Political Agency, Oversight, and Bias: The Instrumental Value of Politicized Policymaking*

Ian R. Turner†

August 28, 2015

Abstract

We develop a political-institutional theory of policymaking between a principal, an agent, and an overseer that can overturn agent regulatory actions. The agent can increase the overall quality of implemented outcomes through costly ex ante effort investments. Oversight of the agent can impact agent effort incentives, but only if the ideological bias of the agent does as well. The principal, in light of the oversight dynamics, chooses whether or not to authorize the agent to make policy on her behalf. We show that anytime oversight is not an effective means of political control, the principal never benefits from authorizing a biased agent relative to an ally. However, when oversight is effective at inducing higher effort investments, the principal always benefits from increased agent bias. That is, in certain institutional environments, based on the dynamics of policymaking oversight, the ally principle fails for instrumental reasons. The principal benefits from trading off agent bias for reduced slack in policy implementation. The results have implications for bureaucratic personnel politics, political appointments, and the efficacy of managerial motivation strategies.

Paper prepared for APSA 2015
Preliminary and incomplete

*I would like to especially thank John Patty, Maggie Penn, Randy Calvert, Justin Fox, Gary Miller, Keith Schnakenberg and Richard Van Weelden for many insightful conversations regarding this work. I would also like to thank Jay Krehbiel, Dalston Ward, Cathy Hafer, Caitlin Ainsley, Stephane Wolton, Rachel Augustine Potter, Ryan Hubert, and Andrea Aldrich for providing exceptionally insightful commentary at different stages. The standard caveat applies with respect to any and all errors of omission or commission.

†Assistant Professor, Department of Political Science, Texas A&M University, Email: irturner@tamu.edu.

1 Introduction

In bureaucratic policymaking delegation is the rule, not the exception. Congress writes authorizing legislation empowering agencies to craft and implement public policy. The President empowers agency heads to pursue the Administration's policy goals. Agency heads direct career civil servants to craft and implement policy objectives.[1] Put simply, the vast majority of public policy in the United States is promulgated by bureaucrats who were empowered to act by some political principal (Warren 2004). There are many reasons posited for why delegation is necessary, including to take advantage of relative expertise due to policy environmental complexities generally (Epstein and O'Halloran 1994; Spence and Cross 2000; Gersen and Vermeule 2012), or to provide incentives for specialization (Gilligan and Krehbiel 1987, 1989), information gathering (Gailmard and Patty 2007, 2013*b*), and to exert costly effort in policy development (Bueno de Mesquita and Stephenson 2007; Hirsch and Shotts 2014).[2]

While delegation may be used to solve several problems it does not come without cost. Once policymaking authority is conferred from a principal to an agent potential problems arise, known as political agency problems. First, agents may be *biased* and prefer policy outcomes that diverge from those of the principal. In this case the agent may attempt to subvert the principal's wishes once authority is transferred by making incongruent policy choices (Gailmard 2002).[3] Second, the agent may *slack* on pursuing high quality outcomes. In this case the principal's worry is that the agent will invest insufficient effort toward effective implementation or enforcement of policy (Bueno de Mesquita and Stephenson 2007; Turner 2014). Both of these potential problems can lead to low quality policy outcomes from the principal's perspective. Put simply, bureaucratic policymakers are tasked with not only crafting the substance of policy, but also implementing it effectively whatever the substantive content (Carpenter 2001; Derthick 1990; Lipsky 1980).

Two commonly proposed solutions to these agency problems are extensive oversight and

[1] See Bendor, Glazer and Hammond (2001) for a comprehensive overview of different theories of delegation, and Gailmard and Patty (2013*a*) for a less comprehensive (with respect to delegation specifically), but more recent, review dealing with the domain of bureaucratic politics.

[2] Normatively, there are those that have argued that delegation to agencies enhances democratic representation through its effects on policy legitimation, for instance (Meier 1997; Meier and O'Toole 2006). Others have argued that delegation is largely undesirable. Most notably, Lowi (1979) argued that delegation was a democratically illegitimate abdication of authority by Congress.

[3] Examples that other scholars have labeled bias or drift include agency capture by regulated or interested groups (Carpenter and Moss 2013; Niskanen 1971), selection practices and career concerns of bureaucrats (Heclo 1988), cognitive and/or institutional biases, or implicit motivations (Gailmard and Patty 2007; Prendergast 2007; Seidenfeld 2002).

delegating to allies. The logic underlying the ally principle is straightforward and has been shown to hold in very general environments (Bendor and Meirowitz 2004). All else equal, a principal prefers to authorize an ally agent because the agent can, rightfully, be expected to take actions in line with the principal's interests regardless of informational or experiential advantages. However, the situation becomes more complicated when effort to improve implementation quality is introduced. Since effort is often costly to the agent alone, he cannot precommit to effort levels. The principal, in this case, must also consider the provision of proper effort incentives, which can be supplied through oversight (Bueno de Mesquita and Stephenson 2007; Turner 2014).

While there are many forms of bureaucratic oversight, one of the most prevalent forms is to subject an agent's decision-making authority to ex post review by a third party that has the power to either accept or reject agent-made policy. Common examples include the federal judiciary and the Office of Information and Regulatory Affairs (OIRA) in the President's Office of Management and Budget (OMB). The former is empowered to act as a bureaucratic overseer through judicial review provisions written into authorizing legislation.[4] Similarly, proposed rules and regulations must be accepted following review by the OIRA before they become binding policy. Essentially, the overseer has the ability to veto agent-made policy. In both cases, a principal has not only authorized a bureaucratic agent to make policy but has also subjected the agent's actions to subsequent review, and possible rejection, by another political actor or institution. The overseer will also often have preferences that diverge from those of the principal, further compounding the political agency problems inherent in the policymaking process. This raises the question of when does a political principal benefit from authorizing an agent to make policy on her behalf given that the agent, once authorized, faces oversight? Moreover, if the principal does benefit from delegating policymaking authority, what type of agent would she prefer?

In this paper we develop a political-institutional theory of policymaking and show that if *bias* and *slack* are both potential problems, the potential solutions — delegating to allies and extensive oversight — interact in unexpected ways. The dynamics of oversight imply that anytime oversight matters with respect to agent effort incentives, the principal is best served by delegating to a biased agent rather than an ally. Oversight and agent bias interact in such a way that neither will influence agent effort incentives unless both do simultaneously. This has upstream effects on both whether the principal benefits from delegation at all and whether the principal is made better off by authorizing an ally or biased agent to make policy.

The basic intuition is that the ally principle is violated anytime oversight is effective at pro-

[4]Additionally, the Administrative Procedures Act (APA) directs courts to engage in so-called hard look review of agency regulations and overturn actions found to be "arbitrary and capricious" (Breyer 1986; Stephenson 2006). Moreover, (Shipan 1997) provide a comprehensive study of the politics surrounding the choice of these provisions.

viding positive effort incentives for the agent due to the fact that the agent sets the substance of policy *and* invests effort to develop the capacity to implement effectively. Crucial to the theory is the fact that this insight only holds because of the intervening influence of ex post oversight. Even if a principal benefits from delegating policymaking authority when oversight is ineffective, she never benefits from a biased agent relative to an ally. Otherwise, when oversight is effective, agent bias serves two interrelated purposes that increase the agent's effort investments in equilibrium. First, it increases the agent's own motivations to invest more effort to ensure that his policies are realized. Second, it increases the stringency of oversight, which in turn increases the effort the agent must invest in order to have his preferred policy realized. The principal can benefit from leveraging this dynamic by delegating policymaking authority to a biased agent. The principal can essentially play the other actors' biases off one another to induce higher levels of effort investment than would be possible if there were no effective oversight *and* if the principal was the overseer. Thus, while there are situations where the principal *will* delegate policymaking authority to an ally agent, she is better off under particular institutional environments delegating that authority to a biased agent. That is, there is instrumental value to the principal for delegating policymaking authority to an agent that will then make policy in a politicized (i.e., ideologically contentious) policymaking environment.

The main insights contribute to literature on political control and delegation directly. In terms of the effect of oversight on policymaking incentives, previous work has uncovered its impact on incentives for politicians to pander in electoral settings (Fox and Stephenson 2011), incentives to exert costly effort (Bueno de Mesquita and Stephenson 2007; Stephenson 2006; Turner 2014), the relationship between legislatures and courts (Rogers 2001; Rogers and Vanberg 2002; Vanberg 2001), and information acquisition and more informed policymaking (Dragu and Board 2013). In addition, extant research has analyzed how different types of oversight may impact policymaking (Staton and Vanberg 2008). While this paper shares the goal of further understanding how institutions like judicial or executive review impacts policymaking generally, we contribute to this literature by going beyond the direct effects of oversight on policymaking incentives. We extend the analysis to speak to how these effects also impact the incentives for political principals to delegate authority and what types of agents are preferable given that delegation itself is desirable.[5]

The particular focus on what types of agents the principal benefits from complements literature examining optimal agent bias. Bendor and Meirowitz (2004) show that principals may prefer biased agents to allies if biased agents are willing to work harder or have some type of valence that is correlated with their bias and beneficial to the principal. In a sense, we provide a distinctly political-

[5]The way we model oversight as ex post review with a veto and how it structures agent incentives is also related to a family of models examining agent retention and career concerns under different combinations of moral hazard and adverse selection (see Banks and Sundaram 1993, 1998, for canonical examples).

institutional microfoundation for why biased agents can be preferable for principals concerned with motivating desirable investments in valence. Specifically, the theory developed here does not assume that biased agents are *per se* better suited to invest effort toward implementation improvement. Rather, agent bias only impacts effort incentives if effective oversight is present in the policymaking environment.[6] The logic we present in this paper is similar to work that has shown that agent bias is useful in delegation situations to induce specialization (Gilligan and Krehbiel 1987, 1989; Krehbiel 1991) or generate bargaining power following delegation (Bertelli and Feldmann 2007; Gailmard and Hammond 2011).

For instance, Gailmard and Hammond (2011) argue that the House of Representatives has incentive to create biased committees to increase House bargaining power relative to the Senate. Essentially, the biased committee represents a tougher veto point that the Senate is subsequently forced to take into account. The authors write that, "an unrepresentative committee is a veto constraint for the other chamber, and can prevent it from making proposals that the committee's parent H would rather not face but cannot commit to reject" (p. 541). In our theory, the principal benefits from a biased agent precisely because she is able to sidestep her own commitment problems by leveraging those of the overseer. A biased overseer represents a "tougher veto point" with respect to agent policymaking, which the agent responds to by investing higher levels of effort to satisfy the overseer. The more divergent the agent and overseer's ideal points, within limits, the more this dynamic intensifies effort incentives. While the logic between our theory and the aforementioned paper are related, the setting is quite different. There are no expertise or effort considerations in Gailmard and Hammond (2011), while both are key to the results in this paper. Put simply, while the focus on incentives is similar, the institutional environments analyzed are quite distinct.

Overall, the theory developed in this paper provides novel insight into the institutional situations and policy environments in which the ally principle fails. The principal, anytime oversight is able to be effectively utilized as a tool of political control, benefits from delegating policymaking authority to a biased agent who will subsequently face a biased overseer in the policymaking game. Without the intervening dynamics of effective oversight, the ally principle holds, but generally political principals derive instrumental value from the dual usage of oversight and agent bias as institutional effort motivators when bias and slack are both concerns.

[6]More generally, several previous studies have shown that, at times, principals prefer biased agents based on divergent beliefs (Che and Kartik 2009), the optimal distribution of tasks between agents and reviewers (Bubb and Warren 2014), the need to induce information disclosure (Dessein 2002), and to reduce rent-seeking by electorally motivated politicians (Van Weelden 2013). We provide results that are similar in the sense that they also imply a rationale for why political principals may prefer a biased agent, however, we diverge from previous work by showing that the institution of oversight is a necessary condition for the ally principle to fail when agent bias and slack are both concerns.

The remainder of the paper proceeds as follows. The next section presents the model. Then we analyze the dynamics of oversight by characterizing the agent-overseer subgame assuming the principal authorizes the agent to make policy. Following that, we turn to characterizing when the principal chooses to delegate policymaking authority and when she benefits from empowering a biased agent relative to an ally. We then discuss several implications with a focus on U.S. bureaucratic politics and the final section concludes.

2 The model

We study a simplified model of policymaking between a principal P, an agent A, and an overseer or reviewer R. Each players' induced preferences over policy depend on their respective "type" or ideal point, denoted by $t_i \in \mathbb{R}$, $i \in \{P,A,R\}$. Each players' ideal point dictates their welfare-maximizing policy outcome relative to a true state of the world, denoted by $\omega \in \Omega = \mathbb{R}$. The state of the world is drawn according to probability distribution F that is symmetric around mean 0 with strictly positive, finite variance V_F. We normalize the principal's ideal point so that $t_P = 0$, which implies that the principal is solely concerned with final outcomes matching the state as closely as possible. We also assume that the overseer's ideal point is to the left of the principal so that $t_R < 0$. The analysis focuses on how oversight, agent authorization, and policymaking incentives vary as t_A varies relative to the other players' ideal points.

The principal first decides whether to delegate authority to the agent or not. This choice is denoted by $a \in \{0,1\}$ where $a = 0$ signifies a choice not to delegate and $a = 1$ means she did authorize the agent. If the principal chooses to authorize the agent she incurs authorization costs $c \geq 0$. This captures the fact that authorizing an agent to make policy requires an investment by the principal. Legislatures must write authorizing legislation and allocate budgetary resources, the President must outline administrative goals or directives, and agency heads must establish policy goals, staff departments, and outline procedures to direct the policymaking actions of bureaucratic subordinates. In all of these cases the principal incurs direct or indirect costs associated with making the choice to authorize an agent to make policy on her behalf.[7] If the principal chooses not to authorize the agent to make policy then she foregoes paying this cost, but she must accept the realization of unregulated outcomes (ω). If she does choose to delegate authority to the agent, then the agent is fully empowered to make policy. Once the agent is authorized, his behavior and the overseer's behavior ultimately dictate outcomes.

Regulated (agent-made) outcomes are dependent on agent effort investments and substantive (spatial) policy choices relative to ω. The agent first chooses an effort investment $e \in [0,1]$. This

[7]Essentially this renders costless authorization $c = 0$ a baseline case with which to compare behavior in the more realistic case of positive authorization costs.

effort investment represents the agents capacity to more effectively implement policy whatever the substantive content. Formally, one component of agent-made policy is an implementation shock denoted by ε. This shock is distributed according to distribution $G_\varepsilon(e)$ that is symmetric around mean 0 with strictly positive, finite variance $V_\varepsilon(e)$. The variance of $G_\varepsilon(e)$ is conditional on the agent's effort investment so that the more effort the agent invests the lower the variance. We assume that $V_\varepsilon(e)$ is strictly decreasing and convex so that $V_\varepsilon(e) < V_\varepsilon(e')$ if and only if $e > e'$.[8] No player observes ε, but the distributional characteristics of $G_\varepsilon(e)$, including how $V_\varepsilon(e)$ decreases in e, are common knowledge. Following the agent's effort investment, he observes ω directly and chooses the substantive content of policy. This choice is denoted by $x \in X = \mathbb{R}$ and represents a substantive policy target set by the agent. It is a target in the sense that realized outcomes also depend on how much effort the agent has invested through its effect on the likely implementation shock ε. By separating these tasks, we take seriously a differentiation in bureaucratic policy tasks embodied in a growing literature on agency capacity (e.g., Carpenter 2001; Huber and McCarty 2004; Ting 2011). For instance, Carpenter (2001) distinguishes between an agency's analytic and programmatic capacities. The former refers to the agency's ability to adequately craft the content of policy based on the technical expertise within the agency. The latter refers the the agency's ability to actually implement or apply policy effectively on the ground.[9] Our conception of agent effort investments and the functional effects they have on policy follow in this tradition.

Following the agent's choices the overseer observes the agent's effort investment and chooses to either uphold or overturn the agent. This choice is denoted by $r \in \{0,1\}$ where $r = 0$ means the overseer upholds and $r = 1$ means the agent is overturned. If the agent is upheld then agent-made policy, contingent on the agents effort and substantive policy choices, is realized. If the agent is overturned then the unregulated outcomes that arise as a product of interactions between private individuals without agent intervention obtain (i.e., ω obtains unencumbered). Accordingly, final policy outcomes are generated by the following function,

$$y = \begin{cases} x - \omega + \varepsilon & \text{if } a = 1 \text{ and } r = 0, \\ -\omega & \text{if } a = 0 \text{ or } a = 1 \text{ and } r = 1. \end{cases} \tag{1}$$

If the principal authorizes the agent ($a = 1$) to make policy and the overseer ultimately upholds ($r = 0$) the agent then agent-made policy is realized. If however, the principal does not delegate authority ($a = 0$) or she does ($a = 1$) but the overseer overturns the agent ($r = 1$) then the true state is realized absent agent intervention. The outcomes for not authorizing the agent and authorizing the

[8] That is, if $e > e'$ then $G_\varepsilon(e)$ second-order stochastically dominates $G_\varepsilon(e')$.

[9] Another way of thinking about this issue is through the lens of "street level bureaucracy" (Lipsky 1980).

agent when he will be overturned are equivalent because in either case the substantive impact is the same: outcomes are a product of the contingencies of the policy environment ω, which are taken to represent the outcome generated by the private, unregulated, interactions between individuals and/or firms. Following this policymaking process, the game ends and payoffs are realized.

Payoffs. The payoffs of the principal, the agent, and the overseer are given by the following expressions, respectively.

$$\begin{aligned} u_P(e,y,r) &= -y^2 - ca, \\ u_A(e,y,r) &= -\beta(y-t_A)^2 - \kappa e - \pi r, \\ u_R(e,y,r) &= -(y-t_R)^2. \end{aligned}$$

The principal is solely concerned with outcomes matching the contingencies of the policy environment ($y = \omega$), but does take into account her potential authorization costs. The overseer decides to uphold or reverse the agent based on policy being realized as close as possible to its ideal point t_R relative to the true state. The agent also desires policy outcomes to be realized as close as possible to his ideal point, but his policy motivations, relative to the other components of his utility and the motivations of the other players, is captured by $\beta > 0$. The larger is β, the more policy motivated the agent is. This can represent stronger "sense of mission" within an agency (Wilson 1989), a higher ratio of zealots to slackers (Gailmard and Patty 2007) or political appointees to career civil servants (Lewis 2008), or simply higher intrinsic policy motivations for the bureaucratic agent (Prendergast 2007). All else equal, players prefer more effective implementation generated through increased agent effort investment, but only the agent bears the costs of investing that effort, denoted by $\kappa > 0$. This effort cost captures intuitive concepts of building bureaucratic capacity like increased staffing, investing time and resources toward streamlining procedures, or expanding enforcement programs. Finally, the agent is also averse to being overturned by the overseer, captured by $\pi > 0$. The higher π, the more averse the agent is to being overturned. While we are agnostic as to the microfoundations of this parameter, it captures intuitive, realistic conceptions based on things like career concerns, e.g., agent reputational losses for looking incompetent, budgetary considerations, etc. It suffices to simply think of π as a reversal cost the agent must internalize if he is overturned. The parameters of the problem are exogeneous and common knowledge.

Information and policymaking. The players are forced to confront the uncertainty inherent in policymaking. This uncertainty is captured in the distributions of ω and ε. Recall that ω is distributed according to F with mean 0 and variance V_F. The principal must choose whether to authorize the agent to make policy or not. She makes this decision based on knowledge of F and $G_\varepsilon(e)$ as well as knowledge of the agent's payoff structure. Thus, the principal makes her delegation decision

based on beliefs over ω, ε, and the agent's policy choice and effort strategies given the incentives produced by the presence of oversight.

The agent observes the realization of ω following its effort investment e. After x is chosen by the agent, ε is realized according to $G_\varepsilon(e)$. No player observes this realization, but all players know that the higher the agent's effort investment, the lower $V_\varepsilon(e)$ becomes, and, therefore, the higher the precision of final agent-made policy. So, the agent, after choosing e, observes ω, chooses x, ε is realized, and then the overseer must decide whether to uphold or reverse the agent. This oversight decision is made by the overseer based on (in equilibrium, correct) beliefs over the agent's policy choice strategy and the level of variance associated with upholding or reversing the agent's actions, which is further conditional on the choice of e. Moreover, the agent's policy bias relative to the overseer is common knowledge. Thus, the overseer does know the choice of e and the level of preference divergence relative to the agent, but does not know ω or ε.[10] The overseer does know F and $G_\varepsilon(e)$ and, thus, also knows V_F and $V_\varepsilon(e)$.

Strategies and equilibrium concept. We utilize perfect Bayesian equilibrium (PBE) in weakly undominated strategies. The principal's strategy consists of an agent authorization choice. Denote this strategy by s_P and the principal's equilibrium authorization choice by $a^* \in \{0,1\}$. The principal also has beliefs over ω and ε, which are represented by μ_P, a cumulative distribution function that represents a probability distribution over ω and ε. The agent's strategy consists of an effort investment choice denoted by s_A^e, and a policy mapping conditional on the realization of ω denoted by $s_A^x(\omega)$. Further denote the agent's equilibrium effort choice as e_A^* and his equilibrium substantive policy choice conditional on ω as $x_A^*(\omega)$. The agent also has beliefs over ε denoted by μ_A. The overseer's review strategy consists of a mapping from the set of agent effort levels and the potential policy outcomes into a review decision. Denote this strategy by $s_R(e)$ that holds for any agent effort level $e \in [0,1]$ and potential policy outcome $y \in \mathbb{R}$. The overseer, like the principal, also has beliefs over ω and ε characterized in the same manner as the principal's beliefs, which are denoted by μ_R. A PBE is a complete profile of strategies and beliefs $\rho = (s_P, \mu_P, s_A^e, s_A^x, \mu_A, s_R, \mu_R)$ such that all players are maximizing their expected payoffs given other players' strategies and, when applicable, beliefs are consistent with Bayes's rule.[11]

3 Oversight, bias, and agent effort investments.

In this section we analyze the subgame interactions between the agent and the overseer assuming that the principal has authorized the agent to make policy. To begin the analysis we note that the agent will always set policy at his ideal point when given the opportunity to make policy. That is, the

[10] The overseer also does not observe x or y.

[11] Given the set-up, these beliefs will always be pinned down by Bayes's rule.

agent's equilibrium substantive policy choice is $x^*(\omega) = \omega + t_A$.[12] This is a weakly dominant strategy for the agent, independent of his effort investment and the overseer's oversight strategy, because the overseer does not observe x directly. Moreover, the agent's effort investment is a sunk cost at the point of the game at which this decision is made, the implementation shock has expectation zero, and the agent's utility is separable in his effort and substantive policy choices. This feature of the equilibrium can be thought of as the agent making *sincere* policy choices (from his point of view). It also helps to isolate the effects of oversight on agent effort investment incentives and the principal's potential ability to exploit agent bias to reduce slack. With this characteristic of the equilibrium in hand we now turn to the overseer's optimal oversight strategy.

The overseer's equilibrium strategy is driven by the desire to minimize the distance between its ideal point and realized outcomes. However, oversight is limited to a veto of agent-made policy, which is in line with the types of oversight discussed in the introduction. Courts, executive reviewers, and intra-agency veto points can often only accept or reject policies rather than supplant them with their own policy (e.g., Bueno de Mesquita and Stephenson 2007). The overseer, upon observation of the agent's effort investment, can only accept the expected losses from upholding agent policy actions or overturn the agent and accept the expected losses from allowing unregulated outcomes to obtain. With this in mind, the overseer's net expected payoff from upholding the agent is given by,[13]

$$\Delta U_R(\text{uphold: } r = 0; t_A, e) = -t_A^2 + 2t_A t_R - V_\varepsilon(e) + V_F.$$

Incentive compatibility implies that the overseer will uphold the agent, given his bias t_A and observed effort investment e, if and only if $\Delta U_R(\text{uphold: } r = 0; t_A, e) \geq 0$. Combining and rearranging the net expected payoff yields the incentive compatibility condition for the overseer to uphold an agent with bias t_A who invested effort e:

$$\underbrace{V_F - V_\varepsilon(e)}_{\text{Increased implementation quality}} \geq \underbrace{t_A^2 - 2t_A t_R}_{\text{Net spatial policy losses}} \tag{2}$$

Equation 2 provides an intuitive condition that must be met for the overseer to uphold the agent. The agent must invest sufficient effort to improve the quality of implementation, relative

[12] While this is out of order in the sense of working backward through the game, making this observation up front aids in simplifying exposition of the overseer's equilibrium strategy. Full proof of this point can be found in the appendix.

[13] We use the notation $U_i(\cdot;\cdot)$, $i \in \{P, A, R\}$ to represent players' expected utility given their proposed action and those of the other players. We also use $\Delta U_i(a;\cdot) \equiv U_i(a;\cdot) - U_i(b;\cdot)$ to represent the net expected payoff for player i taking action a instead of action b given the expected behavior of the other players in equilibrium. Full derivations can be found in the appendix.

to the volatility of the underlying policy environment, enough to offset any spatial policy losses incurred by his bias. The more effort the agent invests to improve implementation the more likely it is equation 2 will be satisfied. Conversely, the more biased the agent is relative to the overseer the less likely it will be satisfied. This implies that the more biased the agent is relative to the overseer, the more stringent is oversight. However, the more volatile the underlying policy environment, the less stringent is oversight. The agent making policy becomes more important the more volatile is the underlying policy environment. This highlights a commitment problem for the overseer: the more the agent is needed to regulate, the less demanding oversight is with respect to effort investments.

Since implementation quality is strictly increasing in agent effort investments ($V_\varepsilon(e)$ is strictly decreasing in e) the overseer's equilibrium strategy is equivalent to an effort threshold. Denote this threshold as $\underline{e}_R(t_A)$. It is the minimum level of effort investment an agent must make in order to be upheld by the overseer given his bias: $\underline{e}_R(t_A) \equiv e$ such that $V_F - V_\varepsilon(e) = t_A^2 - 2t_A t_R$. This implies the following equilibrium oversight strategy for the overseer,

$$s_R^*(e) = \begin{cases} \text{uphold: } r = 0 & \text{if } e \geq \underline{e}_R(t_A), \\ \text{overturn: } r = 1 & \text{otherwise.} \end{cases} \tag{3}$$

The impact of oversight on agent effort investments depends crucially on the agent's bias relative to the overseer. If the agent is too biased then the overseer will never uphold the agent, regardless of effort investment levels. In this case the overseer is perfectly skeptical of regulatory intervention. This environment is one in which even if the agent makes a maximal effort investment, $e = 1$, to improve implementation quality, he cannot offset the spatial policy losses. That is, if t_A is sufficiently extreme relative to t_R so that equation 2 fails to hold even when $e = 1$ then the overseer always prefers unregulated outcomes. Note that the level of agent bias that is *too biased* is decreasing in the volatility of unregulated outcomes, V_F. The more an agent is needed to improve implementation, the more biased he can be before the overseer becomes perfectly skeptical.

In this case the agent responds by never investing positive effort. If an agent with this level of bias makes any positive effort investment, given the overseer will overturn with certainty, he incurs a net utility loss proportional to the cost of that investment κ. Thus, when facing a perfectly skeptical overseer, the agent never invests any positive effort toward implementation quality.

On the other extreme, if the agent is too moderate relative to the overseer he will never be overturned. In this environment we say that the overseer is perfectly deferential to agent policy actions. This is the case anytime spatial policy losses are always offset even when the agent invests zero effort toward implementation. That is, if t_A is sufficiently close to t_R such that equation 2 holds even when $e = 0$ then the overseer can never commit to overturning the agent. The level of bias that can support perfect deference is increasing in unregulated outcome volatility, V_F. All else equal, the

more volatile unregulated outcomes become, the less stringent oversight becomes and the harder it is for the overseer to commit to overturning a relatively low-bias agent. This reveals a pathological limitation of oversight in this model: if the agent *is not biased enough* relative to the overseer then oversight plays no effective role in the provision of agent effort incentives.

It may seem intuitive that in response to perfect deference the agent again never invests positive effort since he will be upheld regardless. However, the agent is intrinsically motivated to improve outcomes. While oversight as an institution does not impact effort investments in this case, the agent's own motivations do. Since the overseer will never overturn the agent, the agent makes effort investments that maximize his utility based solely on his own motivations. Denote this effort choice by,

$$\hat{e}_A(\beta, \kappa) = \arg\max_e -\beta V_\varepsilon(e) - \kappa e. \tag{4}$$

This expression gives the agent's optimal effort investment when the overseer is perfectly deferential. The agent chooses a level of investment as if there were no oversight. In this case the agent's effort investment is greater than the overseer's threshold level of acceptable effort investment: $\hat{e}_A(\beta, \kappa) \geq \underline{e}_R(t_A)$. This follows from the fact that oversight is not stringent enough to bind the agent's investment decision. Intuitively, the agent's investment is increasing in his implicit policy motivations, β, and decreasing in effort costs, κ, in this case.

The final, most interesting, environment is one in which the agent's effort investment *is* affected by oversight. In this case the overseer employs conditional-deference. The agent is biased enough away from the overseer that the agent's unconstrained effort investment $\hat{e}_A(\beta, \kappa)$ is not sufficient to satisfy the overseer's threshold $\underline{e}_R(t_A)$. That is, given the arrangement of t_A and t_R in this environment, the agent's effort investment based on his own motivations is not enough to satisfy the overseer's threshold. However, there is a level of effort the agent could invest that would satisfy this threshold and lead to deference.

Accordingly, the agent responds by deciding if he is better off investing the threshold level of effort required to be upheld or making no effort investment and being overturned.[14] With this in mind, consider the agent's net expected payoff for making an effort investment sufficient to be upheld,

$$\Delta U_A(e \geq \underline{e}_R(t_A); r^*(e) = 0) = \beta(t_A^2 + V_F - V_\varepsilon(e)) - \kappa e + \pi.$$

Incentive compatibility implies that the agent will invest enough effort to be upheld if $\Delta U_A(e \geq$

[14]Note that if it is not incentive compatible for the agent to invest the threshold level of effort to be upheld then he makes zero effort investment because any positive investment that fails to the meet the threshold results in a net utility loss equal to the cost of that investment, as in the perfectly skeptical case.

$\underline{e}_R(t_A); r^*(e) = 0) \geq 0$. Solving the agent's incentive compatibility condition for e so that it holds with equality, and bounding the problem to ensure that a feasible solution always exists, yields the maximum level of effort investment the agent is willing to make in order to be upheld when facing a conditional-deference overseer:[15]

$$e_A^{\max}(t_A) = \max\left[\min\left[\frac{\beta(t_A^2 + V_F - V_\varepsilon(e_A^{\max}(t_A))) + \pi}{\kappa}, 1\right], 0\right]. \tag{5}$$

If the maximum level of effort the agent is willing to invest to be upheld exceeds the threshold required by the overseer then the agent does so. Otherwise, if $e_A^{\max}(t_A) < \underline{e}_R(t_A)$ then the agent invests zero effort and accepts being overturned. Thus, when facing conditional-deference oversight — $\underline{e}_R(t_A) > \hat{e}_A(\beta, \kappa)$ — the agent will make an effort investment exactly equal to the overseer's threshold if and only if $e_A^{\max}(t_A) \geq \underline{e}_R(t_A)$, and make no investment otherwise.

Taken collectively the oversight/effort investment combinations described above imply the following optimal effort investment strategy for the agent,

$$s_A^{e*} = \begin{cases} \hat{e}_A(\beta, \kappa) & \text{if } \hat{e}_A(\beta, \kappa) \geq \underline{e}_R(t_A), \\ \underline{e}_R(t_A) & \text{if } \hat{e}_A(\beta, \kappa) < \underline{e}_R(t_A) \text{ and } e_A^{\max}(t_A) \geq \underline{e}_R(t_A), \\ 0 & \text{if } \hat{e}_A(\beta, \kappa) < \underline{e}_R(t_A) \text{ and } e_A^{\max}(t_A) < \underline{e}_R(t_A), \end{cases} \tag{6}$$

where $\underline{e}_R(t_A)$ is defined as e such that equation 2 holds with equality, $\hat{e}_A(\beta, \kappa)$ is implicitly defined by equation 4, and $e_A^{\max}(t_A)$ is implicitly defined by equation 5.

There are a few aspects of the agent's equilibrium effort investment strategy worth noting further. First, notice that the presence of an overseer can induce higher levels of effort investment from the agent than if there were no oversight. This is the second case of s_A^{e*} in which $\hat{e}_A(\beta, \kappa) < \underline{e}_R(t_A)$ and $e_A^{\max}(t_A) \geq \underline{e}_R(t_A)$. Conversely, the overseer can also induce the agent to invest lower effort than it would otherwise. This is the third case of s_A^{e*} in which $\hat{e}_A(\beta, \kappa) < \underline{e}_R(t_A)$ and $e_A^{\max}(t_A) < \underline{e}_R(t_A)$. In this case the overseer provides a "bail out effect" for the agent. Since implementation effort investment is costly, the agent is deterred from investing any effort at all because the overseer will not allow outcomes to turn out worse than the reversion level of policy precision (V_F), which

[15]The assumptions made are to ensure that a solution exists within the specified effort range, $[0, 1]$. Specifically, since the fraction derived from rearranging the agent's incentive compatibility condition to solve for e can dip below 0 or rise above 1 we bound the problem using max and min operators to rule out this possibility. Thus, by continuity (of the fraction within the expression) and the Intermediate Value Theorem, an effort investment always exists in this case, either on the boundaries of 0 or 1 or on the interior of the unit interval. Derivations and details are provided in the appendix.

in this case (holding the agent's bias constant) is not *bad enough* to induce the agent to invest more effort.[16] Combining all of the analysis above yields the subgame equilibrium for agent and overseer interactions when the agent has been authorized to make policy, embodied in the following proposition.

Proposition 1. *Suppose the agent is authorized to make policy by the principal. Then a perfect Bayesian equilibrium of the agent-overseer subgame is characterized by the following collection of strategies,*

1. *The agent makes effort investments according to s_A^{e*}, given by equation 6,*
2. *The agent always sets policy at his ideal point, $x^*(\omega) = \omega + t_A$,*
3. *The overseer makes review decisions according to $s_R^*(e)$, given by equation 3,*

Figure 1 provides a graphical example of the agent-overseer subgame equilibrium. The y-axis represents agent effort investment level, while the x-axis captures the distance between overseer and agent ideal points. Agent bias relative to the overseer is increasing as we move left to right on the horizontal axis. The dotted (flat) line represents the agent's effort investment when he will always be upheld: $\hat{e}_A(\beta, \kappa)$. When the agent's ideal point is sufficiently close to the overseer's oversight does not bind the agent's effort decisions. In this case, the agent invests effort based on his own motivations without taking into account the overseer since $\hat{e}_A(\beta, \kappa) > \underline{e}_R(t_A)$ (the red line). As the agent becomes more biased there is a point at which the overseer begins to require positive effort investments (the red line begins increasing with agent bias). Past the point at which $\hat{e}_A(\beta, \kappa) = \underline{e}_R(t_A)$ the maximum effort investment the agent is willing to make becomes the operable function. So long as $e_A^{\max}(t_A) \geq \underline{e}_R(t_A)$ the agent invests effort at the overseer's threshold and is upheld. This is the intermediate region of equilibrium effort investment in the figure where the blue line tracks the overseer's threshold. As the agent continues to become more biased, oversight becomes more stringent and the agent becomes willing to invest higher levels of effort to be upheld. However, once the agent becomes too biased, i.e., $e_A^{\max}(t_A)$ drops below $\underline{e}_R(t_A)$, equilibrium effort investments drop to 0. Oversight simply becomes too demanding and the agent no longer finds it incentive compatible to invest enough effort to be upheld. This is the final blue line segment in

[16] This deterrence effect is qualitatively similar to the "bail out effect" provided by judicial review identified in previous theoretical work. Specifically, it is related to the way in which judicial review bails out candidates in an electoral environment identified in Fox and Stephenson (2011) and the dissuading effect of review on the choice to regulate or agency effort exertion as in Bueno de Mesquita and Stephenson (2007) and Turner (2014), respectively. The effect of oversight in this model complements these papers.

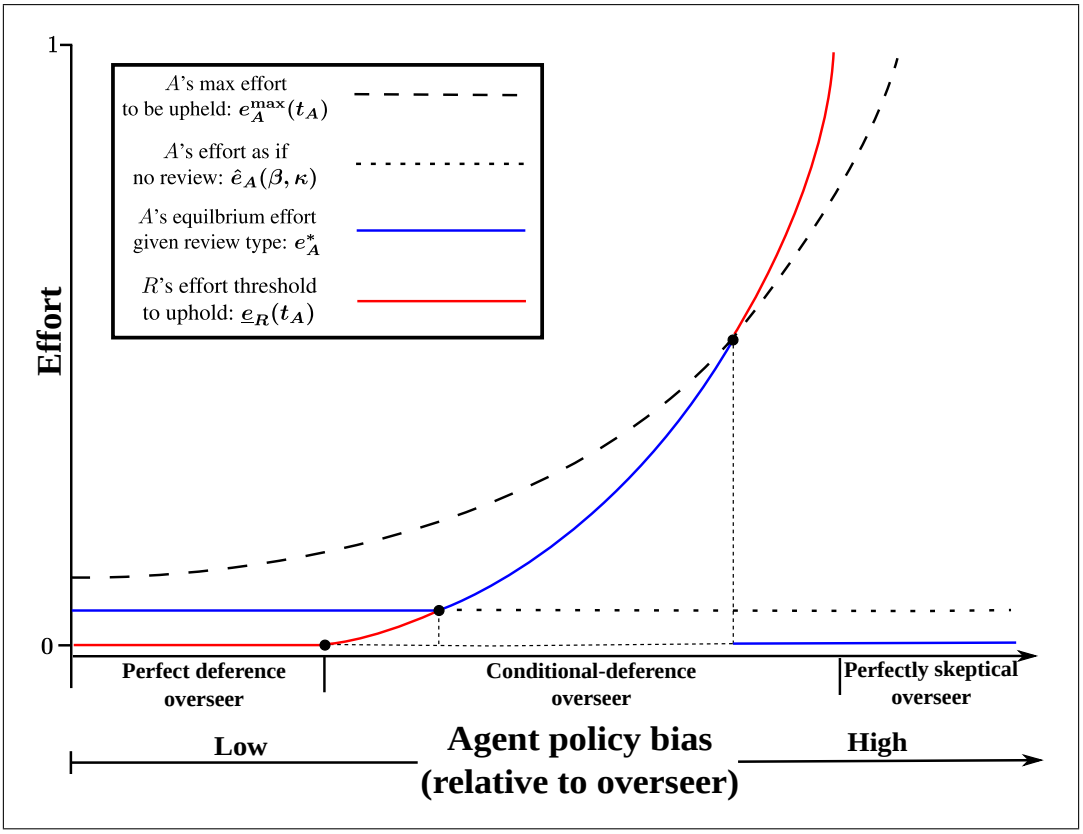


Figure 1: An Example of Equilibrium Effort Investments Conditional on Agent Bias.

the figure. Figure 1 also provides visual intuition for the main result characterizing the relationship between oversight, bias, and agent effort investments, described in the following proposition.

Proposition 2. *In equilibrium, agent bias affects agent effort investment if and only if oversight also affects agent effort investment.*

Proposition 2 presents a central result for the theory developed here. When oversight does not effectively provide effort incentives for the agent neither does agent bias. When the agent is not too biased he simply invests effort based on his own motivations (β and κ). Neither the agent's bias (t_A) nor oversight (through π) play a role in this investment. Similarly, when the agent is too biased, effort investments are also invariant. They are always zero since the agent will always be overturned. However, in the intermediate regions of agent bias, effort investments are increasing in both agent bias and the agent's aversion to being overturned, which only applies when oversight is effective. Thus, an agent's bias affects his effort investments if and only if the presence of oversight also affect those investments. This illustrates a fundamental interdependence between utilizing tools like staffing an agency with zealots rather than slackers (Gailmard and Patty 2007) or presidential appointees rather than careerist civil servants (Lewis 2008, 2011) and institutionalized oversight

to impact agent effort incentives. One is not effective without the other. This raises the question of when, based on the dynamics described in this section, a political principal benefits from authorizing an agent to make policy on her behalf given that he will face oversight of his actions and what type of agent benefits a principal when authorization is the best option.

4 The instrumental value of politicized policymaking

In this section we explore when, and under what circumstances, a political principal benefits from authorizing a biased, non-ally, agent to make policy on her behalf. The dynamics between agent effort investment and oversight described in the previous section play a central role. To start, anytime the principal chooses *not* to authorize the agent to make policy she receives the following expected payoff,

$$U_P(a = 0) = -V_F.$$

Since not authorizing the agent to make policy is equivalent to allowing unregulated outcomes to obtain, the principal can expect to lose utility equal to her expectation of these outcomes.[17] Whether the principal finds it beneficial to authorize the agent depends on the relative locations of agent and overseer ideal points. Once authorized, the agent will invest effort according to the subgame equilibrium characterized above. Thus, we analyze the principal's choices based on the environment that the agent and overseer will face if the agent is authorized: perfectly skeptical, perfectly deferential, or conditional-deference.

The first two cases are situations in which agent effort investments are invariant to both agent bias and oversight: the agent is either always overturned or always upheld. In contrast, the latter case is one in which the agent's effort investments are increasing in the agent's bias relative to the overseer (up to a point). We first analyze the former two cases together and then turn to the final, most interesting, case.

The first case is one in which the agent will always be met with reversal. This is true anytime it is not incentive compatible for the agent to invest sufficient effort to be upheld, which could be because it is impossible to do so — the agent is so biased that the overseer will never uphold independent of effort investment — or because oversight is too stringent and the agent is not willing to invest at the overseer's threshold — $e_A^{\max}(t_A) < \underline{e}_R(t_A)$. In either environment, if the principal does authorize the agent to make policy, knowing that he will ultimately be overturned, she can expect to simply bear the costs of authorization,

$$\Delta U_P(a = 1; r^*(e^*) = 1) = -c.$$

[17] This is because $\mathbb{E}[-\omega^2] = -[\mathbb{E}[0]^2 + V_F] = -V_F$.

In terms of policy, this payoff is equivalent to the principal simply not authorizing the agent. In both instances, final outcomes are predicated on the unregulated actions of private individuals or firms. However, she must also incur the authorization costs to empower the agent. Intuitively anytime the principal would pay a positive cost, $c > 0$, for authorizing an agent to make policy just to see that agent's actions overturned, she is better off simply ending the game by not authorizing. When authorization is costless ($c = 0$) the principal is indifferent between authorizing the agent and not. Since costly authorization is more realistic we do not dwell on breaking this knife-edge case in a particular way.[18] Overall, when authorizing the agent does not impact policy outcomes and delegation is costly to the principal, she never authorizes the agent to make policy in equilbrium.

The second case is when the agent, if authorized to make policy, will always be upheld. In this environment the agent will generally invest positive effort based on his own motivations. So the principal must decide if it is beneficial for her to allow the agent to make policy given that the agent will have unfettered discretion once authority is transferred. In this case the agent's actions will always obtain if he is authorized and the principal's corresponding incentive compatibility condition that must be met to authorize is given by,

$$\underbrace{t_A^2}_{\text{Spatial loss}} \leq \underbrace{V_F - V_\varepsilon(\hat{e}_A(\beta,\kappa))}_{\text{Implementation improvement}} - \underbrace{c.}_{\text{Authorization cost}}$$

This condition is simply a rearrangement of the principal's net expected payoff for authorizing an agent that will be upheld with certainty. Intuitively, the principal benefits from authorizing the agent to make policy in this environment if the agent is not too biased. Specifically, the spatial losses associated with delegating authority to the agent must be outweighed by the implementation improvement induced given that the agent will always invest effort based on his own motivations, $\hat{e}_A(\beta,\kappa)$, less the costs of authorization. The likelihood this condition is met and the principal benefits from agent authorization is unambiguously decreasing in agent bias t_A since this has no bearing on the agent's equilibrium effort investment. Further, because this effort level is invariant to agent bias, the likelihood that this condition will be met is increasing in the agent's intrinsic policy motivations β and the volatility of unregulated outcomes V_F, and decreasing in effort and authorization costs, κ and c, respectively.

Substantively, this result highlights the fact that when oversight is ineffective at structuring

[18] One can also imagine situations in which c could be negative. For instance, if the principal gains by "shifting blame" for policy failures to the agent then you could imagine that even though outcomes will not be appreciably different, the principal gains utility from being able to blame the agent for that failure and avoid external political costs like electoral challenge. While we do not expressly deal with this situation, we open the door to extending the model to incorporate this potentially interesting political situation.

agent effort incentives, the principal benefits from delegation based solely on agent and policy-environmental characteristics. If the agent is highly motivated, or if effort costs are low perhaps due to simple or mundane policy tasks, then it is more likely that agent authorization is beneficial. However, if the policy environment is relatively stable without regulation or the agent is extremely biased, perhaps through a process like agency capture, then it is unlikely that the principal benefits from delegation even with a formal institutional "check" like oversight in place.

The invariance of agent effort investments in these two scenarios, in which oversight does not impact effort, has clear implications for the instrumental value of authorizing a biased agent from the principal's perspective, captured by the following proposition.

Proposition 3. *If, upon being authorized to make policy, the agent will be always overturned or the overseer is perfectly deferential, the principal never benefits from a biased agent relative to an ally agent.*

In the environments in which the agent will always be overturned if he is authorized to make policy, agent bias has no impact on principal utility. Regardless of the agent's bias, policy outcomes are the same: unregulated outcomes will obtain. Therefore, the only thing that impacts the principal's payoff is the cost of authorization. Conversely, in environments where the overseer is perfectly deferential, a biased agent strictly decreases the principal's utility. While there are cases where the principal benefits from authorizing a positively biased agent, it is strictly better for the principal if the agent is an ally. Since the agent's effort investment is invariant to his bias, the principal can only lose utility from positive bias as it only increases the spatial policy losses associated with agent-made policy with no effort benefits. Thus, when oversight does not affect agent effort investment, the principal never benefits from a biased agent relative to an ally, even if delegating policymaking authority is incentive compatible.

Now consider the most interesting case in which oversight does impact agent effort investments. In this case, the environment is characterized by intermediately biased agents (relative to the overseer) in which equilibrium effort investments are at the overseer's threshold level of required effort to uphold (that is, $e_A^{\max}(t_A) \geq \underline{e}_R(t_A)$). The agent, if authorized, will invest effort equal to $\underline{e}_R(t_A)$ and be upheld by the overseer. This, combined with the reversion utility of not authorizing the agent, implies the following net expected payoff of the principal for authorizing the agent,

$$\Delta U_P(a=1; r^*(\underline{e}_R(t_A))=0) = -t_A^2 - V_\varepsilon(\underline{e}_R(t_A)) - c + V_F.$$

That is, since the agent will be upheld for investing effort exactly at the overseer's threshold, the principal can expect to lose utility based on the spatial losses associated with an agent with bias t_A, but gains as the agent's associated effort investment improves implementation quality relative to the volatility of unregulated outcomes. Now, since we know from the agent-overseer subgame that the

agent will make the overseer indifferent, we can reduce this net expected payoff by substituting the value of $V_\varepsilon(\underline{e}_R(t_A))$ when the overseer's incentive compatibility condition (equation 2) holds with equality. This reduces the principal's net payoff in this environment to,

$$\begin{aligned}\Delta U_P(a=1; r^*(\underline{e}_R(t_A))=0) &= -t_A^2 - \left[V_F - t_A^2 + 2t_A t_R\right] - c + V_F, \\ &= -2t_A t_R - c.\end{aligned}$$

By incentive compatibility, the principal will authorize the agent to make policy if $\Delta U_P(a=1; r^*(\underline{e}_R(t_A)) = 0) \geq 0$. This implies that the principal will authorize the agent in this case if,

$$-2t_A t_R \geq c.$$

If authorization costs are positive, $c > 0$, then the principal can only benefit from empowering the agent to make policy if the agent and overseer are on opposite sides of her. That is, the left-hand side of the incentive compatibility condition is only positive if t_A and t_R are oppositely signed. Since by assumption $t_R < 0$ this means that if the principal benefits from authorizing the agent to make policy at all then the agent is positively biased on the opposite side of the principal than the overseer: $t_A > 0$. If authorization is costless then it is possible that the principal can benefit from delgating to an ally agent. However, even in that case the principal's utility is increasing in agent bias. Thus, the only time in this environment that the principal jointly benefits from delegating to the agent *and* having that agent be an ally is the knife-edge case in which if the agent were positively biased at all oversight would become too stringent and the agent would begin investing no effort at all. This leads to the main proposition that characterizes when politicized policymaking is instrumentally valuable to the principal.

Proposition 4. *If, upon being authorized to make policy, the agent will invest the threshold level of effort required by the overseer to be upheld then anytime the principal benefits from authorizing the agent to make policy when authorization costs are positive, she also benefits from a biased agent relative to an ally agent.*

Propositions 3 and 4 provide the basis for the main theoretical insights of this paper. When oversight is ineffective as an institutional check on agent policymaking, agent bias is only detrimental to political principals. However, when oversight is effective at providing positive incentives for agent policymaking, the principal would prefer to have a biased agent to continue to strengthen these incentives. That is, agent effort incentives are increasingly strengthened the more effective is oversight *and* the more biased that agent is. In this way, the principal instrumentally prefers to trade off biased content of policy for reduced slack in implementation so long as oversight is an effective institutional check on agent behavior. By pitting a biased agent against a biased overseer, in particu-

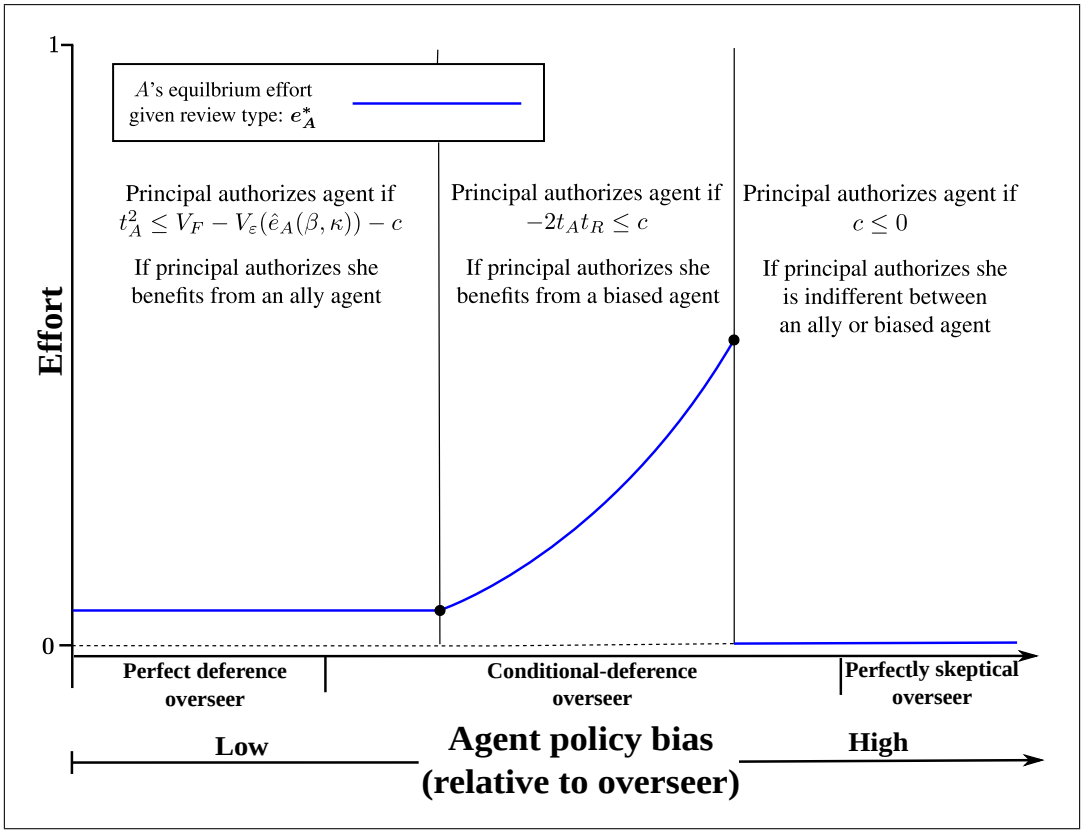
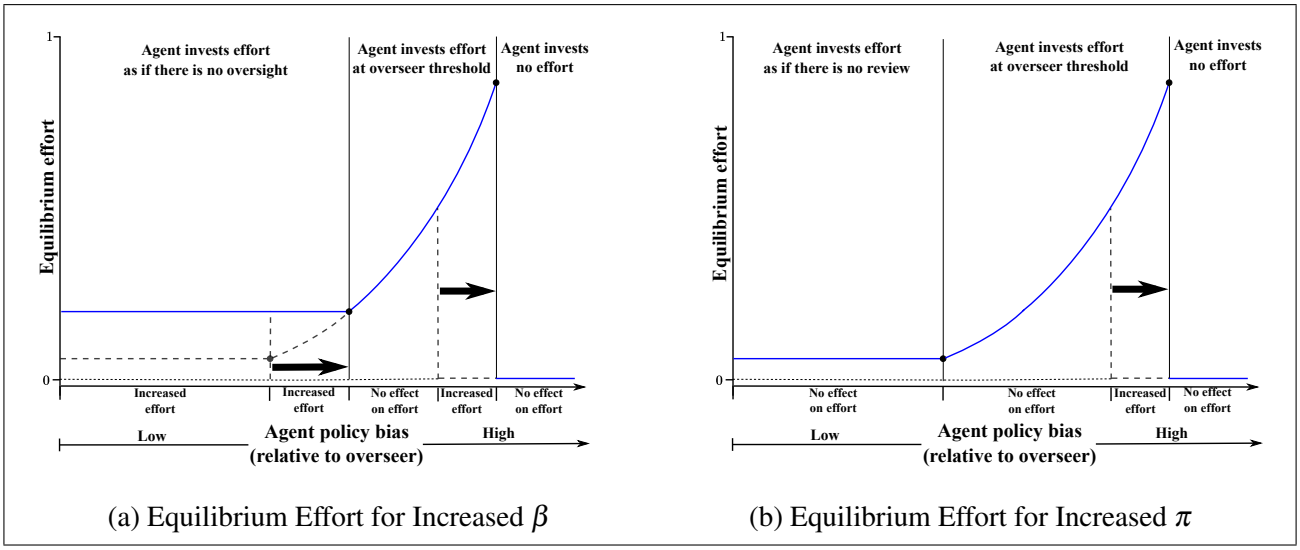


Figure 2: An Example of Equilibrium Authorization and the Instrumental Value of Agent Bias Conditional on Oversight Environment

lar having an agent oppositely biased from the overseer, the principal can benefit from the increased precision induced through agent effort investments. This is true anytime oversight has any impact on agent incentives and anytime the principal would benefit from having an agent make policy at all.

Figure 2 provides a graphical representation of the intuition underlying propositions 3 and 4 using the example equilibrium effort investments from figure 1. The axes are the same as in figure 1, but in this case only the equilibrium effort investment is graphed. When the agent will always be upheld his effort is unresponsive to increasing bias and therefore the principal is only harmed by increased agent bias; she prefers an ally agent. On the other extreme, when the agent is too biased relative to the overseer so that he will always be overturned, positive effort investments never occur in equilibrium. In this case, the agent never invests positive effort and therefore the principal derives no benefit from a biased agent. These two cases — a sufficiently moderate and sufficiently extreme agent — capture the principal's incentives under the environments in proposition 3. However, when the agent is intermediately biased relative to the overseer, he invests effort that exactly matches the overseer's threshold. In this case equilibrium effort is always increasing in agent bias until the point at which the agent becomes too biased. In this case, if the principal benefits from authorizing the



(a) Equilibrium Effort for Increased β

(b) Equilibrium Effort for Increased π

Figure 3: Examples of Comparative Statics for Increased Policy Motivations and Reversal Aversion

agent then she also benefits from increased agent bias relative to authorizing an ally, as in proposition 4. These insights provide several empirically relevant implications that can be applied to many different political environments of interest. These are discussed in the next section.

5 Empirical implications

In this section we apply the insights of the model to bureaucratic politics. We first discuss comparative statics relating to both agent policy motivations and aversion to being overturned. We then turn to discussion of implications for personnel politics and political appointments to the bureaucracy. Finally, we discuss how the results imply limits to the efficacy of managerial strategies.

We focus on two comparative statics of interest and discuss how they relate to different aspects of bureaucratic politics: increased intrinsic policy motivation, β, and increased reversal aversion, π. In both cases, aggregate net levels of equilibrium effort investment increase, but the positive relationship is conditional on what type of oversight is induced. Figure 3 displays examples of these intuitions graphically. In both graphics the gray dashed lines denote previous levels of equilibrium effort investments prior to increasing the parameters of interest. The blue solid lines denote the equilibrium effort investments following an increase in the parameters. Ultimately, the figures illustrate how the impact of these parameter shifts depend on how agent bias (increasing along the x-axis) interacts with oversight.

First, consider a case in which agent policy motivations, β, increase, illustrated in figure 3a. This initially seems unambiguously positive in that it will generally produce a net increase in aggregate effort investments toward implementation quality. However, the relationship is conditional on how oversight impacts agent incentives. When the agent is ideologically close to the overseer

oversight does not impact the agent's effort investment. However, the agent's policy motivations do increase $\hat{e}_A(\beta, \kappa)$ and therefore, effort investments increase proportional to the increase in β. This also expands the range of agent biases in which oversight has no bite with respect to effort. Once the agent becomes moderately biased, oversight does become stringent enough to induce the agent to increase his effort investments to be upheld. In this case the agent invests effort exactly at the threshold as in the baseline case. The increase in β, while it does increase the maximum level of effort the agent *would be willing to invest*, does not effectively alter observed investment levels. However, by increasing $e_A^{\max}(t_A)$, increased policy motivations expand the range of agents that find it incentive compatible to invest sufficient effort to be upheld. These shifts in the ranges of agent biases in which sufficient effort is invested is partly due to the fact that increasing β strengthens agent effort investment incentives but does *not* affect the stringency of oversight. More biased agents now find it beneficial to invest sufficient effort to avoid reversal than under lower levels of policy motivations. This further implies that the principal can benefit from a larger range of more extreme agent biases. Thus, from the principal's perspective, there is a positive correlation between agent policy motivations and agent bias in terms of returns from authorization. Finally, the effort investment levels of extremely biased agents remain unaffected and those agents invest zero effort and accept being overturned.

Now consider what happens as an agent becomes more averse to being overturned, i.e., π increases, illustrated in figure 3b. The impact of doing so is similar to increasing policy motivations in that from a net perspective, effort investments increase, but they are again conditional on the relationship between oversight and agent bias. When agents are ideologically proximate to the overseer, effort investments remain unchanged. This is because $\hat{e}_A(\beta, \kappa)$ does not respond to changes in reversal aversion since there is no risk of being overturned when oversight is ineffective. However, the maximum level of effort the agent is willing to invest to be upheld, $e_A^{\max}(t_A)$, does increase in π while, again, the stringency of oversight does not. Thus, the range of intermediately biased agents that will now invest sufficient effort to be upheld expands as in the policy motivations case. Higher biased agents now switch from investing zero effort and accepting reversal to investing sufficient effort to be upheld. Once again, extremely biased agents still find it incentive incompatible to invest any positive effort and continue to invest no effort. Thus, strengthening an agent's aversion to being overturned will increase observed equilibrium effort investments, but only for a small range of agent biases. From the principal's perspective, agent reversal aversion and agent bias are complemenatary. An increase in aversion increases the maximal agent bias the principal can benefit from.

Taken together, these comparative statics predict positive correlations between effort investments and increased policy motivations and reversal aversion within agencies. However, these relationships are conditioned by the fact that these increases only 'work' at increasing effort for particular ranges of agent biases. In both cases the principal can benefit from a larger range of more extreme

agent biases than before increases in these parameters. Since the principal's utility is increasing in agent bias when oversight affects effort investments (the intermediate case) increasing agent policy motivations and reversal aversion increase the level of biases that benefit the principal. We now turn to applying these insights to particular situations in bureaucratic politics.

One of the most important, and most difficult, tasks a president faces is staffing top positions in the federal bureaucracy (Waterman 1989). It is estimated that presidents must staff approximately 4,000 such positions upon taking office (Lewis 2008, 2011). The theory developed here, while appointments were not modeled explicitly, provides insight into what types of appointees can benefit presidents conditional on the nature of the larger institutional environment, e.g., the nature of oversight. Propositions 3 and 4 have clear implications for how presidents can leverage the institutional system in various ways to provide strong incentives for increased policy quality.

First, if the overseer is conceived of as an external interest group, for instance, then our theory's implications are generally in line with the findings of Bertelli and Feldmann (2007). Appointing a biased agent to offset the interest group's biases can be beneficial insofar as the interest group serves as a fire alarm for the legislature. The divergence between the interests of the group and those of the agency can induce higher quality policies overall through the group's credible threat of "sounding the alarm." The theoretical insights also speak to presidential appointments across insitutions within the Executive branch. The president can simultaneously make appointments to direct agency policymaking (by appointing directors, secreataries, agency heads, etc.) and to shape the nature of oversight (by appointing the head of the OIRA, for example). By appointing an agency head that is oppositely biased the OIRA director the president can put pressure on the agency to more adequately justify policy choices and provide evidence that it is well equipped to implement policies effectively. In particular, the president ought to appoint an agency head that is more pro-regulation (anti-regulation) and an overseer that is more anti-regulation (pro-regulation) than herself to induce the highest effort investments. Moreover, the comparative statics described above suggest that appointing "zealots" that are highly policy motivated, while simultaneously strengthening the role of oversight, actually increases the level of bias that the principal would prefer from the agency. Overall, the results provide an instrumental rationale for why an executive might optimally choose to appoint subordinates that do not share her substantive or ideological views.

Lewis (2011) suggests that Presidents benefit from appointing ideologically distinct agency heads when these appointees have difficulty impacting agency policy outputs in less ideologically friendly agencies (54-55). That is, when the EPA is largely staffed with pro-regulatory ("careerist") bureaucrats that seek to implement stringent environmental protection regulation, above and beyond what the president would prefer, it may be difficult for the head of the EPA to fully temper policy output and direct it back toward less stringent regulation. In this instance, our theory suggests that appointing an agency head (that would serve as a "policy gatekeeper") that prefers less stringent

regulation than even the president will induce the subordinate bureaucrats within the agency to produce higher quality regulatory interventions than if they were led by someone that shared their enthusiasm for stringent regulation. More generally, the results suggest that intra-agency conflict in the form of institutionalized gatekeepers or veto points can strongly incentivize policymaking bureaucrats to work harder than they otherwise would in order to increase the probability that their policy goals may be realized (Feldman 1989). In short, the theory provides an instrumental rationale for organizing agencies to promote a particular type of "internal conflict" in regulatory agencies (West 1988).

The comparative statics of the model also have implications for the efficacy of managerial motivational strategies within the bureaucracy. If altering agencies' preferences over outcomes is prohibitively costly then a savvy principal may wish to attempt to strengthen effort incentives by increasing the policy motivations of bureaucrats or tying stronger penalties to being overturned by overseers, thereby increasing reversal aversion. Increasing policy motivations may be accomplished by streamlining procedures so that there is less "red tape", strengthening hierarchical authority (Moynihan and Pandey 2007), increasing the ratio of zealots to slackers (Gailmard and Patty 2007), or enhancing an agencies 'commitment to mission' through staffing or other means (Wilson 1989). Similary, tying oversight outcomes more strongly to agency budgets, promotional decisions, or something akin to "performance pay," may allow a principal to increase an agency or bureaucrat's aversion to being overturned. In general, both strategies will be effective at increasing net levels of observed effort, but the comparative statics point out important qualifications predicated on the policymaking environment.

The parameter denoting agent policy motivations, β, intuitively captures the effect(s) of increasing intrinsic policy motivations. As the comparative static displayed in figure 3a illustrates, the efficacy of this managerial strategy is conditional on the institutional environment the agency must navigate. It is a strategy that ought to produce net benefits with respect to strengthing effort incentives for agencies that are either moderate relative to the overseer or intermediately biased. In particular, increasing a moderate agency's motivations can serve as a substitute when oversight is ineffective. A larger range of moderate-biased agencies are unaffected by oversight, but there effort investments still increase since their motivations to produce high quality policy increased. The strategy will be ineffective for a middle range and high range of agency biases, but for a range of agencies that were once deterred from investing effort, their investments increase dramatically with an increase in policy motivations. Counter-intuitively, this implies that when a manager wishes to increase the motivations of her subordinates, she would, if given the choice, actually prefer them to become more biased as well since doing so would intensify the effects of the motivational strategy itself. Overall, a managerial strategy of this sort will not always impact observed output, but it can still be used selectively quite effectively.

Similarly, if a principal attempts to strengthen the role of oversight through increasing an agency's reversal aversion net effort increases, but there are caveats. Specifically, this strategy does not work when the agency or bureaucrat has interests that are closely aligned with the overseer. The only increase in effort comes by inducing agencies that once found it incentive incompatible to invest effort to begin investing high levels of effort to pass muster in ex post review. Put another way, strengthening oversight of policymaking agents is only effective when the agents are biased enough away from their potential overseers. Without a sufficient level of divergence the overseer cannot commit to requiring more from the agent. Even though the agent may be more averse to being overturned once a motivational strategy of this sort is applied, that aversion is inconsequential if the overseer cannot credibly commit to sanctioning the agent.[19]

6 Conclusion

In this paper, we developed a theory of delegation and showed that political principals — Presidents, legislatures, agency leaders — can benefit from authorizing biased agents to make policy on their behalf. This potential benefit is due to the recognition that policymakers both craft the substance of policy and invest effort to implement those policies effectively. Due to this duality in policymaking the principal benefits from pitting a biased agent, with full policymaking discretion, against an oppositely biased overseer, empowered to reverse the agent's actions if insufficient effort is invested toward improving outcomes. Institutionalized oversight is only effective as a means of political control if the agent is biased, and leveraging agent bias to induce effort is only a viable route to policy improvements if oversight is an effective means of political control. The characteristics of the agent, the policy environment, and the dynamics of political oversight introduce both opportunities and constraints for principals interested in promoting strong effort incentives for agents they will authorize to make policy on their behalf. The model is flexible enough to be extended to include other important determinants of output such as interest group participation, oversight by multiple institutions, and splitting up policymaking tasks across multiple agents. This paper represents a step toward a fuller understanding of how ubiquitous processes, like bureaucratic policymaking in the shadow of oversight, impact the dynamics of political decisions like personnel decisions and appointments, agency design, and the efficacy of managerial motivational strategies.

[19] In general, these effects are consistent with theories of institutional determinants of public service motivation (see Moynihan and Pandey 2007, for a review).

References

Banks, Jeffrey S. and Rangarajan K. Sundaram. 1993. Adverse Selection and Moral Hazard in a Repeated Elections Model. In *Political Economy: Institutions, Competition, and Representation*, ed. William A. Barnett, Melvin J. Hinich and Norman J. Schofield. New York, NY: Cambridge University Press.

Banks, Jeffrey S. and Rangarajan K. Sundaram. 1998. "Optimal Retention in Agency Problems." *Journal of Economic Theory* 82(2):293 – 323.

Bendor, Jonathan and Adam Meirowitz. 2004. "Spatial Models of Delegation." *American Political Science Review* 98(2):293–310.

Bendor, Jonathan;, Amihai; Glazer and Thomas Hammond. 2001. "Theories of Delegation." *Annual Review of Political Science* 4:235–269.

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

Breyer, Stephen. 1986. "Judicial Review of Questions of Law and Policy." *Administrative Law Review* 38(Fall):363–398.

Bubb, Ryan and Patrick L. Warren. 2014. "Optimal Agency Bias and Regulatory Review." *Journal of Legal Studies* 43(January):95–135.

Bueno de Mesquita, Ethan and Matthew C. Stephenson. 2007. "Regulatory Quality under Imperfect Oversight." *American Political Science Review* 101(3):605–620.

Carpenter, Daniel and David Moss, eds. 2013. *Preventing Regulatory Capture: Special Interest Influence and How to Limit It*. Cambridge University Press.

Carpenter, Daniel P. 2001. *The Forging of Bureacuratic Autonomy: Reputations, Networks, and Policy Innovation in Executive Agencies, 1862-1928*. Princeton, NJ: Princeton University Press.

Che, Yeon-Koo and Navin Kartik. 2009. "Opinions as Incentives." *Journal of Political Economy* 117(5):815 – 860.

Derthick, Martha. 1990. *Agency Under Stress: The Social Security Administration in American Government*. Washington, D.C.: The Brookings Institution.

Dessein, Wouter. 2002. "Authority and Communication in Organizations." *Review of Economic Studies* 69(4):811–838.

Dragu, Tiberiu and Oliver Board. 2013. "On Judicial Review in a Separation of Powers System." *Working paper. New York University.* .

Epstein, David and Sharyn O'Halloran. 1994. "Administrative Procedures, Information, and Agency Discretion: Slack vs. Flexibility." *American Journal of Political Science* 38(3):697–722.

Feldman, Martha. 1989. *Order without Design: Information Production and Policymaking*. Stanford, CA: Stanford University Press.

Fox, Justin and Matthew C. Stephenson. 2011. "Judicial Review as a Response to Political Posturing." *American Political Science Review* 105(2):397–414.

Gailmard, Sean. 2002. "Expertise, Subversion, and Bureaucratic Discretion." *Journal of Law, Economics, & Organization* 18(2):536–555.

Gailmard, Sean and John W. Patty. 2007. "Slackers and Zealots: Civil Service, Policy Discretion, and Bureaucratic Expertise." *American Journal of Political Science* 51(4):873–889.

Gailmard, Sean and John W. Patty. 2013*a*. "Formal Models of Bureaucracy." *Annual Review of Political Science* 15:353–377.

Gailmard, Sean and John W. Patty. 2013*b*. *Learning While Governing: Expertise and Accountability in the Executive Branch*. Chicago, IL: University of Chicago Press.

Gailmard, Sean and Thomas Hammond. 2011. "Intercameral Bargaining and Intracameral Organization in Legislatures." *Journal of Politics* 73(2):535–546.

Gersen, Jacob E. and Adrian Vermeule. 2012. "Delegating to Enemies." *Columbia Law Review* pp. 2193–2238.

Gilligan, Thomas and Keith Krehbiel. 1987. "Collective Decision-Making and Standing Committees: An Informational Rationale for Restrictive Amendment Procedures." *Journal of Law, Economics, and Organization* 3(2):287–335.

Gilligan, Thomas W. and Keith Krehbiel. 1989. "Asymmetric Information and Legislative Rules with a Heterogeneous Committee." *American Journal of Political Science* 33:459–490.

Heclo, Hugh. 1988. "The In-and-Outer System: A Critical Assessment." *Political Science Quarterly* 103(Spring):37–56.

Hirsch, Alexander V. and Kenneth W. Shotts. 2014. "Policy-Development Monopolies: Adverse Consequences and Institutional Responses." *Unpublished manuscript. California Institute of Technology* .

Huber, John D. and Nolan McCarty. 2004. "Bureaucratic Capacity, Delegation, and Political Reform." *American Political Science Review* 98(3):481–494.

Krehbiel, Keith. 1991. *Information and Legislative Organization*. University of Michigan Press.

Lewis, David E. 2008. *The Politics of Presidential Appointments: Political Control and Bureaucratic Performance*. Princeton, NJ: Princeton Univesity Press.

Lewis, David E. 2011. "Presidential Appointments and Personnel." *Annual Review of Political Science* 14:47–66.

Lipsky, Michael. 1980. *Street-Level Bureaucracy*. New York, NY: Russell Sage Foundation.

Lowi, Theodore J. 1979. *The End of Liberalism*. 2nd ed. New York, NY: Norton Press.

Meier, Kenneth J. 1997. "Bureaucracy and Democracy: The Case for More Bureaucracy and Less Democracy." *Public Administration Review* 57(3):193 – 199.

Meier, Kenneth J. and Laurence J. O'Toole, Jr. 2006. "Political Control versus Bureaucratic Values: Reframing the Debate." 66(2):177–192.

Moynihan, Donald P. and Sanjay K. Pandey. 2007. "The Role of Organizations in Fostering Public Service Motivation." *Public Administration Review* 67(1):40–53.

Niskanen, William. 1971. *Bureaucracy and Representative Government*. Chicago: Aldine.

Prendergast, Canice. 2007. "The Motivation and Bias of Bureaucrats." *American Economic Review* 97:180–196.

Rogers, James R. 2001. "Information and Judicial Review: A Signaling Game of Legislative-Judicial Interaction." *American Journal of Political Science* 45(1):84–99.

Rogers, James R. and Georg Vanberg. 2002. "Judicial Advisory Opinions and Legislative Outcomes in Comparative Perspective." *American Journal of Political Science* 46(2):379 – 397.

Seidenfeld, Mark. 2002. "The Psychology of Accountability and Political Review of Agency Rules." *Duke Law Journal* 18(December):1051 – 1095.

Shipan, Charles. 1997. *Designing Judicial Review*. Ann Arbor, MI: University of Michigan Press.

Spence, David B. and Frank Cross. 2000. "A Public Choice Case for the Administrative State." *Georgetown Law Journal* 89:97–142.

Staton, Jeffrey K. and Georg Vanberg. 2008. "The Value of Vagueness: Delegation, Defiance, and Judicial Opinions." *American Journal of Political Science* 52(3):504–519.

Stephenson, Matthew C. 2006. "A Costly Signaling Theory of "Hard Look" Judicial Review." *Administrative Law Review* 58(4):753–814.

Ting, Michael M. 2011. "Organizational Capacity." *Journal of Law, Economics, & Organization* 27(2):245–271.

Turner, Ian R. 2014. "Working Smart *and* Hard? Agency Effort, Judicial Review, and Policy Precision." *Unpublished Manuscript. Texas A&M University* .

Van Weelden, Richard. 2013. "Candidates, Credibility, and Re-election Incentives." *Review of Economic Studies* 80(4):1622–1651.

Vanberg, Georg. 2001. "Legislative-Judicial Relations: A Game-Theoretic Approach to Constitutional Review." *American Journal of Political Science* 45(2):346–361.

Warren, Kenneth F. 2004. *Administrative Law in the Political System*. 4th ed. Boulder, CO: Westview.

Waterman, Richard W. 1989. *Presidential Influence and the Administrative State*. Knoxville, TN: University of Tennessee Press.

West, William F. 1988. "The Growth of Internal Conflict in Administrative Regulation." *Public Administration Review* 48(4):773–782.

Wilson, James Q. 1989. *Bureaucracy: What Government Agencies Do and Why They Do It*. New York, NY: Basic Books.

A Supplemental Appendix

A.1 Agent-overseer subgame

In this section we formally characterize the agent-overseer subgame equilibrium constructed in the text. We begin a bit out of sequence, but in line with the progression of the text, by first showing that the agent will always set substantive policy at his ideal point: $x^*(\omega) = \omega + t_A$. Then we derive the overseer's threshold level of effort investment required to uphold the agent given that the agent sets policy at his ideal point. Finally, we formally derive the agent's effort investment decisions for each oversight regime — perfectly skeptical, perfectly deferential, conditional-deference — which completes characterization of the (subgame) equilibrium.

Agent substantive policy choice. In this section we show that the agent always sets substantive policy at his ideal point given the opportunity to make policy.

Lemma 1. *The agent always sets the substantive content of policy at his ideal point when given the opportunity:* $x^*(\omega) = \omega + t_A$.

Proof of Lemma 1. To show that the agent always sets policy at his ideal point we show that he is always better off by checking deviations in two cases: (1) when the overseer upholds the agent and (2) when the overseer reverses the agent. In both cases let $\delta > 0$ denote the agent's deviation so that if he deviates $x = \omega + t_A + \delta$.

Case 1: Overseer upholds. The agent's expected utility from setting policy sincerely, $x = \omega + t_A$, is given by,

$$U_A(x = \omega + t_A | r = 0) = -\beta V_\varepsilon(e) - \kappa e.$$

The agent's expected utility from deviating and setting policy to $x = \omega + t_A + \delta$ is given by,

$$U_A(x = \omega + t_A + \delta | r = 0) = -\beta(\delta^2 + V_\varepsilon(e)) - \kappa e.$$

These combine to give the agent's net expected payoff from deviating from sincere policymaking:

$$\begin{aligned} \Delta U_A(x = \omega + t_A + \delta | r = 0) &= -\beta(\delta^2 + V_\varepsilon(e)) - \kappa e + \beta V_\varepsilon(e) + \kappa e, \\ &= -\beta\delta^2. \end{aligned}$$

If $\beta = 0$ then the agent is no better off from deviating and if $\beta > 0$ the agent is strictly worse off from deviating. Thus, when the overseer upholds the agent, he, in weakly undominated strategies, does not deviate and chooses policy at his ideal point.

Case 2: Overseer reverses. The agent's expected utility from making policy sincerely given the overseer reverses is given by,

$$U_A(x = \omega + t_A | r = 1) = -\beta(t_A^2 + V_F) - \pi.$$

The agent's expected utility from deviating by δ is given by,

$$U_A(x = \omega + t_A + \delta | r = 1) = -\beta(t_A^2 + V_F) - \pi.$$

The net expected payoff for deviating then, since the agent receives the same payoff from overseer reversal regardless, is zero. Having shown that the agent gains nothing from deviating from the posited equilibrium strategy of sincere policymaking in both cases, the result follows. ■

Overseer threshold. In this section, we derive the overseer's optimal review strategy given correct beliefs regarding the agent's optimal substantive policy choice, $x^*(\omega)$, and observed agent effort investment e.

Lemma 2. *In equilibrium, the overseer plays the following best response strategy,*

$$s_R^*(e) = \begin{cases} \textit{uphold: } r = 0 & \textit{if } e \geq \underline{e}_R(t_A), \\ \textit{reverse: } r = 1 & \textit{otherwise,} \end{cases}$$

Proof of Lemma 2. First, consider the overseer's subjective expected utility for overturning the agent,

$$\begin{aligned} U_R(r=1;\rho_{-R}) &= -(y-t_R)^2 \\ &= -(\omega - t_R)^2, \\ &= -\mathbb{E}[\omega - t_R]^2 - V[\omega], \\ &= -t_R^2 - V_F. \end{aligned}$$

Now, consider the overseer's subjective expected utility for upholding the agent with bias t_A that invested effort e,

$$\begin{aligned} U_R(r=0;\rho_{-R}) &= -(y-t_R)^2, \\ &= -(x^*(\omega) - \omega + \varepsilon - t_R)^2, \\ &= -\mathbb{E}[x^*(\omega) - \omega - t_R]^2 - V[x^*(\omega) - \omega - t_R] - \mathbb{E}[\varepsilon|e]^2 - V[\varepsilon|e], \\ &= -(t_A - t_R)^2 - V_\varepsilon(e). \end{aligned}$$

Define $\Delta U_R(r=0;\rho_{-R}) \equiv U_R(r=0;\rho_{-R}) - U_R(r=1)$ as the overseer's net expected utility for upholding. Then we have,

$$\begin{aligned} \Delta U_R(r=0;\rho_{-R}) &= -(t_A - t_R)^2 - V_\varepsilon(e) + t_R^2 + V_F, \\ &= -t_A^2 + 2t_A t_R - t_R^2 - V_\varepsilon(e) + t_R^2 + V_F, \\ &= -t_A^2 + 2t_A t_R - V_\varepsilon(e) + V_F. \end{aligned}$$

Incentive compatibility implies that the overseer, given ideal point t_R and unregulated policy volatility V_F, will uphold an agent with bias t_A that has invested effort e if and only if $\Delta U_R(r=0;\rho_{-R}) \geq 0$. Thus we have,

$$-t_A^2 + 2t_A t_R - V_\varepsilon(e) + V_F \quad \geq \quad 0.$$

Now, rearranging and simplifying to isolate variances (e.g., policy precision) on one side of the inequality and spatial policy losses based on divergent ideal points, we have:

$$V_F - V_\varepsilon(e) \quad \geq \quad t_A^2 - 2t_A t_R, \tag{A.7}$$

as is presented in-text in equation 2. The increase in implementation quality on the LHS must outweigh the net spatial policy losses based on divergent ideal points on the RHS. Now, by incentive compatibility the overseer's threshold level of required effort investment to uphold the agent is defined as $\underline{e}_R(t_A) \equiv e$ such that equation A.7 holds with equality given agent bias t_A. This yields the optimal overseer review strategy as stated in the result. ■

Agent effort investments. In this section we derive the agent's optimal effort investments for each possible scenario: facing a perfectly skeptical, perfectly deferential, or conditional-deference overseer, ultimately yielding the agent's optimal effort investment strategy given by equation 6 in text.

Lemma 3. *Define* $e_A^{\max}(t_A) = \max\left[\min\left[\frac{\beta(t_A^2+V_F-V_\varepsilon(e_A^{\max}(t_A)))+\pi}{\kappa}, 1\right], 0\right]$. *The agent will never invest effort higher than* $e_A^{\max}(t_A)$ *to be upheld by the overseer.*

Proof of Lemma 3. When the agent faces a conditional-deference overseer his net expected utility from investing the threshold level of effort required to be upheld is given by (we simply use e to represent this level of effort),

$$\Delta U_A(e \geq \underline{e}_R(t_A); \rho_{-A}) = \beta(t_A^2 + V_F - V_\varepsilon(e)) - \kappa e + \pi.$$

Thus, the agent will invest this level of effort if and only if $\Delta U_A(e \geq \underline{e}_R(t_A); \rho_{-A}) \geq 0$. Solving the expression with equality for e gives the maximum level of effort the agent would be willing to invest given t_A in order to be upheld (by incentive compatibility):

$$e = \frac{\beta(t_A^2 + V_F - V_\varepsilon(e)) + \pi}{\kappa}. \tag{A.8}$$

The RHS of Equation A.8 can fall below 0 and rise above 1. So to ensure an effort investment always exists further define:

$$e_A^{\max}(t_A) = \max\left[\min\left[\frac{\beta(t_A^2 + V_F - V_\varepsilon(e_A^{\max}(t_A))) + \pi}{\kappa}, 1\right], 0\right].$$

Given this formulation, $e_A^{\max}(t_A)$ always exists. The RHS of equation A.8 is continuous over the interval $[0,1]$ (and if 0 or 1 is the solution). So, either $e_A^{\max}(t_A)$ is on a boundary (0 or 1, if the RHS of equation A.8 is either negative or above one respectively) or there is an interior solution, which

is implied by (continuity and) the Intermediate Value Theorem. Optimality of this as an agent best response follows from incentive compatibility. ■

Lemma 4. *In equilibrium, the agent invests effort according to the following strategy,*

$$s_A^{e*} = \begin{cases} \hat{e}_A(\beta,\kappa) & \text{if } \hat{e}_A(\beta,\kappa) \geq \underline{e}_R(t_A), \\ \underline{e}_R(t_A) & \text{if } \hat{e}_A(\beta,\kappa) < \underline{e}_R(t_A) \text{ and } e_A^{\max}(t_A) \geq \underline{e}_R(t_A), \\ 0 & \text{if } \hat{e}_A(\beta,\kappa) < \underline{e}_R(t_A) \text{ and } e_A^{\max}(t_A) < \underline{e}_R(t_A), \end{cases}$$

where $\hat{e}_A = \arg\max_e -\beta V_\varepsilon(e) - \kappa e$, $\underline{e}_R(t_A) \equiv e$ *such that* $V_F - V_\varepsilon(e) = (t_A - t_R)^2 - t_R^2$, *and* $e_A^{\max}(t_A) = \max\left[\min\left[\frac{\beta(t_A^2 + V_F - V_\varepsilon(e_A^{\max}(t_A))) + \pi}{\kappa}, 1\right], 0\right]$.

Proof of Lemma 4. To verify that these are best responses for the agent we need to check three cases: (1) the overseer always overturns ($\hat{e}_A(\beta,\kappa) < \underline{e}_R(t_A)$ and $e_A^{\max}(t_A) < \underline{e}_R(t_A)$); (2) the overseer always upholds ($\hat{e}_A(\beta,\kappa) \geq \underline{e}_R(t_A)$); (3) the overseer upholds if and only if the agent invests effort high enough, which is higher than the agent would invest absent oversight ($\hat{e}_A(\beta,\kappa) < \underline{e}_R(t_A)$ and $e_A^{\max}(t_A) \geq \underline{e}_R(t_A)$). These cases are defined by the overseer's best response in Lemma 2 and the maximum effort investment the agent is willing to make to be upheld in Lemma 3.

Overseer always overturns (perfectly skeptical). To see why the agent never invests positive effort in an environment in which it will always be reversed note that the agent's expected payoff for investing positive effort given he will be overturned is:

$$U_A(e > 0 | r = 1) = -\beta(t_A^2 + V_F) - \kappa e - \pi.$$

The agent's expected payoff from investing no effort given it will be overturned is:

$$U_A(e = 0 | r = 1) = -\beta(t_A^2 + V_F) - \pi.$$

These combine to give the agent's net expected payoff from investing positive effort given that he will be reversed by the overseer,

$$\begin{aligned} \Delta U_A(e > 0 | r = 1) &= -\beta(t_A^2 + V_F) + \beta(t_A^2 + V_F) - \kappa e - \pi + \pi, \\ &= -\kappa e. \end{aligned}$$

Thus, if the agent invests positive effort when he will be reversed he simply pays the cost for that effort, and, therefore, optimally invests zero effort.

Overseer always upholds (perfectly deferential). When the agent is unconstrained we simply take the expected payoff for the agent given it will always be upheld.

$$\begin{aligned} u_A(e,y,r) &= -\beta(y-t_A)^2 - \kappa e - \pi r, \\ u_A(e,y,0) &= -\beta(y-t_A)^2 - \kappa e, \\ &= -\beta(x-\omega+\varepsilon-t_A)^2 - \kappa e, \\ &= -\beta(\varepsilon)^2 - \kappa e, \\ U_A(e|r=0) &= -\beta(\mathbb{E}[\varepsilon]^2 + V_\varepsilon(e)) - \kappa e, \\ &= -\beta V_\varepsilon(e) - \kappa e. \end{aligned}$$

The agent seeks to maximize $U_A(e|r=0)$ with his effort choice, which implies that the agent solves the following problem with his effort investment,

$$\hat{e}_A(\beta,\kappa) = \arg\max_e -\beta V_\varepsilon(e) - \kappa e.$$

Moreover, $\hat{e}_A(\beta,\kappa)$ exists since it is the maximum of a continuous function on a compact set and is unique so long as $V_\varepsilon(e)$ is strictly monotone.

Conditional-deference overseer. In this environment $\hat{e}_A(\beta,\kappa) < \underline{e}_R(t_A)$ so the agent is constrained by the overseer. The agent compares his expected utility from investing the threshold level of effort and being upheld by the overseer and his expected utility from investing zero effort and being overturned. These expected payoffs are given by the following expressions, respectively:

$$\begin{aligned} U_A(e=\underline{e}_R;\rho_{-A}) &= -\beta V_\varepsilon(\underline{e}_R) - \kappa \underline{e}_R, \\ U_A(e=0;\rho_{-A}) &= -\beta(t_A^2 + V_F) - \pi. \end{aligned}$$

These combine to give the net expected payoff for investing the threshold level of effort (and being upheld rather than overturned):

$$\begin{aligned} \Delta U_A(\underline{e}_R;\rho_{-A}) &= -\beta V_\varepsilon(\underline{e}_R) - \kappa \underline{e}_R + \beta(t_A^2 + V_F) + \pi, \\ &= \beta(t_A^2 + V_F - V_\varepsilon(\underline{e}_R)) - \kappa \underline{e}_R + \pi. \end{aligned} \tag{A.9}$$

Equation A.9 gives the agent's incentive compatibility condition for investing the threshold level of effort, $\underline{e}_R(t_A)$, rather than $e=0$ and being overturned. As long as this condition is weakly greater than zero the agent, in weakly undominated strategies, will invest the threshold level of effort to be upheld when constrained by the overseer. ■

Proposition 1. *Suppose the agent is authorized to make policy by the principal. Then a perfect*

Bayesian equilibrium of the agent-overseer subgame is characterized by the following collection of strategies,

1. *The agent makes effort investments according to s_A^{e*}, given by equation 6,*
2. *The agent always sets policy at his ideal point, $x^*(\omega) = \omega + t_A$,*
3. *The overseer makes review decisions according to $s_R^*(e)$, given by equation 3,*

Proof of Proposition 1. This follows from a straightforward combination of lemma 1 and lemmas 2, 3, and 4. Lemmas 3 and 4 yield number 1 in the proposition, lemma 1 yields number 2, lemma 2 yields number 3. ■

Proposition 2. *In equilibrium, agent bias affects agent effort investment if and only if oversight also affects agent effort investment.*

Proof of Proposition 2. This follows from the fact that neither agent bias t_A nor the agent's aversion to being overturned π appears in equation 4: the agent's effort investment when oversight is not effective, i.e., when the overseer will always uphold, but both agent bias and the agent's aversion to being overturned appear in the agent's effort investment given by equation 5 when the oversight is effective. ■

A.2 Principal decision-making

Lemma 5. *When the agent will always be overturned by the overseer if he is authorized by the principal, the principal empowers the agent to make policy if $c \leq 0$.*

Proof of Lemma 5. This follows from incentive compatibility for the principal to authorize the agent when he will always be overturned by the overseer. First, the principal's subjective expected payoff for not authorizing the agent is simply,

$$\begin{aligned} U_P(a=0;r=1,e^*) &= -y^2 - ca, \\ &= -\mathbb{E}[\omega]^2 - V[\omega] - c(0), \\ &= -V_F. \end{aligned}$$

Now, since if the principal authorizes the agent he will get overturned the policy payout is the same but she must incur c. So, her subjective expected payoff from authorizing an agent in this environment is given by,

$$U_P(a=1;r=1,e^*) = -V_F - c.$$

Combining these two expected payoffs yields the principal's net expected payoff for authorizing the agent when the overseer is perfectly skeptical and is given by $\Delta U_P(a=1;r=1) = U_P(a=1;r=1) - U_P(a=0)$:

$$\begin{aligned}\Delta U_P(a=1;r=1) &= -V_F - c + V_F, \\ &= -c.\end{aligned}$$

Incentive compatibility implies that the principal, given her net expected payoff for doing so, will only authorize the agent to make policy when the overseer will overturn with certainty if $\Delta U_P(a=1;r=1) \geq 0$, which requires that $-c \geq 0$. Thus, when $c > 0$ authorizing the agent is strictly dominated by *not* authorizing ($a=1$ is dominated by $a=0$). And, if $c=0$ then the principal is indifferent between $a=0$ and $a=1$ and therefore either choice (authorize or not) is optimal. ■

Lemma 6. *When the agent, if authorized to make policy, will be upheld by the overseer with certainty, the principal empowers the agent to make policy if the improvements in policy implementation, given agent effort, and authorization costs outweigh the spatial losses associated with the agent's bias:* $a^*(t_A, e^*) = 1$ *if* $t_A^2 \leq V_F - V_\varepsilon(\hat{e}_A(\beta,\kappa)) - c$.

Proof of Lemma 6. This follows from incentive compatibility for the principal to authorize the agent when the agent will always be upheld. First, the principal's subjective expected payoff when she does not authorize the agent to make policy is again,

$$U_P(a=0) = -V_F.$$

Similarly, given that the principal knows that if she authorizes the agent to make policy then $x^*(\omega) = \omega + t_A$ and $e^* = \hat{e}_A(\beta,\kappa)$, her subjective expected payoff for authorizing the agent to make policy in this environment is given by,

$$\begin{aligned}U_P(a=1;r=0,e^*=\hat{e}_A(\beta,\kappa)) &= -y^2 - ca, \\ &= -(x^*(\omega) - \omega + \varepsilon)^2 - c, \\ &= -(\omega + t_A - \omega + \varepsilon)^2 - c, \\ &= -t_A^2 - \mathbb{E}[\varepsilon|e^*]^2 - V[\varepsilon|e^*] - c, \\ &= -t_A^2 - V_\varepsilon(e^*) - c, \\ &= -t_A^2 - V_\varepsilon(\hat{e}_A(\beta,\kappa)) - c.\end{aligned}$$

Combining these expected payoffs yields the principal's net expected payoff, defined as $\Delta U_P(a=$

$1; r = 0, \hat{e}_A(\beta, \kappa)) = U_P(a = 1; r = 0, e^* = \hat{e}_A(\beta, \kappa)) - U_P(a = 0)$:

$$\Delta U_P(a = 1; r = 0, \hat{e}_A(\beta, \kappa)) \quad = \quad -t_A^2 - V_\varepsilon(\hat{e}_A(\beta, \kappa)) - c + V_F.$$

Incentive compatibility implies that the principal will authorize the agent if and only if $\Delta U_P(a = 1; r = 0, \hat{e}_A(\beta, \kappa)) \geq 0$, which requires that,

$$\begin{aligned} -t_A^2 - V_\varepsilon(\hat{e}_A(\beta, \kappa)) - c + V_F &\geq 0, \\ V_F - V_\varepsilon(\hat{e}_A(\beta, \kappa)) - c &\geq t_A^2. \end{aligned}$$

The LHS of this incentive compatibility constraint consists of the policy precision improvements induced by authorizing the agent to make policy ($V_F - V_\varepsilon(\hat{e}_A(\beta, \kappa))$) minus the authorization cost of doing so. The RHS is simply the agent's bias relative to the principal. Thus, in order for the principal to authorize the agent to make policy in this environment, the LHS must outweigh the RHS, as stated in the result. ■

Proposition 3. *If, upon being authorized to make policy, the agent will be always overturned or the overseer is perfectly deferential, the principal never benefits from a biased agent relative to an ally agent.*

Proof of Proposition 3. First, we consider the case in which the agent, if authorized, will be overturned by the overseer with certainty. In this case the principal's subjective expected utility for not authorizing is given by,

$$\begin{aligned} U_P(a = 0) &= -y^2 - ca, \\ &= -\omega^2 - c(0), \\ &= -\mathbb{E}[\omega]^2 - V[\omega], \\ &= -V_F. \end{aligned}$$

The principal's subjective expected utility for authorizing an agent given perfect skepticism is given by,

$$\begin{aligned} U_P(a = 1; r = 1) &= -y^2 - c(1), \\ &= -\omega^2 - c, \\ &= -V_F - c. \end{aligned}$$

This implies that the principal's net expected utility for authorizing the agent to make policy when he will face a perfectly skeptical overseer, defined as $\Delta U_P(a = 1; r = 1) = U_P(a = 1; r = 1) - U_P(a = 0)$,

is,

$$\begin{aligned}\Delta U_P(a=1;r=1) &= -V_F - c + V_F, \\ &= -c.\end{aligned}$$

Thus, the overseer never authorizes the agent when $c > 0$ (but may if $c = 0$ depending on assumptions made on principal indifference). Since t_A does not appear in the principal's net expected utility, even if the principal benefits from empowering the agent she does not benefit, nor is she hurt by, a biased versus an ally agent. This is because since the overseer always reverses the agent the policy-related component of her payoff is invariant to agent type.

Now we consider the second case: authorizing an agent that will be upheld with certainty (e.g., perfectly deferential overseer). Recall that if authorized, since the overseer will always uphold regardless of agent effort investment, the agent invests effort $\hat{e}_A(\beta,\kappa)$. The principal's expected payoff from authorizing the agent is given by,

$$\begin{aligned}U_P(a=1;r=0,\hat{e}_A(\beta,\kappa)) &= -y^2 - c(1), \\ &= -(x^*(\omega) - \omega + \varepsilon)^2 - c, \\ &= -(t_A + \varepsilon)^2 - c, \\ &= -t_A^2 - \mathbb{E}[\varepsilon|\hat{e}_A(\beta,\kappa)]^2 - V[\varepsilon|\hat{e}_A(\beta,\kappa)] - c, \\ &= -t_A^2 - V_\varepsilon(\hat{e}_A(\beta,\kappa)) - c.\end{aligned}$$

Combining this with the payoff for not authorizing, $-V_F$, yields the principal's net expected payoff for authorizing an agent in this environment, defined as $\Delta U_P(a=1;r=0,\hat{e}_A(\beta,\kappa)) = U_P(a=1;r=1,\hat{e}_A(\beta,\kappa)) - U_P(a=0)$:

$$\Delta U_P(a=1;r=0,\hat{e}_A(\beta,\kappa)) = -t_A^2 - V_\varepsilon(\hat{e}_A(\beta,\kappa)) - c + V_F.$$

This implies that the principal benefits from empowering the agent if $t_A^2 \leq V_F - V_\varepsilon(\hat{e}_A(\beta,\kappa)) - c$. In this case it is clear that the principal's payoff is decreasing in $|t_A|$. The larger is $|t_A|$ the less likely it is that the principal will benefit from authorizing the agent at all and the lower is her utility for doing so. Thus, in this case, the principal is strictly better off with an ally agent, $t_A = 0$, than a biased agent, $|t_A| \neq 0$. ■

Lemma 7. *Suppose the agent, if authorized, will invest effort sufficient to be upheld by the overseer. Further, suppose agent and overseer ideal points are organized such that the agent faces a conditional-deference overseer. Then the principal authorizes the agent to make policy if* $-2t_A t_R \geq c$.

Proof of Lemma 7. This follows from incentive compatibility for the principal to authorize the agent

to make policy when facing a conditional-deference overseer. Note first that we are assuming that if the principal authorizes the agent to make policy then the agent will invest effort equal to the overseer's threshold $\underline{e}_R(t_A)$ and be upheld. The principal's subjective expected payoff if she does not authorize the agent is again,

$$U_P(a=0) = -V_F.$$

Now, given that the principal knows that if she authorizes the agent to make policy the agent will invest effort $e^* = \underline{e}_R(t_A)$, her subjective expected payoff for authorizing the agent is given by,

$$\begin{aligned} U_P(a=1; r=0, e^* = \underline{e}_R(t_A)) &= -y^2 - ca, \\ &= -(x^*(\omega) - \omega + \varepsilon)^2 - c, \\ &= -(\omega + t_A - \omega + \varepsilon)^2 - c, \\ &= -t_A^2 - \mathbb{E}[\varepsilon|e^*] - V[\varepsilon|e^*] - c, \\ &= -t_A^2 - V_\varepsilon(e^*) - c, \\ &= -t_A^2 - V_\varepsilon(\underline{e}_R(t_A)) - c. \end{aligned}$$

Define the principal's net expected payoff from authorizing the agent as $\Delta U_P(a=1; r=0, \underline{e}_R(t_A)) = U_P(a=1; r=0, e^* = \underline{e}_R(t_A)) - U_P(a=0)$:

$$\Delta U_P(a=1; r=0, \underline{e}_R(t_A)) = -t_A^2 - V_\varepsilon(\underline{e}_R(t_A)) - c + V_F.$$

Incentive compatibility implies that the principal will authorize the agent to make policy if $\Delta U_P(a=1; r=0, \underline{e}_R(t_A)) \geq 0$, which requires that,

$$-t_A^2 - V_\varepsilon(\underline{e}_R(t_A)) - c + V_F \geq 0.$$

Now, solving the overseer's incentive condition to uphold (i.e., solving for $V_\varepsilon(e)$) with equality allows us to substitute for $V_\varepsilon(\underline{e}_R(t_A))$:

$$\begin{aligned} -t_A^2 - \left[V_F - t_A^2 + 2t_A t_R\right] - c + V_F &\geq 0, \\ -t_A^2 - V_F + t_A^2 - 2t_A t_R - c + V_F &\geq 0, \\ -2t_A t_R - c &\geq 0, \\ -2t_A t_R &\geq c. \end{aligned}$$

Thus, the principal will authorize the agent to make policy when the overseer is employing conditional-deference and the agent will invest the threshold level of effort required to be upheld if $-2t_A t_R \geq c$, as stated in the result. ■

Proposition 4. *Suppose the agent, if authorized to make policy, faces a conditional-deference overseer. Further, suppose the agent will invest the threshold level of required effort to be upheld. Then the principal benefits from a biased agent anytime she benefits from authorizing the agent to make policy.*

Proof of Proposition 4. By supposition we are in an environment in which the principal will authorize the agent to make policy if $-2t_A t_R \geq c$ (the incentive compatibility constraint to authorize the agent given conditional-deference oversight). Assume that this incentive compatibility constraint holds so that the principal benefits from authorizing the agent. We show that the principal benefits from a biased agent for each cost level.

Case 1: $c > 0$**.** When authorization costs are positive incentive compatibility requires that t_A and t_R be oppositely signed. That is, for $-2t_A t_R \geq c$ to be true when $c > 0$ it must be the case that if $t_R < 0$ (as is assumed) then $t_A > 0$ so that the LHS is positive (and of course, the assumption that the principal benefits from authorization implies that the magnitude of $-2t_A t_R$ is larger than c). Now, the principal's expected utility in this case is given by $U_P(a = 1; r = 0; \underline{e}_R(t_A)) = -2t_A t_R - c$. Since the incentive compatibility constraint implies that $t_A > 0$ and $-2t_A t_R > c$ the principal's utility is increasing in t_A. The larger is t_A the higher is the principal's utility up until the point at which the agent becomes too biased and invests no effort and the overseer overturns. Thus, the principal benefits from a biased agent anytime she benefits from authorization when $c > 0$.

Case 2: $c = 0$**.** In this case the relevant incentive compatibility constraint for the principal to authorize the agent is $-2t_A t_R \geq 0$. This implies that, since $t_R < 0$ by assumption, the principal will still authorize an ally agent, $t_A = 0$. However, the same argument in Case 1 applies here: the principal's utility is increasing in $t_A > 0$ provided that the posited subgame agent-overseer behavior does not break down. Thus, while the principal would still benefit from authorizing an ally agent, she is better off with a positively biased agent when $c = 0$.

Taken together, these cases imply that when the principal benefits from authorizing an agent when that agent will face a conditional-deference overseer, the agent will invest the threshold level of effort required to be upheld, and the overseer will uphold, she also benefits from an agent biased away from her ideal point. ■