition 3.2 For any e, Lemma 3.1 implies that R would either select $x = s$ or $x = 1$ after both signals. The former yields $EU_f^R(e)$, while the latter yields $q_1 b_1^R - c(e)$. The assumptions on $g(e)$ and $c(e)$ imply that, ex ante, the R would like to set either $e = 0$ and $x = 1, \forall s$, or $e = \hat{e}$ and $x = s$. The definitions of types of researchers imply that an unmotivated researcher would prefer the former and a motivated researcher the latter. Ex post, Lemma 3.1 implies that it would select policy consistently with its ex ante preferences. Specifically, an unmotivated researcher has $q_1 b_1^R > \frac{1}{2}(q_0 b_0^R + q_1 b_1^R)$, and a motivated researcher has $EU_f^R(\hat{e}) > q_1 b_1^R$, so that $g(e)(q_0 b_0^R + q_1 b_1^R) > q_1 b_1^R$. ■

Proof of Proposition 3.3 If $\tilde{x}^L = 1$ and $\bar{e}_1$ does not exist, then R's highest possible payoff comes uniquely from $e = 0$ and $x^L = 1$, and, by Lemma B.3, it can assure this outcomewith $\epsilon_L^R = \delta$ after $e = 0$. Otherwise, $\bar{e}_{\tilde{x}^L}$ exists, and further analysis is needed.

First, $e \in (0, e^L)$ cannot occur in equilibrium. By Lemma 3.1, such an equilibrium would require $x^L = \tilde{x}^L$ for both signals. Then R would receive $q_{\tilde{x}^L} b_{\tilde{x}^L}^R - c(e)$, and it would deviate by setting $e = 0$ and $\epsilon_L^R = \delta$ to induce $x^L = \tilde{x}^L$ by Lemma B.3. R would prefer to similarly deviate from any proposed equilibrium in which $e > \bar{e}_{\tilde{x}^L}$ because it would receive less than

$q_{\tilde{x}^L} b^R_{\tilde{x}^L}$, regardless of the policy selected.

If $e^L > \bar{e}_{\tilde{x}^L}$, then $e = 0$ is the only possible equilibrium effort level. Suppose, instead, that $e^L < \bar{e}_{\tilde{x}^L}$. For any $e \in [e^L, \bar{e}_{\tilde{x}^L}]$, an equilibrium would entail $x^L = s$. L would deviate from $x^L = 1 - s$ and from $x^L = 1 - \tilde{x}^L, \forall s$, while R would prefer to deviate from any equilibrium in which $x^L = \tilde{x}^L$ by setting $e = 0$ and $\epsilon^R_L = \delta$. If $\hat{e} \in (e^L, \bar{e}_{\tilde{x}^L})$, R would prefer to deviate from any equilibrium in which $e \neq \hat{e}$ by selecting $e = \hat{e}$ and $\epsilon^R_L = \sigma^R_L = \delta$. (Note that $\hat{e} = \bar{e}_{\tilde{x}^L}$ is ruled out since $EU^R_f(\hat{e}) \neq q_1 b^R_1$ by assumption in Footnote 1.) If, instead, $e^L \in [\hat{e}, \bar{e}_{\tilde{x}^L})$, it would prefer to deviate from any equilibrium in which $e = 0$ by selecting some $e \in (e^L, \bar{e}_{\tilde{x}^L}))$ and $\epsilon^R_L = \sigma^R_L = \delta$, and from any equilibrium in which $e \in (e^L, \bar{e}_{\tilde{x}^L}))$ by setting a lower e in that interval and $\epsilon^R_L = \sigma^R_L = \delta$; however, with no minimum value in the interval, the only permissible equilibrium effort level is e^L. Finally, if $e^L = \bar{e}_{\tilde{x}^L}$, then the two possible equilibrium effort levels are 0 and e^L.

In all these cases, an equilibrium can be constructed in which L sets $x^L = 1 - \tilde{x}^L$ only when $\mathring{e} \geq e^L$ and $\mathring{s} = 1 - \tilde{x}^L$. R receives $q_{\tilde{x}^L} b^R_{\tilde{x}^L}$ for $e = 0$ and $EU^R_f(e)$ for $e \geq e^L$. If $e^L \geq \bar{e}_{\tilde{x}^L}$, $q_{\tilde{x}^L} b^R_{\tilde{x}^L} > EU^R_f(e), \forall e > e^L$, so it optimizes by setting $e = 0$ and $\epsilon^R_L = \sigma^R_L = \delta$. If $e^L \leq \bar{e}_{\tilde{x}^L}$, $EU^R_f(e) > q_{\tilde{x}^L} b^R_{\tilde{x}^L}, \forall e \geq e^L$, and $\arg\max e \in [e^L, \bar{e}_{\tilde{x}^L}] EU^R_f(e) = \max\{e^L, \hat{e}\}$. In this case, she can ensure $x = s$ by setting $\epsilon^R_L = \sigma^R_L = \delta$ after $e = \max\{e^L, \hat{e}\}$. Lemma B.3 implies no deviation by L. ■

Proof of Proposition 3.5 (a): Proposition 3.3 implies that either $e^* = 0$ and $x^{L*} = 1$ when $\tilde{x}^L = 1$, $\tilde{x}^A = 0$, and $e^L \not\leq \bar{e}_1$ or $e^* = \max\{e^L, \hat{e}\}$ and $x^{L*} = s$ otherwise under administration, for a payoff that can be expressed as $EU^L_f(e)$ for some $e \in \{e^L, \hat{e}\}$. Under delegation Corollary 3.4 implies $e^* = e^A$ and $x^{A*} = s$ for a payoff of $g(e^A)(q_0 b^L_0 + q_1 b^L_1) > \max_{e \in \{e^L, \hat{e}\}} EU^L_f(e)$.

(b): Proposition 3.3 implies $e^* = 0$ and $x^{L*} = 1$ under administration, for a payoff of $q_1 b_1^L$, whereas Corollary 3.4 implies $e^* = \hat{e}$ and $x^{A*} = s$ under delegation, for a payoff of $g(\hat{e})(q_0 b_0^L + q_1 b_1^L) > q_1 b_1$ since $\hat{e} > e^L$. ■

Proof of Proposition 3.6 That $e^* \leq e^L$ when $x^{L*} = s$ implies that L does not receive less than her reservation payoff. Also, she cannot exceed her reservation payoff of $q_{\tilde{x}^L} b_{\tilde{x}^L}^L$ if the same policy occurs after each signal or if $x^L = 1 - s, \forall s$. These facts and Lemma B.4 imply that the given equilibrium maximizes her payoff if it exists. Assume that the conditions stated in the proposition hold. It is sufficient to specify that L's strategy includes the rules in Lemma B.3 and $x^L = x^A$ when $\mathring{e}_L = \mathring{s}_L = \emptyset$. A's strategy can be $\epsilon_L^A = \sigma_L^A = \nu$ for any information it receives from R, with $x^A = 1 - \tilde{x}^A$ if any only if $\mathring{e}_A \geq e^A$ and $\mathring{s}_A = 1 - \tilde{x}^A$.

Faced with L and A's strategies, R will receive $q_{\tilde{x}^A} b_{\tilde{x}^A}^R - c(e)$ unless $e \geq e^A \geq e^L$, and it displays both items of information when $s = 1 - \tilde{x}^A$. When $\bar{e}_{\tilde{x}^A} \geq e^A$, it prefers to research at $e \in [e^A, \bar{e}_{\tilde{x}^A}]$, since $EU_f^R(e) \geq q_{\tilde{x}^A} b_{\tilde{x}^A}^R$ for these levels of effort. Since $EU_f^R(e)$ decreases with effort from its maximum at $\hat{e}$ (due to concavity of $g(\cdot)$ and $c(\cdot)$), its best response entails (1) $e = \hat{e}$ if $\hat{e} \geq e^A$ and $e = e^A$ otherwise to exert the least effort needed so that $x^L = s$; and (2), when $s = 1 - \tilde{x}^A$ to set $\epsilon_A^R = \sigma_A^R = \delta$. Given R's effort, A maximizes his utility if he induces $x^L = s$ from L. Since he does so by setting $x^A = s$ and $\epsilon_L^A = \sigma_L^A = \nu$, his strategy is a best response. Finally, since $e \geq e^L$ in equilibrium, L is best-responding: behind either proposal, with no other information, is a signal matching the proposal, supported by enough effort to persuade a leader inclined toward the opposite policy. ■

Proof of Proposition 3.7 As in Proposition 3.6, the facts that $e^* \leq e^L$ when $x^{L*} = s$, so that L does not receive less than her reservation payoff; that she cannot exceed her reservation payoff of $q_{\tilde{x}^L} b_{\tilde{x}^L}^L$ if the same policy occurs after each signal or if $x^L = 1 - s, \forall s$, combine with

Lemma B.4 to imply that the given equilibrium maximizes her payoff if it exists. Assume that the conditions given in the proposition hold. L's strategy can be partially filled in with $x^L = 0$ when $x^A = 0$ and $\mathring{e}_L = \mathring{s}_L = \emptyset$, and $x^L = 1 - \tilde{x}^L$ when $\mathring{e}_L = e^L$ and $\mathring{s}_L = 1 - x^L$ (and is not unique beyond this specification). A's strategy can be $x^A = 1$ and $\epsilon_L^A = \sigma_L^A = \delta$ when $\mathring{e}_A \geq e^A$ and $\mathring{s}_A = 1$, and $x^A = 0$ and $\epsilon_L^A = \sigma_L^A = \nu$ otherwise.

Faced with these two players' strategies, R will receive $q_0 b_0^R$ unless $e \geq \max\{e^A, e^L\}$ and displays both items of information when $s = 1$. Because $\max\{e^A, e^L\} < \bar{e}_0$, it prefers to research at $e \in [\max\{e^A, e^L\}, \bar{e}_0]$, since $EU_f^R(e) \geq q_0 b_0^R$ for these levels of effort. Since $EU_f^R(e)$ decreases with effort from its maximum at $\hat{e}$ (due to concavity of $g(\cdot)$ and $c(\cdot)$), its best response entails (1) $e = \hat{e}$ if $\hat{e} \geq \max\{e^A, e^L\}$ and $e = \max\{e^A, e^L\}$ otherwise to exert the least effort needed so that $x^L = s$; and (2), when $s = 1$ to set $\epsilon_A^R = \sigma_A^R = \delta$. Given R's effort, A maximizes his utility if he induces $x^L = s$ from L. Since he does so by setting $x^A = 1$ and $\epsilon_L^A = \sigma_L^A = \delta$ when he observes the effort and $\mathring{s}_A = 1$ and by choosing $x^A = 1$ and $\epsilon_L^A = \sigma_L^A = \nu$ otherwise, his strategy is a best response. Finally, since $e \geq e^L$ in equilibrium, L is best-responding: by Lemma B.3 when she observes the effort level and signal and because $\mathring{e}_L = \mathring{s}_L = \emptyset$ implies a signal of 0 with $e = \max\{\hat{e}, e^A, e^L\}$, so that either type of leader prefers policy 0. ■

Proof of Theorem 3.8 (a): The conditions in Proposition 3.5 are a subset of the conditions under which the equilibrium in Proposition 3.6 exists: when $\max\{e^L, \hat{e}\} < e^A \leq \bar{e}_{\tilde{x}^A}$, or when $e^A \leq \hat{e}$ and $e^L < \hat{e}$ with $\tilde{x}^A = 0$, $\tilde{x}^L = 1$ and an unmotivated researcher $e^A \leq \bar{e}_{\tilde{x}^A}$ and $e^L \leq \max\{\hat{e}, e^A\}$. (In the latter case $\hat{e} < \bar{e}_0$ establishes that $e^A \leq \bar{e}_{\tilde{x}^A}$.) From Corollary 3.4 and Proposition 3.6, $e^* = \max\{\hat{e}, e^A\}$ and $x^* = s$, which implies that delegation and oversight yield the same payoff. Since Proposition 3.5 is refers to the conditions under which

delegation outperforms administration, it follows that oversight outperforms administration by the same amount under these conditions. Also, since $e^* = \max\{e^L, \hat{e}, e^A\}$ in these cases, Lemma B.4 implies L cannot do any better.

(b)(i): Here, $e^* = e^L$ and $x^{L*} = s$ when $e^L \leq \bar{e}_{\tilde{x}^L}$ or $e^* = 0$ and $x^{L*} = \tilde{x}^L, \forall s$, under administration for her default payoff by Proposition 3.3, whereas $e^* = \max\{\hat{e}, e^A\} < e^L$ and $x^* = s$ under delegation for a payoff less than her default payoff by Corollary 3.4. Under oversight, however, L can achieve her default payoff by setting $x^L = \tilde{x}$ unless $\mathring{e}_L \geq x^L$ and $\mathring{s}_L = 1 - \tilde{x}^L$ as she would under administration. For A it is sufficient to specify that $\epsilon_L^A = \sigma_L^A = \delta$ when $\mathring{e}_A \geq e^L$ and $\mathring{s}_A = 1 - \tilde{x}^L$. If $e < e^L$, there is nothing that R can disclose to A and have him relay to L so that she would prefer $x^L = 1 - \tilde{x}^L$. Then $x^L = \tilde{x}^L, \forall s$ and R will receive $q_{\tilde{x}^L} b^R_{\tilde{x}^L} - c(e)$, which is maximized at $e = 0$. However, if $e \geq e^L$, he can disclose both items, which A will relay at least when $s = 1 - \tilde{x}^L$, for $EU_f^R(e)$. Since $e^L > \hat{e}$, R prefers the lowest in this range, e^L. If $e^L \leq (>) \bar{e}_{\tilde{x}^L}$ then $g(e^L)(q_0 b_0^R + q_1 b_1^R) - c(e^L) \geq (<) q_{\tilde{x}^L} b^R_{\tilde{x}^L}$. When $e = e^L$ (0), $x = s$ $(\tilde{x}^L, \forall s)$. These are the same effort and policy choices as under administration, so L's payoff is the same if the equilibrium exists. A prefers to follow his strategy for $e \geq e^L$: when he observes the signal, his strategy leads L to select the same policy he would after each signal. If he does not observe the signal, any disclosure yields $x^L = 1 - \tilde{x}^L$. P will not defect, since on the equilibrium path she is selecting her preferred policy based on R and A's strategies and on Lemma 3.1. Also, Lemma B.4 implies that L a higher payoff since $e^L = \max\{e^L, \hat{e}, e^A\}$.

(b)(ii): Proposition 3.3 implies that $e = \hat{e}$ and $x = s$, for a payoff of $g(\hat{e})(q_0 b_0^L + q_1 b_1^L)$, but Corollary 3.4 implies $e = 0$ and $x = \tilde{x}^L, \forall s$, for a payoff of $q_{\tilde{x}^L} b^L_{\tilde{x}^L} < g(\hat{e})(q_0 b_0^L + q_1 b_1^L)$. However, L can recover her administration payoff under oversight by setting $x^L = 1 - \tilde{x}^L$ when $\mathring{e}_L = \mathring{s}_L = \emptyset$. Then R, which is motivated and has $\hat{e} \in (e^L, e^A)$, maximizes its payoff by

setting $e = \hat{e}$, disclosing at least s when $s = \tilde{x}^L$, and setting $\epsilon_A^R = \sigma_A^R = \nu$ when $s = 1 - \tilde{x}^L$. Then $\sigma_L^A = \delta$ when $\mathring{s}_A = \tilde{x}^L$ to avoid having $x^L = 1 - \tilde{x}^L$, while A cannot disclose anything when $s = 1 - \tilde{x}^L$. Finally, L is best-responding: by Lemma B.3, $x^L = \tilde{x}^L$ should follow $\mathring{s}_L = \tilde{x}^L$, and when $\mathring{s}_L = \emptyset$, $s = 1 - \tilde{x}^L$, she prefers $x^L = 1 - \tilde{x}^L$ since $e = \hat{e} > e^L$. L cannot achieve a higher payoff with $x^L = x^A$ or $x^L = \tilde{x}^L$ when $\mathring{e}_L = \mathring{s}_L = \emptyset$. Lemma B.3 implies A's best response would entail $\epsilon_L^A = \sigma_L^A = \nu$ whenever $\mathring{e}_A < e^A$ and setting $x^A = \tilde{x}^L$ as needed to induce $x^L = \tilde{x}^L$. Since $e^A > \bar{e}_{\tilde{x}^L}$, R would not be best-responding if $e \geq e^A$ and would prefer to set $e = 0$ and accept $x^L = \tilde{x}^L$, which it can ensure with $e = 0$ and $\epsilon_A^R = \delta$. Meanwhile, if $\mathring{e}_A = \emptyset$, A would not be best-responding if $e < e^A$ and A did not act as needed to ensure $x^L = \tilde{x}^L$. Since $x^L = \tilde{x}^L$ for $e < e^A$ for both signals in any equilibrium L receives her default payoff or less.

(b)(iii): Corollary 3.4 implies that, under delegation, L receives less than her default payoff under administration: When $e^A \not\leq \bar{e}_{\tilde{x}^A}$, $x^* = \tilde{x}^A, \forall s$, for $q_{\tilde{x}^A} b^L_{\tilde{x}^A} < q_{\tilde{x}^L} b^L_{\tilde{x}^L}$. When $e^A \leq \bar{e}_{\tilde{x}^A}$ but $\max\{\hat{e}, e^A\} < e^L$, $e^* = \max\{\hat{e}, e^A\}$ for $\max^L_{e \in \{\hat{e}, e^A\} EU^L_f(e) < q_{\tilde{x}^L} b_{\tilde{x}^L}}$. However, oversight allows L to recover her administration payoff. L's strategy be $x^L = \tilde{x}^A$ if and only if $\mathring{e}_L \geq e^L$ and $\mathring{s}_L = \tilde{x}^A$. Then R maximizes his payoff as follows: If $e^L \leq \bar{e}_{\tilde{x}^L}$, it sets $e = \max\{\hat{e}, e^L\}$ and $\epsilon_A^R = \sigma_A^R = \delta$. When $s = \tilde{x}^A$, A best-responds with $\epsilon_L^A = \sigma_L^A = \delta$ so that L will select $x^L = \tilde{x}^A$. If $e^L \not\leq \bar{e}_{\tilde{x}^L}$, R maximizes by setting $e = 0$, which leads to $x = \tilde{x}^L$ regardless of the disclosures. L is best responding since she selects the right policy after each equilibrium disclosure. With the same efforts and policy choices as in Proposition 3.3, L can obtain the same payoff under oversight as under administration.

L cannot receive a higher payoff with oversight when $\max\{\hat{e}, e^A\} < e^L$ by Lemma B.4. Suppose $\max\{\hat{e}, e^A\} \geq e^L$, but $e^A \not\leq \bar{e}_{\tilde{x}^A}$. First, whenever $\max\{e^A, e^L\} \leq \hat{e}$, and $e^A \not\leq \bar{e}_{\tilde{x}^A}$, $\tilde{x}^A = 1$ and R is unmotivated. Also, $\tilde{x}^L = 0$, in which case $e^* = \hat{e} = \max\{e^L, \hat{e}, e^A\}$ under

administration. Then Lemma B.4 implies that L cannot do better.

Now suppose $\max\{e^L, \hat{e}\} \leq e^A$. For $\mathring{e}_L = \mathring{s}_L = \emptyset$, L must set $x_L = \tilde{x}^L$. Otherwise, Lemma B.3 implies A's best response would entail $\epsilon_L^A = \sigma_L^A = \nu$ whenever $\mathring{e}_A < e^A$ and setting $x^A = \tilde{x}^L$ as needed to induce $x^L = \tilde{x}^L$. If $\tilde{x}^L = 1$, $e^A > \bar{e}_0$, and R is not best-responding if $e \geq e^A$ since it would prefer to set $e = 0$. Meanwhile, if $\mathring{e}_A = \emptyset$, A would not be best-responding if $e < e^A$ and A did not act as needed to ensure $x^L = 0$. Thus, if $x_L = 0$ or $x^L = x^A$, having A and R best-respond entails $e = 0$ and $x^L = 0$, in which case L would defect with $x^L = 1$. If $\tilde{x}^L = 0$ and $x_L = 0$ or $x^L = x^A$ when $\mathring{e}_L = \mathring{s}_L = \emptyset$, then R can guarantee $q_1 b_1^R$ by setting $e = 0$ and $\epsilon_A^R = \delta$, since A will disclose and propose as needed to secure $x^L = 1$. If R is unmotivated, he would just set $e = 0$, and L would defect by setting $x^L = 0$. Even if R is motivated, it would defect if $e^A > \bar{e}_1$ since it could set $e = 0$, and A would defect if $e \leq \bar{e}_1$ and $x^L = s$ by always disclosing and proposing such that $x^L = 1$. That leaves $e \leq \bar{e}_1$ and policy not always matching the signal, so his best payoff is from $e = 0$, which would lead L to defect with $x^L = 0$. Thus, $x_L = \tilde{x}^L$ for $\mathring{e}_L = \mathring{s}_L = \emptyset$. Also, L cannot benefit if $x^L = \tilde{x}^A$ after $\mathring{e}_L = \emptyset$ and $\mathring{s}_L = \tilde{x}^A$. Then R would maximize with $e = \hat{e}$, $\epsilon_A^R = \nu$, $\sigma_A^R(\hat{e}, \tilde{x}^A) = \delta$, and $\sigma_A^R(\hat{e}, \tilde{x}^L) = \nu$.

Suppose $\max\{e^L, \hat{e}\} \leq e^A$ and $e^L \not\leq \bar{e}_{\tilde{x}^L}$. If $\tilde{x}^L = 0$, then R would never set $e \geq e^L > \bar{e}_0$ since it would prefer $e = 0$ followed by any policy. Then any equilibrium with $e < e^L$ cannot yield more than $q_0 b_0^L$, her payoff under administration. If $\tilde{x}^L = 1$, then when $\mathring{e}_L = \mathring{s}_L = \emptyset$, $x^L = 1$. R can guarantee $q_1 b_1^R$ with $e = 0$ and $\epsilon_L^A = \sigma_L^A = \nu$. There is no equilibrium in which R's utility is higher and in which L would be best-responding since $e^L \not\leq \bar{e}_1$. So it must be that $e^* = 0$ and $x^{L*} = 1, \forall s$, which is the same result as under administration.

If $\max\{e^L, \hat{e}\} \leq e^A$, $e^L \leq \bar{e}_{\tilde{x}^L}$ and $\bar{e}_{\tilde{x}}^A \not\leq e^A$, then L may be able to exceed her payoff from administration and delegation with oversight. This will not happen when $\tilde{x}^A = 0$, $\tilde{x}^L = 1$,

and R is unmotivated, in which case R maximizes his payoff with $e = 0$ and $\epsilon_L^A = \sigma_L^A = \nu$. Also, if $\tilde{x}^A = 1$, $\tilde{x}^L = 0$, and $b_0^R < 0$, in which case L must set $x^L = 0$ when $\mathring{e}_L > e^L$ and $\mathring{s}_L = \emptyset$. Otherwise, R always prefers $x^L = 1$ ex post and would set e or slightly above e^L and $\epsilon_A^R = \delta$ and $\sigma_A^R = \nu$, and A would set $\epsilon_L^A = \delta$ and $x^A = 1$ as necessary so that $x^L = 1$. With $x^L = 0$ when $\mathring{e}_L > e^L$ and $\mathring{s}_L = \emptyset$, R can only aim for $x^L = s$, in which case it selects $e = \max\{\hat{e}, e^L\}$, which would occur under administration or delegation.

(c) Under these conditions, Proposition 3.3 implies a payoff of $EU_f^L(\hat{e}) > q_0 b_0^L$ under administration, while Corollary 3.4 implies a payoff of $q_0 b_0^L$ under delegation. Oversight cannot yield L more than delegation. L's strategy after $\mathring{e}_L = \mathring{s}_L = \emptyset$ cannot be $x^L = 1$, or else R would maximize its utility by setting $e = 0$ and withholding both items of information. If instead, the strategy for $\mathring{e}_L = \mathring{s}_L = \emptyset$ is $x^L = 0$ or $x^L = x^A$, A can ensure that $x^L = 0$ for any $\mathring{e}_A < e^A$. Since $e^A > \bar{e}_{\tilde{x}^L}$, R would not be best-responding if $e \geq e^A$ and would prefer to set $e = 0$ and accept $x^L = \tilde{x}^L$, which it can ensure with $e = 0$ and $\epsilon_A^R = \delta$. Meanwhile, if $\mathring{e}_A = \emptyset$, A would not be best-responding if $e < e^A$ and A did not act as needed to ensure $x^L = \tilde{x}^L$. Since $x^L = \tilde{x}^L$ for $e < e^A$ for both signals in any equilibrium L receives no more than her default payoff, which is less than $EU_f^L(\hat{e})$ under administration.

(d)(i): Among cases in this set, Proposition 3.3 and Corollary 3.4 imply that $e^* = \hat{e}$ under both administration and delegation. Then Lemma B.4 implies that there cannot be any equlibria with $e^* > \hat{e}$ with $x^{L*} = s$, and any other equilibria would yield L no more than $EU_f^L(\hat{e})$. The exception is when $\tilde{x}^L = \tilde{x}^A = 1$, and R is unmotivated, in which case $e^* = 0$ in both modes. Then oversight adds nothing since R can set $e = 0$ and $\epsilon_A^R = \delta$, in which case A would set $\epsilon_L^A = \delta$ as needed to ensure $x^L = 1$.

(d)(ii): Proposition 3.3 implies $e^* = e^L$ or $e^* = 0$, yielding $q_{\tilde{x}^L} b_{\tilde{x}^L}$, and Corollary 3.4 entails $e^* = 0$ for the same payoff. L's payoff is the same, but not more, under oversight.

If $\tilde{x}^L = 0$, then $\mathring{e}_L = \mathring{s}_L = \emptyset$ should not lead to $x^L = 1$. Otherwise, R can guarantee $q_1 b_1^R$ through $x^L = 1$ with $e = 0$ and $\epsilon_L^A = \sigma_L^A = \nu$, R would not be best-responding if $e \geq e^A > \bar{e}_0$, and A would not be best-responding if R received more than $q_1 b_1^R$. Thus, if $x^L = 1$ when $\tilde{x}^L = 0$ and $\mathring{e}_L = \mathring{s}_L = \emptyset$, $e = 0$ and $x^L = 1$, causing L to defect with $x^L = 0$. Then if $x^L = 0$ or $x^L = x^A$ when $\mathring{e}_L = \mathring{s}_L = \emptyset$, A will set $\epsilon_L^A = \sigma_L^A = \emptyset$ and set $x^A = 0$ as needed to ensure $x^L = 0$ when $\mathring{e} < e^A$. Again, R would not be best-responding if $e \geq e^A > \bar{e}_0$. Then in any possible equilibrium A must induce $x^L = 0$ if $\mathring{e} = \emptyset$. The result is $e = 0$ and $x^L = 0$, which yields the same payoff as under administration or delegation. If $\tilde{x}^L = 1$ and R is unmotivated, $e = 0$ with $\epsilon_A^R = \epsilon_L^A = \delta$ to ensure that $x^L = 1$. Even if R is motivated, $e^A > \bar{e}_1$, a best response by implies that there is no equilibrium with $e > \bar{e}_1$ (or else R is not best-responding) or with $e \leq \bar{e}_1$ and $x^L = s$ (or else A is not best-responding). Then $e = 0$ and $x^L = 1$, which yields L the same payoff as under administration or delegation. ■

Proof of Proposition 3.9 If oversight is available, Theorem 3.8 implies that oversight dominates delegation and that oversight dominates administration when an agent with $\tilde{x}^L = 0$ and $e^A \leq \bar{e}_0$ is involved. Most of the equilibria in which L benefits from oversight involve $\max\{e^L, \hat{e}\} = e^A = e^* \leq \bar{e}_0$ and $x^{L*} = s$. Among such equilibria, she prefers the greatest value of e^A because her payoff, represented by Equation (3.1), increases with e. The exception is those alluded to Proposition 3.5(b), but even then, L would prefer a higher value of e^A to achieve an equilibrium in the previous category. If oversight is not available, Theorem 3.8 neither delegation nor administration dominates among agents with $\tilde{x}^L = 0$ and $e^A \leq \bar{e}_0$. However, she maximizes by choosing an agent with the greatest value of e^A. Under administration, the e^A is not relevant. Under delegation, however, Corollary 3.4 implies that her utility weakly increases with e^A provided that $e^A \bar{e}_0$. ■

Proof of Proposition 3.10 Theorem 3.8 implies that, when $e^L \leq e^A < \bar{e}_0$ and $\tilde{x}^A = 0$, only cases (a) and (d)(i) apply, in which case L prefers oversight or delegation to administration. As long as $e^L \leq e^A$ after stautory bias-shifting, she will still prefer oversight or delegation. Meanwhile, e^A and $\hat{e}$ not affected by statutory bias-shifting, so the equilibria effort and policy selections in Corollary 3.4 and Propositions 3.6 and 3.7 are unaffected. ■

Proof of Proposition 3.11 L's payoff exceeds her default only when $e > e^L$ and $x^* = s$. Her payoff would be represented by Equation (3.1). Then her maximum possible payoff is $EU_f^L(\bar{e}_0)$ given her choice of agents. Proposition 3.5 implies that L would continue to at least prefer oversight or delegation as long as $\max\{e^L, \hat{e}\} \leq e^A$. Thus, if L only receives $EU_f^L(\bar{e}_0) - \max_{e \in \{e^L, \hat{e}\}} EU_f^L(e) \equiv \Delta_1 V^L$, she would still be willing to use delegation or oversight with an agent who has $e^A = \max\{e^L, \hat{e}\}$.

Proposition 3.3 implies that $\max_{e \in \{e^L, \hat{e}\}} EU_f^L(e)$ is her payoff under administration if $e^L \leq \bar{e}_{\tilde{x}^L}$. Since $e^L < \bar{e}_0$, if $\tilde{x}^L = 0$, then $\Delta_2 V^L = \Delta_1 V^L$. The same is true if $\tilde{x}^L = 1$ and R is motivated. If $e^L \leq \bar{e}_1$, $e^* = \max\{e^L, \hat{e}\}$ under any game form. If $e^L > \bar{e}_1 > \hat{e}$, then her payoff under administration is $q_1 b_1^L = EU_f^L(e^L)$. If, however, $\tilde{x}^L = 1$ and R is unmotivated and $\hat{e} > e^L$, then her payoff under delegation or oversight is $EU_f^L(\hat{e})$, whereas her administration payoff is $q_1 b_1^L < EU_f^L(\hat{e})$. Here, $\Delta_2 V^L = EU_f^L(\bar{e}_0) - q_1 b_1^L < \Delta_1 V^L$ is necessary for L to select administration.

If R is motivated, R prefers $\hat{e}$ but does not receive it if $e^L > \hat{e}$. Then L needs $\Delta_3 V^L = EU_f^L(\bar{e}_0) - EU_f^L(\hat{e}) > EU_f^L(\bar{e}_0) - q_1 b_1^L = \Delta_2 V^L$ to accept R's referred outcome. If R is unmotivated, R prefers $e = 0$ and $x^L = 1$ but does not receive it if $\tilde{x}^L = 0$. Then $\Delta_3 V^L = EU_f^L(\bar{e}_0) - q_1 b_1^L > \max\{EU_f^L(\hat{e}), q_0 b_0^L\}$. ■

Proof of Proposition 3.12 Following Corollary 3.4 and Proposition 3.6 or 3.7, A's utility under delegation or oversight in the two respective cases is $EU_f^A(e^A)$ and $EU_f^A(\check{e}^A)$, for $\Delta V^A = EU_f^A(e^A) - EU_f^A(\check{e}^A)$. With the definition of B_0^A and A's standard of proof, the two respective biases are $B_0^A = 2g(e^A) - 1$ and $\check{B}_0^A = 2g(\check{e}^A) - 1$, for $\Delta B_0^A = 2(g(e^A) - g(\check{e}^A))$. ■

References

Aberbach, Joel D. 1990. *Keeping a Watchful Eye: The Politics of Congressional Oversight.* Washington, DC: Brookings Institution.

Aghion, Philippe, and Jean Tirole. 1997. "Formal and Real Authority in Organizations." *Journal of Political Economy* 105 (February): 1–29.

Ambrus, Attila, Eduardo M. Azevedo, and Yuichiro Kamada. 2013. "Hierarchical Cheap Talk." *Theoretical Economics* 8 (January): 233–61.

Appelbaum, Binyamin. 2011. "Supreme Court Denies a Move to Bar the Details of a Fed Bailout." *New York Times* (March 22): B2.

Asamoah, Afia K., and Joshua M. Sharfstein. 2010. "Transparency at the Food and Drug Administration." *New England Journal of Medicine* 362 (June 24): 2341–43.

Ayres, Ian, and John Braithwaite. 1991. "Tripartism: Regulatory Capture and Empowerment." *Law and Social Inquiry* 16 (Summer): 435–96.

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

Beermann, Jack M. 2006. "Congressional Administration." *San Diego Law Review* 43 (February-March): 61–158.

Bendor, Jonathan, and Adam Meirowitz. 2004. "Spatial Models of Delegation." *American Political Science Review* 98 (May): 293–310.

Bernanke, Ben. 2011 December 6. "Correction of Recent Press Reports Regarding Federal Reserve Emergency Lending During the Financial Crisis." Memorandum. http://www.federalreserve.gov/generalinfo/foia/emergency-lending-financial-crisis-20111206.pdf.

Bertelli, Anthony, and Sven E. Feldmann. 2007. "Strategic Appointments." *Journal of Public Administration Research and Theory* 17 (January): 19–38.

Besley, Timothy. 2006. *Principled Agents? The Political Economy of Good Government.* New York: Oxford University Press.

Bloomberg News Responds to Bernanke Criticism of U.S. Bank-Rescue Coverage. 2011. *Bloomberg News* (December 7). http://www.bloomberg.com/news/2011-12-06/bloomberg-news-responds-to-bernanke-criticism.html.

Bueno de Mesquita, Ethan, and Matthew C. Stephenson. 2007. "Regulatory Quality Under Imperfect Oversight." *American Political Science Review* 101 (August): 605–20.

Callander, Steven. 2008. "A Theory of Policy Expertise." *Quarterly Journal of Political Science* 3 (July): 123–40.

Carpenter, Daniel. 2010. *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA*. Princeton: Princeton University Press.

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.

Carpenter, Daniel P. 2004. "Protection without Capture: Product Approval by a Politically Responsive, Learning regulator." *American Political Science Review* 98 (November): 613–31.

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

Cini, Michelle. 2008. "European Commission reform and the origins of the European Transparency Initiative." *Journal of European Public Policy* 15 (August): 743–760.

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.

Coglianese, Cary, Richard Zeckhauser, and Edward Parson. 2004. "Seeking Truth for Power: Informational Strategy and Regulatory Policymaking." *Minnesota Law Review* 89 (December): 277–341.

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.

Critical Mass Energy Project v. *Nuclear Reguatory Comm'n*. 1992. 975 F. 2d 871. (D.C. Cir.).

Dal Bó, Ernesto. 2006. "Regulatory Capture: A Review." *Oxford Review of Economic Policy* 22 (Summer): 203–25.

Dessein, Wouter. 2002. "Authority and Communication in Organizations." *Review of Economic Studies* 69 (October): 811–38.

Epstein, David, and Sharyn O'Halloran. 1999. *Delegating Powers: A Transaction Cost Politics Approach to Policy Making under Separate Powers*. New York: Cambridge University Press.

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

Gailmard, Sean, and John W. Patty. 2012. "Formal Models of Bureaucracy." *Annual Review of Political Science* 15: 353–77.

Gailmard, Sean, and John W. Patty. 2013a. *Learning While Governing: Information, Accountability, and Executive Branch Institutions*. Chicago: University of Chicago Press.

Gailmard, Sean, and John W. Patty. 2013b. "Stovepiping." *Journal of Theoretical Politics*. Forthcoming.

Gersen, Jacob E., and Anne Joseph O'Connell. 2009. "Hiding in Plain Sight? Timing and Transparency in the Administrative State." *University of Chicago Law Review* (Summer): 1157–1214.

Gibbons, Ann. 1991. "Can David Kessler revive the FDA?" *Science* 252 (April 12): 200–03.

Graber, Doris A. 2003. *The Power of Communication: Managing Information in Public Organizations*. Washington: CQ Press.

Graham, John D., Paul R. Noe, and Elizabeth L. Branch. 2006. "Managing the Regulatory State: The Experience of the Bush Administration." *Fordham Urban Law Journal* 33 (May): 953–1002.

Greenwald, Marilyn and Joseph Bernt, eds. 2000. *The Big Chill: Investigative Reporting in the Current Media Environment*. Ames, IA: Iowa University Press.

Hamilton, James. 2004. *All the News That's Fit to Sell: How the Market Transforms Information into News*. Princeton, NJ: Princeton University Press.

Harris, Gardiner. 2008. "F.D.A. Scientists Accuse Agency Officials of Misconduct." *New York Times* (November 18): A15.

Heald, David. 2006. "Transparency as an Instrumental Value." In *Transparency: The Key to Better Governance?*, ed. Chrostopher Hood and David Heald. New York: Oxford University Press.

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

Hood, Christopher. 2006. "Transparency in Historical Perspective." In *Transparency: The Key to Better Governance?*, ed. Chrostopher Hood and David Heald. New York: Oxford University Press.

Hood, Christopher. 2007. "What Happens When Transparency Meets Blame-avoidance." *Public Management Review* 9 (June): 191–210.

Huber, John D., and Charles R. Shipan. 2002. *Deliberate Discretion? The Institutional Foundations of Bureaucratic Autonomy*. New York: Cambridge University Press.

Ivanov, Maxim. 2010. "Communication via a Strategic Mediator." *Journal of Economic Theory* 145 (March): 869–84.

Ivry, Bob, Bradley Keoun, and Phil Kuntz. 2011. "Secret Fed Loans Gave Banks $13 Billion Undisclosed to Congress." *Bloomberg Markets Magazine* (November 27). http://www.bloomberg.com/news/2011-11-28/secret-fed-loans-undisclosed-to-congress-gave-banks-13-billion-in-income.html.

Kallas, Siim. 2005. "The Need for a European Transparency Initiative." Speech before the European Foundation for Management Development, Nottingham Business School. May 3.

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

Kreps, David M. 1990. "Corporate Culture and Economic Theory." In *Perspectives on Positive Political Economy*, ed. James E. Alt and Kenneth A. Shepsle. Cambridge, UK: Cambridge University Press.

Kwak, James. 2013. "Cultural Capture and the Financial Crisis." In *Preventing Regulatory Capture: Special Interest Influence, and How to Limit It*, ed. Daniel Carpenter and David Moss. New York: Cambridge University Press. Forthcoming.

Laffont, Jean-Jacques, and Jean Tirole. 1991. "The Politics of Government Decision-making: A Theory of Regulatory Capture." *Quarterly Journal of Economics* 106 (November): 1089–1127.

Landis, James M.C. 1938. *The Administrative Process*. New Haven, CT: Yale University Press.

Leaver, Clare. 2009. "Bureaucratic Minimal Squawk Behavior: Theory and Evidence from Regulatory Agencies." *American Economic Review* 99 (June): 572–607.

Lee, Mordecai. 1999. "Reporters and Bureaucrats: Public Relations Counter-Strategies by Public Administrators in an Era of Media Disinterest in Government." *Public Relations Review* 25 (Winter): 451–63.

Lee, Mordecai. 2008. "Media and Bureaucracy in the United States." In *Encyclopedia of Public Administration and Public Policy*. 2d ed. New York: Taylor & Francis.

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.

Lodge, Martin. 2004. "Accountability and Transparency in Regulation: Critiques, Doctrines, and Instruments." In *The Politics of Regulation: Institutions and Regulatory Reforms for the Age of Governance*, ed. Jacint Jordana and David Levi-Faur. Cheltenham, UK: Edward Elgar.

Lubbers, Jeffrey S. 2006. *A Guide to Federal Agency Rulemaking*. Chicago: American Bar Association.

Lurie, Peter, and Allison Zieve. 2006. "Sometimes the Silence Can Be Like the Thunder: Access to Pharmaceutical Data at the FDA." *Law and Contemporary Problems* 69 (Summer): 85–97.

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

McCarty, Nolan. 2013. "Complexity, Capacity, and 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.

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.

McGarity, Thomas O., and Sidney A. Shapiro. 1980. "The Trade Secret Status of Health and Safety Testing Information: Reforming Agency Disclosure Policies." *Harvard Law Review* (March): 837–88.

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.

Morgan, David. 1986. *The Flacks of Washington: Government Information and the Public Agenda*. New York: Greenwood Press.

Neill, Katharine A., and John C. Morris. 2012. "A Tangled Web of Principals and Agents: Examining the Deepwater Horizon Oil Spill through a Principal–Agent Lens." *Politics and Policy* 40 (August): 629–56.

Nownes, Anthony J. 2006. *Total Lobbying: What Lobbyists Want (and How They Try to Get It)*. Cambridge, UK: Cambridge University Press.

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.

O'Neill, Onora. 2006. "Transparency and the Ethics of Communication." In *Transparency: The Key to Better Governance?*, ed. Chrostopher Hood and David Heald. New York: Oxford University Press.

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.

Prendergast, Canice. 2007. "The Motivation and Bias of Bureaucrats." *American Economic Review* 97 (March): 180–96.

Quirk, Paul J. 1981. *Industry Influence in Federal Regulatory Agencies*. Princeton, NJ: Princeton University Press.

Roberts, Alasdair. 2006. "Dashed Expectations: Government Adaptation to Transparency Rules." In *Transparency: The Key to Better Governance?*, ed. Chrostopher Hood and David Heald. New York: Oxford University Press.

Rogoff, Kenneth. 1985. "The Optimal Degree of Commitment to an Intermediate Monetary Target." *Quarterly Journal of Economics* 100 (November): 1169–89.

Rose-Ackerman, Susan. 1999. *Corruption and Government: Causes, Consequences, and Reform*. New York: Cambridge University Press.

Shipan, Charles R. 2004. "Regulatory Regimes, Agency Actions, and the Conditional Nature of Congressional Influence." *American Political Science Review* 98 (August): 467–80.

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.

Thurber, James A. 2011. "The Contemporary Presidency: Changing the Way Washington Works? Assessing President Obama's Battle with Lobbyists." *Presidential Studies Quarterly* 41 (June): 358–74.

Ting, Michael M. 2008. "Whistleblowing." *American Political Science Review* 102 (May): 249–67.

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

Tirole, Jean. 1986. "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations." *Journal of Law, Economics, and Organizations* 2 (Autumn): 181–214.

Urbina, Ian. 2010, May 14. "U.S. Said to Allow Drilling Without Needed Permits." *New York Times*: A1.

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

Weber, Max. 1978. *Economy and Society*. Trans. Ephraim Fischoff et al. Berkeley, CA: University of California Press.

Weil, David, Archon Fung, Mary Graham, and Elena Fagotto. 2006. "The Effectiveness of Regulatory Disclosure Policies." *Journal of Policy Analysis and Management* 25 (Winter): 155-81.

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.

Yackee, Jason Webb, and Susan Webb Yackee. 2010. "Administrative Procedures and Bureaucratic Performance: Is Federal Rule-making Ossified?" *Journal of Public Administration Research and Theory* 20 (April): 261–82.