DESIGNING BUREAUCRATIC ACCOUNTABILITY

ARTHUR LUPIA* AND MATHEW D. MCCUBBINS**

I

INTRODUCTION

A fundamental question in the study of democracy concerns the extent to which the will of the governed, as expressed by their elected representatives, affects the actions of the government. Some scholars observe the complexity of modern policymaking processes, infer that the elected representatives of the people lack the knowledge and skills required to constrain bureaucratic behavior, and conclude that democracy cannot work in the modern world. Others argue that members of representative legislatures can and do adapt to the problems produced by complexity. These scholars conclude that modern representative legislatures have the ability to translate meaningfully the will of the governed into the policy choices of the government. We address this debate by investigating the extent to which legislators can use institutional design to adapt to the challenges presented by the complexity of policymaking. In so doing, we produce new and more general conclusions about the consequences of institutional design for democratic decisionmaking and conclude that, in general, democracy can work.

In his essay "Bureaucracy," Max Weber argued that the general will of the governed is necessarily subverted when legislatures react to complexity by delegating their authority to bureaucrats.[1] He reasoned that "every bureaucracy seeks to increase the superiority of the professionally informed by keeping their knowledge and intentions secret."[2] In essence, he argued that an elected legislature's forfeiture of policymaking authority to the bureaucracy, together with the policy expertise that bureaucrats are alleged to possess, makes the legislative act of delegation equivalent to abdication.

Many scholars agree that the act of delegation combined with the existence of bureaucratic expertise are sufficient conditions for abdication.[3] Others have

---

* Assistant Professor of Political Science, University of California, San Diego.

** Professor of Political Science, University of California, San Diego.

We thank Jeff Banks, Kathy Bawn, Elisabeth Gerber, Will Heller, Jonathan Katz, Susanne Lohmann, Roger Noll, Brian Sala, Pablo Spiller, and Michael Thies for helpful comments.

Professor McCubbins acknowledges the support of the National Science Foundation grant number SES 9022882 and the support of the Ford Foundation.

1. Max Weber, *Bureaucracy, in* FROM MAX WEBER: ESSAYS IN SOCIOLOGY 196, 232-35 (H.H. Gerth & C. Wright Mills eds. & trans., Oxford Paperback 1958).

2. *Id.* at 233.

3. *See, e.g.*, Allen Schick, *Congress and the "Details" of Administration*, 36 PUB. ADMIN. REV. 516 (1976).

pointed out that a legislature's abilities to screen candidates for the bureaucracy and establish bureaucratic budgets and jurisdictions enable elected representatives to influence bureaucratic behavior. Those who take this position also point out that legislators have ample resources with which to reward responsive bureaucrats and discipline unresponsive ones.[4] Another group of scholars further pursues this line of reasoning and argues that legislators can design the structure and process of bureaucratic decisionmaking to ensure that bureaucratic expertise cannot be used against legislative interests.[5]

These two polar views, that delegation must equal abdication, and that legislators possess ample tools with which to discipline their agents, obviously cannot both be correct. A closer examination of the two arguments reveals that neither view may be correct because each conclusion is critically dependent on a potentially flawed premise. The flaw arises because previous researchers have assumed, rather than explored, the manner in which legislators deal with complexity.

Those who argue that delegation inevitably leads to abdication typically beg the central question of the debate: can legislators adapt to complexity? Instead, they merely assume that legislators are unable to overcome the problems that could result from bureaucratic expertise. Those who argue that delegation is unproblematic beg the same question. They assume that because legislators have the capability to overcome potential problems arising from bureaucratic expertise, they do so. They ignore the possibility that a bureaucrat's hidden knowledge, the source of his expertise and potential power over both legislators and citizens, may remain hidden even in the face of a legislator's attempts to uncover it. This point is critical, for if legislators are unable to determine what bureaucrats are doing, they may not know enough to reward helpful bureaucrats or punish those who take obstructive or destructive actions.

Whether a legislator can overcome the problems associated with bureaucratic expertise depends on her ability to obtain information about the consequences of bureaucratic activity. A legislator can obtain this information from three types

---

4. *See, e.g.*, Barry R. Weingast & Mark J. Moran, *Bureaucratic Discretion or Congressional Control? Regulatory Policymaking by the Federal Trade Commission*, 91 J. POL. ECON. 765 (1983).

5. *See, e.g.*, MORRIS P. FIORINA, CONGRESS: KEYSTONE OF THE WASHINGTON ESTABLISHMENT 40-41 (1977); JOSEPH HARRIS, CONGRESSIONAL CONTROL OF ADMINISTRATION (1964); D. RODERICK KIEWIET & MATHEW D. MCCUBBINS, THE LOGIC OF DELEGATION: CONGRESSIONAL PARTIES AND THE APPROPRIATIONS PROCESS 37-38 (1991); BERNARD ROSEN, HOLDING GOVERNMENT BUREAUCRACIES ACCOUNTABLE (1989); John Ferejohn, *On a Structuring Principle for Administrative Agencies*, *in* CONGRESS: STRUCTURE AND POLICY (Mathew D. McCubbins & Terry Sullivan eds., 1987); Mathew D. McCubbins, et al., *Administrative Procedures as Instruments of Political Control*, 3 J.L. ECON. & ORG. 243 (1987) [hereinafter McCubbins, *Procedures*]; Mathew D. McCubbins, et al., *Structure and Process, Politics and Policy: Administrative Arrangements and the Political Control of Agencies*, 75 VA. L. REV. 431 (1989) [hereinafter McCubbins, *Arrangements*]; Mathew D. McCubbins & Thomas Schwartz, *Congressional Oversight Overlooked: Police Patrols vs. Fire Alarms*, 28 AM. J. POL. SCI. 165, 165-79 (1984); Roger C. Noll, *The Behavior of Regulatory Agencies*, 29 REV. SOC. ECON. 15, 15-19 (1971) [hereinafter Noll, *Behavior*]; Roger C. Noll, *The Economics and Politics of Regulation*, 57 VA. L. REV. 1016, 1016-32 (1971); Pablo T. Spiller & Santiago Urbiztondo, Political Appointees vs. Career Civil Servants: A Multiple Principals Theory of Political Bureaucracies (March 1991) (unpublished manuscript, on file with author).

of sources: direct monitoring of a bureaucrat's activities, a bureaucrat's own report of bureaucratic activity, or the report of a knowledgeable third party. While each of these methods can provide a legislator with valuable information, all have serious drawbacks.

The primary drawback of direct monitoring is that it consumes large quantities of time and effort that could be expended towards other, perhaps more valuable, activities. Direct monitoring also defeats one of the main justifications for delegation: the efficiency gains possibly resulting from specialization and division of labor. If direct monitoring is prohibitively costly, then a legislator who wants to influence bureaucratic activity is forced to rely on someone else for information.

The advantage of relying on a bureaucrat's own report is that the bureaucrat is likely to have the information the legislator desires. The drawback of this strategy is that the bureaucrat may be reluctant to reveal valuable private information. When the bureaucrat is reticent and direct monitoring is prohibitively costly, a legislator's ability to learn about a bureaucrat's hidden knowledge depends solely on the legislator's ability to obtain information from an informed third party.

The advantage of relying on an informed third party is that the legislator does not have to bear the cost associated with direct monitoring. The drawback is that once a legislator provides an informed third party an opportunity to report on bureaucratic activity, the legislator also provides that person an opportunity to pursue his possibly distinct self-interest by misrepresenting bureaucratic actions in an attempt to mislead the legislator. In sum, legislators may be unable to acquire useful information if direct monitoring is prohibitively costly and they must rely on either bureaucrats or informed third parties to report on bureaucratic activities. As a result, bureaucratic knowledge may remain hidden despite attempts to uncover it.

While an individual who possesses hidden knowledge always has an opportunity to misrepresent what he knows when communicating with a relatively uninformed legislator, the incentive to do so does not always exist. Therefore, if a legislator wants to learn from the knowledge of others, she must possess some knowledge about the information provider's incentives. It follows that the question of interest to democratic theorists—namely, can legislators influence bureaucratic actions?—reduces to the question can legislators design contracts and other institutional features that affect the motives of information providers. If the answer to the latter question is yes, then legislators may be able to learn enough about the bureaucracy's hidden knowledge to manage delegated policymaking authority successfully. Otherwise, delegation and abdication will be equivalent.

We address the last question by developing a model of legislative-bureaucratic interaction. We first use the model to identify conditions under which a legislator can learn about a bureaucrat's hidden knowledge. We then use the model to show how legislators can create structures and processes that affect bureaucratic accountability. In so doing, we produce new, general conclusions

about the consequences of institutional design on democratic decisionmaking. We then review the actual practices of the U.S. Congress. We find that many congressional actions create the conditions for learning and, as a result, increase the likelihood that legislators can distinguish bureaucratic activities that are consistent with legislative interests from those that are not. Since Congress has many resources to create the conditions for learning, we conclude that even in the face of significant bureaucratic hidden knowledge, legislators can manage delegated authority.

The remainder of the article proceeds as follows: part II provides a definition of "learning" and explains how learning affects legislators' reactions to complexity; part III presents a model that highlights the relationship between the ability to learn, institutional design, and the consequences of delegation; part IV discusses ways in which Congress has designed institutions that help its members make more informed evaluations of agency policy proposals; and part V concludes. The appendix contains the technical foundations of our model, the derivation of our results, and a numerical example.

II

EXPERTISE AND LEARNING

Throughout this article we discuss decisionmaking under uncertainty. We define *uncertainty* as the inability to distinguish which of multiple possible "states of the world" is the true one. We are fairly certain that some uncertainty characterizes nearly all human decisions. Despite this uncertainty, people make decisions almost every waking moment of their lives. It follows that if people are uncertain about the consequences of their actions, they must make decisions based on their beliefs about the relationship between the actions they can take and the consequences of those actions. For the purpose of analysis, we define *beliefs* as a set of probabilities (summing to one) that an individual assigns to each of the conceivable states of the world. A *state of the world* is a complete specification of all relevant events that have occurred or will occur at a specified time, or a set of values for all relevant stochastic parameters at a specified time. We discuss two types of beliefs: *prior* beliefs, what an individual believes before observing an event, and *posterior* or updated beliefs, what an individual believes after observing an event.[6] The type of event on which we focus is signaling.

In our analysis, we distinguish between learning and knowledge. We define *knowledge* as the ability to assign a probability of one to a particular state of the world and a probability of zero to all other states of the world. In addition, we call the process by which an individual moves from prior beliefs to posterior beliefs *learning*. The ways that players learn, in our model, abide by Bayes's

6. We make the common technical assumption that all beliefs are "consistent," where consistency requires that the individual's beliefs assign a positive probability to the true state of the world. *See* David M. Kreps & Robert Wilson, *Sequential Equilibria*, 50 ECONOMETRICA 863 (1982).

Rule,[7] which is a method for rationally updating beliefs. Thus, learning is impossible in the absence of an event. Notice that learning does not necessarily impart knowledge; for instance, when an individual's prior and posterior beliefs are identical, we say the individual has learned nothing.

If all the signals sent from one person to another are known to be truthful, the story of how legislators overcome the effects of their relative lack of knowledge would be quite short. Legislators would design institutions that give other persons incentives to become informed about bureaucratic activity and to report their information to the legislature.[8] If all signals are not truthful, however, the construction of these institutions will not be sufficient for legislators to overcome the problems associated with their lack of knowledge.

The resolution of the disagreement about the effect of delegation depends on the answer to the question, are signals truthful? To answer this question, we begin with the premise that if a statement is not true, it is a lie. We then assert that there are two necessary conditions for lying: an opportunity and a motive. The opportunity to lie is ubiquitous in the act of communication. The above question then simplifies to another: do information providers have a motive to lie?

If it is known that an information provider has no motive for lying, even if she has the opportunity, we can conclude that the content of a signal from that information provider is truthful. If, by contrast, we are either uncertain about the information provider's motives or know that an information provider has a motive to lie, we cannot be certain about a signal's veracity. Under certain conditions, people can learn from both potential liars and actual liars. To better understand this and other implications of the relationship between a legislator's ability to learn about bureaucratic hidden knowledge, institutional design, motives for lying, and the consequences of delegation, we now move to the description and analysis of our model.

III

A MODEL OF OVERSIGHT AND THE CONSEQUENCES OF DELEGATION

Our main purpose is to identify both the consequences of delegation for legislators and the methods of institutional design that legislators can use to increase the likelihood that these consequences are beneficial for them.[9] We

---

7. This rule says that the posterior belief that a particular "state of the world," call it X, is the true "state of the world," given the observation of a particular event, call it E, equals "the probability E occurs given that X is the true state of the world times the prior probability that X is the true state of the world" divided by "the sum of the probabilities of each state of the world that includes event E." *See, e.g.*, IRVING H. LAVALLE, AN INTRODUCTION TO PROBABILITY, DECISION, AND INFERENCE 84-89 (1970).

8. *See* McCubbins, *Procedures*, *supra* note 5; McCubbins & Schwartz, *supra* note 5.

9. For simplicity, we treat the legislature as an individual actor. In effect, we assume that individual legislator preferences and the existing legislative institution have already interacted to produce a single legislative preference ordering over the possible alternatives. Our "legislative interest," then, is a generalization of what is often called the "median legislator's (or median committee member's)

begin with the observation that the consequences of delegation seem most bleak when, in delegating, the legislature abdicates. When legislative abdication occurs, the ties that bind the will of the governed and the actions of government are severed.

When delegation and abdication are equivalent, the consequences of delegation for the legislature depend entirely on whether the bureaucratic agent to whom the policy choice was delegated has policy desires similar to those of the legislators. If legislative and bureaucratic interests are similar, the consequences of delegation are likely to be beneficial to legislators. Otherwise, the consequences of delegation could be quite negative for legislators. By contrast, when delegation and abdication are distinct, the consequences of delegation for the legislature are determined not only by the overlap of legislative and bureaucratic interests, but also by the legislature's ability to reward beneficial bureaucratic actions and punish harmful ones. The greater this ability, the more likely it is that the consequences of delegation will be beneficial.

With few exceptions,[10] legislatures do not formally or intentionally abdicate their authority. In fact, a general characteristic of legislatures is that they retain some ability to affect bureaucratic policymaking. For instance, legislatures generally retain the right to reject bureaucratic policy initiatives through legislation.[11] Therefore, in general, if delegation and abdication are equivalent, some factor besides intent is responsible.

As long as the legislature retains the ability to reward and punish, abdication can occur only if the legislature delegates its authority and does not know whether rewarding or punishing is the appropriate response to a particular bureaucratic action. It follows that a necessary and sufficient condition for the equivalence of delegation and abdication is that the bureaucratic agent possess so much hidden knowledge about the consequences of its actions that relatively ignorant legislators are unable to distinguish beneficial bureaucratic activities from harmful ones. Thus, the key to understanding the consequences of delegation is to understand the conditions under which the legislature can learn enough to approve bureaucratic activities that have beneficial consequences for legislators and to reject those activities that have detrimental consequences. This, in our opinion, is where most previous research on delegation has gone astray.

In contrast, we use a model of oversight to identify the consequences of delegation. "Police patrols" and "fire alarms"[12] describe the modes of oversight

---

preferences."

10. *See* BRIAN LOVEMAN, THE CONSTITUTION OF TYRANNY: REGIMES OF EXCEPTION IN SPANISH AMERICA (1993). Loveman claims that several Latin American "constitutions of exception" limit a legislature's lawmaking powers, especially with respect to the authority of the military.

11. Of course, legislatures may have other, less costly ways of rejecting, or even amending, bureaucratic initiatives. For instance, explicit legislative approval might be necessary for bureaucratic proposals to become law (as is the case with the legislative veto), so that the legislature, through inaction, rejects the agent's proposal.

12. McCubbins & Schwartz, *supra* note 5, at 166.

available to legislators. Our model includes both types. Police-patrol oversight, or direct monitoring, is a situation where legislators buy knowledge of the true state of the world and become informed first parties. In this case, legislators pay the costs to monitor bureaucratic activities directly. Fire-alarm oversight is a situation where legislators receive signals from informed parties. These informed parties can be either the bureaucrats themselves (informed second parties) or constituents who have an interest in and information about bureaucratic activities (informed third parties).

The effects of hidden knowledge and learning on the consequences of delegation can be identified in a situation where a single bureaucratic agent, who may or may not have the same policy preferences as the legislature, can use its previously delegated authority to propose an alternative to an existing policy. After the agent has made such a proposal, the game we model begins. A legislative principal can use police-patrol oversight, fire-alarm oversight, neither, or both before it decides to accept or reject the proposal. If the principal obtains a sufficient amount of information from oversight activities, it can influence the consequences of delegation. Otherwise, delegation is equivalent to abdication.

A. A Description of the Model

We model a multi-stage, single-shot game between two players: an information-providing *fire alarm* and a legislative *principal.* The legislative principal's task is to render a judgment whether a previously offered bureaucratic proposal, denoted o, or the existing policy of the government, called the *status quo* and denoted sq, should prevail. These two policies are represented as points on the unit interval $[0, 1]$. Also represented as a point on this interval is an *ideal point* for each player. We assume that each player has single-peaked preferences, which means that neither player strictly prefers an outcome that is relatively far from its ideal point to an outcome that is relatively close to it.[13] Thus, each player's objective is to obtain the policy, o or sq, that is closest to its ideal point. Whether the principal and the fire alarm prefer the same policy or different policies is a critical variable within the model.

Another relevant characteristic of this game is that each player's actions may be costly to it. For example, both the fire alarm and principal know that the agent has paid cost $c_a \geq 0$ for the specific purpose of proposing o. The fire alarm and principal can also take costly actions (lying and direct monitoring, respectively) that are later described in greater detail. Unless stated otherwise, we assume that the value of each parameter in the game, such as the location of

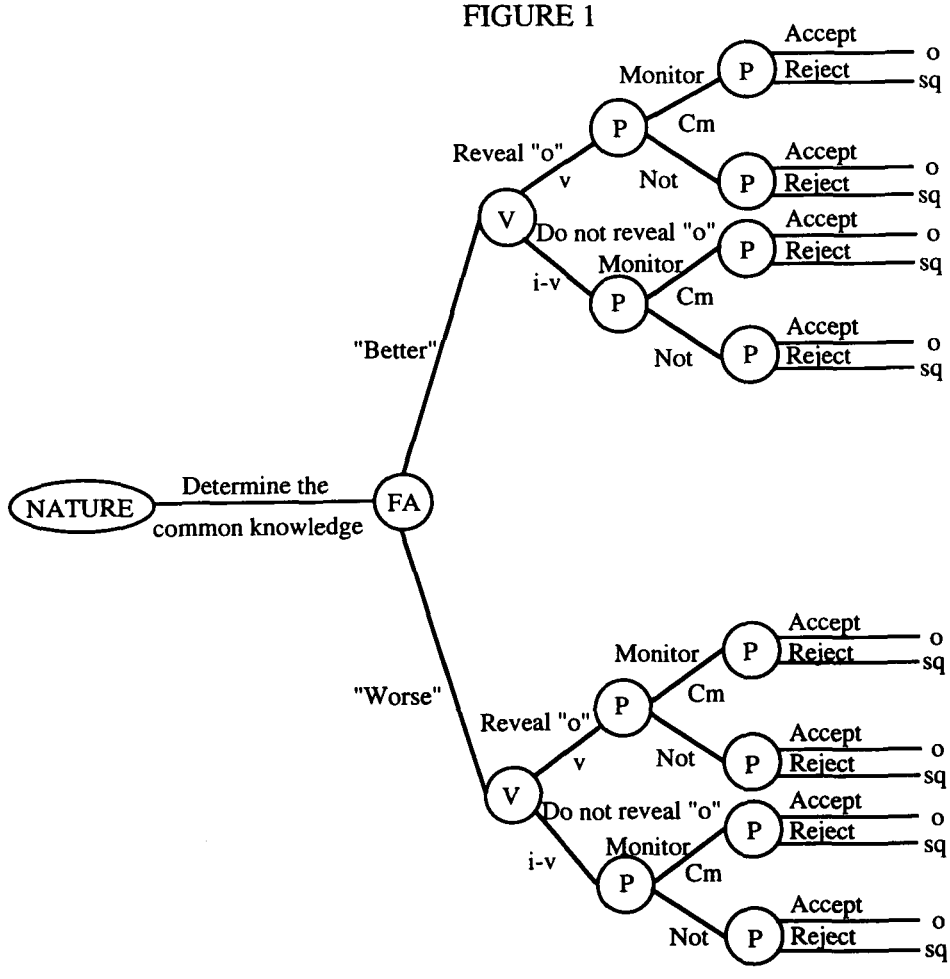
13. We assume that the principal's utility function is $-|X - P|$ and the fire alarm's utility function is $-|X - F|$, where P is the principal's ideal point, F is the fire alarm's ideal point, and $X \in \{o,sq\}$. Our results are dependent on neither unidimensionality nor the utility functions stated. These particular assumptions are used to explain the logic in relatively simple terms. It is easy to see that our results also prevail under the assumptions that o and sq are points in n-dimensional space (n finite) and utility functions are single-peaked. In addition, nothing precludes embedding in o and sq commonly held expectations of future play.

the status quo, the location of the principal's ideal point, and the magnitude of all costs, is common knowledge.

The exceptions to the common-knowledge assumption are that the locations of the proposal *o* and the fire alarm's ideal point may be the private information of the fire alarm. To say that "the fire alarm has private information about the spatial location of *o*" is equivalent to saying that the fire alarm knows something about the consequences of the principal's judgment of the proposal that the principal does not know (that is, whether *o* or *sq* is closer to the principal's ideal point). Similarly, to say that the fire alarm has private information about the location of its own ideal point is equivalent to saying that the fire alarm knows something that the principal does not know about the fire alarm's incentives.

Another element of the model is the structure of the interaction between the principal and the fire alarm. Simply stated, there are two significant actions: what the fire alarm says and what the principal does in response. To isolate the factors that lead these actors to take particular actions, we model the principal-fire alarm interaction as a series of five events, depicted in figure 1.

FIGURE 1



First, the values of several parameters are determined exogenously to the play of the game. These values are the location of the status quo (*sq*) and the bureaucratic policy proposal (*o*), the location of the principal and fire alarm ideal points, the magnitude of all costs, and the prior information about the locations of the fire alarm ideal point and *o*.[14] Second, the fire alarm decides to send one of two messages to the principal. The fire alarm can either tell the principal *o* is closer to your ideal point than *sq* (*o* is *better*) or *o* is at least as far from your ideal point as *sq* (*o* is *worse*). The fire alarm has the option of telling the truth or lying. Third, an exogenous third party called the verifier may reveal the true location of *o* to the principal. It is common knowledge that the verifier will reveal *o* to the principal with probability *v* and will reveal no new information with probability *1-v*. Fourth, after observing the signals, both costly actions and messages, sent by the fire alarm and verifier, the principal chooses whether to monitor the agent directly. If the principal chooses direct monitoring, it must pay a cost c_m, which represents the resources expended to learn the true location of *o*.[15] Fifth, and finally, the principal either accepts *o* or rejects it in favor of *sq*.

Several additional characteristics of this game complete its description. In considering the signal it will send, the fire alarm accounts for three factors: (1) the influence that its message can have on the outcome of the game; (2) the presence of a *penalty for lying*; and (3) the possibility that the veracity of its message will be revealed to the principal before the principal makes its moves. The "penalty for lying" in our model is $t \geq 0$. We examine the case where a dissembling fire alarm pays a penalty for lying only if the verifier verifies.[16] In effect, the expected penalty for lying is $t \times v$.

We introduce the penalty for lying and the possibility of message verification to demonstrate how an information provider's concern for its reputation and/or institutional incentives for providing truthful information can affect an information provider's credibility and the consequences of delegation. Although we do not model the verifier as a player whose strategies are determined endogenously to the play of the game, we include the possibility of verification in order to capture part of the dynamics produced by the presence of multiple fire alarms.[17] To demonstrate simply the effect of multiple fire alarms, we

---

14. In Figure 1, we follow game-theoretic custom and attribute these exogenous determinations to a player called "nature."

15. The case where direct monitoring is successful with some exogenously determined probability less than one follows straightforwardly.

16. An equivalent conceptualization is that the dissembling fire alarm pays t_h if the verifier reveals *o* and $t_l < t_h$, otherwise. We adopt the former for its expositional simplicity.

17. The effect of introducing a second fire alarm on the actions of an existing fire alarm can be explained in few words. If the second fire alarm can verify the message of the first, the first fire alarm's incentives for truth-telling are likely to be altered. If, by contrast, the second fire alarm lacks the ability to alter the principal's beliefs about the location of *o*, the second fire alarm is likely to have no impact on the first. Assumptions similar to the one we use are instrumental in the communication models of Milgrom and Roberts and Okuno-Fujiwara, Postlewaite and Suzumura. *See* Paul Milgrom & John Roberts, *Relying on the Information of Interested Parties*, 17 RAND J. ECON. 18 (1986); Masahiro Okuno-

introduce a verifier that appears with probability v. In short, high values of v represent cases in which the principal is likely to have another source that allows verification of the fire alarm message, and low values of v represent cases where verification is unlikely.

The final unique characteristic of this model is what the principal knows about the fire alarm. In addition to the principal's beliefs about the fire alarm's preferences over outcomes (that is, prior beliefs about the location of the fire alarm's ideal point), the principal knows that the fire alarm faces a penalty for lying and that the verification probability is v. As we will show, these pieces of information determine the principal's beliefs about the fire alarm's credibility. In addition, though the fire alarm is restricted in the type of message it can send, the intuition provided by examining this type of communication is quite general. Depending on the fire alarm's credibility, the principal can use "better than" and "worse than" messages to learn about the location of o and can make more accurate "how much better than" and "how much worse than" inferences. Thus, our model allows a relatively rich description of the principal's ability to learn about bureaucratic hidden knowledge.

B. The Conditions for Learning

Since we have established that the consequences of delegation depend on understanding the conditions under which the principal can learn about the bureaucrat's hidden knowledge, we begin by describing the conditions for learning that the model allows us to identify. Since we assume the principal can learn from police-patrol oversight, we will describe conditions for learning with respect to fire-alarm oversight. The conditions we derive in the appendix are the existence of positive penalties against the fire alarm for lying; the existence of observable and costly action by an informed person; the degree of similarity between fire alarm and principal preferences; and a possibility that the fire alarm's message will be verified. With the exception of a few extreme cases, none of the individual conditions is necessary or sufficient for learning; however, the principal will never be able to learn from fire-alarm oversight in the absence of all four conditions. In addition, comparative statistics demonstrate that the amount the principal can learn from fire-alarm oversight is nondecreasing in the size of the penalty for lying, the degree of preference similarity, the size of observable action costs, and the probability of verification.

The first condition for learning is the existence of a penalty for lying. This condition is a straightforward application of the economic concept of opportunity costs to the context of communication. When the marginal cost of lying is positive, lying will be worthwhile only when the expected benefit outweighs the expected cost. In the context of our model, we show that a sufficient condition for learning is that $t \times v$ is large enough so that some locations of o that the

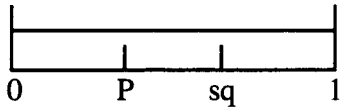
---

Fujiwara et al., *Strategic Information Revelation*, 57 REV. ECON. STUD. 25 (1990).

principal originally believed to be possible are known to be either less likely or impossible in the principal's posterior beliefs.

FIGURE 2

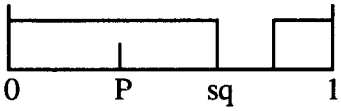
PRIOR BELIEF ABOUT THE LOCATION OF "o".



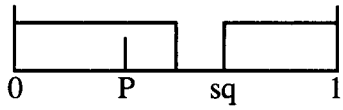
POSTERIOR BELIEFS ABOUT THE LOCATION OF "o".

Penalties for Lying (t x v > 0)

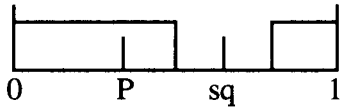
IF "Better"



IF "Worse"

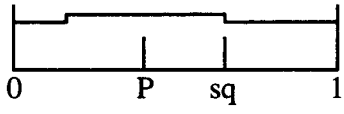


Costly Action



Similarity of Preferences

Low



High

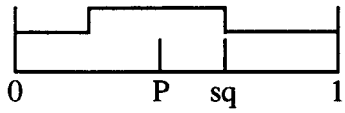


Figure 2 illustrates the spatial implications of the existence of a positive expected penalty for lying.[18] As the top portion of Figure 2 demonstrates, if the fire alarm sends the message "*o* is closer to your ideal point than is *sq*" and $t \times v > 0$, then the principal can correctly infer that *o* cannot be both close to *sq* and a little worse for the principal than *sq*. The principal can make such an inference because this set of potential locations of *o* does not offer the fire alarm sufficient benefit, relative to telling the truth and having the principal choose *sq*, to make lying worthwhile. Similarly, if the fire alarm sends the message "*o* is at least as far from your ideal point as is *sq*" and $t \times v > 0$, then the principal can correctly infer that *o* cannot be both close to *sq* and a little better for the principal than *sq*. Of course, the larger the expected penalty for lying, *ceteris paribus*, the less likely it is that the fire alarm would find it worthwhile to lie and the more likely it is that the fire alarm's message is true.

In short, if the fire alarm sends a particular message in the presence of a positive expected penalty for lying, then the principal can correctly infer that either the message must be true or the message is false and the fire alarm believed it was worthwhile to pay the penalty. Thus, what the principal learns in the presence of a penalty for lying is not necessarily that the fire alarm has told the truth, but rather that particular states of the world cannot be true. With this knowledge, the principal can use the content of the fire alarm's message to make more accurate inferences about the consequences of the bureaucratic proposal (the location of *o*).

The second "condition for learning" is the presence of costly action. The logic underlying this condition closely follows the old adage, "actions speak louder than words." In short, one person can learn about a second person's hidden knowledge by observing the choices that the second makes when some of the second person's actions are costly. While the logic that drives the effect of costly entry also underlies Michael Spence's finding about the ability of an employer to distinguish unskilled job applicants from skilled job applicants,[19] placing costly action in the context of delegation provides its own unique inference. The middle portion of Figure 2 depicts the effect of the agent's cost of proposing the policy *o* on the principal's beliefs about the location of *o*. If the principal knows or believes that the agent, like the principal and the fire alarm, has a single-peaked utility function and desires a policy that is closest to its own ideal point, the principal can infer that *o* must be sufficiently distinct from *sq* to make worthwhile the agent's payment of the proposal cost. Given that a proposal is made, it follows that the larger the proposal cost, the larger the likely difference between *o* and *sq*. While the existence of observable, costly actions does not allow the principal to learn all of the bureaucrat's hidden knowledge, it does allow the principal to approximate with relative accuracy the minimum

18. The case illustrated in Figure 2 is the case where the principal's prior beliefs about the location of *o* take the shape of a uniform distribution. While our results hold for any prior distribution, we illustrate the uniform case because it is the easiest to draw.

19. *See* Michael Spence, *Job Market Signaling*, 87 Q.J. ECON. 355 (1973).

possible difference between *o* and *sq*; this information can be quite valuable if the principal feels very differently about small and large changes to the existing status quo policy.

The third condition for learning is the degree of similarity between principal and fire-alarm preferences over outcomes. Underlying this insight is the principle that people can learn from others whose preferences are known to be similar to their own, since people with similar preferences have little incentive to mislead each other. The condition we identify here follows that originally derived by Crawford and Sobel[20] and applied insightfully in the context of lobbying by Austen-Smith[21] and in the context of committee-floor relationships by Gilligan and Krehbiel.[22] However, differences in the communication context between the Crawford-Sobel type models and our model allow us to draw unique inferences that are of particular usefulness for the problem at hand.[23]

The bottom portion of Figure 2 shows the spatial implications of preference similarity in our model. In short, the more likely it is that the fire alarm and the principal prefer the same outcome, the greater the weight the principal assigns to the fire alarm's claim being true. When the principal is certain that the fire alarm shares its preferences over outcomes, the fire alarm's message can be treated as though it were true. In contrast, when the principal is certain that the fire alarm has different preferences over outcomes, the fire alarm's message can be treated as though it were uninformative, regardless of whether it is actually true of false. When the principal is uncertain about the similarity of its and the fire alarm's preferences over outcomes, the principal's posterior beliefs about the location of *o* depend on the likelihood that the fire alarm shares the same preferences over outcomes. When the likelihood is high and the fire alarm signals "better," the principal's posterior beliefs place more weight on states of

---

20. Vincent Crawford & Joel Sobel, *Strategic Information Transmission*, 50 ECONOMETRICA 1431 (1982).

21. David Austen-Smith, *Information and Influence: Lobbying for Agendas and Votes*, 37 AM. J. POL. SCI. 799 (1993); David Austen-Smith, *Interested Experts and Policy Advice: Multiple Referrals under Open Rule*, 5 GAMES AND ECONOMIC BEHAVIOR 3 (1993) [hereinafter Austen-Smith, *Interested Experts*].

22. Thomas W. Gilligan & Keith Krehbiel, *Collective Decision Making and Standing Committees: An Informational Rationale for Restrictive Amendment Procedures*, 3 J.L. ECON. & ORG. 287, 308 n.28 (1987); Thomas W. Gilligan & Keith Krehbiel, *Organization of Information Committees by a Rational Legislature*, 34 AM. J. POL. SCI. 531 (1989) [hereinafter Gilligan & Krehbiel, *Organization of Information*].

23. To see this, we briefly review the differences between our model and Crawford and Sobel's. Crawford & Sobel, *supra* note 20. The two modeling approaches differ in three important ways. Two of the differences make our model more general than Crawford and Sobel's. First, unlike the information receiver in the Crawford-Sobel model, our principal is uncertain about the location of the fire alarm's ideal point. Second, unlike the information provider in the Crawford and Sobel model (called the sender), our fire alarm can send untruthful information. The third difference makes Crawford and Sobel's model more general than ours. Their information provider chooses its message from an infinite vocabulary, while our fire alarm can say only "better" or "worse." We believe that while both approaches are useful, ours is particularly well suited for the problem of delegation where legislators often have access to several types of information about the fire alarm and very little time to obtain detailed information on a particular issue.

the world where "better" is true and less weight on states of the world where "better" is false.

The fourth and final condition for learning is the possibility of verification. The primary effect of an increase in the likelihood of verification is an increase in the probability that the principal will be able to choose the policy (*o* or *sq*) that is closest to its ideal point. While this primary relationship is irrelevant to the immediate discussion, a secondary effect is less obvious and quite important. If the fire alarm believes its message is likely to be verified, dissembling is less likely to get the fire alarm the outcome it desires. As a result, the fire alarm is more likely to provide truthful information as the likelihood of verification increases. This dynamic is especially pronounced if the payment of the penalty for lying depends on the likelihood of message verification, as it is in our model. In essence, the possibility of verification allows the principal to learn from a fire alarm who otherwise lacks those characteristics that are likely to enable the principal to learn from the fire alarm's message.

In sum, if the expected penalty for lying, the costs of observable actions, the probability that principal and fire alarm preferences are similar, or likelihood of verification are relatively large, then the principal is relatively likely to be able to use fire-alarm oversight to learn about the consequences of the bureaucrat's policy choice. When these conditions for learning are absent, the principal will have no such ability. Thus, delegation and abdication will be equivalent when police-patrol oversight is prohibitively costly and the conditions for learning from fire alarms are absent. As a result, a legislator who wants to use institutional design to improve the consequences of delegation will experience greater success if she can use her design capabilities to create the conditions for learning from fire-alarm oversight.

C. Designing Bureaucratic Accountability

We now answer the following question: how can legislators increase the likelihood that delegation has beneficial consequences for them? It should be obvious that legislators can affect the consequences of delegation by either creating the conditions under which they can learn about bureaucratic hidden knowledge or by changing the nature of the authority they have delegated. It remains for us to show which institutional alterations will be most effective. We begin by describing designs that can affect the principal's ability to learn from oversight. We then describe designs that can affect the incentives of a bureaucratic agent.

1. *Designing Institutions to Facilitate Learning.* A legislator who wants to learn about bureaucratic hidden knowledge would like to avoid paying the costs associated with such an education. As McCubbins and Schwartz argue,[24] this is why a fire-alarm oversight system is likely to be favored over direct monitoring

24. McCubbins & Schwartz, *supra* note 5.

when both are equally informative. How does a legislator design an informative fire-alarm system? One way is to implement penalties for lying. When these penalties can be established, fire alarms who would otherwise pursue their self-interests by misleading legislators may find that truth-telling is more rewarding. To be effective, penalties for lying need neither be large nor applied with certainty. However, the product of (1) the probability that liars will be caught, (2) the probability that caught liars will be punished, and (3) the size of the penalty, must be positive. Ensuring that each of these requirements is present may be difficult, however. This is particularly true of the first requirement.

Our model relies on an exogenous verification device. In reality, a legislator will have to create a verifier or an interested third party will have to convince the legislator it is qualified to be a verifier. How easy is it to create or become a verifier? To transform a regular fire alarm into someone who can play the role of the verifier in our model, it is necessary to impose on the fire alarm a large penalty for lying or to assume it is common knowledge that a second fire alarm disagrees about which outcome should be chosen. The necessary and sufficient condition is that these two factors, in combination or alone, remove an informed third party's incentive to lie. When this condition is satisfied, the principal can learn from a fire alarm's message even if the fire alarm is not subject to a penalty for lying, cannot take costly action, and is not perceived to have the same interests as the principal. In the absence of a verifier, the presence of one of these other conditions is necessary for learning.

One way that a legislator can create a verifier is to induce competition among adversarial fire alarms. For fire alarms who care about outcomes and their reputations as valuable providers of information, the presence of an adversary who would like to damage the fire alarm's reputation for veracity should induce the fire alarm to reveal only information that it believes to be true. Thus, the types of issues for which this design strategy is most likely to be effective are those whose consequences are important to many informed interests.

Beyond establishing the possibility of verification, legislators have other methods to create effective oversight systems. A well-known method is screening. From the preference similarity dynamics discussed earlier, it follows that legislators should find fire alarms that share their policy preferences. However, a legislator who relies on screening alone adopts a risky strategy when her ability to perceive fire alarm preferences accurately is limited; that is, people can be misled by others they trust. In contrast, screening combined with penalties for lying and with verification results in an oversight system with a fire alarm whom the principal believes is unlikely to have a reason to dissemble, as well as institutional characteristics that discourage dissembling in the event that the screening process was imperfect.

2. *Designing Institutions to Influence Bureaucratic Incentives.* A legislature wants to dissuade bureaucrats from taking actions that conflict with legislative interests. When the legislature creates the conditions for learning, it increases the likelihood that it will be able to distinguish bureaucratic actions with

beneficial consequences from those with negative consequences. When the legislature credibly threatens to become more discriminating, it affects the agent's incentive structure. If there are opportunity costs associated with taking actions and all else is constant, a bureaucrat prefers to take actions that will not be reversed. If legislators who can learn are more likely to reverse actions with negative consequences and less likely to reverse others, the bureaucrat, if all else is constant, will have a greater incentive to make proposals that have beneficial consequences for the legislature.

If creating the conditions for learning is either difficult or relatively costly, a legislature may be able to improve the consequences of delegation by changing the amount of authority it delegates. For instance, the legislature could raise the "price" to the bureaucracy of taking particular actions.[25] However, this design strategy involves a tricky trade-off. On the one hand, increasing the price of bureaucratic action decreases the likelihood of bureaucratic action. In fact, setting the price high enough is equivalent to not delegating at all. On the other hand, should a bureaucrat take action, the legislature is likely to be less uncertain about the action's consequences. The legislature's knowledge of the requirement of costly action by the bureaucrat allows a relatively accurate approximation of the minimum possible difference between the bureaucrat's proposal and the existing status quo. If the legislature desires bureaucratic actions that result in large changes to the status quo, would otherwise be unable to distinguish between small and large changes, and is relatively skilled at distinguishing large beneficial changes from large negative changes, then raising the price of bureaucratic activity would be an effective institutional alteration. If, by contrast, the principal is relatively certain that the agent shares her policy preferences, then restricting bureaucratic activity in this way might produce a net loss.

IV

MAKING BUREAUCRATS ACCOUNTABLE

In constructing an institution that allows a legislator to learn about a bureaucrat's hidden knowledge, the legislator must be able to affect the motives for lying that accompany the opportunity to communicate. A legislator's beliefs about the types of motivations that information providers are likely to have should affect her beliefs about what types of institutional design are likely to be effective. Whether legislators can and will design such institutions is the subject of our final discussion.

We turn first to the issue of cost. No bureaucratic agency can take action without cost to itself. Every agency has limited resources in terms of budget and staffing. Hence, bureaucrats must make choices as to how these resources will be expended. It follows that in choosing any particular action, bureaucrats send

25. This action would be an example of what McCubbins, Noll, and Weingast might call "stacking the deck" against particular actions. McCubbins, *Procedures*, *supra* note 5, at 261-64.

a signal that changing this particular policy will be so beneficial to their interests that they have found it worthwhile to bear these costs.

Agency actions necessarily fall in one of two categories: (1) rulemaking, or statements of policy; and (2) adjudication, or applications of general rules and policies to specific situations. Congress has imposed costs on agencies in taking actions of either sort. The broadest imposition of such costs arises from the Administrative Procedure Act of 1946 ("APA").[26] The APA established general criteria that administrative agencies must satisfy when creating new policies or writing rules of general applicability. Agencies must give public notice announcing their intentions to make policy of a specific sort.[27] They then must solicit comments from interested groups and individuals who wish to express their views on how such a new policy should be written.[28] After drafting a proposed rule, they must expose that draft to public criticism.[29] Each of these stages in the rulemaking process consumes time and resources. At the very least, bureaucrats must maintain a paper trail.

Beyond the provisions of the APA, Congress affects agency rulemaking with statutes authorizing regulatory activities and through administrative mandates. These statutes and mandates affect the costs faced by agencies that want to challenge the existing policy and the incentives to those who would provide information about bureaucratic activity.

A. Affecting the Costs of Bureaucratic Action

The Constitution empowers Congress to regulate numerous activities, from interstate commerce to bankruptcy law to the use of federal roadways to the establishment and protection of intellectual property rights through patent and copyright laws.[30] Federal law requires many business activities to obtain federal or state permits or licenses: for example, the construction and operation of nuclear power plants,[31] coal mining,[32] radio and television broadcasting.[33] Obtaining a license is a cost of business, and from the agency's perspective, issuing a license is both a costly action (because the agency must conform to the general rules it previously promulgated) and a message that subjects the agency to penalties for lying.

Issuing a license is, in essence, a statement by the agency that the licensee has met and will continue to meet the substantive and procedural requirements of the relevant laws and regulations. The Surface Mining Control and Reclamation Act of 1977,[34] for example, requires strip-mining companies to

---

26. Pub. L. No. 404, 60 Stat. 237 (codified as amended in scattered sections of 5 U.S.C.).
27. 5 U.S.C. § 553(b) (1988).
28. *Id.* § 553(c).
29. *Id.* § 553(e).
30. U.S. CONST. art. I, § 8.
31. 42 U.S.C. § 2131 (1988).
32. 30 U.S.C. § 1252 (1988).
33. 47 U.S.C. § 301 (1988).
34. Pub. L. No. 95-87, 91 Stat. 445 (codified as amended in scattered sections of 18 and 30 U.S.C.).

submit to federal or state regulatory authorities plans for protecting land and water resources adjacent to mines from toxic wastes generated during mining and plans for reclaiming the mine site after the mine is played out.[35] Each state that wishes to establish jurisdiction over the regulation of surface coal mining is required to establish standards that all mining permit applications must meet.[36] These state-established standards are subject to federal agency supervision[37] and review by the courts.[38]

Congress also chooses the level of costs to impose on agencies delegated the responsibility to enforce specific laws by establishing the number and range of regulatory decisions subject to review by other agencies or courts. An example of this sort of costly action is toxic chemical regulation by the Environmental Protection Agency ("EPA") under the Toxic Substances Control Act of 1976.[39] The Act requires the EPA to regulate substances found toxic to human life. Pursuant to this goal, the EPA must propose testing requirements for determining whether a substance is harmful to health or the environment before it can promulgate a rule to regulate the substance.[40] Thus, if it wants to regulate a chemical, the EPA must undertake two costly actions: designing a test rule and writing a regulation after the tests are done.

Another tool often used to raise or lower the cost of agency action is the definition of evidentiary standards to be used in courts. Evidence law prescribes rules for determining the burden of producing evidence and the burden of persuasion.[41] The burden of production determines which party must present evidence (bear costs) in order to proceed.[42] The burden of persuasion describes the tests a party must meet in order to carry an issue.[43]

Congress also has found numerous ways to raise or lower costs of taking action for potential fire alarms and has established penalties for lying. One of the best known costs of taking action for a fire-alarm is obtaining standing to sue an agency for actions it has taken.[44]

In general, to "make a federal case" out of a regulatory proceeding, an aggrieved citizen must show actual damages to his own interests, not merely

---

35. 30 U.S.C. §§ 1257-1258 (1988).
36. *Id.* § 1253(a).
37. *Id.* § 1253.
38. *Id.* § 1276.
39. Pub. L. No. 94-469, 90 Stat. 2003 (codified as amended in scattered sections of 15 U.S.C.).
40. 15 U.S.C. § 2603 (1988).
41. ARTHUR E. BONFIELD & MICHAEL ASIMOW, STATE AND FEDERAL ADMINISTRATIVE LAW 574-75 (1989).
42. *Id.*
43. *Id.* at 575.
44. For the nonlegal reader, an explanation of the concept of standing is in order. Under Article III of the Constitution, federal courts may hear cases only in which there exists a controversy between at least two parties, each of whom has a sufficient stake in the outcome to justify court action. Hence, for example, if the parents of black public-school students wished to sue the Internal Revenue Service for recognizing the tax-exempt status of racially discriminatory private schools because that recognition caused "stigmatizing injury" to their children, federal courts would deny standing to the plaintiffs unless the plaintiffs' children were personally denied equal treatment. Allen v. Wright, 468 U.S. 737 (1984). *See also* BONFIELD & ASIMOW, *supra* note 41, at 683-703.

concerns for the general welfare. But in some statutes, such as the APA, the National Environmental Policy Act of 1969 ("NEPA"),[45] or the Atomic Energy Act of 1954,[46] standing to challenge regulatory proceedings is defined broadly. The APA states that "[a] person suffering legal wrong because of agency action, or adversely affected or aggrieved by agency action . . . is entitled to judicial review thereof."[47] The courts have held that this provision allows citizens to challenge agency regulations for a variety of economic and noneconomic reasons.[48] Environmental legislation such as NEPA and the Endangered Species Act of 1973[49] has extended citizens' ability to challenge agency decisions to questions of general environmental quality and the interests of wildlife species classified as endangered or threatened.[50] In some cases, federal agencies have set up offices of consumers' counsel to assist these litigants financially by providing them with the resources to challenge agency proposals in court.[51] These provisions reduce costs of entry for fire alarms and increase penalties on the agencies for lying.

B. Verification and Competition

When interested constituents fail to win satisfaction in the courts or agencies directly, they often take their case to Congress. This is, of course, the classic example of fire-alarm behavior. How does Congress dissuade these constituents from exaggerating their claims or telling outright lies?

It is commonly observed that affluent interest groups have considerable access to Congress. They contribute heavily to congressional reelection campaigns and, in turn, members of Congress find time in their hectic schedules to listen to them. But a member's time is finite. Committee meetings, floor debates, and trips back home to shake hands and kiss babies clutter their schedules, leaving only narrow windows of opportunity for interest groups. Since there are thousands of interest groups lobbying Congress for every conceivable cause, the competition between interest groups for members' time is fierce. The wise interest group, therefore, guards its access jealously by providing legislators with accurate, succinct information on its favored issues, because once a legislator's trust has been broken by an overzealous lobbying effort, there may be little opportunity to win it back. Interest groups, in other words, may compete to play the role of verifier for legislators.

Legislators also have constructed adversarial fire-alarm systems and "verifier" agencies to monitor the actions of other actors. The most famous is the Office of Management and Budget ("OMB"), created in the 1921 Budget and

45. 42 U.S.C. §§ 4321, 4331-4335, 4341-4347 (1988).
46. Pub. L. No. 703, 68 Stat. 919 (codified as amended in scattered sections of 42 U.S.C.).
47. 5 U.S.C. § 702 (1988).
48. Japan Whaling Ass'n v. American Cetacean Soc., 478 U.S. 221, 230 n.4 (1986).
49. Pub. L. No. 93-205, 87 Stat. 884 (codified as amended in scattered sections of 7 and 16 U.S.C.).
50. 16 U.S.C. § 1540(g) (1988).
51. Arthur Earl Bonfield, *Representation for the Poor in Federal Rulemaking*, 67 MICH. L. REV. 511, 538 (1969).

Accounting Act[52] to help the president compile and submit an executive budget.[53] The OMB and the president are authorized to suggest to Congress changes to existing law if such changes are accompanied by detailed justifications. The Budget and Accounting Act also created the General Accounting Office ("GAO"), a special agent of Congress independent of the executive branch.[54] The GAO acts as a trustee for legislators. It is Congress's auditor and accountant, examining agencies' books at the end of the fiscal year, and its comptroller, checking the flow of funds to agencies throughout the year against what has been authorized and appropriated by law.[55] In addition, the GAO performs special investigations of agency policy performance under standing authority and by special request of individual members of Congress.

V
CONCLUSION

We have shown that, under general circumstances, legislators can uncover at least some of the bureaucracy's hidden knowledge. This ability allows legislators to mitigate, at least in part, the deleterious effects of bureaucratic expertise. Our results offer a challenge to those who argue that the specialized knowledge of bureaucrats puts them in a dominant position relative to the legislature. Although legislators may never possess the specialized information of the bureaucracy, they can use institutional design to learn enough to make decisions that sometimes are equivalent to the decisions they would have made if they possessed all of the bureaucracy's hidden knowledge.

Legislators have a box of tools they can use to improve the consequences of delegation for themselves. That they use this box of tools is apparent from an examination of statutory law, which is replete with measures intended to alter the costs and benefits of particular bureaucratic actions. Therefore, the simultaneous appearance of delegation and bureaucratic expertise need not be equivalent to abdication.

52. Pub. L. No. 13, 42 Stat. 20 (codified as amended in scattered sections of 31 U.S.C.).
53. 31 U.S.C. § 501 (1988).
54. *Id.* § 702.
55. *Id.* §§ 711-720.

APPENDIX

The purpose of this appendix is to describe a model of the consequences of delegation. In it, we provide a formal definition of every aspect of our model and derive the results upon which the conclusions drawn in the text are based. The appendix is organized as follows: in part A, we describe all of the premises upon which the model is based; in part B, we draw conclusions from these premises about both the conditions under which a legislative principal can learn about the consequences of accepting or rejecting a particular bureaucratic proposal and the nature of the learning process; in part C, we use the basic premises and the conclusions about learning to describe the consequences of delegation; in part D, we describe a simplifying assumption; and in part E, we provide a numerical example.

A. Basic Premises

Two players, called the principal and the fire alarm, play a single-shot game. Unless otherwise stated, all aspects of this game are common knowledge. The purpose of the game is to choose one of two exogenously determined points, o and sq, on the line segment $[0, 1]$. This choice determines a payoff in utils for each player; each player's objective is to maximize his or her own utility. The principal's utility function is $-|X - P|$ and the fire alarm's utility function is $-|X - F|$, where P is the principal's ideal point, F is the fire alarm's ideal point, and $X \in \{o, sq\}$. For expositional simplicity, we discuss the case where $P < sq$. Our results are without a loss of generality to the case $P > sq$, which is equivalent, and the case $P=sq$, which is trivial.

The single exception to the common knowledge assumption is that the locations of o and F may be known only to the fire alarm. We assume that the locations of o and F are the results of single draws from the distributions O and Γ, respectively. O has density O', Γ has density Γ' and each has support on known, but undenoted, subsets of $[0, 1]$. In effect, we assume that O and Γ are common knowledge and that only the fire alarm observes the result of the draw from each distribution. If either distribution has mass at more than one point, then the fire alarm has private information. For expositional simplicity, we examine the case where $O'(sq) = 0$. It may also be known that o's proposer, an agent who is assumed to have taken actions prior to the play of this game, had the same-shaped utility function as the principal and fire alarm and paid $c_a \geq 0$ for the privilege of proposing o.

The fire alarm makes the game's first move when it decides to send one of two messages, $M_F(F, sq, o, t, P, O, c_m, \Gamma, v, c_a) \in \{B, W\}$. B (better than sq for the principal) means that $o \in (sq - 2x(sq - P), sq)$. W (worse than sq for the principal) means that $o \in [0, sq - 2x(sq - P)] \cup (sq, 1]$. The fire alarm is not restricted to the transmission of a truthful message, but may have to pay an additional penalty for lying, t, if it chooses to dissemble. Whether a dissembling fire alarm has to pay the penalty for lying depends on the actions of a third player called the verifier. The verifier is a player whose actions are determined exogenously to the play of

this game. After the fire alarm has signaled, the verifier reveals the true location of o to the principal with probability v and reveals no new information (signals the distribution O) with probability $1 - v$ ($M_V(v, o, sq) \in \{O, o\}$). If the fire alarm has dissembled and the verifier reveals the true location of o, the fire alarm pays the penalty for lying, otherwise it does not. After receiving messages from the fire alarm and the verifier, the principal can choose to pay c_m to learn the location of o ($MON(P, sq, (O, o), c_m, c_\varphi, M_P, t, v, M_V) \in \{Y, N\}$). The principal then makes the game's final move by choosing either o or sq ($APP(P, sq, (O, o), c_m, c_\varphi, M_P, t, v, M_V) \in \{Y, N\}$).

This article's equilibrium concept is a variant of the sequential equilibrium concept of Kreps and Wilson.[56] A sequential equilibrium consists of strategies that players believe to be the best responses to the chosen strategies of others, prior beliefs that are consistent, and an updating procedure that is based on Bayes's Rule.[57] Consistency implies that player beliefs assign positive probability to the true state of the world.

The variation we introduce is that we assume the principal utilizes an exogenously determined algorithm to decide whether or not to condition her beliefs on her knowledge of the fire alarm's strategy. We introduce this concept to simplify the formal statement of the model and the exposition that follows. The algorithm suggests that a principal with limited cognitive resources will opt to consider the fire alarm's statement if she expects, without explicitly considering all possible outcomes of the game, that doing so will increase the probability that she makes the same decision she would have made had she known the location of o. An algorithm with these characteristics is proposed in section D. The remainder of our analysis focuses on the case in which the algorithm directs the principal to use information about the fire alarm and the fire alarm's strategy to update her prior beliefs about the location of o.[58] The validity of our results relies on the validity of this concept, since we do not examine the consequences of play that strays from the equilibrium path. In the description of this model's equilibria, we also employ the following tie-breaking rules: (1) If the expected benefit of an action (that is, proposing, dissembling, and monitoring) is not strictly positive, then this action is not taken. (2) If sq and o provide the principal with the same expected utility, then the principal chooses sq.

B. Interim Steps

We now describe the factors that enable the principal to learn about the location of o. For each factor, we detail the minimum inference that can be

56. Kreps & Wilson, *supra* note 6.

57. *See supra* note 7.

58. Stated another way, we assume that the principal uses information about the fire alarm and the fire alarm's strategy to update her prior beliefs about the location of o.

drawn given that the algorithm directs the principal to consider information about the fire alarm.

1. Penalty for Lying and Verifiability. Let τ be the smallest distance from the point *sq* for which the fire alarm could find the payment of the expected penalty for lying *(t x v)* to be worthwhile. Since *sq, t, v* and the shape of the fire alarm's utility function are common knowledge, so is τ.

Lemma 1: In the presence of a penalty for lying *t* and verifier *v*, truth telling is a dominant partial strategy for the fire alarm when $o \in [sq - \tau, sq + \tau]$.

Proof: When $o \in (sq, sq + \tau]$, a fire alarm that signals *B* expects to pay penalty *(t x v)*. A fire alarm that signals *W* when $o \in [sq - \tau, sq)$ has a similar expectation. In each case, the definition of τ implies that the maximum possible benefit to the fire alarm of affecting the outcome by dissembling cannot possibly be higher than the penalty for lying. Therefore, truth-telling is an undominated partial strategy in the cases described.

In effect, if the principal observes *B* in the presence of expected penalty for lying *t* x *v*, she learns that *o* cannot be located in the interval $(sq, sq + \tau]$. Similarly, if she observes *W* in the presence of expected penalty for lying *t* x *v*, she learns that $o \notin [sq - \tau, sq)$. The following propositions follow straightforwardly from Lemma 1, and are offered without proof.

Proposition 1: *Learning from a penalty for lying.* The density of *O* at *o* (or a closed interval of small and positive length with endpoints that are equidistant from *o*) in the principal's posterior beliefs minus the density of *O* at that point (or interval) in the principal's prior beliefs is nondecreasing in *t*.

Proposition 2: *Learning from the presence of a verifier.* The density of *O* at *o* (or a closed interval of small and positive length with endpoints that are equidistant from *o*) in the principal's posterior beliefs minus the density of *O* at that point (or interval) in the principal's prior beliefs is strictly increasing in *v*.

These propositions state that the presence of a penalty for lying and the presence of a verifier allow the principal to make more accurate inferences about the spatial location of *o*. More accurate inferences are possible because the presence of positive *t* and *v* are sufficient to allow the principal to identify certain locations of *o* as impossible. Also, the statement of proposition 2 is stronger than the statement of proposition 1 because, while both *t* and *v* lead to the same size increase in the expected penalty for lying, only an increase in *v* directly increases the probability that the principal observes *o*.

2. Costly Entry. We assume that, prior to the beginning of play in this game, a bureaucratic agent proposed the alternative *o*. While the agent's actions are not

explicitly modeled in this article, we utilize findings from related models[59] whose logic transfers straightforwardly to describe what the principal can learn from her knowledge about the magnitude of the agent's proposal costs. If the approximate shape of the proposer's single-peaked utility function—the fact that the proposer paid c_a for the privilege of making a proposal and the fact that all of the returns to making a proposal—accrue at the moment that the principal either accepts or rejects the proposal are common knowledge, then the principal can use information about the proposer's costs to form more accurate beliefs about the spatial location of o. These relatively accurate beliefs can be formed because the principal can identify a range on $[0, 1]$ for which the maximum benefit from making an accepted proposal in this range could not be greater than the proposal cost. Let ε be a nondecreasing function of c_a that equals the maximum distance for which the benefit of making a proposal could not possibly be larger than c_a. We have shown that the principal can use information about proposal costs to make the following inference.[60]

Proposition 3: *Learning from costly entry.* If the principal observes that an offer was made in the presence of proposal cost c_a, then she can infer that $o \notin [sq - \varepsilon, sq + \varepsilon]$.

3. Perceived Similarity of Preferences and Simultaneous Effects. We now move to the relationship between the fire alarm's incentives for truth-telling and the similarity of fire alarm and principal preferences. We begin by making a preliminary claim whose proof is straightforward and follows the same logic as that found in Lemma 6 of Crawford and Sobel.[61]

Lemma 2: When it is common knowledge that $-|o - F| > -|sq - F|$, $-|o - P| > -|sq - P|$, then the fire alarm should send B and the principal should treat the message as though it were true. Similarly, when it is common knowledge that $-|o - F| \leq -|sq - F|$ and $-|o - P| \leq -|sq - P|$, then the fire alarm should send W and the principal should treat the message as though it were true. When it is common knowledge that $(t\ x\ v) = 0$ and either $-|o - F| \leq -|sq - F|$ and $-|o - P| > -|sq - P|$ or $-|o - F| > -|sq - F|$ and $-|o - P| \leq -|sq - P|$, then the principal should disregard the content of the fire alarm's message.

By applying the logic of Lemma 1, Proposition 3, and Lemma 2, we can now describe the simultaneous impact of proposal costs, penalties for lying for the fire alarm, the presence of a verifier, and the perceived similarity of fire alarm and principal preferences on the principal's beliefs about the location of o. Let s_β be

59. Arthur Lupia, *Busy Voters, Agenda Control and the Power of Information*, 86 AM. POL. SCI. REV. 390 (1992); Arthur Lupia & Mathew D. McCubbins, *Learning from Oversight: Police Patrols and Fire Alarms Reconstructed*, 10 J.L. & ECON. PROBS. 96 (1994); Thomas Romer & Howard Rosenthal, *Political Resource Allocation, Controlled Agendas, and the Status Quo*, 33 PUBLIC CHOICE 27 (1978); Spence, *supra* note 19.

60. Lupia & McCubbins, *supra* note 59.

61. Crawford & Sobel, *supra* note 20.

the probability that o is better than sq for both the principal and the fire alarm and let s_W be the probability that o is worse than sq for both the principal and the fire alarm. Let s_B be the probability that o is better than sq for both the principal and the fire alarm and let s_W be the probability that o is worse than sq for both the principal and the fire alarm. If $\varepsilon \geq 2\ x\ (sq - P)$, *then* $s_B = 0$. Otherwise, $s_B =$

$[O(sq-\varepsilon)-O(sq-(2x(sq-P))]$ x $prob[o:-|F\text{-}o|>-|F\text{-}sq$ and $o \in (sq-(2x(sq-P)),sq-\varepsilon]$.

If $\varepsilon < 2$ x $(sq - P)$, then $s_W =$

$[1-O(sq+\varepsilon)+O(sq-(2x(sq-P))]x$

$prob[o:-|F\text{-}o|\leq-|F\text{-}sq$ and $o \notin (sq-(2x(sq-P)),sq+\varepsilon)]$.

Otherwise, $s_W =$

$[1-O(sq+\varepsilon)+O(sq-\varepsilon)]$ x $prob[o:-|F\text{-}o|\leq-|F\text{-}sq$ and $o \notin (sq-\varepsilon,sq+\varepsilon)]$.

Notice that since P is common knowledge, it must be the case that $s_B = 0$ and/or $s_W = 0$ and that $s_B + s_W \leq 1$. Also notice when $\varepsilon \geq 2\ x\ (sq - P)$, o cannot be better for the principal than is sq.

Let d_B be the common prior probability that the principal and the fire alarm have different preferences over the set o, sq when $o \in [sq - \tau, sq - \varepsilon)$ and $\varepsilon < \tau$. Let d_W have an equivalent definition for the case $o \in [sq + \varepsilon, sq + \tau)$ and $\varepsilon < \tau$. d_B and d_W are the probabilities that the penalty for lying is large enough to persuade a fire alarm, who would otherwise find dissembling worthwhile, to send a truthful message. If $\tau > \varepsilon$, then:

$$d_B \in [0,1] = (O(sq-\varepsilon)-O(sq-\tau))x$$
$$[prob(o:-|F\text{-}o|>-|F\text{-}sq|,-|o\text{-}P\leq-|sq\text{-}P| \text{ if } o\in[sq-\tau,sq-\varepsilon))$$
$$+prob(o:-|F\text{-}o|\leq-|F\text{-}sq|,-|o\text{-}P|>-|sq\text{-}P| \text{ if } o\in[sq-\tau,sq-\varepsilon))]$$
$$d_W \in [0,1] = (O(sq+\tau)-O(sq+\varepsilon))x$$
$$[prob(o:-|F\text{-}o|>-|F\text{-}sq|,-|o\text{-}P|\leq-|sq\text{-}P| \text{ if } o\in[sq+\varepsilon,sq+\tau))$$
$$+prob(o:-|F\text{-}o|\leq-|F\text{-}sq|,-|o\text{-}P|>-|sq\text{-}P| \text{ if } o\in[sq+\varepsilon,sq+\tau))]$$

If $\varepsilon \geq \tau$, then $d_B = d_W = 0$.

It follows that $1 - s_B - s_W - d_B - d_W$ is the common prior probability that the fire alarm is a type that could find it profitable to lie. (For those familiar with the argument presented by Lupia and McCubbins,[62] it is worthwhile to point out that the value of the s and d terms used here are a function of ε, while the equivalent terms in our previous work have no such affiliation.)

From Lemma 1, Proposition 3, and Lemma 2, it follows that when $\varepsilon \leq \tau \leq 2\ x\ (sq - P)$ (that is, lying can be profitable) and $M_V = O$, the principal's posterior beliefs ($O'\ (o\,|B, s, d, \tau, \varepsilon)$) are related to her prior beliefs (O') in the following manner:

(The case where B could have been sent by a fire alarm who either has the same preferences over outcomes or is attempting to mislead.)

---

62. Lupia & McCubbins, *supra* note 59.

$$O'(o|B)= \left(\frac{s_B}{1-(s_W+d_W)}\times\frac{O'}{o(sq-\tau)-o(sq-2xOsq-P))}\right) + \left(\frac{1-s_W-d_W-s_B-d_B}{1-(s_W+d_W)}\times\frac{O'}{1-O(sq+\tau)+O(sq-\tau)}\right) \quad o\in[sq-2\times(sq-P),sq-\tau)$$

$$O'(o|B)= \left(\frac{d_B+s_B}{1-(s_W+d_W)}\times\frac{O'}{O(sq-\varepsilon)-(1-O(sq-\tau))}\right) o\in[sq-\tau,sq-\varepsilon)$$

(The case where B could have been sent by a fire alarm who either has the same preferences over outcomes or faces a large penalty for lying.)

$$O'(o|B)= \qquad O \qquad o\in(sq-\varepsilon,sq+\tau]$$

(The case where if o were in this range, B would not be sent.)

$$O'(o|B)=\left(\frac{1-s_B-d_B-s_W-d_W}{1-(s_W+d_W)}\times\frac{O'}{1-O(sq+\tau)+O(sq-\tau)}\right) o\in[0,sq-2\times(sq-P))\cup(sq+\tau,1]$$

(The case where B would be sent regardless of the truth.)

$$O'(o|W)= \left(\frac{s_W}{1-(s_B+d_B)}\times\frac{O'}{1-O(sq+\tau)+O(sq-2\times(sq-P))}\right) + \left(\frac{1-s_B-d_B-s_W-d_W}{1-(s_B+d_B)}\times\frac{O'}{1-O(sq+\tau)+O(sq-\tau)}\right) \quad o\in[0,sq-2\times(sq-P))\cup(sq+\tau,1]$$

(The case where W could have been sent by a fire alarm who either has the same preferences over outcomes or is attempting to mislead.)

$$O'(o|W) = \left(\frac{d_W+s_W}{1-(s_B+d_B)}\times\frac{O'}{O(sq+\tau)-(1-O(sq+\varepsilon))}\right) o\in(sq+\varepsilon,sq+\tau]$$

(The case where W could have been sent by a fire alarm who either has the same preferences over outcomes or faces a large penalty for lying.)

$$O'(o|W)= \qquad O \qquad o\in[sq-\tau,sq+\varepsilon)$$

(The case where either o could not be in this range or W would not be sent.)
(The case where W could be sent regardless of o's true location.)

$$O'(o|W) = \left(\frac{1-s_B-d_B-s_W-d_W}{1-(s_B+d_B)} \times \frac{O'}{1-O(sq+\tau)+O(sq-\tau)}\right) \quad o\in(sq-2\times(sq-P),sq-\tau)$$

It is easy to verify that this updating scheme renders the content of the fire alarm's message uninformative when $s_B = s_W = t = v = 0$ and perfectly credible when either $s_B + s_W = 1$ or when $(t \times v)$ is sufficiently high. It is also apparent that the updating scheme depends on the relative size of τ, ε and $2 \times (sq - P)$. From Proposition 1 it follows that if $\varepsilon \geq 2 \times (sq - P)$, then it is common knowledge that o cannot be better for the principal than sq; therefore, the principal will not accept o. Therefore, we can present the updating schemes for the two remaining cases where the fire alarm's signal can affect the principal's decision: $\tau \geq 2 \times (sq - P) > \varepsilon$ (if o is better, the fire alarm will not find it profitable to dissemble):

$$O'(o|B) = \left(\frac{s_B+d_B}{1-(s_W+d_W)} \times \frac{O'}{O(sq-\varepsilon)-O(sq-2\times(sq-P))}\right) \quad o\in[sq-2\times(sq-P),sq-\varepsilon)$$

$$O'(o|B) = O \quad o\in[sq-\tau,\ sq-2\times(sq-P)]\cup(sq-\varepsilon,\ sq+\tau]$$

$$O'(o|B) = \left(\frac{1-s_B-d_B-s_W-d_W}{1-(s_W+d_W)} \times \frac{O'}{1-O(sq+\tau)-O(sq+\tau)}\right) \quad o\in[0,\ sq-\tau)\cup(sq+\tau,1]$$

$$O'(o|W) = \left(\frac{d_W s_W}{1-(s_B+d_B)} \times \frac{O'}{O(sq+\tau)-(1-O(sq+\varepsilon))+O(sq-2\times(sq-P))-O(sq-\tau)}\right) \quad o\in[sq-\tau,sq-2\times(sq-P)]\cup(sq+\varepsilon,sq+\tau]$$

$$O'(o|W) = O \quad o\in[sq-2\times(sq-P),sq+\varepsilon)$$

$$O'(o|W) = \left(\frac{1-s_B-d_B-d_W}{1-(s_B+d_B)} \times \frac{O'}{1-O(sq+\tau)+O(sq-\tau)}\right) o\in[O,sq-\tau)\cup(sq,+\tau,1]$$

By contrast, if $2 \times (sq-P) > \varepsilon \geq \tau$, then the penalty for lying is meaningless and learning takes the following form:

$$O'(o|B) = \left(\frac{s_B}{1-s_W} \times \frac{O'}{O(sq-\varepsilon)-O(sq-2\times(sq-P))}\right) + \left(\frac{1-s_W-s_B}{1-s_W} \times \frac{O'}{1-O(sq+\varepsilon)+O(sq-\varepsilon)}\right) \quad o\in[sq-2x(sq-P),sq-\varepsilon)$$

$$O'(o|B) = O \quad o\in[sq-\varepsilon,sp+\varepsilon]$$

$$O'(o|B) = \left(\frac{1-s_B-s_W}{1-s_W} \times \frac{O'}{1-O(sq+\varepsilon)+O(sq-\varepsilon)}\right) \quad o\in[0,sq-2x(sq-P))\cup(sq+\varepsilon,1]$$

$$O'(o|W)= \left(\frac{s_W}{1-s_B}\times\frac{O'}{1-O(sq+\varepsilon)+O(sq-2\times(sq-P))}\right)$$
$$+\left(\frac{1-s_B-s_W}{1-s_B}\times\frac{O'}{1-O(sq+\varepsilon)+O(sq-\varepsilon)}\right) \quad o\in[0,sq-2\times(sq-P))\cup(sq+\varepsilon,1]$$
$$O'(o|W)= \quad 0 \quad o\in[sq-\varepsilon,sq+\varepsilon)$$
$$O'(o|W)= \left(\frac{1-s_B-s_W}{1-(s_B+d_B)}\times\frac{O'}{1-O(sq+\varepsilon)+O(sq-\varepsilon)}\right) \quad o\in(sq-2\times(sq-P),sq-\varepsilon)$$

C. Equilibrium

We now use a proposition to describe behavior and outcomes in our model of the consequences of delegation. To make this result accessible, we first present this proposition in words; we then present it using formal mathematics.

Proposition 4: *In equilibrium, o is the outcome if and only if one of the following statements are true*: (1) The verifier reveals that *o* is better for the principal. (2) The verifier does not reveal the true location of *o* and the principal believes that one of the following, mutually exclusive cases is true:

(a) *o* could be better for the principal than *sq*, the fire alarm could find it profitable to dissemble (the width of the range of alternatives that the principal prefers to *sq* is greater than τ which, itself is at least as great as ε) and one of statements i-v, given below, is true.
(b) *o* could be better for the principal than *sq*, the expected penalty for lying is sufficiently high that it is common knowledge that the fire alarm could not find it worthwhile to dissemble when *o* is better for the principal than *sq* (τ is greater than the width of the range of alternatives that the principal prefers to *sq* which, itself, is larger than ε) and one of statements i-iv is true.
(c) *o* could be better for the principal than *sq*, and the expected penalty for lying is sufficiently small that it is common knowledge that it will not restrict fire alarm behavior (the width of the range of alternatives that the principal prefers to *sq* is larger than ε which, itself greater than τ) and one of statements i-iv is true.

(i) *o* is better than *sq* for both players, and the fire alarm is sufficiently credible (that is, some elements of the set $(s_B, s_W, d_B, d_W, t, v)$ are large enough to cause prior and posterior beliefs to diverge by such a degree that the principal's strategy depends on the content of the fire alarm's message) that he can persuade the principal to either monitor or choose *o* without monitoring and if the principal monitors, she will learn that *o* is better for her than *sq*.

(ii) *o* is worse than *sq* for the principal and is better than *sq* for the fire alarm, and the fire alarm is sufficiently credible that he can persuade the

principal to choose o without monitoring even though o is actually worse for her (ex post) than is sq.

(iii) the fire alarm is not sufficiently credible to affect the principal's behavior and regardless of the fire alarm's action, the principal will accept o without monitoring.

(iv) o is better than sq for the principal, the fire alarm is not sufficiently credible to affect the principal's behavior, and regardless of the fire alarm's action, the principal will monitor and learn that o is better.

(v) o is better than sq for the principal, o is not necessarily better than sq for the fire alarm, the expected penalty for lying faced by the fire alarm is larger than the maximum possible benefit from lying, the fire alarm is sufficiently credible that he can persuade the principal to either monitor or choose o without monitoring, and if the principal monitors, she will learn that o is better for her than sq.

Formal Equivalent: In equilibrium, o is the outcome if one of the following statements is true:

(1) $M_V = o, o \in (sq - 2 \times (sq - P), sq)$.

(2) $2 \times (sq - P) > \tau \geq \varepsilon$ and one of the following:

(a) $o \in (sq - 2 \times (sq - P), sq - \tau)$, $min[- \int |o - P| \, d\,O'(o|B),$ $- \int_{o \in (sq - 2 \times (sq - P), sq + \tau)} |sq - P| \, dO'(o|B) - \int o \in (sq - 2 \times (sq - P),$ $sq - \varepsilon) \, |o - P| \, dO'(o|B) - c_m] > - |sq - P|$ and $-|o - F| > -|sq - F|$.

(b) $o \in [0, sq - 2 \times (sq - P)] \cup (sq + \tau, 1]$, $- \int |o - P| \, d\,O'(o|B) \geq -$ $\int_{o \in (sq - 2 \times (sq - P), sq + \tau)} |sq - P| \, dO'(o|B) - \int_{o \in (sq - 2 \times (sq - P), sq - \varepsilon)} |o -$ $P| \, dO'(o|B) - c_m$, $- \int |o - P| \, d\,O'(o|B) > - |sq - P|$ and $- |o -$ $F| > - |sq - F|$.

(c) $\forall \{B, W\} \in M_F$: $- \int |o - P| \, d\,O'(o|M_F) >$ $\max[- \int_{o \in (sq - 2 \times (sq - P), sq + \tau)} |sq - P| \, dO'(o|M_F) - \int_{o \in (sq - 2 \times (sq - P), sq -}$ $_{\varepsilon)} |o - P| \, dO'(o|M_F) - c_m$, $- |sq - P|]$.

(d) If $o \in (sq - 2 \times (sq - P), sq - \tau)$, then $\forall B, W \in M_F$: $- \int_{o \in (sq - 2 \times (sq - P), sq + \tau)} |sq - P| \, dO'(o|M_F) - \int_{o \in (sq - 2 \times (sq - P), sq - \varepsilon)}$ $|o - P| \, dO'(o|M_F) - c_m]$ $> max[- \int |o - P| \, d\,O'(o|M_F), - |sq - P|]$.

(e) $o \in [sq - \tau, sq)$ and $min[- \int |o - P| \, d\, O'(o|B), - \int_{o \in (sq - 2\,x\,(sq - P),\, sq + \tau)} |sq - P| \, dO'(o|B) - \int_{o \in (sq - 2\,x\,(sq - P),\, sq - \varepsilon)} |o - P| \, dO'(o|B) - c_m] > - |sq - P|$.

(3) $\tau \geq 2\,x\,(sq - P) > \varepsilon$ and one of the following:

(a) $o \in (sq - 2\,x\,(sq - P),\, sq - \varepsilon)$, $min[- \int |o - P| \, d\, O'(o|B), - \int_{o \in (sq - 2\,x\,(sq - P),\, sq + \tau)} |sq - P| \, dO'(o|B) - \int_{o \in (sq - 2\,x\,(sq - P),\, sq - \varepsilon)} |o - P| \, dO'(o|B) - c_m] > - |sq - P|$ and $-|o - F| > -|sq - F|$.

(b) $o \in [0, sq - \tau) \cup (sq + \tau, 1]$, $- \int |o - P| \, d\, O'(o|B) \geq - \int_{o \in (sq - 2\,x\,(sq - P),\, sq + \tau)} |sq - P| \, dO'(o|B) - \int_{o \in (sq - 2\,x\,(sq - P),\, sq - \varepsilon)} |o - P| \, dO'(o|B) - c_m$, $- \int |o - P| \, d\, O'(o|B) > - |sq - P|$ and $- |o - F| > - |sq - F|$.

(c) $\forall\, B, W \in M_F$: $- \int |o - P| \, d\, O'(o|M_F) > [max[- \int_{o \in (sq - 2\,x\,(sq - P),\, sq + \tau)} |sq - P| \, dO'(o|M_F) - \int_{o \in (sq - 2\,x\,(sq - P),\, sq - \varepsilon)} |o - P| \, dO'(o|M_F) - c_m, - |sq - P|]$.

(d) If $o \in (sq - 2\,x\,(sq - P),\, sq - \varepsilon)$, then $\forall\, B, W \in M_F$: $- \int_{o \in (sq - 2\,x\,(sq - P),\, sq + \tau)} |sq - P| \, dO'(o|M_F) - \int_{o \in (sq - 2\,x\,(sq - P),\, sq - \varepsilon)} |o - P| \, dO'(o|M_F) - c_m > max[- \int |o - P| \, d\, O'(o|M_F), - |sq - P|]$.

(4) $2\,x\,(sq - P) > \varepsilon \geq \tau$ and one of the following:

(a) $o \in (sq - 2\,x\,(sq - P),\, sq - \varepsilon)$, $min[- \int |o - P| \, d\, O'(o|B), - |sq - P| \, x \int_{o \in (sq - 2\,x\,(sq - P),\, sq - \epsilon)} dO'(o|B) - \int_{o \in (sq - 2\,x\,(sq - P),\, sq - \varepsilon)} |o - P| \, dO'(o|B) - c_m] > - |sq - P|$ and $-|o - F| > -|sq - F|$.

(b) The remaining conditions are the same as conditions 2c-2e, with τ replaced by ε.

Proof: We first discuss the case where $M_V = o$. Statement 1 of the proposition is obvious as is the fact that o will not be chosen if $o \in (sq - 2\,x\,(sq - P),\, sq)$. It remains to discuss the case where $M_V = O$.

From Proposition 1 it follows that if $\varepsilon \geq 2\,x\,(sq - P)$; therefore, it is common knowledge that o cannot be better for the principal than sq; therefore the principal should not accept o. There remain three cases where it is possible that o could be chosen: $M_V = O$ and $2\,x\,(sq - P) > \tau \geq \varepsilon$, $M_V = O$ and $\tau \geq 2\,x\,(sq - P) > \varepsilon$ and $M_V = O$ and $2\,x\,(sq - P) > \varepsilon \geq \tau$.

> Recall that we are examining the case where the algorithm directs the principal to use (as opposed to ignore) the content of the fire alarm's message, the existence and magnitude of the penalty for lying, and the probability that the fire alarm and the principal have the same preference ordering over *o, sq*. (The case where the algorithm directs the principal to ignore this information is equivalent to the case where $s_B = s_W = d_B = d_W = t = v = 0$.)

For each of the three remaining cases, Bayes's Rule, the assumption of consistent beliefs, Lemma 1, Proposition 3, and Lemma 2 establish the updating procedure described.

Since the principal moves last in this game, it is relatively easy to describe her equilibrium behavior. When the principal monitors, her decision to accept or reject o is straightforward. At the time that she can make her first move in the game, the principal can choose to take one of three actions: reject o without monitoring, accept o without monitoring, or pay c_m to monitor. The expected utility from rejecting o without monitoring is $-|sq - P|$. If $\tau \geq \varepsilon$, then the expected utility from accepting o without monitoring is either $-\int_{o \in (sq,\, sq+\tau]} |o - P|\, dO'(o|B)$ or $-\int_{o \in [sq-\tau,\, sq)} |o - P|\, dO'(o|W)$. If $\varepsilon \geq \tau$, then the utility from accepting o without monitoring is $-\int_{o \in [sq-\varepsilon,\, sq+\varepsilon)} |o - P|\, dO'(o|M_F)$. If $\varepsilon \geq \tau$, then the expected utility from monitoring is: $-[-[|sq - P| \times \int_{o \in (sq-2\times(sq-P),\, sq-\varepsilon)} dO'(o|M_F)] - [\int_{o \in (sq-2\times(sq-P),\, sq-\varepsilon)} |o - P|\, dO'(o|M_F)] - c_m$. If $\tau > 2 \times (sq - P) > \varepsilon$, then the expected utility from monitoring is either: $-[|sq - P| \times \int_{o \in (sq-2\times(sq-P),\, sq+\tau)} dO'(o|B)] - [\int_{o \in (sq-2\times(sq-P),\, sq-\varepsilon)} |o - P|\, dO'(o|B)] - c_m$ or $-|sq - P|$ if W is signaled. If $2 \times (sq - P) > \tau \geq \varepsilon$, then the expected utility from monitoring is either: $-|sq - P| \int_{o \in (sq-2\times(sq-P),\, sq+\tau)} dO'(o|B) - \int_{o \in (sq-2\times(sq-P),\, sq-\varepsilon)} |o - P|\, dO'(o|B) - c_m$ or $-|sq - P| \int_{o \in (sq-2\times(sq-P),\, sq+\varepsilon)} dO'(o|W) - \int_{o \in (sq-2\times(sq-P),\, sq-\tau)} |o - P|\, dO'(o|W) - c_m$. From the assumption of expected utility maximization and the validity of the updating procedure, whether the principal chooses to accept without monitoring, reject without monitoring, or monitor depends on which of the preceding values is highest given the relative size of ε, τ and $2 \times (sq - P)$.

Now we turn to the strategy of the fire alarm. From the validity of the updating scheme it follows that if $s_B > 0$ or $s_W > 0$ then, all else constant, the likelihood that the principal chooses to accept o if B is signaled cannot be less than the likelihood that the principal chooses to accept o if W is signaled. Thus, if the conditions stated in case 2a are true, then signaling B is a unique best response for the fire alarm given his beliefs and "accept if B is signaled" is a unique best response for the principal given her beliefs. If case 2b is true, lying is a dominated strategy for the fire alarm, and therefore the same best responses as stated in case 1 apply here. If case 2c is true, the fire alarm finds it worthwhile to persuade the principal to choose her least preferred outcome, and given the principal's beliefs, she maximizes her interim utility by choosing the

strategy "accept if B is signaled." In this case, the principal is made worse off than if she had not listened to the fire alarm; however, this could only have occurred because the low probability realization (fire alarm dissembled) occurred. If either case 2d or case 2e is true, the fire alarm cannot affect the principal's choice of strategy. Thus, the principal responds to her prior beliefs by either choosing o (case 2d) or monitoring (case 2e). Cases 3 and 4 follow the same logic as case 2. In short, if either player chooses any strategy other than that stated in the situation identified, they play a strategy that provides lower expected utility. The cases where o is rejected follow straightforwardly. QED.

We use similar logic to demonstrate that, in equilibrium, increases in the perceived similarity of fire alarm and principal preferences, the probability of message verification, and the magnitude of the expected penalty for lying each lead to either an increase or no change in the likelihood that the principal's choice from the set $\{o, sq\}$ is the same that she would have made had she known the location of o when making her choice.

Corollary 1: If prior beliefs are consistent, then the likelihood that the principal chooses the element of $\{o, sq\}$ and that it would have chosen had it known the location of o is nondecreasing in $s_B + s_W$, v, and $(t \times v)$.

Proof: It is sufficient to show that as either $s_B + s_W$, v or $(t \times v)$ increase, then the likelihood that o is chosen if $o \in (sq - 2 \times (sq - P), sq)$ is nondecreasing. If beliefs are consistent, then the density of O at o (or a closed interval of small and positive length with endpoints that are equidistant from o) in the principal's posterior beliefs minus the density of O at that point (or interval) in the principal's prior beliefs is nondecreasing in s_B and s_W. Propositions 1 and 2 provided a similar statement for the effect of a penalty for lying and the presence of a verifier. If the true location of o is closer to P than is sq, then it follows from Lemma 1, Lemma 2, and Bayes's Rule that the probability that the message B is sent and the magnitude of the probability mass placed on o (or a finite interval that has o as its center and boundaries within $(sq - 2 \times (sq - P), sq)$) are nondecreasing in s_B, s_W, or t. As the probability mass on this interval increases, so does the likelihood that $-\int_{|o-P|} dO'(o|M_F) > -|sq - P|$, and so does the likelihood that o is chosen. QED.

D. An Algorithm That Determines the Principal's Willingness to Hear a Fire Alarm

One of the problems faced in modeling signaling games is that the probability that the message receiver reacts to a message in a particular way is dependent on the actions of the message sender, which themselves are dependent on the probability that the message receiver reacts to a message in a particular way. This type of problem often requires modelers to make special assumptions in order to obtain useful results. Examples of these assumptions are the consistency

requirements implicit in the definition of the Sequential Equilibrium[63] and in several refinements of the concept.[64] Our response to this problem is to invoke an algorithm that we believe is a good representation of how people deal with this type of situation. The algorithm suggests that a principal with limited cognitive resources will opt to consider the fire alarm's statement if she expects, without explicitly considering all possible outcomes of the game, that doing so will increase the probability that she makes the same decision she would have made had she known the location of o. This algorithm's invocation allows for the relatively simple formal statement of the model.

The algorithm's first inputs are the principal's prior beliefs about the similarity of her and the fire alarm's preferences. For this purpose, we utilize s_B, the common prior probability that the principal and the fire alarm have the same preferences over the set $\{o, sq\}$ when $o < sq$, and s_W, which has an equivalent definition for the case $o > sq$.

The algorithm's next input is the principal's prior beliefs about the extent to which the fire alarm could benefit from making an untruthful statement. Let $qlie(sq, o, F, t, v) = 1$ if $|(|sq - F| - |o - F|)| > t \times v$ and 0 otherwise. *Qlie* tells whether a fire alarm with ideal point f who observes o could find it profitable to make an untruthful statement. All else constant, the likelihood that $qlie = 1$ is increasing in $|(|sq - F| - |o - F|)|$, which is the maximum potential benefit from lying, and is decreasing in the magnitude of the expected penalty for lying.

If the principal knew F and o she would know the value of *qlie*. However, her information about F and o are limited to her knowledge of the distributions Γ and O. Let $Qlie(sq, O, \Gamma, t, v) \in [0, 1]$ be the principal's prior belief about the probability that the fire alarm could find it profitable to make an untruthful statement, where

$$Qlie(sq, O, \Gamma, t, v) = \iint qlie\ (sq, o, F, t, v)\ dO'\ d\Gamma'.$$

Let $h(s_L, s_R, Qlie)$ be an exogenously determined correspondence that is everywhere nondecreasing in s_B and s_W and everywhere nonincreasing in *Qlie*. h denotes the principal's (common knowledge) expectation about the relationship between the content of the statement and the actual location of o. Let $\underline{h}$ be an exogenously determined constant. We say that a principal of type p chooses to condition her inferences on the fire alarm's actions if and only if

$$h(s_L, s_R, Qlie) > \underline{h}.$$

We have chosen to examine the case where this threshold is surpassed. Alternatively, the rule of thumb might dictate that the principal either discount or ignore information about the fire alarm. Fortunately, the case where the principal chooses to ignore this information is equivalent to the case where the fire alarm's entry costs are prohibitively high. In effect, we examine that case as well. The case where the principal discounts information in a systematic manner can be equivalent to an analysis of the present model exchanging current fire

63. Kreps & Wilson, *supra* note 6.
64. For a review, see JEFFREY S. BANKS, SIGNALING GAMES IN POLITICAL SCIENCE (1991).

alarm prior beliefs about the fire alarm's ideal point with relatively diffuse priors or by decreasing the value of t. Since the rule of thumb is solely a function of the common knowledge, we assume that the principal's inference technique is also common knowledge.

E. An Example

To provide additional intuition about the dynamics that are at work in the relationship between the legislative principal and a bureaucratic agent, we provide an example based on the model just presented, that shows the effect of certain types of institutional design, and is computationally simple. We add one simple variation to the game to demonstrate its utility: we assume that a player called the agent makes the game's first move when he decides whether or not to pay $c_a \geq 0$ to propose $o \in 0, 1$. If the agent does not participate, the game ends with sq as the point that determines player payoffs. If the agent participates, the game described in the appendix is played. We assume that the agent has utility function $- |A - X|$, where $A = 0.1$ is the principal's ideal point and $X \in \{o, sq\}$ is the outcome of the game. We also assume that it is common knowledge that the agent's ideal point was drawn from a distribution that makes O uniformly distributed over $[0, 1]$.

Let $P = 0.5$, $sq = 0.2$, so unlike the case discussed in sections A-C, $P > sq$. First, we introduce a variation of the model with a minimal institutional structure. For this first analysis, we assume that $c_a = v = 0$ and that there is no fire alarm. In equilibrium, the agent locates o at his ideal point, 0.1. The agent chooses his ideal point because it gives him higher utility than any other point on $[0, 1]$ and because he knows that the principal will not be able to learn enough to reward her agent for choosing o closer to 0.5, the principal's ideal point. The principal's uniform prior beliefs about the location of o lead her to accept o as it provides greater expected utility than does sq ($-.25 > -.3$). Ironically, the utility that the principal actually receives from choosing o (-0.4) is less than the utility she would have received had she chosen sq. While maximizing her expected utility, she did not make the same choice she would have made had she known the location of o. In this example, the principal paid the price for her uncertainty as her control over policy has effectively been abdicated to the agent.

We now add one change to the institutional structure under which the principal-agent relationship takes place: we increase c_a from 0 to 0.05. Any change in player beliefs and outcomes can be directly attributed to the increased cost of entry. In equilibrium, the agent again chooses his ideal point as the location of o. Though the agent's action is identical to the previous case, the principal's beliefs about this action differ in an important way. The agent's positive cost of entry can provide the principal with knowledge that she did not otherwise possess, since the principal can infer positive costs will not be paid for sufficiently small returns. From Proposition 3, we know that the principal can infer that $o \in [0.15, 0.25]$. (Alternatively, if we had set $c_a = 0.1$ the principal

could make the same inferences about the interval *[0.1, 0.3]*.) Thus, at the time she chooses her strategy, the principal now believes that the expected value of *o* is - *.238*, which is still higher than the - *0.3* offered by *sq*. Therefore, the principal chooses *o*. So, while the agent's cost of entry allows the principal to reduce her uncertainty, in this particular example the new information is not sufficient, given her prior beliefs, to lead her to choose the alternative that is actually closer to her ideal point. In contrast, if we had raised c_a to *0.1* or higher, there would be no point in *[0, 1]* that, if it were to become the outcome, could provide the agent with enough utility to make challenging *sq* worthwhile. Therefore, when $c_a \geq 0.1$, the agent will not propose an alternative and *sq*, the principal's preference relative to agent's ideal point, will be the outcome.

We add one further change to show the effect of fire alarm oversight: we add a fire alarm whose ideal point is *0.1*. The fire alarm's ideal point is private information to him. We assume that the principal believes that the fire alarm's ideal point was drawn from a distribution that placed the fire alarm's ideal point at *0.1* 20 percent of the time and at *0.5* 80 percent of the time. For now, we assume that $t \times v = 0$.

In the absence of a penalty for lying on the fire alarm, the extent to which the principal is willing to condition her beliefs and behavior on the content of the fire alarm's signal depends on the similarity of their preferences. The principal prefers *o* to *sq* if and only if *o* is between *0.2* and *0.8*. The principal knows that this preference is shared by a fire alarm whose ideal point is located at 0.5. By contrast, when the fire alarm's ideal point is *0.1*, the principal believes that he will prefer *o* to *sq* if and only if *o* is between *0* and *0.2*.

Whether the principal can learn from the fire alarm's message depends, in the absence of a penalty for lying, on her beliefs about the truthfulness of the fire alarm's message. In the example, the principal believes that a fire alarm at 0.1, in sending a message, is likely to have an incentive to dissemble. That is, when *o* is between 0 and 0.2, the fire alarm wants the principal to believe "better" when in fact "worse" is true. Similarly, when *o* is between 0.2 and 0.8 the fire alarm wants the principal to believe "worse," when, in fact, "better" is true. In contrast, when $o \in [0.8, 1]$, both the fire alarm and the principal prefer *sq*. Thus, the principal can infer that a fire alarm whose ideal point is located at *0.5* will always have an incentive to send a truthful message while a fire alarm whose ideal point is located at *0.1* will have an incentive to send a truthful message with only a probability of 20 percent (when $o \in [0.8, 1]$). Given the principal's prior beliefs about the location of *F* ($\Gamma(0.1) = 0.2$, $\Gamma(0.5) = 0.8$), the principal can infer that there is an 84.4 percent chance ($s_B = .489$, $s_W = .355$) that the fire alarm shares her preferences over outcomes.

When the principal knows that $c_a = 0.5$ and observes *B*, she infers that there is an approximately 75.8 percent chance ($\frac{s_B}{1-s_W}$) that $o \in (.25, .8)$ and an approximately 24.2 percent chance that $o \in [0, 0.15] \cup [0.25, 1]$. Similarly, upon observing *W*, she concludes that there is an approximately 69.5 percent chance that $o \in [0.15] \cup [.8, 1]$ and an approximately 30.5 percent chance that $o \in [0, 0.15] \cup$

[0.25, 1]. In equilibrium, the agent proposes his ideal point, the fire alarm sends the untruthful message *B* and the principal chooses *o* because the utility she expects to receive from *o* (- *.167*) is greater than the utility she will receive from choosing *sq*. In this case, adding a fire alarm who the principal believed (but did not know for certain) had preferences similar to her own was not sufficient to help the principal choose the alternative that will make her better off.

Fortunately, fire-alarm oversight systems can be designed to increase the likelihood that legislators can learn about bureaucratic actions. We show this by making one change to the previous example: we increase $t \times v$ from *0* to *0.1*. In this case, it is no longer worthwhile for the fire alarm to dissemble when $o = 0.1$ and $sq = 0.2$. In equilibrium, there is no point that the agent could propose that would provide the fire alarm with an incentive to dissemble. As a result, the agent knows that the message *W* would be sent and that the relatively high values to t, v, s_B and s_W would lead the principal to reject the offer after observing *W*. Therefore, *sq* is the outcome and the principal is made better off than in the previous examples.