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

Ian R. Turner†

January 24, 2019

Article forthcoming at
Journal of Law, Economics, & Organization

Abstract

We develop a theory of policymaking between an agent and an overseer, with a principal whose welfare is affected by agent-overseer interactions. The agent can increase the quality of policy outcomes through costly capacity investments. Oversight and agent bias jointly determine optimal agent capacity investments. We show that when oversight improves agent investment incentives the principal always benefits from an agent with biases opposite the overseer. Competing agent-overseer biases translate into higher quality policy outcomes than the principal could induce were she monitoring the agent. Effective oversight is necessary for these incentive effects. The results imply that political principals ought to consider the nature of the broader policymaking environment when appointing agents to make policy on their behalf and when designing managerial strategies aimed at motivating agents.

Keywords: Policymaking; Bureaucracy; Oversight; Political agency; Formal theory
JEL: D73; D82; H11

*I would like to especially thank the anonymous reviewers and editor Kevin Quinn, as well as John Patty, Maggie Penn, Randy Calvert, Justin Fox, Gary Miller, and Keith Schnakenberg for many insightful conversations regarding this work. I would also like to thank Jay Krehbiel, Dalston Ward, Richard Van Weelden, Brian Rogers, Caitlin Ainsley, Stephane Wolton, Rachel Augustine Potter, Ryan Hübert, and Andrea Aldrich for providing exceptionally insightful commentary at different stages. All errors are mine. Early work on this project was supported by NSF Grant DGE-1143954.

†Assistant Professor, Department of Political Science, Yale University, Email: ian.turner@yale.edu.

The vast majority of public policy is developed and implemented by bureaucratic agencies whose authority to do so was delegated from a political principal.[1] Delegation, however, introduces the potential for political agency problems: the ability of agents to subvert the wishes of their principal(s) and pursue their own goals. Agents to whom authority has been delegated may be biased and pursue policy outcomes that diverge from those preferred by the principal, thereby subverting the principal's wishes (Gailmard 2002). In addition, the agent may invest insufficient effort toward the production of high quality policy outcomes, sometimes referred to as slack (Bueno de Mesquita and Stephenson 2007; Turner 2017*b*). Both possibilities can lead to low quality policy outcomes from the principal's perspective. In short, bureaucracies are tasked with not only crafting the substance of policy, which allows for the introduction of bias, but also investing in capacity to effectively implement policy in practice, which introduces the potential for slacking.[2]

Two commonly proposed solutions to these agency problems are extensive oversight and delegating to allies. The logic underlying the ally principle is straightforward and has been shown to hold in diverse environments (Bendor and Meirowitz 2004).[3] All else equal, a principal prefers an agent with preferences closer to her own because the agent is then more likely to take actions in line with the principal's interests regardless of expertise advantages (Gailmard and Patty 2013*a*, p.4-5). However, the situation becomes more complicated when the agent must also invest in capacity that improves the overall quality of outcomes. The principal, in this case, must also consider the provision of incentives to spur this investment, which can be supplied through oversight.

In the United States federal government perhaps the most famous example of this oversight

[1]The advantages of delegation include the ability to exploit the superior expertise contained in agencies (Epstein and O'Halloran 1994; Spence and Cross 2000), and the provision of incentives to specialize (Gilligan and Krehbiel 1987), gather information (Gailmard and Patty 2013*b*), and improve efficiency or reliability (Ting 2002, 2003). See Bendor, Glazer and Hammond (2001) for a comprehensive overview of theories of delegation, and Gailmard and Patty (2013*a*) for a treatment specific to bureaucratic politics.

[2]Of course, capacity within agencies can refer to many things including manpower, procedural development, technology to aid in processing, and the like. We use the term capacity as a catch-all term to capture the idea that being able to implement policy effectively — accurately process permits, conduct inspections, enforce compliance — requires investment in the ability to execute on the ground. Specifically, higher capacity leads to more precise implementation. This is closely related to what Carpenter (2001) calls "programmatic capacity," and can also be viewed through the lens of "street-level bureaucracy" (Lipsky 1980).

[3]For example, the principal prefers to choose a "clone" of herself as her agent (Gailmard and Patty 2013*a*) and grants more discretion to an ally agent (Epstein and O'Halloran 1994) when she wants to induce expertise acquisition (Gailmard and Patty 2007), or when uncertainty is high (Bendor, Glazer and Hammond 2001; Moe 2012).

is judicial review of bureaucratic agencies. Judicial review provisions are written into authorizing legislation when an agency is empowered to take action.[4] Essentially, the overseer is empowered to reverse (or, veto) agency policy actions. When a bureaucratic agent makes policy he is also often subjected to subsequent review, and possible rejection, of his actions by another political institution. Overseers such as courts also often have preferences that diverge from those of the principal, further compounding the political agency problems inherent in the policy process. This raises the question at the heart of this article: Are there environments in which political principals benefit from biased agents making policy on their behalf given that they make policy in the shadow of oversight?

In this paper we develop a political-institutional theory of policymaking and show that if both bias and slack are potential problems, the potential solutions – ally agents and extensive oversight – interact in unexpected ways. Specifically, we show that when oversight impacts agent capacity investment incentives the principal can benefit from agents with preferences *further* from her own. Oversight and agent bias interact in such a way that neither strengthen investment incentives unless both do simultaneously. This has upstream effects on whether the principal prefers a policymaking agent with preferences closer or further from her own.

The ally principle is weakened whenever oversight is effective at providing positive investment incentives for the agent due to the fact that the agent sets the substance of policy *and* invests in policy-improving capacity. Crucial to the theory is the fact that this insight only holds because of the intervening influence of oversight. When oversight has no impact on agent capacity incentives the principal never benefits from agent bias. She would be better off with an agent with preferences closer to her own in those situations. Otherwise, when oversight is effective at providing incentives, agent bias serves two interrelated purposes that increase the agent's capacity investments in equilibrium. First, it intensifies the agent's own motivations to acquire more capacity to ensure that his policies are realized. Second, it increases the stringency of oversight, which

[4]Shipan (1997) provides a comprehensive study of the politics surrounding the choice of these provisions (see also McCann, Shipan and Wang 2016). Additionally, the Administrative Procedure Act (APA) directs courts to engage in so-called hard look review of agency regulations and overturn actions found to "arbitrary and capricious" (Breyer 1986).

in turn increases the capacity the agent must develop in order to have his preferred policy realized. The principal can benefit from this dynamic. Agent-overseer biases induce higher levels of capacity investment than would be possible if there were no effective oversight or if the principal engaged in oversight herself. Thus, while there are situations where the principal prefers agents closer to her – in line with the ally principle – there are also environments in which she benefits more as agent preferences move further from her own – in contrast to the ally principle. This latter dynamic only obtains when oversight is an effective institutional check on agent behavior. Thus, agent bias is instrumentally valuable to the principal in politicized (i.e., ideologically contentious) policymaking environments.

We add to existing literature by shifting attention away from the role that oversight plays in constraining agents' substantive policy choices and toward how oversight structures agent incentives to improve the implementation of policy. Previous work has focused heavily on oversight's impact on disciplining the substance of agent policy choices to be more closely aligned with their principal (e.g., Epstein and O'Halloran 1999; Patty and Turner 2017; Shipan 1997).[5] In contrast, our argument tracks the empirical reality that many oversight institutions like courts have moved heavily toward procedural, rather than substantive, review of administration actions in recent times (Kagan 2001). Specifically, courts are often most concerned with the way(s) in which agencies implement or apply policy, rather than the specific content of the policies being challenged.

Concrete empirical motivations for this include the Federal Emergency Management Agency (FEMA) being taken to court based on the processes in place to allocate housing assistance following Hurricanes Katrina and Rita (see *ACORN v. FEMA*, 463 F. Supp. 2d 26 (D.D.C. 2006)) and the Social Security Administration being similarly challenged in the 1960s and 70s due to the ways in which aid was terminated for families with dependent children (see *Goldberg v. Kelly*, 392

[5]More generally, previous work has highlighted how oversight impacts incentives for politicians to pander in electoral settings (Fox and Stephenson 2011) or to acquire information (Dragu and Board 2015), and induces more ideologically desirable policy (Wiseman 2009), as well as the invaluable insight provided by previous work examining signaling dynamics and their effect on substantive policy choice (see e.g., Boehmke, Gailmard and Patty 2006; Carpenter and Ting 2007; Gailmard and Patty 2007, 2013*a*,*b*,*c*; Gilligan and Krehbiel 1987, 1989; Gordon and Hafer 2005; Patty 2009; Stephenson 2006; Ting 2008). Finally, see Bueno de Mesquita and Stephenson (2007), Stephenson (2006), and Turner (2017*b*) for recent work that highlights the ways in which oversight impacts effort incentives.

U.S. 254 (1970) and Derthick (1990), specifically 132-135). In the former case the agency had developed a computer program to process applications that led to erroneous housing assistance decisions. In the latter case the agency had failed to develop procedures allowing for individuals whose public assistance benefits were terminated to adequately appeal the agency's decision. In both cases, the overseer ruled that the agencies' interests in keeping administrative costs low were not sufficient to excuse the existing (lack of) capacity the agencies had in place, and the right to due process was of central concern rather than the substantive content of policy itself (i.e., assistance standards). These examples highlight the importance of agencies developing capacity to implement policy effectively on the ground, which is the central focus of our argument.[6]

To that end, we develop a theory that holds agent bias fixed. We structure the model so that the agent invests in capacity to reduce outcome uncertainty prior to learning the nature of the policy environment.[7] Then the agent learns about the environment and sets the (possibly biased) substance of policy spatially. The overseer, meanwhile, only reviews the agent's investment decision to ensure he has sufficient capacity to produce high quality policy in practice. Substantively, this turns the focus to the effects of oversight on the incentives for agencies to develop effective procedures to more accurately reach permitting and licensing decisions, properly and effectively allocate public assistance and government benefits, conduct adequate inspections and improve enforcement, and generally implement policy well. This setup also implies that the agent in our model has full discretion to set the substance of policy, which depends on his bias in equilibrium.

More generally, the model speaks to policy areas in which overseers (e.g., courts) are either explicitly directed to ignore substantive content and instead focus on the agent's ability to effectively implement policy on the ground or policy areas in which agents are simply not engaged in much technical policymaking relative to implementation or enforcement activities (e.g., FEMA). Similarly, the model also captures environments in which the agent's discretion to control content

[6]Similarly, Huber (2007) provides a comprehensive study of the Occupational Health and Safety Administration (OSHA), part of which highlights areas where the lack of agency capacity for on site inspections led to low quality regulatory enforcement.

[7]The way in which capacity investments improve the precision of policy realizations is similar to the set-up in Huber and McCarty (2004).

of policy is non-controversial, which elevates equal protection and due process related concerns with implementation and/or enforcement (e.g., FEMA, OSHA). Finally, this investment can also be understood as the agent's ex ante investment in understanding how different policy instruments map into likely outcomes in his policy area – a slightly different understanding of capacity – which is known to be an important factor in judicial and executive review processes (Gailmard and Patty 2017).[8] Overall, our analysis complements existing work focused on ameliorating the potential bias in the content of policy by directly addressing the relatively understudied effect of oversight institutions on the incentives for agents to develop capacity that leads to higher quality policy outcomes. Indeed, in our model bias can prove to be instrumentally valuable for the principal precisely because it can be coupled effectively with oversight institutions to strengthen agent investment incentives.[9]

This latter point also speaks to literature examining optimal agent bias. Bendor and Meirowitz (2004) show that principals may prefer biased agents if they are willing to work harder or have some type of beneficial valence that is correlated with their bias and benefits the principal. We provide a distinctly political-institutional rationale for why biased agents can benefit principals concerned with motivating desirable investments in valence. The theory developed here does not assume that biased agents are *per se* better equipped to produce high quality policy. Rather, agent bias only motivates capacity investments if effective oversight is present. Our results most closely resemble studies that show that agent bias is useful to incentivize specialization (Gilligan and Krehbiel 1987, 1989) or generate bargaining power (Gailmard and Hammond 2011).[10]

[8]While full analysis of an alternative information structure in which the overseer judges policy content is beyond the scope of this article, two other pieces of research address this question in similar political environments. Patty and Turner (2017) provide formal treatment of agent incentives when policy content is reviewed and Turner (2017*a*) compares two models by varying whether or not policy content is observed during review.

[9]This insight also complements recent work on competitive policy development (Hirsch and Shotts 2015, Forthcoming) and some of the issues studied in Huber and McCarty (2004) and Ting (2011).

[10]More generally, several previous studies have shown that principals may 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 for information disclosure (Dessein 2002), to incentivize costly investment in policy development (Hirsch and Shotts 2015), and to reduce rent-seeking by elected politicians (Van Weelden 2013). We provide results that are similar in that they also show why political principals may prefer a biased agent. However, we diverge from previous work by analyzing an environment in which the institution of oversight is a necessary condition for biased agency to be beneficial.

For instance, Gailmard and Hammond (2011) argue that the House of Representatives creates biased committees to increase House bargaining power relative to the Senate. The authors write that, "an unrepresentative committee is a veto constraint for the other chamber..." (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 in more capacity to satisfy the overseer. The more divergent agent and overseer ideal points are, within reasonable limits, the more this dynamic intensifies these incentives. While the logic between our theory and this body of work are related, we extend and complement it. First, as noted above, we incorporate both the substantive setting of policy and investments to improve outcomes in one framework. Both are key to our results. Second, in our theory the presence of effective oversight is a necessary condition for agent bias to strengthen investment incentives.

Overall, the theory developed in this paper provides novel insight into the institutional and policy environments in which biased policymaking can benefit political principals. When oversight is an effective tool for political control, the principal benefits from a biased agent who will subsequently face a biased overseer in the policymaking game. That is, political principals derive instrumental value from the dual usage of oversight and agent bias as institutional motivators when bias and slack are both concerns.

1 The model

We study a non-cooperative policymaking game involving a principal (P), an agent (A), and an overseer (R). The principal is a non-strategic player whose welfare is affected by agent-overseer interactions.[11] The agent possesses private policy-relevant information and is directed to make policy. The overseer is empowered to either uphold or reverse agent-made policy. If the overseer upholds the agent then agent-made policy is realized and if the overseer overturns the agent then an unregulated, status quo, outcome is realized. This represents an environment in which the agent

[11] In an extension below we also allow for endogenous delegation.

has been previously empowered to make policy on behalf of the principal, either by herself or a past principal, but the interactions between the agent and overseer do affect principal welfare. For instance, the EPA develops and implements regulations based on authority previously authorized by the Clean Air Act. Those actions are sometimes challenged and reviewed by courts, and the outcomes of those interactions affect the utility of the current political principal (e.g., Congress or the president).

At the beginning of the game Nature draws the *state of the world*, $\omega \in \Omega = \mathbb{R}$, that is distributed according to F which has mean zero and variance $0 < V_F < \infty$. This state variable captures the contingencies of the policy environment, which can be understood as outcomes that will arise from private interactions between individuals and firms without further agent intervention (i.e., the status quo). The characteristics of F are common knowledge but ω is only observed by the agent.

Prior to learning ω, the agent makes an ex ante investment in implementation capacity, $e \in [0,1]$. This investment directly affects an *implementation shock*, $\varepsilon \in \mathbb{R}$, that is distributed according to $G_\varepsilon(e)$ with mean zero and variance $0 < V_\varepsilon(e) < \infty$. Specifically, the variance of $G_\varepsilon(e)$ is continuously strictly decreasing and convex in e. This ensures that $V_\varepsilon(e) < V_\varepsilon(e')$ if and only if $e > e'$.[12] The likely magnitude of the implementation shock is decreasing in agent capacity investments. The characteristics of $G_\varepsilon(e)$ are common knowledge but no player observes ε directly. It is drawn by Nature following review and affects outcomes by potentially shifting policies away from their intended spatial location.

Once the agent has made his investment he learns about the policy environment by observing ω. In this sense, the agent is an expert. Upon observing ω the agent sets a substantive policy target, denoted by $x \in X = \mathbb{R}$. This is a target since the implementation shock may ultimately lead to agent-made outcomes that diverge from x.[13] Thus, both *biased choices* and *insufficient investment* can lead to inefficient agent-made policy outcomes. Insufficient ex ante capacity investment can produce poor outcomes by increasing $V_\varepsilon(e)$ *even when the agent targets the principal's pre-*

[12] That is, for all $e > e'$, $G_\varepsilon(e)$ second-order stochastically dominates $G_\varepsilon(e')$.

[13] This capacity investment set-up is reminiscent of the way that Huber and McCarty (2004) models capacity.

ferred policy. Similarly, agent bias can lead to the substantive content of policy being distorted away from the principal's ideal, which negatively impacts principal welfare *even when the agent has invested maximally in capacity*. For a principal concerned with policy outcomes matching the true state, both bias and insufficient capacity are omnipresent concerns.

Following the agent's choices, the overseer reviews the agent by observing the agent's capacity investment, denoted by $r(e) \in \{0,1\}$. This represents a form of procedural review, which, for instance, has increased in judicial oversight in recent years (Kagan 2001; Stephenson 2006).[14] If the overseer reverses the agent ($r = 1$) then ω obtains, the game ends, and payoffs are realized. If instead the overseer grants the agent deference ($r = 0$) then agent-made policy is implemented, the game ends, and payoffs are realized. Accordingly, final policy outcomes, y, are given by,

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

Each players' induced preferences over policy depend on their respective "type" or ideal point, denoted by $t_i \in \mathbb{R}$, $i \in \{P,R,A\}$. Each players' ideal point dictates their welfare-maximizing policy outcome relative to ω. 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. The main results pertain to the principal's welfare given agent-overseer interactions. 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 and policymaking incentives vary as t_A varies relative to the other players' ideal points. The payoffs of the principal, the overseer, and the agent are given by the following expressions, respectively:

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

[14] As noted above, Patty and Turner (2017) and Turner (2017*a*) explore oversight in which x is observable.

As noted above, the principal wants outcomes that match ω. The overseer seeks to minimize the distance between realized policy y and its ideal point t_R. The agent also desires policy outcomes to be realized as close as possible to his ideal point t_A, but his policy motivations, relative to the other components of his utility and the motivations of the other players, is captured by $\beta > 0$. Agent policy motivations increase in β.[15] All else equal, all players prefer more effective implementation generated through increased capacity, but only the agent bears the costs of that investment, denoted by $\kappa > 0$. This cost captures intuitive concepts of building bureaucratic capacity like increased staffing, investing time and resources toward streamlining procedures, or expanding enforcement programs (Huber 2007). Finally, the agent is also averse to being reversed by the overseer, captured by $\pi > 0$. The agent becomes more averse to being overturned as π increases. This cost captures intuitive, realistic concepts based on general career concerns such as reputational losses, budgetary considerations, or the like. The parameters are exogeneous and common knowledge.

We utilize perfect Bayesian equilibrium (PBE) in weakly undominated strategies. A PBE, which we denote with ρ, is a complete profile of strategies and beliefs such that all players are maximizing their subjective expected payoffs given other players' strategies and, when applicable, beliefs are consistent with Bayes's rule. In the analysis we will sometimes use the notation ρ_{-i} to denote the set of equilibrium strategies and beliefs for all players except player i.

2 Oversight, bias, and capacity

In this section we analyze the interactions between the agent and the overseer. To begin, in equilibrium the agent will always set substantive policy at his ideal point: $x^*(\omega) = \omega + t_A$. This strategy is weakly dominant for the agent, independent of his capacity investment and the overseer's oversight strategy, because the overseer does not observe x directly. This feature of the equilibrium can be thought of as the agent making *sincere* policy choices (from his point of view). It also isolates the

[15] These policy motivations 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).

effects of oversight on agent capacity investment incentives and how the interaction of agent bias and investment incentives interact to affect principal welfare.

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. Courts, executive reviewers, and intra-agency veto points can often only accept or reject policies rather than supplant them with their own policy (Bueno de Mesquita and Stephenson 2007). The overseer, upon observation of the agent's capacity e, 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,[16]

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

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

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

Equation 2 provides an intuitive condition that must be met for the overseer to uphold the agent. The agent must invest enough in capacity to improve the precision of policy outcomes, relative to the volatility of the underlying policy environment, to offset any spatial policy losses incurred by his bias. The more capacity the agent develops to improve outcomes 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 and the more stringent oversight becomes. However, the agent making

[16]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 $\Delta U_i(a; \rho_{-i}) \equiv U_i(a; \rho_{-i}) - U_i(b; \rho_{-i})$ to represent the net expected payoff for player i taking action a instead of action b given the expected behavior of the other players, $-i$, in equilibrium, ρ_{-i}.

policy becomes more important the more volatile the underlying policy environment becomes. This highlights a commitment problem for the overseer. The more the agent is needed to regulate, the less demanding oversight is with respect to capacity investments.

Since the precision of policy outcomes is strictly increasing in agent capacity investments, the overseer's equilibrium strategy is equivalent to an investment threshold. Denote this threshold as $e_R^{\min}(t_A) \coloneqq e$ such that $V_F - V_\varepsilon(e) = t_A^2 - 2t_A t_R$.[17] This threshold is the minimum level of capacity investment an agent must make in order to be upheld by the overseer, given his bias. This yields the following equilibrium oversight strategy,

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

The impact of oversight on agent capacity investments depends crucially on the agent's bias relative to the overseer, which leads to three regimes of review. There are two extreme cases: when the agent has relatively low bias and extremely high bias relative to the overseer. The most interesting case is when the agent is intermediately biased away from the overseer.

Low bias. When agent-overseer ideal points are relatively aligned oversight does not affect the agent's investment decision because he will never be overturned. In this environment we say that the overseer is perfectly deferential to the agent. This is the case any time spatial policy losses are offset even when the agent invests nothing in capacity. 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. 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 moderate agent. This reveals a pathological limitation of oversight in this model: if the agent is *not biased enough* then oversight plays no effective role in the provision of agent investment incentives.

[17] A capacity investment that solves equation 2 with equality may not always be feasible. In cases where there is no $e \in [0,1]$ that solves equation 2, the overseer either always overturns or always upholds the agent. We discuss these scenarios in greater detail below. For the remainder of the analysis we focus on the more interesting cases in which the agent can invest in capacity to satisfy the overseer's threshold as defined.

It may seem intuitive that in response to perfect deference the agent never invests in capacity since he will be upheld regardless. However, the agent is intrinsically motivated to improve outcomes. While oversight does not impact capacity investments in this case, the agent's own motivations do. Since the overseer will never overturn the agent, the agent makes capacity investments based solely on his own motivations. Denote this choice by,[18]

$$e_A^u(\beta, \kappa) \in \arg\max_e -\beta V_\varepsilon(e) - \kappa e. \tag{4}$$

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 capacity investment is greater than the overseer's threshold level of acceptable capacity investment: $e_A^u(\beta, \kappa) \geq e_R^{\min}(t_A)$. Oversight is not stringent enough to bind the agent's investment decision. Intuitively, the agent's investment in this case is increasing in his intrinsic policy motivations, β, and decreasing in costs, κ.

Extreme bias. On the other extreme, if the agent is too biased then the overseer will never uphold the agent, regardless of 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 capacity investment, $e = 1$, to improve implementation quality, he cannot offset spatial policy losses. 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 increasing in the volatility of unregulated outcomes, V_F. The more an agent is needed to improve policy outcomes, the more biased he can be before the overseer becomes perfectly skeptical.

In this case the agent responds by never making positive capacity investments. If an agent with this level of bias makes any positive capacity 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 in capacity.

[18] We use the superscript, u, to denote the agent's 'unconstrained' (by oversight) investment: $e_A^u(\beta, \kappa)$.

Intermediate bias. The final, most interesting, environment is one in which the agent's capacity 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 capacity investment based on his own motivations is not sufficient to satisfy the overseer's threshold: $e_A^u(\beta,\kappa) < e_R^{\min}(t_A)$.

Accordingly, the agent responds by deciding if he is better off making the threshold capacity investment required to be upheld or making no capacity investment and being overturned.[19] Consider the agent's net expected payoff for a capacity investment sufficient to be upheld,

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

Incentive compatibility implies that the agent will make a capacity investment sufficient to be upheld if $\Delta U_A(e \geq e_R^{\min}(t_A); r^*(e) = 0) \geq 0$. The likelihood that this condition is satisfied is increasing in the agent's policy motivations β, his reversal aversion π, and his bias t_A, and is decreasing in effort costs κ and policy imprecision given capacity investment, $V_\varepsilon(e)$. Solving the agent's incentive compatibility condition for e so that it holds with equality yields the maximum level of capacity investment the agent is willing to make to be upheld when facing a conditional-deference overseer, which we label $e_A^{\max}(t_A) := e$ such that $\Delta U_A(e \geq e_R^{\min}(t_A); r^*(e) = 0) = 0$.[20] That is, $e_A^{\max}(t_A)$ denotes the maximal level of capacity investment the agent would be willing to make to avoid being reversed by a conditional-deference overseer.

If the maximum level of capacity investment the agent is willing to make to be upheld, $e_A^{\max}(t_A)$, exceeds the threshold required by the overseer, $e_R^{\min}(t_A)$, then the agent invests at the threshold to be upheld. If instead $e_A^{\max}(t_A) < e_R^{\min}(t_A)$ then the agent invests nothing in capacity and accepts being overturned. Thus, when facing conditional-deference oversight $\left(e_R^{\min}(t_A) > e_A^u(\beta,\kappa)\right)$ the agent will make a capacity investment equal to the overseer's threshold if and only if $e_A^{\max}(t_A) \geq$

[19] Note that if it is not incentive compatible for the agent to invest the threshold level to be upheld then he makes zero capacity 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.

[20] Formal details can be found in the appendix, Lemma 3.

Notation	Description
e_A^*	Equilibrium agent capacity investment
$e_A^{\max}(t_A)$	Maximum level of capacity the agent would ever invest in to be upheld, conditional on agent bias
$e_A^u(\beta, \kappa)$	Agent capacity when unconstrained by oversight, conditional on agent policy motivations and investment costs
$e_R^{\min}(t_A)$	Minimum level of capacity required for overseer to uphold, conditional on agent bias

Table 1: Summary of capacity notation

$e_R^{\min}(t_A)$, and invests nothing otherwise.

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

$$e_A^* = \begin{cases} e_A^u(\beta, \kappa) & \text{if } e_A^u(\beta, \kappa) \geq e_R^{\min}(t_A), \\ e_R^{\min}(t_A) & \text{if } e_A^u(\beta, \kappa) < e_R^{\min}(t_A) \text{ and } e_A^{\max}(t_A) \geq e_R^{\min}(t_A), \\ 0 & \text{if } e_A^u(\beta, \kappa) < e_R^{\min}(t_A) \text{ and } e_A^{\max}(t_A) < e_R^{\min}(t_A), \end{cases} \tag{5}$$

where $e_R^{\min}(t_A)$ is defined as e such that equation 2 holds with equality, $e_A^u(\beta, \kappa)$ is implicitly defined by equation 4, and $e_A^{\max}(t_A)$ is the maximum level of effort the agent would ever be willing to invest to avoid reversal. Table 1 provides a summary of capacity investment notation for reference.

There are a few aspects of the agent's equilibrium capacity investment strategy worth noting further. First, notice that the presence of an overseer can induce higher levels of capacity investment from the agent than if there were no oversight. This is the second case of e_A^* in which $e_A^u(\beta, \kappa) < e_R^{\min}(t_A)$ and $e_A^{\max}(t_A) \geq e_R^{\min}(t_A)$. Second, the overseer can also induce the agent to invest less than he would otherwise. This is the third case of e_A^* in which $e_A^u(\beta, \kappa) < e_R^{\min}(t_A)$ and $e_A^{\max}(t_A) < e^{\min}(t_A)$. In this case the overseer provides a "deterrence effect" for the agent. Since capacity investments are costly, the agent is deterred from investing anything because the overseer

Review regime	**Perfect deference**	**Conditional-deference**	**Perfect skepticism**
Equilibrium agent investment: e_A^*	$e_A^u(\beta,\kappa)$	$e_R^{\min}(t_A)$ if $e_A^{\max}(t_A) \geq e_R^{\min}(t_A)$, 0 if $e_A^{\max}(t_A) < e_R^{\min}(t_A)$	0
Effect of bias t_A *on* e_A^*	no effect	increasing in t_A if $e_A^* = e_R^{\min}(t_A)$, no effect if $e_A^* = 0$	no effect

Table 2: Summary of equilibrium agent investment conditional on review regime

Note: *Equilibrium agent investment* displays each e_A^* for the given review regime. The *effect of bias* lists how increasing t_A affects $e_A^*(t_A)$ conditional on being in a given review regime.

will not allow outcomes to turn out worse than the reversion level of policy precision (V_F), which in this case is not *bad enough* to induce the agent to invest more.[21] Combining all of these cases yields our first result.

Proposition 1. *In equilibrium, (1) the agent makes capacity investments according to* e_A^**, given by equation 5; (2) the agent always sets policy at his ideal point,* $x^*(\omega) = \omega + t_A$*; and (3) the overseer makes review decisions according to* $s_R^*(e)$*, given by equation 3.*

Table 2 summarizes equilibrium agent investments conditional on the nature of oversight. As the agent becomes more biased relative to the overseer equilibrium investments are fixed at $e_A^u(\beta,\kappa)$ until $e_R^{\min}(t_A) > e_A^u(\beta,\kappa)$. At that point oversight becomes more demanding and the agent now invests $e_R^{\min}(t_A)$ to be upheld and investments increase in $|t_A|$ up until $e_A^{\max}(t_A) > e_R^{\min}(t_A)$. Past that point the agent invests zero because the overseer always reverses. The key insight is that equilibrium agent capacity only increases as a function of agent bias if oversight also affects investment.

Proposition 2. *In equilibrium, agent bias strengthens agent capacity investment incentives if and only if oversight also strengthens these incentives.*

Proposition 2 presents a central result for our theory. When oversight does not effectively

[21] This deterrence effect is qualitatively similar to the "bail out effect" provided by judicial review identified in previous theoretical work (e.g., Bueno de Mesquita and Stephenson 2007; Fox and Stephenson 2011; Turner 2017*b*).

strengthen agent investment incentives neither does agent bias. When the agent is not too biased he simply invests 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, capacity investments are also invariant. They are always zero since the agent will always be overturned. In this case investment incentives are *weakened* and the agent is deterred from developing any capacity. However, in the intermediate range of agent bias, capacity 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 induces higher capacity investments if and only if oversight does also.

This illustrates a fundamental interdependence between utilizing tools like appointing biased agents (or, "zealots") to direct agencies (Gailmard and Patty 2007) and institutionalized oversight to impact capacity investment incentives. One is not effective without the other. Based on the dynamics characterized in this section, this raises the question: When does the principal benefit from a biased agent? That is, under what circumstances does the principal benefit from agents with preferences more extreme than her own?

3 The instrumental value of politicized policymaking

In this section we explore under what circumstances a political principal benefits from a biased agent making policy on her behalf from the perspective of ex ante welfare. The dynamics between capacity investments and oversight described in the previous section play a central role. In particular, the effects on principal welfare depend on the locations of agent and overseer ideal points relative to one another because this dictates whether and how oversight affects agent capacity investment. We consider each case in turn.

First, consider the environment in which the agent faces a perfectly skeptical overseer so he is always met with reversal. This is true whenever it is not incentive compatible for the agent to make capacity investments sufficient to be upheld, which could be because it is impossible to do so – the agent is so biased that the overseer will never uphold – or because oversight is too stringent and the agent is not willing to invest at the overseer's threshold – $e_A^{\max}(t_A) < e_R^{\min}(t_A)$. In

either environment, the agent is always overturned and therefore the state obtains without agent intervention. This implies that the principal's welfare is not affected by agent bias since realized policy outcomes are invariant. Thus, when the overseer is perfectly skeptical the principal does not benefit from agent bias.

The second case is when the agent receives perfect deference. In this environment the agent develops capacity based on his own motivations: $e_A^* = e_A^u(\beta, \kappa)$. Since the agent's policy will always obtain the principal's ex ante expected welfare is given by,

$$-t_A^2 - V_\varepsilon(e_A^u(\beta, \kappa)).$$

The agent's capacity investment does not respond to t_A so $e_A^u(\beta, \kappa)$ is fixed, but principal welfare is strictly decreasing as agent bias $|t_A|$ becomes larger. As the agent becomes more biased substantive policy moves further from the principal, which harms her welfare, with no concomitant policy precision improvements. Thus, the principal is strictly better (worse) off as the agent moves toward (away from) her ideal point. Taken together, the preceding analysis yields the following result.

Proposition 3. *If the agent will either always be overturned or always be upheld then the principal weakly benefits when the agent's ideal point is closer to her own.*

In environments in which agent capacity investments are invariant to the agent's bias the principal never benefits from an agent further from her ideal point. When the agent is always overturned the agent's bias has no bearing on principal welfare. When the agent is always upheld increased agent bias strictly decreases the principal's welfare. Overall, if oversight has no impact on agent capacity investments the principal is always better off when the agent has an ideal point closer to her own, which is consistent with the general spirit of the ally principle.

Now consider an environment in which the agent faces conditional-deference oversight. In this case, the environment is characterized by intermediately biased agents (relative to the overseer) and the agent invests in capacity at the overseer's threshold, $e_R^{\min}(t_A)$, and targets policy at his ideal point, which is subsequently upheld by the overseer. This implies the following ex ante expected

welfare for the principal:

$$-t_A^2 - V_\varepsilon(e_R^{\min}(t_A)).$$

Since, in equilibrium, the agent will invest capacity to make the overseer indifferent we can reduce this expression by substituting the value of $V_\varepsilon(e_R^{\min}(t_A))$ when the overseer's incentive compatibility condition (equation 2) holds with equality:

$$-2t_A t_R - V_F.$$

Given that $t_R < 0$, the principal's welfare is increasing in t_A. When the agent is on the same side of the principal as the overseer the principal benefits from t_A closer to her ideal point and when the agent's ideal point is opposite the overseer the principal benefits from t_A *further from her ideal point*, which leads immediately to the following result.

Proposition 4. *If the agent will make capacity investments at the threshold level required by the overseer then the principal benefits when the agent's ideal point is closer to her own when* $t_A < 0$ *and benefits when the agent's ideal point is further from her own when* $t_A > 0$.

Agent capacity investments are predicated on the agent's ideal point relative to the overseer, but do not respond to agent-principal preference disagreement. Thus, the substantive upshot from the welfare effects depend on whether the agent is on the same side of the principal as the overseer or not. If the agent and overseer are on the same side then the principal's welfare improves as the agent moves further from the overseer and toward her ideal point (zero). This is again consistent with the spirit of the ally principle: The principal benefits from agents closer to her ideal point. However, the principal's welfare continues to increase as the agent crosses over her ideal point. The principal benefits from increasing agent bias on the opposite side of her ideal point from the overseer provided he continues to invest in capacity sufficient to be upheld. This runs counter to the ally principle: *The principal prefers agents further from her ideal point.*

Propositions 3 and 4 provide the basis for the main theoretical insights of this article. When oversight is ineffective at strengthening agent incentives agent bias is only detrimental to political

Review regime	**Perfect deference**	**Conditional-deference**	**Perfect skepticism**
Equilibrium agent investment: e_A^*	$e_A^u(\beta, \kappa)$	$e_R^{\min}(t_A)$ or 0	0
Effect of bias on principal welfare	decreasing in $\lvert t_A \rvert$	increasing in t_A	no effect

Table 3: Principal welfare effects given agent-overseer interactions

Note: Review regimes and equilibrium agent investments are defined as in Table 2. *Effect of bias on principal welfare* summarizes how t_A affects principal welfare.

principals. However, when oversight is effective at providing positive incentives for agent policymaking, the principal would often prefer to have a biased agent to continue to strengthen these incentives. That is, agent capacity investment incentives are increasingly strengthened the more effective oversight is *and* the more biased the agent. In this way, the principal instrumentally prefers to trade off biased content of policy for increased capacity when oversight is an effective institutional check on agent behavior. By pitting a biased agent against an oppositely biased overseer, the principal can benefit from the increased precision induced through agent capacity investments.

Table 3 summarizes the effects on principal welfare under each scenario. When the agent faces perfect deference his investments are unresponsive to increasing bias and therefore the principal is only harmed by increased agent bias; she prefers agents closer to her ideal point. On the other extreme, when the agent faces perfect skepticism, positive capacity investments never occur in equilibrium. In this case, the principal again derives no benefit from agent bias. However, when the agent faces conditional-deference, he makes capacity investments that exactly match the overseer's threshold. In this case, equilibrium capacity is increasing in agent bias until the point at which he becomes too biased (and drifts into the perfect skepticism environment). In this scenario the principal prefers agents closer to her ideal point if $t_A < 0$ and prefers agents further from her ideal point if $t_A > 0$. More generally, the principal's welfare is increasing in t_A, which implies that she benefits from agents that are biased away from her ideal point in the opposite direction of the

overseer.

Thus far we have focused on environments in which the agent already has the authority to make policy on behalf of the principal. While this is a reasonable assumption in many policy areas – e.g., many policies have been developed and implemented under existing authority conferred in the Clean Air Act – the results also have implications for principal delegation decisions. Even if the agent is acting under previously specified regulatory authority the principal always has the option to 'shut the agent down.' Therefore, one can think of each instance of agent policymaking being the product of a decision by the principal to allow the agent to (continue to) do so. The next section briefly discusses delegation dynamics in light of the results presented above.

3.1 Delegation

In this section we analyze an extension to the model above. The game is exactly the same except that at the beginning of the game the principal can choose to authorize the agent to make policy ($a = 1$) or not ($a = 0$). For simplicity, we assume that if the principal does not delegate authority then ω obtains unencumbered by agent intervention. The outcome, then, is the same as when the principal delegates but the agent is subsequently overturned by the overseer. This is an environment in which the principal does not have the requisite capacity or information to make policy on her own. Thus, it is a classic environment for delegation: The agent has an expertise advantage that the principal can utilize if she chooses, but it may come at the cost of biased policy or insufficient investment in high quality implementation.

To analyze when the principal will delegate to the agent we need only compare the principal's welfare for each environment analyzed above to her reservation utility for not delegating. If the principal chooses not to delegate 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 loses utility equal to her expectation of these outcomes. Whether the principal

finds it beneficial to authorize the agent depends on the relative locations of agent and overseer ideal points. If authorized, the agent and overseer behave according to the equilibrium characterized above. We analyze the principal's choices based on which environment would obtain following delegation: perfectly skeptical, perfectly deferential, or conditional-deference.

When agent-overseer ideal points are organized so that the overseer is perfectly skeptical the agent will invest zero in capacity and will always be reversed. In this case, the principal's utility from delegating and not delegating is equivalent: $-V_F$. In terms of policy, outcomes do not vary whether the agent is authorized to act or not. In both instances, final outcomes are predicated on the unregulated actions of private individuals or firms. Thus, the principal is indifferent between delegating or not when the agent-overseer environment is one in which the agent will always be overturned.[22] Overall, when authorizing the agent does not impact policy outcomes the principal has no incentive to authorize the agent to make policy in equilibrium.

Now consider an environment in which the agent, if authorized to make policy, receives perfect deference. In this case the principal loses utility based on the distance between her ideal point and the agent's ideal point, but equilibrium capacity is fixed since the agent always invests $e_A^u(\beta, \kappa)$. 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. Combining the principal's welfare from above and her reservation payoff for not delegating yields the incentive compatibility condition that must be met for her to delegate to the agent in this environment,

$$t_A^2 \leq V_F - V_\varepsilon(e_A^u(\beta, \kappa)).$$

Intuitively, the principal benefits from delegating to the agent in this environment if he is not too biased. Specifically, the spatial losses associated with delegating authority to the agent must be outweighed by the improvement in policy precision induced given that the agent will always invest in capacity based on his own motivations, $e_A^u(\beta, \kappa)$. The likelihood this condition is met and

[22] It is worth noting that the introduction of any arbitrarily small cost associated with delegation – for instance, due to the need to write authorizing legislation – would break this indifference and the principal would strictly prefer not to delegate to the agent.

the principal benefits from agent authorization is unambiguously decreasing in agent bias t_A since this has no bearing on the agent's equilibrium capacity investment. Further, because $e_A^u(\beta, \kappa)$ 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 capacity costs, κ.

Substantively, this highlights the fact that when oversight is ineffective at strengthening capacity investment incentives, the principal benefits from delegation based solely on agent and policy-environmental characteristics. If the agent is highly motivated, or if capacity costs are low, then it is more likely that delegation is beneficial. However, if the policy environment is relatively stable without agent policy intervention 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.

Finally, consider the case when the agent, if authorized, faces conditional-deference oversight. When the principal delegates the agent targets policy at his ideal point and capacity investments are at the overseer's threshold so that $e_A^* = e_R^{\min}(t_A)$.[23] In response, the overseer upholds the agent in equilibrium. The principal's decision to delegate or not is then dependent on whether it is better to allow the agent to set policy given his capacity investment incentives, which are a function of the relative distance between t_A and t_R. Accordingly, the principal's net expected utility for delegating to the agent is the difference from her welfare from the previous section ($-2t_A t_R - V_F$) and her reservation utility for not delegating ($-V_F$):

$$\Delta U_P(a = 1; r^*(e^*) = 0) \quad = \quad -2t_A t_R - V_F + V_F \quad = \quad -2t_A t_R.$$

Incentive compatibility implies that the principal will authorize the agent to make policy if and only if $-2t_A t_R \geq 0$. The principal only benefits from delegating to the agent if the agent and overseer are on opposite sides of her (i.e., t_A and t_R are oppositely signed). Since by assumption

[23] This assumes $e_A^{\max}(t_A) \geq e_R^{\min}(t_A)$ so that the agent will invest enough in capacity to satisfy the overseer. If instead $e_A^{\max}(t_A) < e_R^{\min}(t_A)$ then the agent invests zero and is overturned. If that is the case then the analysis is the same as when the overseer is perfectly skeptical. Thus we focus on the case in which $e_A^* = e_R^{\min}(t_A)$ in this section.

$t_R < 0$ this means that if the principal benefits from delegation at all then the agent is biased on the opposite side of the principal than the overseer: $t_A \geq 0$. Since the principal's welfare is increasing in t_A and she would only delegate when $t_A \geq 0$, any time she benefits from delegation at all she prefers to delegate to agents further from her ideal point.

This strengthens the observation from the previous section that in a conditional-deference environment the ally principle can fail to hold. If the principal benefits from delegating to the agent at all then she benefits from increasing agent bias.[24] That is, when $t_R < 0$ and $t_A > 0$ the increased capacity investments from increasing t_A outweigh the spatial losses of more biased substantive policy, thereby increasing principal welfare.[25] This follows from the fact that a (negatively) biased overseer can demand more capacity investment from a (positively) biased agent than the principal could (with ideal point zero) if she were the one monitoring. Therefore, the agent invests more in capacity than would be necessary to offset the principal's losses from biased substantive policy. Combining this with the analysis above yields the main result characterizing principal delegation decisions.

Proposition 5. *In equilibrium, the principal delegates as follows. When the agent will always be overturned following delegation the principal is indifferent between delegating and not, implying that agent bias has no effect. When the agent will always be upheld following delegation the principal delegates only if* $V_F - V_\varepsilon\left(e_A^u(\beta,\kappa)\right) \geq t_A^2$*, implying that delegation is less likely as agent bias increases. When the agent faces conditional-deference and will invest sufficient effort to be*

[24] It is possible that the principal can benefit from delegating to a perfect ally agent ($t_A = 0$). However, even in that case the principal's utility is increasing in agent bias. The only time the principal prefers $t_A = 0$ to $t_A > 0$ is when agent and overseer ideal points are arranged such that if $t_A > 0$ then the overseer would become perfectly skeptical–oversight would become too stringent–and the agent would respond by investing nothing in capacity and accept being overturned. Formally, this requires that $e_A^* = e_A^{\max}(t_A) = e_R^{\min}(t_A)$ when $t_A = 0$, which is a restrictive, knife-edge scenario.

[25] This dynamic is reminiscent of Wiseman (2009), where Congress may benefit from delegating to a biased agency that may subsequently face OIRA executive review because OIRA review leads the agency to set policy more beneficial to Congress than a world without OIRA review. That result assumes that the agency is located, ideologically, between the principal (Congress) and the overseer (OIRA) and only speaks to the ideological location of policy. In contrast, due to the multidimensional nature of policy – content and implementation – in our model the principal only benefits from delegating to an agent on the opposite side of her from the overseer. She benefits in this case because the overseer can credibly require higher capacity investment from the agent than the principal could herself. In equilibrium, this makes the principal strictly better off when oversight induces increased capacity investments while the overseer is made indifferent. While the results are complementary, our model extends the dynamic to include incentives for endogenous capacity to improve policy outcomes, a feature absent from Wiseman (2009).

upheld following delegation the principal delegates only if $-2t_A t_R \geq 0$*, implying that the principal only delegates when* $t_A \geq 0$*.*

Proposition 5 reinforces the insights from propositions 3 and 4. In the two cases in which agent capacity investments are not responsive to oversight or bias – perfect skepticism and perfect deference – the principal may benefit from delegation, but always prefers agents closer to her ideal point. However, when oversight does help structure agent investment incentives – conditional-deference – so does agent bias, which opens the door for the principal to benefit from agents with preferences further from her own. Indeed, when delegation is endogenous the principal only benefits from delegating to the agent when he is biased in the opposite direction of the overseer and, from proposition 4, her welfare is higher the further the agent is from her ideal point. Importantly, this insight only holds in environments in which oversight is an effective means of political oversight. When oversight does not incentivize agent capacity investment neither does agent bias, which implies that the principal can not prefer agents further from her ideal point. These theoretical insights provide several implications for bureaucratic politics.

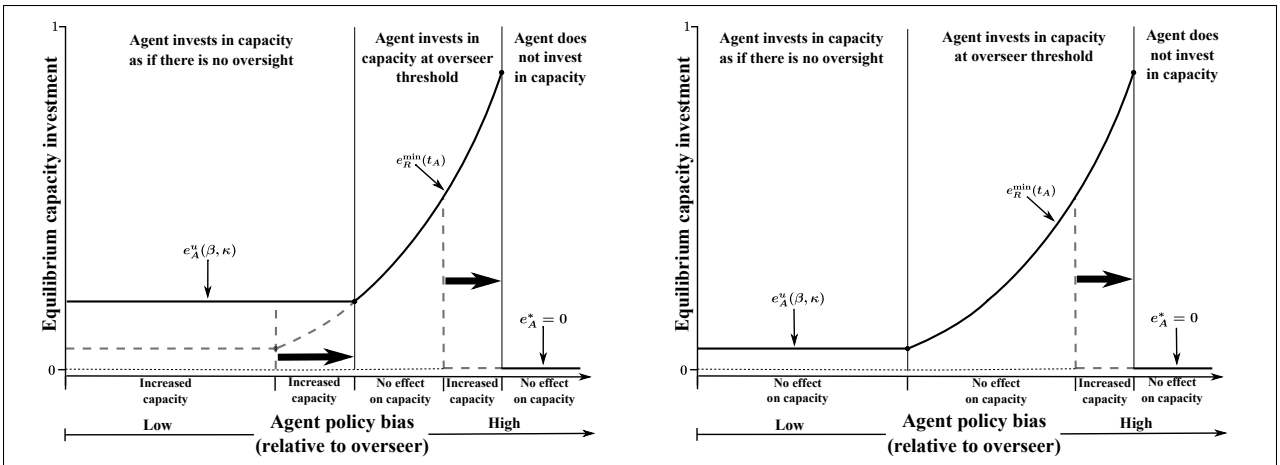
4 Implications

In this section we apply the insights of the model to bureaucratic politics. Specifically, we discuss both normative and empirical implications. 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 capacity investment increase, but the positive relationship is conditional on what type of oversight is induced. Figure 1 displays examples of these intuitions graphically. In both graphics the gray dashed lines denote previous levels of equilibrium capacity investments prior to parameter increases. The black solid lines denote the equilibrium capacity investments following the increases. Ultimately, the figures illustrate how the impact of these parameter shifts depend on how agent bias (increasing left-to-right along the x-axis) interacts with oversight.

First, consider a case in which agent policy motivations, β, increase, illustrated in figure

1a. This initially seems unambiguously positive in that it will generally produce a net increase in aggregate capacity investments. However, the relationship is conditional on how oversight impacts agent incentives. When the agent is ideologically close to the overseer oversight does not increase the agent's investment. However, the agent's policy motivations do increase $e_A^u(\beta, \kappa)$ and therefore, capacity investments increase proportional to the increase in β. This also expands the range of agent biases in which the agent invests as if there is no oversight. Once the agent becomes moderately biased, oversight does become stringent enough to induce the agent to increase his capacity investments to be upheld. The increase in β, while it does increase the maximum investment the agent *would be willing to make*, does not effectively alter observed investment levels. However, by increasing $e_A^{\max}(t_A)$, increased policy motivations expand the range of agents that invest enough to be upheld. These shifts in the range of agent biases in which investments are sufficient follows from the fact that increasing β strengthens agent incentives but does not affect the stringency of oversight. More biased agents now find it beneficial to invest enough 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, there is a positive correlation between agent policy motivations and agent bias in terms of principal welfare gains. Finally, capacity investment levels of extremely biased agents remain unaffected and those agents invest nothing and accept being overturned.

Now consider what happens as an agent becomes more averse to being overturned, illustrated in figure 1b. Similar to increasing policy motivations, increasing reversal aversion leads to a net increase in capacity investment, but this is again conditional on the relationship between oversight and agent bias. When agents are ideologically proximate to the overseer capacity remains unchanged. This is because $e_A^u(\beta, \kappa)$ does not respond to changes in reversal aversion. However, the maximum level of investment the agent is willing to make to be upheld, $e_A^{\max}(t_A)$, does increase in π while the stringency of oversight does not. The range of intermediately biased agents that will now invest sufficiently to be upheld expands, as in the previous case. Higher biased agents now switch from investing nothing and accepting reversal to investing enough to be upheld. Once



(a) Equilibrium Capacity for Increased β

(b) Equilibrium Capacity for Increased π

Figure 1: Examples of Comparative Statics for Increased Policy Motivations and Reversal Aversion
Note: The *y*-axis denotes agent capacity investments and the *x*-axis captures agent bias, relative to the overseer. The gray dashed lines denote previous levels of equilibrium capacity investments prior to parameter increases. The black solid lines denote the equilibrium capacity investments following parameter increases. Arrows illustrate the increased range of agent biases following parameter shifts.

again, extremely biased agents still find it incentive incompatible to make positive investments. Thus, strengthening an agent's aversion to reversal will increase observed equilibrium capacity investments, but only for a small range of agent biases. Agent reversal aversion and agent bias are complementary. An increase in aversion increases the maximal agent bias the principal can benefit from.

Taken together, these comparative statics predict positive correlations between capacity and increased policy motivations and reversal aversion within agencies. However, these relationships are conditioned by the fact that these increases only 'work' at increasing capacity investment for particular ranges of agent biases. In both cases the principal can benefit from a larger range of more extreme agent biases than before. Since the principal's welfare is increasing in agent bias when oversight strengthens investment incentives (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 posi-

tions 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 broader 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 threat of "sounding the alarm" (McCubbins and Schwartz 1984). The theoretical insights also speak to presidential appointments across institutions within the Executive branch. The president can simultaneously make appointments to direct agency policymaking (by appointing directors, secretaries, etc.) and to shape the nature of oversight (by appointing the head of the OIRA, for example).[26] By appointing an agency head that is oppositely biased from 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 capacity 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 agency bias that the principal would prefer. Overall, the results provide an instrumental ratio-

[26]One example in this vein is President Obama's appointments of Lisa Jackson – a self-described environmental protectionist – to head the EPA and Cass Sunstein – a staunch supporter of strict cost-benefit analysis – to head the OIRA. It was thought that the two appointees had radically different priorities when it came to identifying optimal environmental policy, which, in line with our theory, may have benefited President Obama in the ways described above (see also Bubb and Warren (2014) on the Jackson/Sunstein dynamics from an optimal agency bias perspective). Another, more general, example is the "team of rivals" dynamic often discussed in light of President Lincoln's cabinet choices. The theory developed here provides an instrumental rationale for understanding this type of appointment strategy.

nale for why an executive might optimally choose to appoint subordinates that do not share her substantive goals.

Similarly, from an intra-agency perspective, Lewis (2011) suggests that Presidents benefit from appointing ideologically distinct agency heads when these appointees have difficulty affecting agency policy outputs in less ideologically friendly agencies (54-55). For example, suppose 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 EPA director to fully temper policy output and direct it back toward less stringent regulation. In this instance, our theory suggests that appointing an agency head as a "policy gatekeeper" that prefers less stringent regulation than the President will induce 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 bureaucrats to work harder than they otherwise would in order to increase the probability that their policy goals are realized (Feldman 1989). The theory provides an instrumental rationale for bureaucratic organization that promotes a particular type of conflict both within and across regulatory agencies (Farber and O'Connell 2017; West 1988).[27]

The theory also has implications for the efficacy of managerial motivational strategies. If altering agency bias 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" or strengthening hierarchical authority (Moynihan and Pandey 2007), increasing the ratio of zealots to slackers (Gailmard and Patty 2007), or enhancing agencies' commitment to mission through staffing or other means (Wilson 1989). Tying oversight outcomes more strongly to agency budgets, promotional decisions, or the like may allow a principal to increase a bureaucrat's aversion to being

[27]Other interesting connections lie with the "agencies as adversaries" view comprehensively described in Farber and O'Connell (2017) and the related intra-agency dynamics described in Nou (2015).

overturned. Similarly, making clear that agency reputations are dependent on outcomes (Carpenter 2001) or reducing agency independence (e.g., if one takes π to be inversely related to agency independence) may enhance an agency's aversion to reversal. Both strategies will be effective at increasing net levels of observed implementation capacity, 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 figure 1a 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 strengthening capacity 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 their investments still increase since policy motivations 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 in capacity, their investments increase dramatically with an increase in policy motivations. Counterintuitively, 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 capacity investments increase. However, this strategy does not work when the agency or bureaucrat has interests that are closely aligned with the overseer. The only increase in capacity comes by inducing agencies that once found it incentive incompatible to invest to begin investing at high levels to pass muster in ex post review. Put another way, strengthening oversight penalties 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.[28]

5 Conclusion

We developed a theory of bureaucratic policymaking in the shadow of oversight and showed that political principals can benefit from 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 in capacity to ensure those policies are implemented effectively. Due to this duality in policymaking the principal benefits from a biased agent, with full policymaking discretion, interacting antagonistically with an oppositely biased overseer, empowered to reverse the agent's actions. Institutionalized oversight is only effective as a means for strengthening incentives if the agent is biased, and leveraging agent bias to induce higher capacity is only a viable route to improve outcomes 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 incentives for agents that 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 allocation of 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 bureaucratic appointments, agency design, and managerial motivation strategies.

References

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(1): 235–269.

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

Bertelli, Anthony and Sven Feldmann. 2007. "Strategic Appointments." *Journal of Public Administration Research and Theory* 17(1): 19–38.

Boehmke, Frederick, Sean Gailmard and John Patty. 2006. "Whose Ear to Bend? Information Sources and Venue Choice in Policy Making." *Quarterly Journal of Political Science* 1(2): 139–169.

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 Michael M. Ting. 2007. "Regulatory Errors with Endogenous Agendas." *American Journal of Political Science* 51(4): 835–852.

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. 2015. "On Judicial Review in a Separation of Powers System." *Political Science Research and Methods* 3(3): 473–492.

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.

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

Farber, Daniel A. and Anne Joseph O'Connell. 2017. "Agencies as Adversaries." *California Law Review* 105(1): 1375–1469.

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 John W. Patty. 2013*c*. "Stovepiping." *Journal of Theoretical Politics* 25(3): 388–411.

Gailmard, Sean and John W. Patty. 2017. "Participation, Process, & Policy: The Informational Value of Politicized Judicial Review." *Journal of Public Policy* 37(3): 233–260.

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

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(2): 459–490.

Gordon, Sanford C. and Catherine Hafer. 2005. "Flexing Muscle: Corporate Political Expenditures as Signals to the Bureaucracy." *American Political Science Review* 99(2): 245–261.

Hirsch, Alexander V. and Kenneth W. Shotts. 2015. "Competitive Policy Development." *American Economic Review* 105(4): 1646–1664.

Hirsch, Alexander V. and Kenneth W. Shotts. 2018. "Policy-Development Monopolies: Adverse Consequences and Institutional Responses." *Journal of Politics* 80(4): 1339–1354.

Huber, Gregory A. 2007. *The Craft of Bureaucratic Neutrality: Interests and Influence in Government Regulation of Occupational Safety*. New York, NY: Cambridge University Press.

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

Kagan, Elena. 2001. "Presidential Administration." *Harvard Law Review* 114(8): 2245–2385.

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.

McCann, Pamela J., Charles R. Shipan and Yuhua Wang. 2016. "Congress and Judicial Review of Agency Actions." *Unpublished Manuscript. University of Southern California.*

McCubbins, Mathew D. and Thomas Schwartz. 1984. "Congressional Oversight Overlooked: Police Patrols versus Fire Alarms." *American Journal of Political Science* 28(1): 165–179.

Moe, Terry. 2012. "Delegation, Control, and the Study of Public Bureaucracy." *The Forum* 10(2).

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

Nou, Jennifer. 2015. "Intra-Agency Coordination." *Harvard Law Review* 129(1): 421–490.

Patty, John W. 2009. "The Politics of Biased Information." *Journal of Politics* 71(2): 385–397.

Patty, John W. and Ian R. Turner. 2018. "Ex Post Review and Expert Policymaking: When Does Oversight Reduce Accountability?" *Unpublished Manuscript. Yale University. Presented at the 2016 Annual Meeting of the American Political Science Association*.

Prendergast, Canice. 2007. "The Motivation and Bias of Bureaucrats." *American Economic Review* 97(1): 180–196.

Shipan, Charles. 1997. *Designing Judicial Review: Interest Groups, Congress, and Communications Policy*. 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(2000-2001): 97–142.

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

Ting, Michael M. 2002. "A Theory of Jurisdictional Assignments in Bureaucracies." *American Journal of Political Science* 46(2): 364–378.

Ting, Michael M. 2003. "A Strategic Theory of Bureaucratic Redundancy." *American Journal of Political Science* 47(2): 274–292.

Ting, Michael M. 2008. "Whistleblowing." *American Political Science Review* 102(2): 249–267.

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

Turner, Ian R. 2017*a*. "Reviewing Procedure vs. Judging Substance: The Scope of Review and Bureaucratic Policymaking." *Unpublished Manuscript. Yale University*.

Turner, Ian R. 2017*b*. "Working Smart *and* Hard? Agency Effort, Judicial Review, and Policy Precision." *Journal of Theoretical Politics* 29(1): 69–96.

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

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.

Wiseman, Alan E. 2009. "Delegation and Positive-Sum Bureaucracies." *Journal of Politics* 71(3): 998–1014.

A Supplemental appendix

A.1 Agent-overseer subgame

Agent substantive policy choice.

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

Proof of Lemma 1. To show that the agent always sets policy at his ideal point we show that he is always weakly better off doing so 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 $x = \omega + t_A$ is given by,

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

The agent's expected utility from deviating 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:

$$\Delta U_A(x = \omega + t_A + \delta | r = 0) \quad = \quad -\beta\delta^2.$$

Since $\beta > 0$ and $\delta > 0$ the agent is strictly worse off from deviating.
Case 2: Overseer reverses. When the agent is reversed the outcome does not vary regardless of the agent's choice of x. Thus, the agent is indifferent between setting policy faithfully at his ideal point and deviating to another policy. In both cases the agent's expected payoff is the same.

Taken together these two cases imply that the agent weakly prefers setting policy at his ideal point, as stated in the result. ■

Optimal oversight.

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

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

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

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

Now, consider the overseer's subjective expected utility for upholding the agent,

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

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

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

Incentive compatibility implies that the overseer will uphold if and only if $\Delta U_R(r=0;\rho_{-R}) \geq 0$. Thus we have,

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

Rearranging we have:

$$V_F - V_\varepsilon(e) \quad \geq \quad t_A^2 - 2t_A t_R, \tag{A.1}$$

as is presented in-text in equation 2. The increase in policy precision 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 capacity investment to uphold the agent is defined as $e_R^{\min}(t_A) \equiv e$ such that equation A.1 holds with equality given agent bias t_A, assuming such an e exists. ■

Agent capacity investments.

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 make capacity investments 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 in capacity at the threshold level required to be upheld is given by,

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

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

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

The RHS of Equation A.2 can fall below 0 and rise above 1. So to ensure a capacity investment always exists further define:

$$e_A^{\max}(t_A) = \max\left[\min\left[\frac{\beta(t_A^2+V_F-V_\varepsilon(e_A^{\max}(t_A)))+\pi}{\kappa},1\right],0\right]. \tag{A.3}$$

Given this formulation, $e_A^{\max}(t_A)$ always exists. The RHS of equation A.2 is continuous over the interval $[0,1]$. So, either $e_A^{\max}(t_A)$ is on a boundary or there is an interior solution. ■

Lemma 4. *In equilibrium, the agent makes capacity investments according to the following strategy,*

$$s_A^{e*} = \begin{cases} e_A^u(\beta,\kappa) & \text{if } e_A^u(\beta,\kappa) \geq e_R^{\min}(t_A), \\ e_R^{\min}(t_A) & \text{if } e_A^u(\beta,\kappa) < e_R^{\min}(t_A) \text{ and } e_A^{\max}(t_A) \geq e_R^{\min}(t_A), \\ 0 & \text{if } e_A^u(\beta,\kappa) < e_R^{\min}(t_A) \text{ and } e_A^{\max}(t_A) < e_R^{\min}(t_A), \end{cases}$$

where $e_A^u(\beta,\kappa) = \arg\max_e -\beta V_\varepsilon(e)-\kappa e$, $e_R^{\min}(t_A) \equiv e$ *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 upholds ($e_A^u(\beta,\kappa) \geq e_R^{\min}(t_A)$); (2) the overseer always overturns ($e_A^u(\beta,\kappa) < e_R^{\min}(t_A)$ and $e_A^{\max}(t_A) < e_R^{\min}(t_A)$); (3) the overseer upholds if and only if the agent makes a large enough capacity investment, which is higher than the agent would invest absent oversight ($e_A^u(\beta,\kappa) < e_R^{\min}(t_A)$ and $e_A^{\max}(t_A) \geq e_R^{\min}(t_A)$). These cases are defined by the overseer's best response in Lemma 2 and the maximum capacity investment the agent is willing to make to be upheld in Lemma 3.

Overseer always upholds (perfectly deferential). The agent's expected payoff given he will be upheld is given by,

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

The agent seeks to maximize $U_A(e|r=0)$ with his choice of e, which implies the following capacity investment,

$$e_A^u(\beta,\kappa) \in \arg\max_e -\beta V_\varepsilon(e) - \kappa e.$$

Moreover, $e_A^u(\beta,\kappa)$ 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.

Overseer always overturns (perfectly skeptical). To see why the agent never makes positive capacity investments in an environment in which he will always be reversed note that the agent's expected payoff for making positive capacity investments 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 nothing given he 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 making positive capacity investments 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 makes positive capacity investments when he will be reversed he simply pays the cost for that investment, and, therefore, optimally invests nothing.

Conditional-deference overseer. In this environment $e_A^u(\beta,\kappa) < e_R^{\min}(t_A)$ so the agent is constrained by the overseer. The agent compares his expected utility from investing in capacity at the threshold level and being upheld by the overseer and his expected utility from investing nothing ($e=0$) and being overturned. These expected payoffs are given by the following expressions, respectively:

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

These combine to give the net expected payoff for making capacity investments at the threshold (and being upheld rather than overturned):

$$\begin{aligned}\Delta U_A(e_R^{\min};\rho_{-A}) &= -\beta V_\varepsilon(e_R^{\min}) - \kappa e_R^{\min} + \beta(t_A^2 + V_F) + \pi, \\ &= \beta(t_A^2 + V_F - V_\varepsilon(e_R^{\min})) - \kappa e_R^{\min} + \pi. \end{aligned} \tag{A.4}$$

Equation A.4 gives the agent's incentive compatibility condition for investing the threshold level, $e_R^{\min}(t_A)$, rather than $e = 0$ and being overturned. As long as this condition is weakly greater than zero the agent, in weakly undominated strategies, will make capacity investments at the threshold level required to be upheld when constrained by the overseer. ■

Proposition 1. *In equilibrium, (1) the agent makes capacity investments according to e_A^*, given by equation 5; (2) the agent always sets policy at his ideal point, $x^*(\omega) = \omega + t_A$; and (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 lemmas 1, 2, 3, and 4. Lemmas 3 and 4 yield number 1 in the proposition, lemma 1 yields number 2, and lemma 2 yields number 3. To complete the characterization of the PBE define the overseer's beliefs about ω given observation of e as

$$f(\omega|e) = \frac{f(\omega)\sigma(e|\omega)}{\int f(\omega)\sigma(e|\omega)d\omega}$$

where $\sigma(e|\omega)$ denotes the agent's capacity investment strategy. On the path of play $\sigma(e|\omega) = 1$ trivially since e is chosen prior to learning ω and $\int f(\omega)\sigma(e|\omega)d\omega = 1$. Thus, $f(\omega|e) = \frac{f(\omega)}{1}$ on the path of play, implying that the overseer simply retains her prior regarding ω, $f(\omega)$. Finally, fix the overseer's beliefs off the path of play to also be the prior, $f(\omega)$. Any other off-path beliefs would imply that the overseer thinks actions reveal information that the agent did not possess, so retention of the prior is the only reasonable choice. ■

Proposition 2. *In equilibrium, agent bias strengthens agent capacity investment incentives if and only if oversight also strengthens these incentives.*

Proof of Proposition 2. This follows from the fact that neither agent bias t_A nor the agent's aversion to being overturned π appear in equation 4, but both t_A and π appear in the agent's capacity investment given by equation A.3. ■

A.2 Principal welfare

Proposition 3. *If the agent will either always be overturned or always be upheld following delegation the principal never benefits from agent bias.*

Proof of Proposition 3. First, assume that the agent is always overturned. Then the principal's expected utility is given by,

$$\begin{aligned} U_P &= \mathbb{E}[-y^2|r^*=1], \\ &= -\mathbb{E}[\omega]^2 - V[\omega], \\ &= -V_F. \end{aligned}$$

Clearly agent bias t_A plays no role in principal welfare in this case, implying that agent bias does not benefit the principal.

Now assume that the agent is always upheld. In this case the agent invests effort $e_A^u(\beta,\kappa)$ and sets $x^*(\omega)=\omega$ in equilibrium. Thus, the principal's expected utility in this case is given by,

$$\begin{aligned} U_P &= \mathbb{E}[-(x-\omega+\varepsilon)^2|x^*,e_A^*], \\ &= -(\omega+t_A-\omega)^2 - \mathbb{E}[\varepsilon|e_A^*]^2 - V[\varepsilon|e_A^*], \\ &= -t_A^2 - V_\varepsilon(e_A^u(\beta,\kappa)). \end{aligned}$$

In this case principal welfare is decreasing in agent bias t_A, implying that the principal does not benefit from agent bias. ■

Proposition 4. *If the agent will make capacity investments at the threshold level required by the overseer then principal welfare is increasing in agent bias, t_A.*

Proof of Proposition 4. The principal's expected utility in this case is given by,

$$\begin{aligned} U_P &= \mathbb{E}[-(x-\omega+\varepsilon)^2|x^*,e_A^*], \\ &= -(\omega+t_A-\omega)^2 - \mathbb{E}[\varepsilon|e_A^*]^2 - V[\varepsilon|e_A^*], \\ &= -t_A^2 - V_\varepsilon(e_R^{\min}(t_A)). \end{aligned}$$

We can reduce the principal's expected utility by solving the overseer's incentive compatibility to uphold with equality and plugging in $V_\varepsilon(e_R^{\min}(t_A))$:

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

Since $t_R < 0$ by assumption and V_F is fixed the principal's expected utility is increasing in t_A. ■

A.2.1 Delegation

Proposition 5. *In equilibrium, the principal delegates as follows. When the agent will always be overturned following delegation the principal is indifferent between delegating and not, implying that agent bias has no effect. When the agent will always be upheld following delegation the principal delegates only if* $V_F - V_\varepsilon\left(e_A^u(\beta,\kappa)\right) \geq t_A^2$*, implying that delegation is less likely as agent bias increases. When the agent faces conditional-deference and will invest sufficient effort to be upheld following delegation the principal delegates only if* $-2t_A t_R \geq 0$*, implying that the principal only delegates when* $t_A \geq 0$*.*

Proof of Proposition 5. Each case corresponds to one agent-overseer environment outlined above. I consider each in turn. First note that any time the principal does not delegate her payoff is given by,

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

I now turn to considering each case and the principal's expected payoff should she choose to delegate instead given the agent-overseer environment.

The agent is always overturned. If the agent will always be overturned by the overseer following delegation then the principal's payoff, for any delegation decision a, is always simply, $-V_F$, since the agent's policy choice will never be realized. Thus, her net expected payoff is simply zero regardless of her delegation choice, rendering her indifferent.

The agent is always upheld. Given that the principal knows that if she authorizes the agent to make policy then $x^*(\omega) = \omega + t_A$ and $e^* = e_A^u(\beta,\kappa)$, her subjective expected payoff for authorizing the agent to make policy in this environment is given by,

$$\begin{aligned} U_P(a=1;r=0,e_A^u(\beta,\kappa)) &= -t_A^2 - \mathbb{E}[\varepsilon|e^*]^2 - V[\varepsilon|e^*], \\ &= -t_A^2 - V_\varepsilon(e_A^u(\beta,\kappa)). \end{aligned}$$

Combining this with the principal's expected payoff for not delegating from above yields the principal's net expected payoff, defined as $\Delta U_P(a=1;r=0,e_A^u(\beta,\kappa)) = U_P(a=1;r=0,e_A^u(\beta,\kappa)) - U_P(a=0)$:

$$\Delta U_P(a=1;r=0,e_A^u(\beta,\kappa)) = -t_A^2 - V_\varepsilon(e_A^u(\beta,\kappa)) + V_F.$$

Incentive compatibility implies that the principal will authorize the agent if and only if $\Delta U_P(a =$

$1; r = 0, e_A^u(\beta, \kappa)) \geq 0$, which requires that,

$$\begin{aligned} -t_A^2 - V_\varepsilon(e_A^u(\beta,\kappa)) + V_F &\geq 0, \\ V_F - V_\varepsilon(e_A^u(\beta,\kappa)) &\geq t_A^2, \end{aligned}$$

as stated in the proposition.

Agent upheld if and only if $e^* \geq e_R^{\min}(t_A)$. Assume the agent will invest sufficient e to be upheld, which implies that $e^* = e_R^{\min}(t_A)$. Further, $x^*(\omega) = \omega + t_A$ in this case. Thus, the principal's subjective expected payoff for authorizing the agent is given by,

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

Combining this with the principal's expected payoff for not delegating yields the principal's net expected payoff for delegating in this environment. Define the principal's net expected payoff from authorizing the agent as $\Delta U_P(a = 1; r = 0, e_R^{\min}(t_A)) = U_P(a = 1; r = 0, e^* = e_R^{\min}(t_A)) - U_P(a = 0)$:

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

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

$$-t_A^2 - V_\varepsilon(e_R^{\min}(t_A)) + V_F \geq 0.$$

Now, solving the overseer's incentive compatibility condition to uphold with equality for $V_\varepsilon(e)$ allows us to substitute $V_\varepsilon(e_R^{\min}(t_A))$ as follows:

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

This implies the principal will delegate if and only if $-2t_A t_R \geq 0$, as stated in the result. ■