From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Economics

2025

The Informational Content of Policymaking

Sayed, Hassan

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01c821gp22z

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

The Informational Content of Policymaking

Hassan Sayed

A Dissertation

Presented to the Faculty

Of Princeton Univeristy

In Candidacy for the Degree

of Doctor of Philosophy

Recommended for Acceptance by

the Department of

Economics

Advisers: Leah Boustan and Pietro Ortoleva

May 2025

© Copyright by Hassan Sayed, 2025.

All Rights Reserved

Abstract

This dissertation studies how policy choices generate information in political and organizational settings, and how these informational externalities influence leaders' willingness to implement economic reform, propose partisan public policies, or abuse authority.

Chapter 1 argues that under policy uncertainty, policy choices generate a tradeoff between payoffs and information. Leaders can implement their preferred policy, but if the outcome is bad, constituents learn this and threaten office removal. I show that when leaders face this tradeoff in a repeated setting, their policy choices are nonmonotone in beliefs about policy efficacy. At extremal beliefs, they experiment with the most extreme policy that keeps them in office. At middle beliefs, they implement intermediate "learning trap" policies that halt learning about the optimal policy when it would be most useful. I apply the model to the post-1861 reforms ending serfdom in Imperial Russia, arguing that by combining elements of liberalism and serfdom to obfuscate inference about policy efficacy, these reforms represented a learning trap.

In Chapter 2, I study a model where a political executive's policy agenda generates information about her ability. Since policies favoring a single party are harder to pass, the partisanship of an agenda influences what success or failure of passage communicates about ability. The model delivers a U-shape relationship between an executive's ex-ante winning chances and her agenda's partisanship. I use these insights to analyze the partisanship of U.S. governors' policy proposals from 1990-2020, showing that while partisanship has risen on average, there is sizeable variation at the governor level. I interpret these variations as responses to electoral incentives by testing the model and showing a U-shaped relationship between partisanship and approval ratings.

Chapter 3, joint with Kim Sarnoff, develops an organizational model where employees can report harmful behavior — or "transgressions" — committed by managers, but managers face uncertainty about which of their actions employees actually consider transgressions. We show that policies that discourage managers from committing transgressions may harm employees either by motivating more managers to commit harmful actions that employees do not want to report or inducing managers to opt out of interacting with employees altogether.

Acknowledgements

The work of this dissertation would not have been possible without the support of friends, family, and colleagues both inside and outside the academy. I would like to thank the community of faculty, peers, and officemates across both Economics and Politics who provided intellectual support, thoughtful conversations, and steadfastness — including Narek Alexanian, Manuel Pérez Archila, Pietro Buri, Francesco Fabbri, Allie Green, Quan Le, Dan McGee, George Nikolakoudis, Sebastian Roelsgaard, Rafael Schwalb, Carol Shou, Martin Vaeth, Jing Wu, Elaine Yao, Chansik Yoon, Qiyuan Zheng; Roland Bénabou, Ellora Derenoncourt, Germán Gieczewski, Faruk Gul, Matias Iaryczower, Gleason Judd, Atif Mian, Xiaosheng Mu, Abdoulaye Ndiaye, Mehdi Shadmehr; and many others. Thank you to my coauthor Kim Sarnoff for your intellectual insights and banterous wit across our projects; and Meghana Mungikar for the encouragement and chitchat through years of coffee breaks in our department. Our Economics staff Laura Hedden and Stu Sternbach were always there when I had an administrative query, provided comfort through the most mundane conversations, and provided a healthy dose of chocolate when I needed a pick-me-up. Financial support for this dissertation was provided by the National Science Foundation, the William S. Dietrich II Economic Theory Center, and the SPIA Research Program in Political Economy.

The friendships I cultivated in New York City provided social escape and comfort away from the desolation of Princeton. I would like to thank everyone who offered a living space to stay in when I needed it — Naib Mian and Shakeeb Asrar, Nasir Malim and Sara Abdelhalim, Rasheed Abiose, Ali Imam, Arish Jamil, Fatima Khan, and Zia Mohammad. Thank you to the staff at Little Flower Cafe and Villager NY for their provision of caffeine, treats, and hospitality as I worked away at my research under their roofs.

The foundation for this journey was laid in my undergraduate days at Northwestern. Thank you to Mark Witte — who made me realize ten years ago that I wanted to do a PhD in Economics, Jeff Ely — who cultivated my deeper interest in economic theory, and Bob Gordon — my coauthor

of seven years, from whom I inherited my ability to dissect patterns in aggregate data and my intellectual pessimism.

My deepest thanks go to my committee — Leah Boustan, Ilyana Kuziemko, and Pietro Ortoleva — who consistently pushed me intellectually, encouraged me to think creatively, and were there at a moment's notice whenever I needed them. Pietro was the first faculty member I ever spoke to at Princeton. I am grateful for his steady commitment to troubleshooting my models, his willingness to lend an ear — or occasional focaccia — during moments of intellectual or emotional turbulence, his help sharpening my theoretical agenda, and his sheer enthusiasm for entertaining even my wildest ideas. Leah's commitment to the "big picture" of empirical research has been a continual source of inspiration, from our biweekly discussions of economic history tomes in the summer of 2020 to our monthly reading groups and her thoughtful feedback on my work's historical dimensions. Ilyana pushed me to think deeply about the narrative arc of my research — how to connect my theoretical insights to the data and how to anticipate and address alternative explanations. But she also went out of her way to keep her door open when I needed it, connecting me with other scholars and offering her time at a moment's notice. Above all, the three of them have been supportive, thoughtful, and kind — not just as mentors, but as people. I never doubted that they had my best interests in mind, and I always felt comfortable bringing them my questions and concerns. I can only hope to carry these qualities forward in my own scholarly life.

Last of all, thank you to my family — Abu, Mama, Farah, and all of our furry friends — for pouring blood, sweat, and tears into the past 23 years of my educational journey. You once said I had to be a doctor, but never specified which kind.

This dissertation examines how organizations and their leaders are held accountable — and how institutional failures of accountability culminate in regression. It is dedicated to the memory of students who dared to defy, and who shouldered the burden of accountability when their institutions would not.

Contents

Abstract **3**

Acknowledgements **4**

1 Learning Traps **9**

1.1 Introduction 9

1.2 Model 15

1.3 Analysis 18

1.3.1 Analysis of Baseline Model 18

1.3.2 Model with Nonmyopia and Adjustment Costs 22

1.3.3 Unanticipated Information Shocks 26

1.4 Application: The 1861 Reforms as a Learning Trap in Imperial Russia 27

1.4.1 First Equilibrium: Serfdom Pre-1861 28

1.4.2 Second Equilibrium: Learning Trap, 1861-1905 30

1.4.3 Third Equilibrium: Stolypin Reforms, Post-1905 34

1.4.4 Alternative Interpretations for the Learning Trap 35

1.5 Anticipated Information Revelation 38

1.5.1 Information Revelation 38

1.5.2 Political Fragmentation in Europe and China 39

1.6 Conclusion 41

1.7 Proofs of Main Results 43

1.8 Additional Results 54

2 Persuasion with Partisanship **62**

2.1 Introduction . . . 62

2.2 Theoretical Model . . . 70

2.2.1 Preliminaries . . . 70

2.2.2 Equilibrium . . . 73

2.2.3 Welfare . . . 76

2.2.4 Persuasion Model and Data Predictions . . . 79

2.2.5 Extensions . . . 83

2.3 Data . . . 85

2.3.1 U.S. Governors and State of the State Addresses . . . 85

2.3.2 Sourcing and Pre-Processing . . . 89

2.3.3 Speech Processing with BERT . . . 89

2.3.4 Additional Data Sources . . . 91

2.4 Documenting Partisanship . . . 91

2.4.1 Calculating Partisanship . . . 91

2.4.2 Aggregate Behavior of Partisanship . . . 93

2.4.3 Heterogeneity . . . 96

2.5 Testing the Model . . . 99

2.5.1 Estimation . . . 99

2.5.2 Estimating Model's Predictions . . . 101

2.5.3 Presidential Speech . . . 107

2.6 Conclusion . . . 108

2.7 Proofs and Theoretical Extensions . . . 110

2.7.1 Main Proofs . . . 110

2.7.2 Voter Preference Microfoundation . . . 119

2.8 Auxiliary Tables and Figures . . . 121

2.9 Hand-Coding Guide . . . 131

3 Transgressions in Organizations **135**

3.1 Introduction . . . 135

3.2 Setup 141
3.3 Static Equilibrium 145
3.4 Welfare 150
3.5 Dynamics and Non-Learning Steady State 158
3.6 Conclusion 162
3.7 Proofs 165

References **179**

Chapter 1

Learning Traps*

1.1 Introduction

The policies of incumbent leaders — autocrats, CEOs, or elected officials — are often constrained by threats of being ousted by their constituents — elites, boards of directors, or the electorate. When the optimal policy for constituents is certain, a leader simply implements her most preferred policy that avoids *immediate overthrow*. However, when there is uncertainty about the effects of different policies, policy choices generate both payoffs and information about policy efficacy. If a leader experiments with her preferred policy today and the outcome is good, she can continue to remain in office because constituents will have no incentive to oust her. But if the resulting outcome is bad, constituents may take this as a sign that their preferences are misaligned with their leader's and threaten overthrow, forcing a leader to implement undesirable policies in the future. Uncertainty generates an additional *threat of information revelation* that constrains leaders' willingness to implement certain policies.

This paper develops a repeated model of policy uncertainty where an ideologically-motivated "leader" experiments with different policies, but may be overthrown if she implements policies that harm constituents. It argues that leaders' policy experimentation is nonmonotone in beliefs about whether their preferred policy favors their constituents. At extremal beliefs, when constituents are relatively sure what the optimal policy is, the leader experiments, pursuing extreme

*Materials from this chapter were presented at the 2022 Young Economists Symposium and the 2023 North American Summer Meeting of the Econometric Society.

policies that prevent overthrow. At intermediate beliefs, when there is greater uncertainty about the optimal policy, the leader halts learning, precisely when it would be most useful. I show that extreme policies — namely, leaders' most and least preferred policies — generate information about policy efficacy, while spatially intermediate "learning trap" policies generate *no information* but modest payoffs for leaders. This creates a tradeoff between information revelation and policy preferences. At high beliefs, information favors leaders, so they pursue their most preferred policy. At intermediate beliefs, information may reveal leaders' preferred policies are harmful, so they implement "learning traps" and forever receive a modest payoff. At low beliefs, constituents threaten overthrow, forcing leaders to pursue their least preferred policies.

I illustrate an application of the model's findings by studying the 1861 economic reforms ending serfdom and codifying the peasant commune in Imperial Russia. Prior to 1861, Russian peasants (serfs) worked noble land as part of communes. Russia's 1856 defeat in the Crimean War represented a shock that generated policy uncertainty, forcing the Tsar to reconsider serfdom. Many in the Tsar's government believed liberal reforms — like private property and freeing labor to move to the cities — could stimulate economic growth and expand its fiscal capacity. However, in the ensuing reforms, the government mixed elements of private and communal property and restricted labor mobility, hampering the economic growth it so desired. I show how this arrangement illustrates the logic of a learning trap: the government feared unrest if liberal policies backfired, but by mixing liberal and communal institutions, distorted labor and production decisions and halted learning about which policy was better. I situate the choice of an intermediate policy amidst policy uncertainty against the two other equilibria of the model: prior to the uncertainty of the 1850s, the government maintained an extreme policy of serfdom, while after shocks favoring liberal reform in 1905, the government experimented with private property.

In my model, a leader prefers a higher policy $x \in [0, 1]$. A representative agent for "the people" wants to maximize an outcome $y \in \mathbb{R}$. There are two states of the world: one where the relationship between x and y is positive or single peaked — allowing the leader to implement her preferred policy — and one where it is negative, forcing the leader to implement policies undesirable to her. The leader and people share a common prior. As different policies are implemented, information is revealed. Agents use histories of policy-outcome data (y, x) to infer the state. The key feature is that data on policy efficacy are generated endogenously: by setting x, the leader also controls

inference. The people can overthrow the leader at any point, disciplining the policies a leader implements.

Information revolves around a spatially intermediate "learning trap" policy x_{LT} which, when implemented, generates no information about policy efficacy. x_{LT} exists if the people's maximal payoffs in each state of the world are equal. Extreme policies farther from x_{LT} generate more information about the optimal policy, corresponding to policy experimentation.

I show that equilibrium experimentation is nonmonotone in beliefs that the leader's preferred policy benefits the people. At high beliefs, the leader pursues high x, generating favorable flow utility and information. As beliefs decrease, information is less likely to favor her in expectation, forcing her to balance the value of information with policy preferences. Because pursuing x_{LT} obfuscates information, for an intermediate range of beliefs, forever pursuing x_{LT} and achieving a moderate payoff is preferable to pursuing an extreme policy today that risks a lower payoff in the future. This range expands as patience increases, policies become more informative, or the distaste for low x grows. Finally, at low beliefs, the people are certain their preferences are misaligned with the leader's, forcing her to implement low x lest she be overthrown.

In the baseline setting, leaders' policy decisions are the only sources of inference. I extend the model with anticipated and unanticipated *external information revelation*. Unanticipated information — like wars or disasters that "throw back the veil" on policy efficacy — reveals asymmetry in policy variation: leaders switch to experimentation only when large shocks favor their preferred policy. When information is *anticipated*, leaders pursue learning trap policies for a smaller range of beliefs: learning can no longer be completely shut down, so leaders hope experimentation with their preferred policy will favor them. Anticipated shocks can be thought of as political "contagion," where groups of countries with similar historical institutions may learn from each other. I show how these insights can help us understand phenomena like the effect of political fragmentation on differences in experimentation with indirect rule between 19th-century Europe and China.

Reform in Imperial Russia I show how the model can help us understand the tepidness of large-scale economic reform by studying the end of serfdom in Imperial Russia, showing how the Tsar's 1861 emancipation reforms can be seen as a "learning trap" when juxtaposed with pre-1861

serfdom and post-1906 experiments with private property. Prior to 1861, Russian peasants worked state or noble-owned land as members of communes; production and labor choices were managed by seigneurs (Nafziger, 2010). This status quo represents one extreme policy, serfdom. Russia's defeat in the 1853 Crimean War revealed a need to modernize its economy (Starr, 2015). I interpret the war as a shock to beliefs that cast uncertainty about optimal economic policy, marked by heightened discourse about whether serfdom or liberal reform was better. The Tsar's government believed liberal institutions could reduce the power of the nobility (Moon, 2014) jumpstart an industrial revolution (Pereira, 1980), and increase fiscal revenue (Dennison, 2020,2). However, it also worried they could backfire, leading to exploitation, abandonment of land, fiscal straits, and overthrow (Dennison, 2014; Polunov, Owen, and Zakharova, 2015).

The model predicts that, in response to uncertainty, the government should shift to an intermediate policy that minimizes learning. Indeed, the Tsar mixed serfdom with Western European institutions like private property. While seigneurs no longer oversaw peasants, communes remained, and many household production, property, and labor allocation decisions required communal consent. Most agricultural land was held communally instead of privately, distorting investment and work incentives (Dennison, 2020). Over many decades, the government used the commune to enforce mobility restrictions, inhibiting urban migration and industrial growth (Nafziger, 2010). I argue that these forces minimized Russians' capacity to discern the benefits of liberal reform. Finally, I interpret political events at the beginning of the 20th century as generating another shock to beliefs in favor of liberal economic institutions. The model predicts that this should cause a radical shift to an extreme policy. Indeed the government began experimenting with private property and other liberal economic institutions as part of Pyotr Stolypin's reforms.

I also address the two traditional explanations of the reform's tepidness: pressures from the nobility and gentry (Khristoforov and Gilley, 2016) and state capacity (Dennison, 2020). While both forces undoubtedly influenced the reforms, the bargaining power of elites and gentry was relatively weak and slowly waned, and neither force explains why the government did not even gradually reform over four decades. Viewing the reform's intermediacy as a learning trap — in contrast to pre-1861 serfdom and post-1905 reform — provides a fuller picture of its shape in tandem with existing theories.

I introduce the model in section two and solve it in section three. Section four applies the find-

ings of the model to reform in Russia. Section five solves the model with exogenous information revelation with applications to political fragmentation in Europe and China. I conclude in section six. The remainder of the section addresses relevant theoretical literature; literature on Russia is contained in section four and on political fragmentation in section five.

Literature This paper's central insight that leaders pursue obfuscatory policies only at *intermediate* beliefs about policy efficacy contrasts with a political reputation result known as "gambling for resurrection," where leaders pursue their preferred policy at low beliefs and obfuscatory policies at high beliefs (Dewan and Hortala-Vallve, 2019; Dur, 2001; Fu and Li, 2014; Izzo, 2024; Majumdar and Mukand, 2004; Prendergast and Stole, 1996; Tomasi, 2023). This stark difference is driven by the present paper's repeated environment of policy uncertainty. In finite-time settings, leaders pursue their preferred policies in the last period; at high beliefs, obfuscatory policies allow leaders to receive a moderate payoff today and a high payoff tomorrow. However, in repeated settings, leaders must deal with the consequences of their actions for the rest of time. They face a stationary problem where they either choose between a learning trap — perpetually receiving a moderate payoff that never reveals information — or a one-off extreme policy that reveals information. Information revelation will only be preferred to a moderate payoff at high beliefs. Second, leaders in "gambling for resurrection" models can be replaced by an outside option — for example, when there is uncertainty over a politician's ability. [1] At low beliefs, leaders can then only be retained if they experiment and generate a strong signal. By contrast, under policy uncertainty, constituents can threaten overthrow only when they are sure the leader's policy is bad and, even then, the leader can retain office by keeping them indifferent to overthrow. The present paper's insights are conceptually related to partisan traps (de Mesquita and Dziuda, 2024), where voters perpetually elect partisan ideologues who never pursue common value policies; however, "partisan traps" arise due to information asymmetries between politicians and voters that obfuscate the feasibility of non-partisan policies.

This paper relates to microeconomic studies of reputation in repeated environments. Holmström (1999) shows that managers oversupply effort when firms set future wages based on infer-

[1] The learning trap policy in this paper delivers the same utility in all states of the world. If constituents prefer a learning trap to a (constant) outside option at one belief, they prefer it at all beliefs, meaning the paper's result could not hold with an outside option.

ences about their ability, although this simply slows learning.[2] Non-learning often depends on asymmetries in costs (Manso, 2011), information (Ely and Välimäki, 2003), or signal interpretation (Aghion and Jackson, 2016). In the latter, constituents threaten to replace incumbents until a high-ability leader emerges which forces learning in equilibrium; however, this mechanism is driven again by a "gambling for resurrection phenomenon" which cannot be replicated in the present paper. This paper finds more patient leaders prefer to slow learning since the salience of *future* constraints on policies increases, the opposite of Besley and Case (1995) and Banks and Sundaram (1998).

This paper contributes to research on inference using policy-outcome data that argues agents inefficiently learn (Eliaz and Spiegler, 2020; Levy and Razin, 2021a; Montiel Olea, Ortoleva, Pai, and Prat, 2022; Schwartzstein and Sunderam, 2021; Spiegler, 2016). These data may hamper learning about policies or politician ability. Voters learning from spatial platforms may converge on inefficient policies (Callander, 2011). Frequentist inference or bounded memory can cause polarization (Izzo, Martin, and Callander, 2021; Levy and Razin, 2021b). Less information about policies can be better for voters (Prat, 2005) or politicians (Kartik, Squintani, Tinn, et al., 2015). Substitutability between politician effort and ability (Ashworth, De Mesquita, and Friedenberg, 2017) can hamper learning about ability. Empirical work also documents inferences about policy efficacy, including trust in government (Chen and Yang, 2019), external validity of policy experiments (Wang and Yang, 2021), and policy contagion (Buera, Monge-Naranjo, and Primiceri, 2011; Mukand and Rodrik, 2005).

A common theme throughout models of autocratic power-balancing is that rulers self-regulate their exploitation in order to avoid overthrow; they implement the most extreme policies keeping constituents or elites indifferent to overthrowing them (Acemoglu and Robinson, 2000,0; Boix and Svolik, 2013; De Mesquita and Smith, 2010; Dower, Finkel, Gehlbach, and Nafziger, 2018; Li, Gilli, et al., 2014).[3] This paper argues that leaders *moderate* their actions when policies generate information. It also relates to gradualism in "divide-the-dollar" settings like land reform[4], as discussed in Roland (2002) and Acemoglu and Robinson (2008), showing that under uncertainty about the op-

[2]Non-learning in their paper is driven by an evolving state.

[3]The last paper looks at authority given to peasant communes given the frequency of disruptions in 1860s Russia, the period of this paper's case study.

[4]These considerations are important in the context of the Russian case study application, and a microfoundation addressing their specific role is constructed in Section 1.8.

timal form of land ownership, "partial reforms" mixing different property regimes can obfuscate which reform is better.

1.2 Model

Policies Consider an infinite-horizon model beginning at $t = 1$. Each period, a policy $x_t \in X = [0, 1]$ is implemented which determines an outcome $y_t \in \mathbb{R}$ realized later that period. For example, y can represent profits, income, or the value of a public good.

Agents The first agent in the model is the "leader" — an entrenched incumbent, autocrat, or CEO. The second agent — "the people" — constrains the leader's power and can be interpreted as a representative agent for an electorate, board of directors, or elite group. Both agents are infinitely lived.

The leader discounts the future at rate δ. Her flow utility is $(1 - \delta)u_\ell(x_t)$ for $u_\ell(\cdot)$ strictly increasing, differentiable, and weakly concave.[5] Preference for higher x_t could reflect an ideological bent; reputational concerns; or the size of rents the government is able to extract in a setting like land reform.[6]

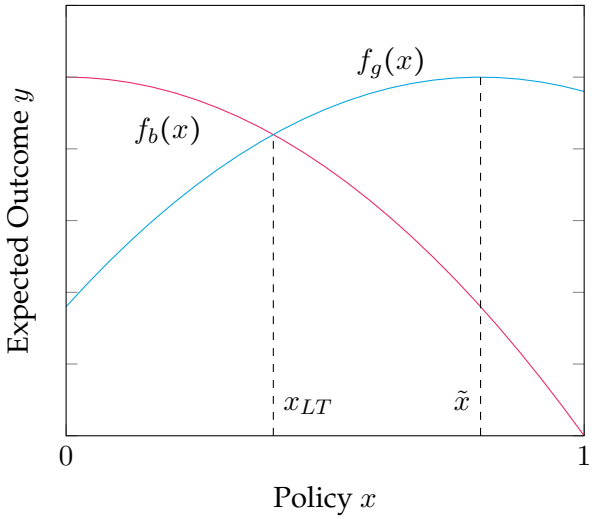
The people's flow utility is y_t. They discount the future at rate 0. This assumption simplifies the model to illuminate its key tradeoffs, but the insights are identical with nonmyopia or other frictions, studied in section 3.2.

Actions At $t = 1$, the leader holds power. At the beginning of each period the leader is in power, the people decide whether to incur a one-time cost $c > 0$ to overthrow the leader. If they overthrow, the people set x_t from t onwards. Overthrow succeeds with probability 1; the leader receives a loss $-L < u_\ell(0)$ and no further flow utility. If the leader is not overthrown, she sets x_t for that period and t moves to $t + 1$.

[5] We assume concavity for tractability; all results go through with $u_\ell(\cdot)$ strictly increasing using a piecewise concavification of $u(\cdot)$.

[6] Section 1.8 explores a fiscal microfoundation in the setting of the paper's Russian case study. Specifically, x represents both the degree of involvement of the gentry in intermediating taxation and access to private property. Increasing x reduces the degree of gentry intermediation in revenue collection, improving the share the government can keep for itself.

Figure 1.1: $f_g(x)$ and $f_b(x)$ graphed



States Agents possess uncertainty over two states of the world governing the relationship between x and y. In the "pro-leader" state, increases in x cause increases in y up to a single-peak $1 \geq \tilde{x} > 0$. In the "anti-leader" state, increases in x cause decreases in y:

$$y_t = -(x_t - \tilde{x})^2 + \epsilon_t \equiv f_g(x_t) + \epsilon_t \qquad \text{Pro-Leader}$$
$$y_t = -x_t^2 + \epsilon_t \equiv f_b(x_t) + \epsilon_t \qquad \text{Anti-Leader}$$

$\epsilon_t \sim \mathcal{U}[-\sigma, \sigma]$ is i.i.d. every period, making inference potentially imperfect; Proposition 7 of Section 1.8 shows a version of the paper's main result for general distributions of ϵ_t. f_g and f_b are graphed in Figure 1.1. [7]

The expected value of y conditional on $x = \tilde{x}$ is higher in the pro-leader than the anti-leader state, and $x = 0$ is better in the anti-leader than the pro-leader state. In particular this means that $f_g(x) - f_b(x)$ is single crossing at some $x_{LT} \in (0, 1)$ with $y_{LT} \equiv f_g(x_{LT})$.[8]

Inference Both agents follow Bayes' Rule and use histories $\{y_\tau, x_\tau\}_{\tau \leq t}$ to update a common belief over the two states. Denote q_{t-1} the belief in the pro-leader state at the beginning of time t. During

[7] All results hold if f_g and f_b are linear with positive and negative slope, respectively; as well as under moderate asymmetry, which we analyze later.

[8] We can also allow u_ℓ to depend on the state of the world; for the main results to go through, we simply need that the utility in the anti-leader state of the world is worse than x_{LT}, which is worse than in the pro-leader state of the world.

period t, agents observe a policy-outcome pair (y_t, x_t) and use it to update q_{t-1} to q_t, to be carried into $t+1$. Denote $\Delta(x) = \min\{|f_g(x) - f_b(x)|, 2\sigma\}$ the *effective difference* between the expectation of y conditional on x in each state. Bayes' Rule implies that the distribution of posteriors conditional on x_t and q_{t-1} is:

$$q_t|q_{t-1}, x_t = \begin{cases} 1 & \text{with prob. } \frac{\Delta(x_t)}{2\sigma} q_{t-1} \\ q_{t-1} & \text{with prob. } 1 - \frac{\Delta(x_t)}{2\sigma} \\ 0 & \text{with prob. } \frac{\Delta(x_t)}{2\sigma}(1 - q_{t-1}) \end{cases}$$

When $x_t = x_{LT}$, $f_g = f_b$, meaning $\Delta(x_{LT}) = 0$. x_{LT} shuts down information revelation ($q_{t-1} = q_t$ with probability 1) and we refer to it as a **learning trap policy**. [9] While extreme policies reveal more information and are preferable when uncertainty is resolved, x_{LT} is a "muddied policy" that mixes elements of these opposing extremes to hamper inference. Finally, we assume $c < x_{LT}^2$ so that the overthrow threat binds in equilibrium.

Sequential Game We describe the game recursively.

1. At time t, the people decide whether to overthrow the leader.
   (a) If the people decide to overthrow the leader:
      i. The people incur the cost of overthrow c. The leader loses power, receives $-L$, and receives no further flow utility.
      ii. The people set x_t, y_t realizes, and they receive utility y_t.
      iii. The people update q_{t-1} to q_t. Return to step 1(a)ii at $t+1$.
   (b) If the people decide not to overthrow the leader:
      i. The leader sets x_t, y_t realizes, and the people receive utility y_t. The leader in power receives $(1-\delta)u_\ell(x_t)$.

[9] The existence of x_{LT} only depends on the single-crossing property, and generalizes beyond uniform noise. If $\epsilon_t \sim p(\epsilon_t)$ for some distribution p :

$$q_t = \frac{p(y_t - f_g(x_{LT}))q_{t-1}}{p(y_t - f_g(x_{LT}))q_{t-1} + p(y_t - f_b(x_{LT}))(1-q_{t-1})} = \frac{p(y_t - f_g(x_{LT}))q_{t-1}}{p(y_t - f_g(x_{LT}))q_{t-1} + p(y_t - f_g(x_{LT}))(1-q_{t-1})}$$
$$= \frac{p(y_t - f_g(x_{LT}))q_{t-1}}{p(y_t - f_g(x_{LT}))} = q_{t-1}$$

ii. Agents update their belief q_{t-1} to q_t. Returns to step 1 at $t+1$.

We focus on pure-strategy Markov Perfect Equilibria with respect to the belief q_{t-1}. $x_\ell(q) \in [0,1]$ indicates the leader's policy plan. The people's strategy is given by $\{p(q), x_p(q)\}$. $p(q) \in \{\text{overthrow}, \text{nooverthrow}\}$ denotes the overthrow decision. $x_p(q) \in [0,1]$ indicates their policy plan once in power. $x_\ell(q)$ and $\{p(q), x_p(q)\}$ must best respond to each other in equilibrium.

1.3 Analysis

We first analyze the baseline model, highlighting the forces predicting nonmonotonic experimentation and policy moderation at intermediate beliefs. We then relax the people's myopia and introduce costly experimentation to illustrate how the model's insights generalize.

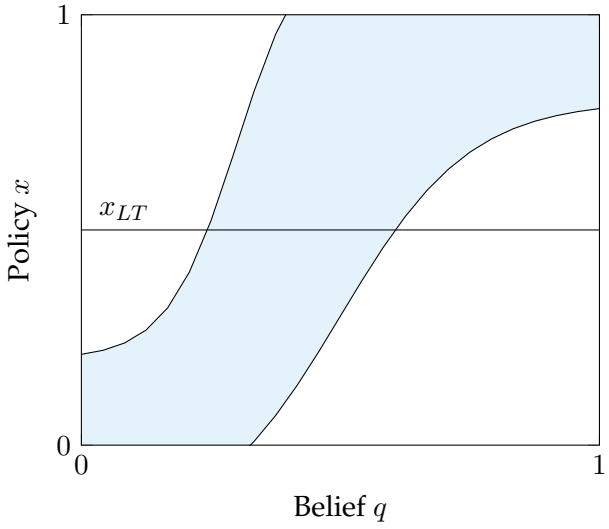
1.3.1 Analysis of Baseline Model

People The people's optimal policy, conditional on being in power, is $x_p(q) = \arg\max_x \mathbb{E}[y|x]$; expectations are taken with respect to q. The people have a strict incentive to overthrow if and only if

$$\max_x \mathbb{E}[y|x] - c > \mathbb{E}[y|x_\ell(q)],$$

i.e. if the best policy they can achieve is better than the leader's, less the cost of overthrow; in equilibrium, they overthrow if and only if the inequality above is satisfied strictly. For each q, we define $\text{NR}(q)$ as the set of policies that weakly prevents overthrow: $\text{NR}(q) = \{x_\ell(q) \in [0,1] : \max_x \mathbb{E}[y|x] - c \leq \mathbb{E}[y|x_\ell(q)]\}$. $\text{NR}(q)$ is monotone in q, in the sense that its minimal and maximal elements are increasing in q. Because x_{LT} delivers the same utility in both states of the world, there exist a range of interior beliefs where x_{LT} is contained in the interior of $\text{NR}(q)$. Its graph as a function of q is shaded below.

Figure 1.2: No-Overthrow Constraint $\mathrm{NR}(q)$



Leader's Solution Denote by $\overline{x}(q) \equiv \max\{\mathrm{NR}(q)\}$ the most extreme policy avoiding overthrow. At $q = 0$ or 1, policies have no information externalities, so the leader simply plays this extreme policy for all time. Denote $u_\ell(\overline{x}(0)) = \underline{u}$ and $u_\ell(\overline{x}(1)) = \overline{u}$, which hence represent the value of the leader's problem at these two points. Denote $u_\ell(x_{LT}) = u_{LT}$ and note that, by the assumptions on c, $\underline{u} < u_{LT} < \overline{u}$. This means that the leader's utility at $q = 1$ is preferred to implementing x_{LT} for all time, but that implementing x_{LT} for all time is *better* than being stuck at $q = 0$.

Since L is large, the leader will always want to avoid overthrow on the equilibrium path. This means that, for interior q, the Bellman Equation describing $x_\ell(q)$ is:

$$V(q) = \max_{x_\ell(q) \in \mathrm{NR}(q)} \underbrace{(1-\delta)u_\ell(x)}_{\text{Flow utility}} + \delta\Bigg(\underbrace{\frac{\Delta(x)}{2\sigma}\Psi(q)}_{\text{Truth revealed}} + \underbrace{(1 - \frac{\delta\Delta(x)}{2\sigma})V(q)}_{\text{Truth not revealed}} \Bigg),$$

where $\Psi(q, x') = q\overline{u} + (1-q)\underline{u}$ is the *expected value from revealing the truth*. When the leader implements a policy x, this policy potentially generates information, in addition to flow utility. With probability $q\Delta(x)/2\sigma$, $q \to 1$. With probability $(1-q)\Delta(x)/2\sigma$, $q \to 0$. In the former case, the leader's value function is $\overline{u}$ and in the latter it is $\underline{u}$. Otherwise, q is invariant.

The following proposition characterizes $x_\ell(q)$ for patient leaders.

Proposition 1. *Suppose $\sigma(1-\delta)$ is small or $\underline{u}$ is sufficiently low. There exists a threshold $\underline{q} \in (0,1)$ such that:*

1. *If $q \leq \underline{q}$, $x_\ell(q) = \min\{x_{LT}, \overline{x}(q)\}$. The leader plays the learning trap or the policy closest to it.*

2. *If $q > \underline{q}$, $x_\ell(q) = \overline{x}(q)$. The leader plays the maximal policy preventing overthrow.*

Proof. For $q \in (0,1)$, consider a problem where the leader is constrained at $q = 0$ or 1 but can otherwise implement any policy in $[\underline{x}(0), \overline{x}(1)]$, which can be rewritten as:

$$\max_{x \in [\underline{x}(0), \overline{x}(1)]} \frac{2\sigma(1-\delta)}{2\sigma(1-\delta) + \delta\Delta(x)} u_\ell(x) + \frac{\delta\Delta(x)}{2\sigma(1-\delta) + \delta\Delta(x)} \Psi(q)$$

This expression is a convex combination of flow utility $u_\ell(x)$ and the value $\Psi(q)$ of the truth. If the leader pursues a strategy where she implements x_{LT} for all time, it delivers a value $u_\ell(x_{LT}) = u_{LT}$. As x moves away from x_{LT}, weight shifts away from flow utility $u_\ell(x)$ and onto the expected value of revealing information $\Psi(q)$.

Suppose q is low, so that $\Psi(q) < u_{LT}$. If the leader implements $x > x_{LT}$, this results in higher flow utility relative to u_{LT}. But it also reveals information, whose expected value $\Psi(q)$ is worse than u_{LT}. If $\sigma(1-\delta)$ is small, the effect of flow utility is small, meaning that pursuing the learning trap for all time and receiving a modest payoff is better than gaining a bit of flow utility that reveals unfavorable information, up to a point $\underline{q}$.

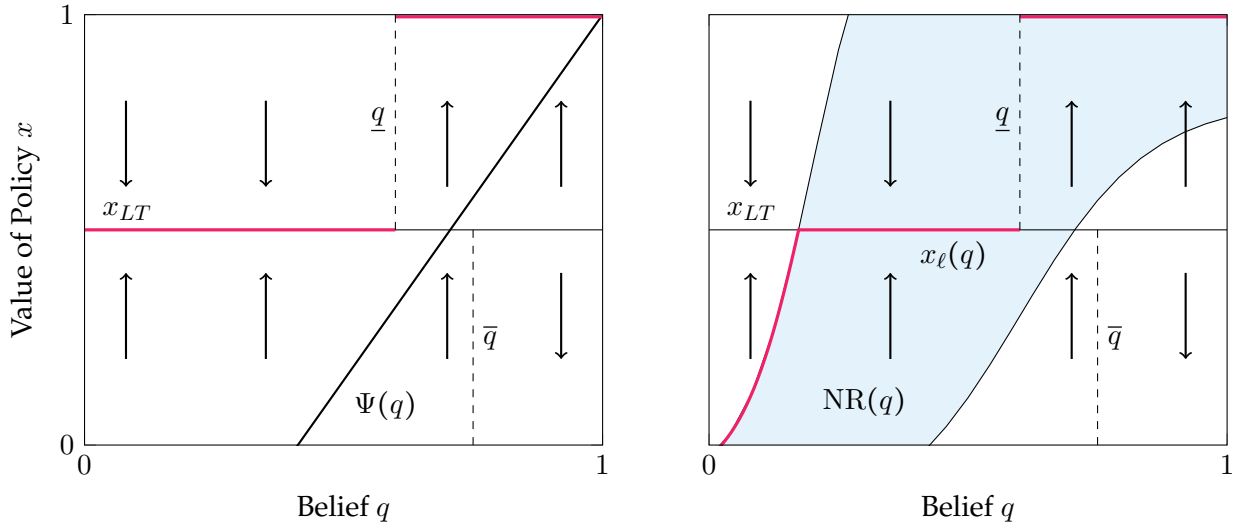
If $\Psi(q) \geq u_{LT}$, information is better in expectation than shutting down learning and receiving a moderate payoff. Pursuing $x > x_{LT}$ both increases flow utility and the weight on the revelation of favorable information. For high q, because information is favorable, $x < x_{LT}$ may counterintuitively be preferred to x_{LT} after a point $\overline{q}$; while this edge case does not bind in the baseline, it may in a more general setting.

The solution to this *unconstrained* problem is shown in the left panel of Figure 1.3 in red. In particular, we normalize $\overline{u} = 1$, $u_{LT} = x_{LT}$, and let $\underline{u} < 0$ so that $\Psi(q)$ is below 0 for low q. This panel also illustrates the thresholds $\underline{q}$ and $\overline{q}$. Below $\underline{q}$, the leader prefers the learning trap for both flow utility and informational purposes; between $\underline{q}$ and $\overline{q}$, prefers a high policy for flow utility purposes; and above $\overline{q}$, prefers any extreme policy to x_{LT} for informational purposes.

The right panel transposes the constraint set $\mathrm{NR}(q)$ (shaded) onto the utility space of the leader. By overlaying the forces from the unconstrained problem onto this constraint set, we are able to trace out the leader's solution $x_\ell(q)$ in the constrained problem (bold), which forces the leader to

pursue $x < x_{LT}$ for low q.

Figure 1.3: Forces Driving Optimal Policy in Unconstrained Leader's Problem Problem (Left) vs. Constrained Leader's Problem (Right)



□

The result also yields the following comparative statics.

Proposition 2. *Suppose σ decreases, δ increases for the leader, or $\underline{u}$ decreases. Then $\underline{q}$ increases.*

As σ decreases, for each x, the probability of information revelation increases. As δ increases, the leader deemphasizes flow utility.[10] In both cases, the leader is worried more about how her actions may reveal the truth, increasing attraction to learning traps. Finally, as $\underline{u}$ decreases, for all q, the value of revealing information decreases.

Comments Proposition 1 highlights a nonmonotonicity in policy experimentation. When the optimal policy is relatively certain, leaders pursue extreme policies. For high q, they implement their most preferred policies, and for low q, the highest policy keeping the people indifferent to overthrow. If the optimal policy is uncertain, leaders moderate policies to tide a threat of information leading to a *future* threat of overthrow — precisely when learning about the optimal policy would be most valuable. Notably, the people have a strict incentive to *retain* the leader at intermediate q, a contrast to dictator games where leaders always keep the people indifferent to overthrow.

[10] $(1-\delta)$ could represent the length of the leader's office tenure and the frequency of evaluations.

A repeated setting is crucial for the result. In a two-period model, a leader would always implement her preferred policy in the second period, meaning she would be overthrown only at low q. This would generate the opposite effect: leaders would implement x_{LT} at high q so as not to "rock the boat," and $x = 1$ at low q to "gamble for resurrection." The result would also not hold if the people received a constant outside option upon overthrow, such as in a setting with uncertainty over a leader's skill. Suppose y_{LT} were preferred to the outside option at $q = 1/2$. At $q = 0$, y_{LT} would still be preferred to the outside option. The leader could then achieve utility at least u_{LT} at $q = 0$ by repeatedly implementing x_{LT}, eroding the need to implement x_{LT} at higher beliefs.

If $\mathrm{NR}(q)$ is increasing in q and its graph convex, a version of this proposition can be shown for arbitrary distributions of ϵ; this is addressed in Proposition 7 of Section 1.8. In practice, and particularly without myopia, this set may not be convex or even connected, meaning standard solution techniques from dynamic programming cannot be used to analyze the problem.

1.3.2 Model with Nonmyopia and Adjustment Costs

We now allow the people in our model to be infinitely lived and discount the future at rate δ. The leader discounts the future at rate δ^ℓ. The people bear adjustment costs $\kappa(x_t, x_{t-1}) = \kappa|x_t - x_{t-1}|$, capturing frictions from institutional reform or costly experimentation.[11] Their utility is $u_p(y_t, x_t, x_{t-1}) = (1-\delta)y_t - (1-\delta)\kappa|x_t - x_{t-1}|$. Let $k \equiv (1-\delta)\kappa$.

We allow for asymmetry in the value of y_t between states:

$$y_t = \beta_g - (x_t - \tilde{x})^2 + \epsilon_t \equiv f_g(x_t) + \epsilon_t \qquad \text{Pro-Leader}$$

$$y_t = \beta_b - x_t^2 + \epsilon_t \equiv f_b(x_t) + \epsilon_t \qquad \text{Anti-Leader}$$

We assume:

1. $2k > |f_g'(x_{LT})|, |f_b'(x_{LT})| > k$;

2. $|(\beta_g - \beta_b)/2\tilde{x}| < k$;

[11] A model where adjustment costs do not depend on the size of an adjustment delivers identical results. Myopia, risk aversion, or replacing the leader with one who wants to minimize x_t could be used as alternative frictions. Forcing the leader to bear costs only strengthens the result.

3. $\tilde{x}^2/4 > 3|\beta_b - \beta_g| + 3(\beta_b - \beta_g)^2/\tilde{x}^2$;

4. $k \in \left(\frac{2\tilde{x} - \sqrt{\tilde{x}^2/4 - 3|\beta_b - \beta_g| - 3(\beta_b - \beta_g)^2/\tilde{x}^2}}{3}, \frac{2\tilde{x} + \sqrt{\tilde{x}^2/4 - 3|\beta_b - \beta_g| - 3(\beta_b - \beta_g)^2/\tilde{x}^2}}{3} \right)$.

These assumptions bound payoff asymmetry between the two states, and suggest policy adjustment should not be *so* costly that policy never moves, but not so cheap that the people excessively prefer experimentation. When $\beta_b = \beta_g$, this assumption reduces to $2k > \tilde{x} > 5\tilde{x}/6 > k$.

Markov Perfect Equilibrium is calculated with respect to the prior q and the previous period's status quo x'. We write $x_p(q, x')$ as the people's optimal policy conditional on overthrow. Full proofs for this section are provided in Section 1.7.

Solution to People's Problem We first solve for $x_p(q, x')$. We then use this to characterize the overthrow decision and $x_\ell(q, x')$. $x_p(q, x')$ is the policy function of the following Bellman equation:

$$W(q, x') = \max_{x \in [0,1]} \underbrace{(1-\delta)\mathbb{E}[y|x] - k|x - x'|}_{\text{Flow utility}} + \delta \Bigg(\underbrace{\frac{\Delta(x)}{2\sigma}\big(qW(1,x) + (1-q)W(0,x)\big)}_{\text{Truth revealed}} + \underbrace{\big(1 - \frac{\Delta(x)}{2\sigma}\big)W(q,x)}_{\text{Truth not revealed}} \Bigg)$$

The first term represents expected flow utility $qf_g(x) + (1-q)f_b(x)$ less costs $k|x - x'|$. The next terms represent continuation utility. With probability $\frac{\Delta(x)}{2\sigma}$, the truth is revealed. Conditional on this, the belief moves to 1 with probability q and 0 with probability $1-q$. With probability $1 - \frac{\Delta(x)}{2\sigma}$, beliefs remains at q; Today's x becomes tomorrow's status quo policy.

The following proposition describes the value function associated with $x_p(q, x')$, which captures a tradeoff between information and policy flexibility.

Proposition 3. *$x_p(q, x')$ is the solution to the following Bellman equation:*

$$W(q, x') = \max_{x \in [0,1]} \frac{2\sigma(1-\delta)\mathbb{E}[y|x] + \delta\Delta(x)\Phi(q,x)}{2\sigma(1-\delta) + \delta\Delta(x)} - k|x - x'|$$

where $\Phi(q, x) = qW(1, x) + (1-q)W(0, x)$. For intermediate values of c, there exists a nonempty open set $LT \subseteq [0,1] \times X$ with $LT \ni (1/2, x_{LT})$ such that for all $(q, x') \in LT$, $W(q, x') \leq y_{LT} + c \leq$

$W(0,x'), W(1,x')$.

As in the leader's problem in our benchmark, the people's utility is a convex combination of flow utility $\mathbb{E}[y|x]$ and revealing the truth $\Phi(q,x)$, less adjustment costs. Implementing the learning trap x_{LT} generates a constant value y_{LT}.

For intermediate q, the people face a tradeoff: extreme x generates information, but because there is uncertainty about whether high or low x is optimal, they may occur excessive adjustment costs by experimenting in one direction, learning that the optimal policy is in another direction, and then reversing course. This tradeoff means that the difference between $W(q,x')$ and the value of shutting down information y_{LT} is smaller at intermediate beliefs. At extremal q, there is no tradeoff: there is relative certainty about the effects of policies, meaning backtracking costs are not likely to be incurred in the future.

The last part of the proposition establishes bounds on c such that, at $q = 1/2$, if the leader pursues x_{LT}, she strictly prevents overthrow, while at $q = 0$ or 1, she cannot implement x_{LT} without being overthrown. This result suggests intuitive comparative statics: increasing policy adjustment or overthrow costs expands the set of beliefs and status quo policies at which it is not credible to overthrow a leader continually implementing x_{LT}.

Proposition 4. *Let $LT(c,k)$ parametrize the set of values which the learning trap is permissible. Fixing all other parameters, suppose either c or k increases to c' or k'. Then, $LT(c,k) \subsetneq LT(c',k)$ and $\subsetneq LT(c,k')$.*

Overthrow Decision Define $\tilde{W}_{x_\ell(q,x')}$ recursively as follows:

$$\begin{aligned}\tilde{W}_{x_\ell(q,x')}(q,x') &= (1-\delta)\mathbb{E}[y|x_\ell(q,x')] - k|x_\ell(q,x') - x'| + \frac{\delta\Delta(x)}{2\sigma}\tilde{\Phi}_{x_\ell(q,x')}(q,x_\ell(q,x')) \\ &\quad + \delta(1-\frac{\Delta(x)}{2\sigma})\tilde{W}_{x_\ell(q,x')}(q,x_\ell(q,x')) \\ \tilde{\Phi}_{x_\ell(q,x')}(q,x) &= q\tilde{W}(1,x_\ell(1,x)) + (1-q)\tilde{W}(1,x_\ell(0,x))\end{aligned}$$

This represents the *people*'s value from accepting a leader's policy plan $x_\ell(q,x')$. Lemmata 7 and 8 in Section 1.7 provide a closed form for this expression and show that, in equilibrium, this strategy has the property that $x_\ell(q,x) = x_\ell(q,x_\ell(q,x))$. If the leader's strategy is only to implement x_{LT} for all time, the value of this objective is $y_{LT} - k|x_{LT} - x'|$.

Let $\mathrm{NR}(q, x')$ be the set of policies that weakly prevents credible overthrow:

$$\mathrm{NR}(q, x') = \left\{x \in [0,1] : \tilde{W}_x(q, x') \geq W(q, x') - c\right\}.$$

The people overthrow the leader if and only if $x_\ell(q, x') \notin \mathrm{NR}(q, x')$.

Leader's Problem Define $\underline{x}_{LT}(q, x')$, $\overline{x}_{LT}(q, x')$, $\underline{x}(q, x')$, and $\overline{x}(q, x')$ as follows:

$$\underline{x}_{LT}(q, x') = \max\{x \in \mathrm{NR}(q, x') : x \leq x_{LT}\} \qquad \overline{x}_{LT}(q, x') = \min\{x \in \mathrm{NR}(q, x') : x \geq x_{LT}\}$$

$$\underline{x}(q, x') = \min\{x \in \mathrm{NR}(q, x')\} \qquad \overline{x}(q, x') = \max\{x \in \mathrm{NR}(q, x')\}$$

$\underline{x}_{LT}$ and $\overline{x}_{LT}$ correspond to the *closest* policies to the learning trap preventing overthrow. $\overline{x}$ and $\underline{x}$ are the lowest and highest policies preventing overthrow altogether.

Precisely as in the benchmark, at $q = 0$ or 1, the leader will pursue the most extreme policy preventing overthrow, i.e. $\overline{x}(q, x')$. Using the assumption on c, we again have:

$$\underline{u}(x') \equiv u_\ell(x_\ell(0, x')) < u_\ell(y_{LT}) < u_\ell(x_\ell(1, x')) \equiv \overline{u}(x')$$

The Bellman describing $x_\ell(q, x')$ is then:

$$V(q, x') = \max_{x_\ell(q,x') \in \mathrm{NR}(q,x')} (1 - \delta^\ell) u_\ell(x) + \frac{\delta^\ell \Delta(x)}{2\sigma} \Psi(q, x) + \delta\left(1 - \frac{\delta^\ell \Delta(x)}{2\sigma}\right) V(q, x)$$

where $\Psi(q, x') = q\overline{u}(x') + (1-q)\underline{u}(x')$. The following theorem characterizes $x_\ell(q, x)$ for patient leaders: the leader is either attracted to policies as close to x_{LT} as possible — minimizing information revelation — or an experimental policy that reveals information.

Theorem 1. *Suppose $\sigma(1 - \delta^\ell)$ is small or $\underline{u}(x')$ is sufficiently low. Then, there exist thresholds $\underline{q} < \overline{q} \in (0, 1)$ such that the following hold:*

1. *If $q \leq \underline{q}$, then $x_\ell(q, x') = \underline{x}_{LT}(q, x')$ or $\overline{x}_{LT}(q, x')$. The leader plays the* closest policy *to the learning trap.*

2. *If $q \in [\underline{q}, \overline{q}]$, then $x_\ell(q, x') = \underline{x}_{LT}(q, x')$ or $\overline{x}(q, x')$. The leader plays either the lowest policy below the learning trap or her highest feasible policy.*

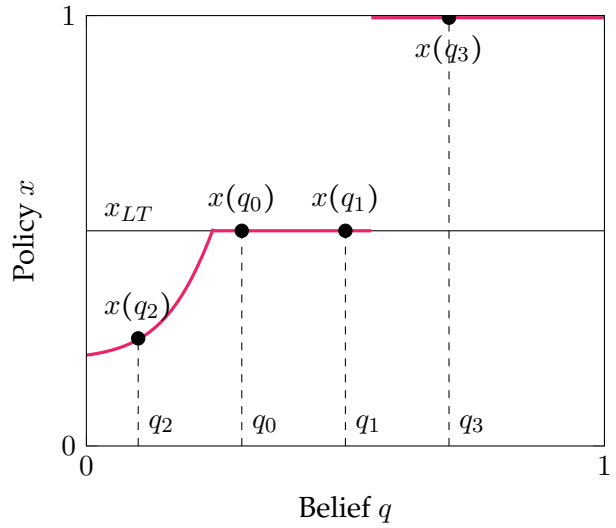
3. *If $q > \overline{q}$, = then $x_\ell(q, x') = \overline{x}(q, x')$ or $\underline{x}(q, x')$ (and it is necessary that $\Delta(\underline{x}(q, x')) > \Delta(\overline{x}(q, x'))$). The leader either plays her highest feasible policy or policies maximizing information revelation.*

This general formulation shows that the central forces of the model are identical to the baseline: leaders prefer learning traps at intermediate and low beliefs and experiment at high beliefs. At some beliefs, an inability to pursue learning traps without overthrow may generate additional nonmonotonicities. Comparative statics with respect to σ, δ^ℓ, and $\underline{u}(x')$ follow as in the baseline.

1.3.3 Unanticipated Information Shocks

Comparative statics with respect to q illustrate the effect of *unanticipated* shocks to beliefs in the model — which may represent sudden events like earthquakes, pandemics, or conflicts that test institutions and force leaders to experiment, thereby revealing information. Figure 1.4 shows how shocks to q at the end of every period affect policy when the status quo is $x' = x_{LT}$, shown below in Figure 1.4 as q_0 moves either to q_1, q_2, or q_3. With a slight abuse of notation, we write $x(q) = x_\ell(q, x_{LT})$.

Figure 1.4: Effect of Unanticipated Shock to q on Policy



If q_0 moves slightly to the right to q_1, there is no policy change even though this information favors the leader. If q_0 drops to q_2, the leader is still attracted to x_{LT}, but must play a lower policy to prevent overthrow. If q_0 jumps to q_3, policy experiences a discontinuous jump from x_{LT} to 1, since information revelation now swings in the leader's favor. Hence, beginning at intermediate

beliefs, policy variation is minimal under smaller or negative shocks to low or intermediate beliefs, but discontinuously shifts to experimentation when large shocks favor the leader.

1.4 Application: The 1861 Reforms as a Learning Trap in Imperial Russia

This section illustrates how the model's insights can be used to understand large-scale economic reform by applying its insights to the institutionalization of the peasant commune in Late Imperial Russia. Prior to 1861, the Russian economy operated under "serfdom": an agricultural land tenure system where peasant farmers worked nobles' land as part of communes overseen by a seigneur. However, faced with a need to modernize its economy in the 1860s, the Russian government considered liberal, Western European style economic reform. By instituting private property rights and allowing peasants freedom to move from their lands, a successful liberal reform would increase the government's fiscal revenue, accelerate Russia's industrial revolution, and encourage economic growth. However, it worried liberal reform was inappropriate to Russia's historical institutions and could backfire, fiscally damaging the Empire and leading to its overthrow.

I argue that, in response to this policy uncertainty, the Russian government pursued a middle ground that mixed elements of both serfdom and liberal institutions. I argue that this intermediate policy can be seen as a "learning trap" that, by mixing these two extremes, obfuscated information about the effects of liberal institutions to tide threats of overthrow. I contrast this learning trap with two other equilibria of the model corresponding to policy experimentation under extremal beliefs — serfdom pre-1861 and experiments with liberal reform after 1905.

Model Mapping To map the model to Imperial Russia, we identify policies and players. One extreme policy is serfdom: The other is a liberal, private property system where peasants own their own land and control individual labor and production decisions. The "people" are a combination of the Russian peasantry and elites. Outcomes y represent aggregate economic wellbeing. The "leader" is the Russian Tsar's government and, for *fiscal* reasons detailed in section 4.1, prefers liberal policies.

In one state of the world, liberal reform is better for Russia's economy than communal serf-

dom; endowing peasants with land and allowing freedom of movement and production decisions is better than serfdom. The leader's utility is higher: liberal reform increases fiscal revenue. In the other state, the government would be better off under serfdom, concentrating administrative power with elites and limiting freedom, since liberal reform would spread the government thin between compensating the gentry for land expropriation and funding state capacity. The government's utility — measured as fiscal revenue — would be worse under serfdom than under successful liberal reform.

The first subsection describes how a pre-1861 status quo of serfdom was questioned after informational shocks in the 1850s created uncertainty. The second subsection shows how a middle ground mixing serfdom and liberal reform from 1861-1905 represented a learning trap in response to this uncertainty. The third subsection argues that, after another large shock in 1905, the government experimented with liberal reform, illustrating a third case of the model. The final subsection addresses two prominent alternative explanations for the learning trap — elite power and state capacity.

1.4.1 First Equilibrium: Serfdom Pre-1861

Prior to 1861, most Russian peasants provided services or payments to the state or a noble in exchange for use of arable land (peasants working noble land were "serfs").[12] Officials oversaw a "commune" of households in charge of repartitioning land, solving disputes, and managing resources.[13] Many peasants possessed limited control over production decisions, had few rights, and could be tied to their land in the event of sale[14] (Vasudevan (1988, 209), Nafziger (2010, 382), Nafziger (2016, 775-776) Dennison (2014, 252-257)).

While there were doubts about serfdom's viability, they had not "challenged the basic structure of local institutions" prior to the 1850s (Starr, 2015, 52). A minister of internal affairs contrasted Russia's "uninterrupted tranquility" under serfdom with the "[i]nternal discord and revolts" plaguing Europe's experiments with liberalism (Polunov et al., 2015, 87); moreover, the

[12] Peasants rarely owned land, but both before and after 1861 owned small pieces of private property. There was heterogeneity in this arrangement; Nafziger (2012) shows how peasants on State-owned lands possessed some more individual property rights than serfs.

[13] Additionally, a failure of one commune member to pay taxes could result in punishment for others; or, the purchase of new land for the commune would make everyone collectively liable for paying mortgages on this land.

[14] For example, if noble A sold land to noble B, the serfs working that land would then be under the tutelage of noble B.

Russian countryside was not beset by an endemic economic crisis (Moon, 2014, 19-22). Russia hence represented an extremal policy (serfdom) under lower policy uncertainty.

Crimean War and Policy Uncertainty Russia's defeat in the 1853 Crimean War crystallized a need to modernize Russia's economy; many argued Russia lost the war due to economic backwardness. I interpret the defeat as a shock that generated uncertainty about whether serfdom was inferior to Western European economic reform. Debates about economic reform circulated, for the first time, in public discourse (Starr, 2015, 57). Many educated Russians believed private economic activity was superior to the commune (Starr, 2015, 53-63), while other nobles favored a Manchesterian arrangement where individual peasants would rent and work nobles' land "on the basis of free competition" (Khristoforov, 2009, 65). Views that communal property was superior to private property due to "Russian exceptionalism" were derided by some economists (Kingston-Mann, 1991, 23-24).

Others maintained that serfdom and the commune were superior. Slavophilic ideologues romanticized the peasant as a self-sufficient "bearer of his own culture" (Khristoforov, 2009, 62) and argued the commune should be left untouched. Some believed private property could lead to a countryside marred with disorder, requiring seigneurial intervention to prevent exploitation of commons (Pravilova, 2014). Others contended that peasants were not aware of issues with the pre-1861 system (Field, 1976, 53), even arguing that peasants preferred communal land (Kingston-Mann, 1991, 34).

Regime's Preferences The Russian regime, headed by Tsar Alexander II, likely preferred a liberal path to modernization. A widening military gap with its neighbors was attributed to the stifled emergence of Russian industry due to the incentive structure of serfdom (Zenkovsky (1961, 284), Emmons (1966, 48-50)), as well as an inability to draw up army rseerves from the peasantry (Moon, 2014, 53-54). Abolishing serfdom could increase worker mobility and productivity, generating industrial growth that could fuel military prowess (Pereira, 1980, 108). Crucically, the Tsar believed the government should "begin eliminating serfdom from above [rather] than to wait until it begins to eliminate itself from below" (Pereira, 1980, 104).

A core motivation for preferring a liberal system was its fiscal efficiency and capacity to accel-

erate the development of state institutions (Dennison, 2023, 642). Under serfdom, tax collection was inefficiently outsourced to nobles who siphoned surpluses (Dennison, 2020,2). Liberal reform would weaken the nobility's bite, increasing revenue for the State. (Moon, 2014, 33).[15] If liberal reform could *substantially* increase surpluses, the government would be in a better place, even if it had to compensate elites for dispossessing their land. Finally, reformers hoped to use fiscal revenues to increase state capacity and fund the expansion of corporate forms and military ventures (Dennison, 2023, 642). The "Additional Results" section develops a microfoundation for these fiscal preferences.

1.4.2 Second Equilibrium: Learning Trap, 1861-1905

This section shows how the post-1861 emancipation reforms and the codification of the commune can be viewed as a learning trap. I show how the government feared liberal reform might backfire, and how an intermediate policy between serfdom and liberal reform was conceived as a response to this fear, distorting economic incentives and halting learning about policy efficacy. Indeed, the relative intermediacy of the reform can even be thought of through the lens of unanticipated information revelation in section 3.3: the Crimean shock favored liberal reform, but since Russia began in a situation where serfdom was relatively accepted, the Tsar's policy did not radically swing in the direction of liberal reform.

Fears of Information Revelation Despite ideological or fiscal preferences for granting peasants economic freedom, Alexander II's government worried liberal policies could fail to *sufficiently* grow the Russian economy, viewing "communal land tenure and mobility restrictions [as] necessary precautions" in the face of this uncertainty (Dennison, 2014, 62). External validity arguments that Westernization was inappropriate for Russia accompanied fears of "the rise of an industrial proletariat, ruthless competition, unemployment, and other evils of industrial capitalism" (Polunov et al., 2015, 61).

Policy uncertainty was linked to fiscal uncertainty. Successful reform would fuel state capacity and the economy. In one state of the world, economic freedom would not only increase peasant income — via increases in agricultural productivity and the takeup of manufacturing jobs — but

[15] Section 4.4 more carefully examines the nobility's impact on the reform process.

also ease revenue collection, generating fiscal growth even after compensating the nobility. However, the government was concerned that, in a non-ideal state of the world, alienation of rural land could limit its ability to collect the revenue necessary to compensate the nobility for expropriation, stressing state capacity (Dennison, 2023, 644). One of Alexander II's committees concluded that "the Order of the State might be shaken" by an emancipation that did *not* tie peasants to land (Field, 1976, 75), suggesting that the commune's institutionalization was seen as a solution to the ambiguous fiscal effects of reform.

Terms of Emancipation The Tsar's emancipation reforms beginning in 1861 ended serfdom and freed peasants from the coercive status-quo of seigneur-peasant agriculture. The reform impacted serfs working private lands more than state peasants, whose institutional constraints arguably changed less post-1861 (Nafziger, 2012). While broadening civil rights (Dennison, 2014), the reform did *not* remove the peasant commune and counterintuitively used it to constrain household labor mobility and capacity to acquire private property.

Nobles were compensated by the State for relinquishing land that was then redistributed between members of peasant communes. Ex-serfs made mortgage-like redemption payments to the State over 49 years to finance nobles' compensation. After payments were made, ex-serfs would privately own their allotted land but were otherwise collectively responsible for redemption and tax payments. Households could not legally sell, lease, transfer, alienate, or use land as collateral *until* redemption payments were complete, restricting freedom to abandon agriculture (Dower and Markevich (2019, 234), Nafziger (2012, 5-6), Nafziger (2016, 776), (Dennison, 2014, 259)).

Intermediate Policy The 1861 reforms mixed communal serfdom with liberal institutions. An external seigneur no longer regulated economic decisions. Peasants could privatize communal land by paying off redemption obligations (Nafziger, 2010, 383). Peasants owned houses, gardens, and rights to output (Kingston-Mann, 1991, 35). Communes (and sometimes households) even engaged in market land transactions with formal credit backing (Nafziger, 2010, 384). However, the persistence of the commune and its codification by the State clashed with individual economic incentives. Land allotments, repartitions, and redemption obligations were enacted communally, sometimes against households' wills. Peasants owned scattered strips of land, at best

allowing for diversification, at worst increasing travel times and disincentivizing technological experimentation, slowing agricultural productivity growth (Williams (2013, 52-53, 65-66), Dower and Markevich (2019, 243)).

Mobility Restrictions The commune restricted labor mobility and supply by controlling the issuance of documents to pursue outside work (Nafziger, 2012, 7). Because households were communally responsible for redemption payments, those with larger obligations were incentivized to enforce mobility restrictions and free-ride on others. Nafziger (2010) notes: "Income-maximizing peasant households may have been unable to freely allocate resources between agricultural production and non-agricultural activities if they were subject to collective decision-making... By potentially inhibiting land transfers... and restricting the allocation of labor outside the village, the communal system may have introduced wedges between the shadow and market values for these factors of production" (384). Notably, these frictions stifled the flow of labor to cities, dampening the emergence of Russia's industrial sector and contradicting one of the original motives for reform.[16]

Sustained Uncertainty Over the decades, elites' and the State's inability to conclude whether the commune would have been better than individualized agriculture suggests that the post-1861 middle-ground sustained policy uncertainty. Despite the commune's apparent discouragement of technological adoption, Kingston-Mann (1991) documents that when innovation did occur, it *spread* more quickly in districts with greater communal intensity. While many in the government still believed in the pursuit of a liberal, Western path, one sources cites a government official,"an enthusiastic critic of the commune" (Kingston-Mann, 1991, 49), who conceded that the commune's ability to provide mutual aid during emergencies was a benefit that could not be realized under private property. Decades later, official contended that limited cases where peasants were permitted to sell land caused exploitation of peasant decision making (Wcislo, 2014, 89), casting doubt about the appropriateness of markets that would accompany liberal reform. Crucially, the *general equilibrium* effects of reform likely differed from the partial equilibrium experience of individual villages toying with private incentives. Regional heterogeneity in the enforcement of commu-

[16]Indeed, Nafziger (2010) argues that weaker communal structures around Moscow allowed this area to develop a stronger industrial sector than other parts of Russia.

nal institutions allowed some households to observe the power of private incentives or accumulate wealth by participating in land rental markets or full-time non-agricultural labor (Nafziger, 2010, 385). Nevertheless, the contemporary remarks above suggest high-level uncertainty about whether liberal reform was surely better, even decades after 1861.

Implications Many contemporaries felt the Tsar's reforms did not go far enough. Finkel, Gehlbach, and Olsen (2015), for example, show a marked increase in peasant disturbances after 1861 due to the tepidness (relative to expectation) of the reform. More puzzlingly, in practice,"[f]ew surpluses could be wrung from peasants who were already obligated to repay the nobility for their freedom and their land and who continued to face obstacles to the most efficient allocation of their resources" (Dennison, 2023, 644-5). These realities make the choice to maintain the commune odd; if liberal reform could certainly result in large economic gains, the State would have been better off fiscally than in a middle ground where both peasant surplus and the ability to pay the nobility seemed constrained. But under policy uncertainty, fears that information revelation could strain the government's coffers and reputation were it to backfire give an informational explanation for this middle-ground.

From early on, the Tsar's government explicitly believed that a system that ending many of serfdom's abuses but still tying peasants to agricultural land could mitigate informational threats posed by more liberal reform. By sustaining uncertainty, a learning trap minimizes information revelation that could lead to overthrow. Not only did the Tsar's government hold power for decades, but except for the spike in disturbances following the Emancipation reform, reported peasant disturbances remained low in the following years (Finkel et al., 2015, 1000). In this sense, an "intermediate" policy, despite being inefficient in the interim, was successful in tiding threats of future overthrow from either radical liberal reform or allowing serfdom to remain.

However, this arrangement may have led to persistent and negative side-effects. Beyond stifling Russia's industrial sector and potentially slowing agricultural adoption in the medium term, Markevich and Zhuravskaya (2018) go as far as to argue that the allocation of property rights to communes instead of households dampened Russian economic growth into the 20th century. [17]

[17]Some historians point to redemption payments as the first order inefficiency impeding livelihood in post-1861 ex-serf communities. Interestingly, Gerschenkron (1962) suggests that the commune *worsened* the bite of redemption payments. Households with particularly large debt obligations would have an incentive to free-ride on other members

1.4.3 Third Equilibrium: Stolypin Reforms, Post-1905

Few radical modifications to the post-1861 arrangement occurred over the ensuing decades until the reforms headed by Prime Minister Pyotr Stolypin beginning in 1906, which represent experiments with private property.

These reforms were motivated by two forces which represent a large shock to beliefs in favor of liberal reform. The first was the culmination of investigations into the affairs of the peasantry led by Sergei Witte, the Russian Minister of Finance, whose findings — along with independent recommendations of the Ministry of Internal Affairs — favored the privatization of property and "individualized farming" (Yaney, 1964, 278). The second was the Russian Revolution of 1905 — peasant uprisings that led to the establishment of (limited) consitutional monarchy (Polunov et al., 2015, 223) and the suspension of households' redemption payments, making the reformation of social order — changing the relationship of the peasantry to the land — necessary to complement these new political institutions; indeed, the reforms existed "within the context of other legislative and organizational measures that Nicholas II's government introduced" (Pallot, 1999, 3). The confluence of these forces meant that "a sufficient number of officials in Nicholas II's government shared this conviction [that individualized farming should be introduced into the countryside] for it to find its way into official state policy" (Pallot, 1999, 4). This shock can also be viewed through the lens of unanticipated information shocks, as in section 3.3: a large shock in favor of liberal reform finally motivated a sharp shift in experimentation.

Stolypin's 1906 Reform gave households the choice between holding property in private or communal tenure. Households could leave the commune while maintaining allotments or an equivalent parcel of land, overseen by a state-mediated bargaining process. The dissolution of the commune's power was accompanied by a reduction in the enforcement of mobility restrictions. Stolypin indeed believed that peasants who found themselves landless would find industrial work in the cities. From 1906 until the outbreak of the First World War, over one third of peasant households withdrew from the commune (Polunov et al., 2015, 231-32).

Stolypin's reforms were seen as experiments. The means of achieving land enclosure and alignment of rural social order with politically liberal institutions evolved as officials learned

of the commune, affecting commune enforcement of labor mobility and production incentives which undoubtedly affected households' very ability to make payments.

about peasants' willingness to take up the reform and grappled with administration (Yaney, 1964, 275-276, 283). The culmination of the process would, ideally, result in economic growth for Russia, and "[o]nce enlightened through education and example to the possibilities of organizing their farms differently, it was thought that the majority of Russia's peasants would greet the reform enthusiastically" (Pallot, 1999, 5), creating "an independent, prosperous husbandman, a stable citizen of the land" (1).

The Stolypin reforms represent a third and final case of the model, where experimentation with liberal reform occurs only after a shock sufficiently favors liberal reform. While the eventual course of the reform was cut short by the First World War, many argued that these reforms put Russia on the path towards economic development in imitation of the Western European model. Dower and Markevich (2019) argues that, indeed, the reform had a positive effect on the productivity of land-use in its short lifespan.

1.4.4 Alternative Interpretations for the Learning Trap

While Stolypin's experiments and pre-1862 serfdom can easily be seen as extreme, informative policies, the literature discusses two main *non-informational* arguments motivating the Tsar's 1861 reforms. First, this mixture may have balanced the desires of elites with those of the state. Second, Russia's lack of state capacity may have prevented more radical reform. In this section, I show that while these factors undoubtedly affected the shape of the emancipation reforms, the threat of informational revelation still affected the post-1861 policy outcome.

Elite Power The landed nobility's commitment to serfdom constrained previous attempts to abolish serfdom. Could the post-1861 reforms have simply been an outcome of bargaining between a liberal State and pro-serfdom nobles?

The commune may have been logistically preferable for dealing with civil and property disputes (Khristoforov and Gilley, 2016, 12). Labor allocation and rental payments may have been easier for landlords to organize when serfs were held collectively responsible (Dennison, 2014, 257). Landlords often disapproved of arrangements forcing them to give land to peasants, and extensively bargained over the size of government compensation during the years of the reform (Moon, 2014, chaps. 7-8). Nevertheless, the constraint binding the government *to the commune* was

far weaker in the 1850s than decades prior.

Moreover, nobles, especially in the 1850s, were ideologically diverse, and may not have overwhelmingly wanted to keep the commune. In fact, many nobles' preferences aligned with the liberal aspirations of the Tsar. Most educated Russians believed that "the sanctity of private property was the basis of the political and social order" (Field, 1976, 58). Some liberals supported abolishing serfdom and endowing peasants with land. Even some conservative nobles wanted peasants to rent and work noble land in a competitive market. Many critics of the post-1861 system — at times more critical of serfdom than the government's own reformers — contrasted Russia's system of land tenure with the more desirable British case, "where large-scale land tenure and local self-government controlled by the aristocracy allegedly guaranteed economic prosperity and political stability" (Khristoforov, 2009, 58, 63-65). Khristoforov (2022) goes as far as to argue that some form of "private property individualism" was in the interest of both landlords and the government.[18]

Finally, a sufficient bloc of the nobility may not have wanted to challenge the Tsar; they "did not want political change... They wanted the tsar's favor... At each stage in the development of the government's program, the nobility swallowed its objections or stated them obliquely" (Field, 1976, 362).

State Capacity State capacity undoubtedly constrained the Tsar's reform ambitions. Emancipating and endowing peasants with property, compensating nobles, setting up a cadastre, and securing property rights was expensive. Depressed stock prices and banking crises in the 1850s drained coffers (Dennison, 2020, 197), meaning the government had to look to the gentry for help. Landlords then played a large role in commune politics after emancipation (Dennison, 2020); outsourcing dispute resolution and tax collection to the commune was logistically easier than protecting individual household rights (Khristoforov (2022, S163), Dennison (2011)).

Puzzles still arise from ascribing the intermediacy of the reform solely to weak administration. First, before and after the 1850s, the government could have relaxed the commune's power on State lands, where elites had a much weaker hold. Instead, the institutional experience of the State peasantry faced few changes before and after 1861, and outcomes of ex-Serf and State peasants

[18] One reason was that a liberal system could oppose socialism.

converged by 1900 (Nafziger, 2012). Second, the State made few attempts to demarcate property. Officials wanted to tie peasants to the land, knew private property was easier to abandon through sale or mortgage, and in fact refused to separate the tax liabilities of some peasants who paid off their redemption obligations early (Khristoforov, 2007, 30). Dennison (2006) shows how the post-1861 commune sustained a pre-serfdom informal economy of social networks and patronage, even though the state could have slowly introduced "universal enforcement of contracts and property rights" (89).

The state's nebulous attitude towards demarcation emanated in part from an informational desire to limit market activity. The results of a project attempting to construct a land cadastre prior to 1861 under Kiselev were never clearly revealed or publicly accessible, likely because the results may not have favored rationalization of property (Khristoforov and Gilley, 2016, 9, 11). Fourth, while the Tsar's committees "recognized [the commune] as a temporary fact" (Khristoforov, 2007, 30), Alexander II never took further steps towards dissolution of the commune, and his son infamously strengthened its enforcement powers [19] (Wortman, 1989, 21). Finally, the reforms were intended to *increase* fiscal revenue and expand state capacity. If liberal reform were truly beneficial for peasant income, maintaining the commune seems surprising. However, the government's belief that liberal reform could have shrunk agricultural output and caused land alienation suggests that uncertainty played a crucial role in the post-1861 mixture.

The power of a critical mass of the nobility, idiosyncratic financial issues, and administrative costs partially bound Russia's 1861 reforms, but the first was relaxed by the late 1850s, the second does not explain the uniformity of the State peasant experience, and the third does not explain why Russia experienced little liberal movement for decades, especially if the reform had fiscal objective. These limitations can be addressed by viewing the 1861 reforms as being in part driven by the risk of revealing information about policy efficacy.

[19] For example, "[a]dditional legislation in 1893 explicitly forbade all sales of allotment land to non-peasants" and required a two-thirds majority of communal assembly votes to redeem or release landholdings and increased the portion of communal assembly votes (to two- thirds) required for a household to obtain release from their share or to redeem their land individually (Nafziger, 2010, 393).

1.5 Anticipated Information Revelation

This section extends the baseline model to show that, under threats of *anticipated* external information revelation, experimentation increases from both a direct channel (exogenous information) and an indirect channel (the leader herself is more likely to prefer experimentation, increasing learning). I show how this insight allows us to understand the interplay of policy contagion — where information about the effects of policies spreads from one polity to another — and political fragmentation. Specifically, I interpret 19th century Europe as a region where the probability of external information revelation was high and China as a region where it was low, and argue that the model can explain why European nations experimented with indirect rule and experienced high political turnover in the process, while China did not. Proofs of this section, as well as an additional extension showing the robustness of the model to multiple states of the world, are provided in Section 1.8.

1.5.1 Information Revelation

Information revelation may be *anticipated* when leaders are consciously aware of external sources of inference. Officials in regions plagued by earthquakes or hurricanes may foresee that these disasters will eventually occur and reveal information about disaster relief policies. Another application is policy contagion, where a group of countries with similar institutions can learn from their policy experiences. Suppose that all these countries are stuck in a learning trap except for one, which commits to implement informative policies. Neighboring countries use these informative policies as an additional source of policy inference, meaning their leaders now face an additional threat of information revelation from external sources.

I model anticipated information shocks as follows. Let $\eta \in (0,1)$ be small. With probability η at the end of every period, the true state of the world is exogenously revealed. An unconstrained version of the leader's problem, anticipating the exogenous revelation of information, can be written as:

$$\max_{x\in[0,1]} \frac{2\sigma(1-\delta)u_\ell(x) + \delta((1-\eta)\Delta(x) + 2\sigma\eta)\Psi(q)}{2\sigma(1-\delta) + \delta((1-\eta)\Delta(x) + 2\sigma\eta)}.$$

Note that when $x = x_{LT}$, the leader's value is now

$$\frac{(1-\delta)u_{LT} + \delta\eta\Psi(q)}{(1-\delta)+\delta\eta}$$

Implementing x_{LT} is no longer a "safe option" for the leader that shuts down learning. Anticipating that she can no longer *fully* control information, she may take extreme actions to enjoy more flow utility. This intuition is reflected in the following proposition.

Proposition 5. *Suppose η increases. Then, $\underline{q}$ decreases and $\overline{q}$ increases. Fixing x', the leader plays the learning trap for fewer values of q.*

Anticipated information revelation improves policy learning compared to the unanticipated or baseline model. While there is mechanically more information revelation due to the realization of shocks, because $\underline{q}$ decreases, the leader's commits to experimental policies for a larger set of values, meaning learning also increases through an *endogenous* channel. As learning increases, q is more likely to move to 0. This means that fragmented regions plagued by policy contagion will see more governments conceding to their people — e.g. movements from absolute to constitutional monarchy — or, insofar as lower x is also a shorthand for increased external threats of overthrow, leaders' terms will be shorter, corresponding to more frequent political turnover.

1.5.2 Political Fragmentation in Europe and China

The model's prediction — politically fragmented regions of the world exhibit larger policy variation, more policy experimentation, and more frequent political turnover — is consistent with stylized historical facts differentiating Western Europe and China. It explains how the spread of constitutional monarchies and representative democracies in fragmented Western Europe contrasted with the consistent stability of consolidated Imperial China. Because systems of indirect rule were not widely experimented with by Eurasian states until the 1700s, we view our "uncertain policy" as the optimal degree of democracy or autocracy. The policy space corresponds to direct rule (e.g. monarchy) on one end and indirect rule (e.g. republican democracy) on the other end. Constitutional monarchies can be thought of as situations where autocrats begrudgingly cede power to the people but remain in office.

European countries around 1800 had a high η; Europe was populated by dozens of sovereign states — on average 85 between the years 1000 and 1799 (Dincecco and Wang, 2018)) — and thus possessed rich sources of policy information. England's "Glorious Revolution" at the end of the 17th century marked Europe's first constitutional monarchy (Kurian, 2011) and one of the first sources of data on indirect rule. After the French Revolution, rulers began to actively worry about citizens' desires for indirect rule (Evans and Von Strandmann, 2002, 10) as demands for freedom of speech, press, and disposal of private property became more prominent (Sperber, 2005, 66-67). By the first half of the 19th century, both Britain (constitutional monarchy) and America (representative democracy) were providing information on the efficacy of indirect rule. As these informational pressures strengthened, by 1850, most European states shifted to some form of government partially controlled by an elected legislature (Tilly, 1989). Some governments peacefully ceded power, while others had power wrested from them as part of the "Age of Revolutions." A threat of external information revelation forced governments to experiment with indirect rule, increased learning via convergence to indirect rule, and led to political upheaval.

In contrast, China was ruled by a single, unified State uninterrupted for a millennium. No states shared China's scale or political structure, suggesting it had low η in 1800. China possessed relatively integrated economic institutions that allowed its central government to oversee trade and a tightly-managed bureaucracy composed of strong elite networks (Rosenthal and Wong, 2011). The state's Confucian ideology centered political continuity and stability, contrasting with the "disruptive progress" of Western Europe. The entrenched power of the emperor coupled with a shared ideology of "stability" among the state and elites arguably prevented the emergence of ideologies promoting indirect rule, which was only strengthened by the lack of external information (Mokyr, 2016, ch. 16). China's high (although stagnant) economic standards in the late 1700s evidence a degree of policy *moderation* that was not seen in Western Europe (Pomeranz, 2021). Rare threats of overthrow came from regional elites and seemingly strengthened — rather than eroded — the collective power of the emperor and elites (Dincecco and Wang, 2018).

The "high η" environment of Western Europe in the 18th and 19th centuries was accompanied by greater policy variation, concessions to liberal demands, and political transitions, in contrast to the consolidated "low η" environment of China, which experienced little variation and minimal turnover. This prediction is consistent with Mokyr (2016), who argues that "there were repressive

and reactionary regimes galore in Europe, but the interstate competitiveness constrained their ability to enforce a specific orthodoxy" (317). This prediction also complements a literature arguing that Europe's "political fragmentation" drove the Great Divergence by fostering competition over technology (Diamond, 2005), intellectual networks (Mokyr, 2016), and institutional innovation (Cox, 2017)

1.6 Conclusion

This paper studies a repeated environment where leaders set policy in the face of uncertainty. Agents learn about the effects of policies as they are implemented. Hence leaders' policies not only generate payoffs but also information which affects what policies they can implement in the future. Leaders facing threats of deposition are hence constrained by a threat of information revelation, generating an equilibrium nonmonotonicity in policy experimentation. When policies' effects are uncertain, leaders pursue spatially intermediate "learning trap policies" — shutting down information revelation when it would be most useful. They pursue extreme policies — keeping constituents indifferent to overthrow — only at extremal beliefs, when they are relatively certain about what the optimal policy is. Because learning traps deliver moderate payoffs, forever implementing a learning trap and revealing no information delivers a sure moderate payoff. Experimentation is preferred to these moderate payoffs only when beliefs that leaders' preferred policies favor their constituents are high. At intermediate beliefs, information is less likely to favor leaders in expectations; uncertainty about the optimal policy means they are able to shut down information revelation without being overthrown. At low beliefs, leaders would prefer to shut down information, but constituents are relatively certain about what the opitmal policy is and threaten overthrow if leaders do not implement it.

I use three equilbria of the model to study land reform in Imperial Russia. I argue that, when the Russian government was unsure about whether liberal economic institutions were preferable to serfdom, they implemented a "learning trap" — codifying the peasant commune and mixing elements of communal and liberal institutions. Even though the Russian government had ideological and fiscal preferences for liberal economic reform, they were uncertain about whether these policies would work in Russia due to cultural and institutional differences with Western Europe.

However, by pursuing this intermediate policy, they distorted labor and production decisions, restricting economic, fiscal, and industrial growth but simultaneously obfuscating information about policy efficacy that minimized an informational threat of overthrow. I contrast this "learning trap" with Russia's adoption of pure serfdom prior to 1861 — in a setting where there was less policy uncertainty — and the liberal Stolypin experiments with private property after 1906 — representing experimentation with the government's preferred policy after a favorable informational shock.

The model's findings are robust to allowing for exogenous information arrival. Unanticipated information generates minimal policy variation unless a large amount of information favors the leader. Anticipated revelation generates greater policy variation since leaders cannot fully control information revelation. By interpreting exogenous information as the result of policy experimentation in an externally valid country, the latter insight can explain why Europe experienced a great deal of political upheaval and convergence to systems of indirect rule in the 19th century, while China did not, complementing "fragmentation" hypotheses of the Great Divergence.

Because learning traps arise at intermediate beliefs, learning about the optimal policy is shut down precisely when it would be most useful. This finding suggests that leaders who either lack term limits or prioritize keeping ideologically similar leaders in office will seldom experiment with potentially beneficial policies unless they are absolutely sure they will work in their favor. Future work will probe mechanisms that encourage efficient policy experimentation, as well as the impacts of polarization or alternative tenure systems, in such repeated settings.

1.7 Proofs of Main Results

Proof of Solution to People's Problem

Proposition 3. *$x_p(q, x')$ is the solution to the following Bellman equation:*

$$W(q, x') = \max_{x \in [0,1]} \frac{2\sigma(1-\delta)\mathbb{E}[y|x] + \delta\Delta(x)\Phi(q,x)}{2\sigma(1-\delta) + \delta\Delta(x)} - k|x - x'|$$

where $\Phi(q, x) = qW(1, x) + (1 - q)W(0, x)$. For intermediate values of c, there exists a nonempty open set $LT \subseteq [0, 1] \times X$ with $LT \ni (1/2, x_{LT})$ such that for all $(q, x') \in LT$, $W(q, x') \leq y_{LT} + c \leq W(0, x'), W(1, x')$.

Lemma 1. *Suppose x^* is a solution to the functional equation defining $W(q, x')$. Then, x^* is also a solution to the functional equation defining $W(q, x^*(q, x'))$.*

Proof. This proof utilizes the linearity of adjustment costs. Suppose by contradiction that x^{**} solves the functional equation associated with $W(q, x^*)$ but not $W(q, x')$. Since both x^* and x^{**} are always feasible, this implies

$$\begin{aligned}
&(1-\delta)\mathbb{E}[y|x^{**}] - k|x^{**} - x^*| + \frac{\delta\Delta(x^{**})}{2\sigma}(qW(1, x^{**}) + (1-q)W(0, x^{**})) + \delta(1 - \frac{\Delta(x^{**})}{2\sigma}W(q, x^{**}) \\
&> (1-\delta)\mathbb{E}[y|x^*] + \frac{\delta\Delta(x^*)}{2\sigma}(qW(1, x^*) + (1-q)W(0, x^*)) + \delta(1 - \frac{\Delta(x^*)}{2\sigma})W(q, x^*)
\end{aligned}$$

which also implies

$$\begin{aligned}
&(1-\delta)\mathbb{E}[y|x^{**}] - k|x^{**} - x^*| - k|x^* - x'| + \frac{\delta\Delta(x^{**})}{2\sigma}(qW(1, x^{**}) + (1-q)W(0, x^{**})) \\
&+ \delta(1 - \frac{\Delta(x^{**})}{2\sigma})W(q, x^{**}) \\
&> (1-\delta)\mathbb{E}[y|x^*] - k|x^* - x'| + \frac{\delta\Delta(x^*)}{2\sigma}(qW(1, x^*) + (1-q)W(0, x^*)) + \delta(1 - \frac{\Delta(x^*)}{2\sigma})W(q, x^*)
\end{aligned}$$

However, since $k|x^{**} - x'| \leq k|x^{**} - x^*| + k|x^* - x'|$, this implies

$$
\begin{aligned}
&(1-\delta)\mathbb{E}[y|x^{**}] - k|x^{**} - x'| + \frac{\delta\Delta(x^{**})}{2\sigma}(qW(1,x^{**}) + (1-q)W(0,x^{**})) + \delta(1 - \frac{\Delta(x^{**})}{2\sigma})W(q,x^{**}) \\
&\geq (1-\delta)\mathbb{E}[y|x^{**}] - k|x^{**} - x^*| - k|x^* - x'| + \frac{\delta\Delta(x^{**})}{2\sigma}(qW(1,x^{**}) + (1-q)W(0,x^{**})) \\
&+ \delta(1 - \frac{\Delta(x^{**})}{2\sigma})W(q,x^{**}) \\
&> (1-\delta)\mathbb{E}[y|x^*] - k|x^* - x'| + \frac{\delta\Delta(x^*)}{2\sigma}(qW(1,x^*) + (1-q)W(0,x^*)) + \delta(1 - \frac{\Delta(x^*)}{2\sigma})W(q,x^*)
\end{aligned}
$$

contradicting that x^* solved the functional equation associated with $W(q,x')$. □

Lemma 2. *There exist* $0 \leq \underline{x} < x_{LT} < \overline{x} < \tilde{x} \leq \overline{\overline{x}} \leq 1$ *such that*

$$
W(1,x') = \begin{cases} f_g(\overline{x}) - k(\overline{x} - x') & x' \notin (\overline{x}, \overline{\overline{x}}) \\ f_g(x') & x' \in (\overline{x}, \overline{\overline{x}}) \end{cases}
$$

$$
W(0,x') = \begin{cases} f_b(\underline{x}) - k(x' - \underline{x}) & x' \geq \underline{x} \\ f_b(x') & x' < \underline{x} \end{cases}
$$

Proof. Suppose $q = 1$ and $x' \leq \tilde{x}$. $W(1,x') = \max_{x\in[0,1]}(1-\delta)f_g(x) + \delta W(1,x) - k|x - x'|$ The previous lemma allows us to simplify this to $\max_{x\in[0,1]} f_g(x) - k|x - x'|$, which is solved by $\overline{x} = \tilde{x} - k/2$. A reverse argument shows $\overline{\overline{x}} = \tilde{x} + k/2$. That these points are to the right of x_{LT} follows from $f_g'(x_{LT}) > k$. □

Lemma 3. *Suppose* x^* *is a solution to the functional equation defining* $W(q,x')$ *and* $q \in (0,1)$ *Then:*

$$
W(q,x^*) = \frac{2\sigma(1-\delta)\mathbb{E}[y|x^*] + \delta\Delta(x^*)\Phi(q,x^*)}{2\sigma(1-\delta) + \delta\Delta(x^*)}
$$

where $\Phi(q,x) = qW(1,x) + (1-q)W(0,x)$.

Moreover, $W(q,x') = \max_x \frac{2\sigma(1-\delta)\mathbb{E}[y|x] + \delta\Delta(x^*)\Phi(q,x)}{2\sigma(1-\delta) + \delta\Delta(x)} - k|x - x'|$.

Proof. Since x^* solves the functional equation associated with $W(q,x')$, by the previous lemmata

it also solves the one associated with $W(q, x^*)$. Hence:

$$W(q,x^*) = (1-\delta)\mathbb{E}[y|x^*] + \frac{\delta\Delta(x^*)}{2\sigma}\Phi(q,x^*) + \delta(1-\frac{\Delta(x^*)}{2\sigma})W(q,x^*)$$
$$= \frac{(1-\delta)\mathbb{E}[y|x^*] + \frac{\delta\Delta(x^*)}{2\sigma}\Phi(q,x^*)}{1-\delta(1-\frac{\Delta(x^*)}{2\sigma})} = \frac{2\sigma(1-\delta)\mathbb{E}[y|x^*] + \delta\Delta(x^*)\Phi(q,x^*)}{2\sigma(1-\delta)+\delta\Delta(x^*)}$$

$W(q, x')$ can then be written as:

$$\begin{aligned}
W(q,x') =& \max_{x^*}(1-\delta)\mathbb{E}[y|x^*] + \frac{\delta\Delta(x^*)}{2\sigma}\Phi(q,x) \\
&+ \delta(1-\frac{\Delta(x^*)}{2\sigma})\frac{2\sigma(1-\delta)\mathbb{E}[y|x^*] + \delta\Delta(x^*)\Phi(q,x)}{2\sigma(1-\delta)+\delta\Delta(x)} - k|x^*-x'| \\
=& \max_x \mathbb{E}[y|x]\big((1-\delta) + \delta(1-\frac{\Delta(x)}{2\sigma})\frac{2\sigma(1-\delta)}{2\sigma(1-\delta)+\delta\Delta(x)}\big) \\
&+\delta\Phi(q,x)\big(\frac{\Delta(x)}{2\sigma} + (1-\frac{\Delta(x)}{2\sigma})\frac{\delta\Delta(x)}{2\sigma(1-\delta)+\delta\Delta(x)}\big) - k|x-x'| \\
=& \max_x \mathbb{E}[y|x]\big((1-\delta)\big(1+\frac{2\sigma\delta-\delta\Delta(x)}{2\sigma(1-\delta)+\delta\Delta(x)}\big)\big) \\
&+\delta\frac{\Delta(x)}{2\sigma}\Phi(q,x)\big(1+\frac{2\sigma\delta-\delta\Delta(x)}{2\sigma(1-\delta)+\delta\Delta(x)}\big) - k|x-x'| \\
=& \max_x \frac{2\sigma(1-\delta)\mathbb{E}[y|x] + \delta\Delta(x)\Phi(q,x)}{2\sigma(1-\delta)+\delta\Delta(x)} - k|x-x'|
\end{aligned}$$

□

Lemma 4. *Suppose $x' \in [\underline{x}, \overline{x}]$. Then $x_p(q, x') \in [\underline{x}, \overline{x}]$.*

Proof. Note that our problem can be written as:

$$\begin{aligned}
W(q,x') = \max_{x\in[0,1]} &\frac{2\sigma(1-\delta)}{2\sigma(1-\delta)+\delta\Delta(x)}\big(\mathbb{E}[y|x] - k|x-x'|\big) \\
&+ \frac{\delta\Delta(x)}{2\sigma(1-\delta)+\delta\Delta(x)}\big(\Phi(q,x) - k|x-x'|\big)
\end{aligned}$$

First, since $\frac{\partial^2}{\partial q\partial x}\mathbb{E}[y|x] = f_g'(x) - f_b'(x) > 0$, since $\arg\max f_g(x) - k|x-x'| = \overline{x}$, and since $\arg\max f_b(x) - k|x-x'| = \underline{x}$, we have: $\arg\max_{\hat{x}\in[0,1]} \mathbb{E}[y|\hat{x}] - k|\hat{x}-x'| \in [\underline{x}, \overline{x}]$. All $x < \underline{x}$ are dominated by $\underline{x}$ and all $x > \overline{x}$ are dominated by $\overline{x}$.

Next, I claim $\arg\max_{\tilde{x}\in[0,1]} W(q,\tilde{x}) - k|\tilde{x}-x'| \in [\underline{x}, \overline{x}]$. Suppose $x < \underline{x}$, meaning $f_b(x)$ is concave and $f_b'(x) > -k$ for $x < \underline{x}$. Then, the derivative with respect to x here is: $qk + (1-q)f_b'(x) + k =$

$(1-q)(f_b'(x)+k) > 0$, meaning $\underline{x}$ dominates all options to its left. Similarly, for $x \in (\overline{x}, \overline{\overline{x}})$, $|f'(x)|$ is necessarily $< k$ and the derivative is $qf'(x) - (1-q)k - k < 0$. Finally, for $x \geq \overline{\overline{x}}$, the derivative is $-k - k < 0$ meaning $\overline{\overline{x}}$ dominates in this range, but since the left and right derivatives at $\overline{\overline{x}}$ are then both negative, we have that $\overline{x}$ dominates all options to its right.

Finally, suppose by contradiction that $[\underline{x}, \overline{x}] \not\ni \hat{x}$, where $\hat{x}$ is the maximizer associated with $W(q, x')$. If $\hat{x} < \underline{x}$, we know that $\mathbb{E}[y|\hat{x}] - k|\hat{x} - x'| < \mathbb{E}[y|\underline{x}] - k|\underline{x} - x'|$ and $\Phi(q, \hat{x}) - k|\hat{x} - x'| < \Phi(q, \underline{x}) - k|\underline{x} - x'|$ Any convex combination of $\mathbb{E}[y|\underline{x}] - k|\underline{x} - x'|$ and $\Phi(q, \underline{x}) - k|\underline{x} - x'|$ is strictly larger than any convex combination of $\mathbb{E}[y|\hat{x}] - k|\hat{x} - x'|$ and $\Phi(q, \hat{x}) - k|\hat{x} - x'|$, meaning playing $\overline{x}$ dominates $\hat{x}$ in the original program. A symmetric argument shows that playing any $\hat{x} > \overline{x}$ is always strictly worse than $\overline{x}$. □

Lemma 5. *Under Assumptions 1 and 2′, there exists $c > 0$ such that: $f_g(\overline{x}) - k(\overline{x} - x_{LT}) \geq y_{LT}$, $f_b(\underline{x}) - k(x_{LT} - \underline{x}) + c > y_{LT}$ and $W(1/2, x_{LT}) < y_{LT} + c$. Hence, the set $LT = \{(q, x') : W(q, x') < y_{LT} - c\}$ has nonempty interior.*

Proof. We constructing an upper bound for $W(1/2, x_{LT})$. First, I claim:

$$\frac{1}{2}\big(f_g(\overline{x}) - k(\overline{x} - x_{LT})\big) + \frac{1}{2}\big(f_b(\underline{x}) - k(x_{LT} - \underline{x})\big) - k(\overline{x} - x_{LT}) < y_{LT}$$
$$\frac{1}{2}\big(f_g(\overline{x}) - k(\overline{x} - x_{LT})\big) + \frac{1}{2}\big(f_b(\underline{x}) - k(x_{LT} - \underline{x})\big) - k(x_{LT} - \underline{x}) < y_{LT}.$$

Sufficient conditions for the first equation are: $f_g(\overline{x}) - 2k(\overline{x} - x_{LT}) < y_{LT}$ and $f_b(\underline{x}) - k(\overline{x} - \underline{x}) < y_{LT}$. The first inequality follows immediately from the fact that $f_g'(x_{LT}) < 2k$, contained in Assumption 2. Next, note that $x_{LT} = \frac{\beta_b - \beta_g + \tilde{x}^2}{2\tilde{x}}$. Then, using the fact that $\underline{x} = k/2$ and $\overline{x} = \tilde{x} - k/2$, $f_b(\overline{x}) - k(\overline{x} - \underline{x}) < y_{LT}$ if and only if:

$$\beta_b - \frac{k^2}{4} - k(\tilde{x} - k) < \beta_b - \left(\frac{\beta_b - \beta_g + \tilde{x}^2}{2\tilde{x}}\right)^2 \iff \frac{3}{4}k^2 - \tilde{x}k + \left(\frac{\tilde{x}^2}{4} + (\beta_b - \beta_g) + \frac{(\beta_b - \beta_g)^2}{4\tilde{x}^2}\right) \quad < 0$$

This holds if $k \in \left[\frac{2\tilde{x} - \sqrt{\tilde{x}^2 - 3\tilde{x}^2/4 - 3(\beta_b - \beta_g) - 3(\beta_b - \beta_g)^2/\tilde{x}^2}}{3}, \frac{2\tilde{x} + \sqrt{\tilde{x}^2 - 3\tilde{x}^2/4 - 3(\beta_b - \beta_g) - 3(\beta_b - \beta_g)^2/\tilde{x}^2}}{3}\right]$ and $k \in \left[\frac{2\tilde{x} - \sqrt{\tilde{x}^2/4 - 3(\beta_b - \beta_g) - 3(\beta_b - \beta_g)^2/\tilde{x}^2}}{3}, \frac{2\tilde{x} + \sqrt{\tilde{x}^2/4 - 3(\beta_b - \beta_g) - 3(\beta_b - \beta_g)^2/\tilde{x}^2}}{3}\right]$. This is satisfied by Assumptions 2 and 2′, which also subsume a similar condition with the second inequality: $k \in \left[\frac{2\tilde{x} - \sqrt{\tilde{x}^2/4 + 3(\beta_b - \beta_g) - 3(\beta_b - \beta_g)^2/\tilde{x}^2}}{3}, \frac{2\tilde{x} + \sqrt{\tilde{x}^2/4 + 3(\beta_b - \beta_g) - 3(\beta_b - \beta_g)^2/\tilde{x}^2}}{3}\right]$.

We now construct our upper bound for $W(1/2, x_{LT})$. $W(1/2, x_{LT})$ =

$$\begin{aligned}
&\max_{x\in[\underline{x},\overline{x}]} \frac{2\sigma(1-\delta)\mathbb{E}[y|x] + \delta\Delta(x)\Phi(1/2,x)}{2\sigma(1-\delta)+\delta\Delta(x)} - k|x - x_{LT}| \\
&= \max_{x\in[\underline{x},\overline{x}]} \frac{2\sigma(1-\delta)}{2\sigma(1-\delta)+\delta\Delta(x)}\big(\mathbb{E}[y|x] - k|x - x_{LT}|\big) + \frac{\delta\Delta(x)}{2\sigma(1-\delta)+\delta\Delta(x)}\big(\Phi(1/2,x) - k|x - x'|\big) \\
&\leq \max_{x\in[\underline{x},\overline{x}]} \frac{2\sigma(1-\delta)}{2\sigma(1-\delta)+\delta\Delta(x)}\big(\max_{\breve{x}} \mathbb{E}[y|\breve{x}] - k|\breve{x} - x_{LT}|\big) + \frac{\delta\Delta(x)}{2\sigma(1-\delta)+\delta\Delta(x)}\big(\Phi(1/2,x) - k|x - x'|\big) \\
&= \max_{x\in[\underline{x},\overline{x}]} \frac{2\sigma(1-\delta)}{2\sigma(1-\delta)+\delta\Delta(x)} y_{LT} + \frac{\delta\Delta(x)}{2\sigma(1-\delta)+\delta\Delta(x)}\big(\Phi(1/2,x) - k|x - x_{LT}|\big) \equiv \max_{x\in[\underline{x},\overline{x}]} U(x)
\end{aligned}$$

$U(x)$ will be the objective associated with the upper bound. The third line holds since

$$\partial/\partial x\Big|_{x=x_{LT}} \mathbb{E}[y|x] = \frac{1}{2}2(\tilde{x} - x_{LT}) - \frac{1}{2}2x_{LT} = \tilde{x} - 2x_{LT} = \tilde{x} - 2\frac{\tilde{x}^2 + (\beta_b - \beta_g)}{2\tilde{x}} = \frac{\beta_b - \beta_g}{2\tilde{x}}$$

and by Assumption 2′, $|\frac{\beta_b-\beta_g}{2\tilde{x}}||\frac{\beta_b-\beta_g}{\tilde{x}}| < |\frac{\beta_b-\beta_g}{\tilde{x}-k}| < k$.

Note that Φ does not depend on x at $q = 1/2$: $1/2(f_g(\overline{x}) - k(\overline{x} - x)) + 1/2(f_b(\underline{x}) - k(x - \underline{x})) = 1/2(f_g(\overline{x}) + f_b(\underline{x}) - k(\overline{x} - \underline{x}))$. Hence:

$$\begin{aligned}
\Phi(1/2, x_{LT}) =& 1/2(f_g(\overline{x}) - k(\overline{x} - x_{LT})) + 1/2(f_b(\underline{x}) - k(x_{LT} - \underline{x})) > y_{LT} \\
>& 1/2(f_g(\overline{x}) - k(\overline{x} - x_{LT})) + 1/2(f_b(\underline{x}) - k(x_{LT} - \underline{x})) - k(\overline{x} - x_{LT}), \\
& 1/2(f_g(\overline{x} - k(\overline{x} - x_{LT})) + 1/2(f_b(\underline{x}) - k(x_{LT} - \underline{x})) - k(\underline{x} - x_{LT})
\end{aligned}$$

Let $\hat{x}_1$ and $\hat{x}_2$ be defined by the following:

$$\begin{aligned}
y_{LT} &= 1/2(f_g(\overline{x} - k(\overline{x} - x_{LT})) + 1/2(f_b(\underline{x}) - k(x_{LT} - \underline{x})) - k(\hat{x}_1 - x_{LT}) \qquad \hat{x}_1 \in (x_{LT}, \overline{x}) \\
y_{LT} &= 1/2(f_g(\overline{x} - k(\overline{x} - x_{LT})) + 1/2(f_b(\underline{x}) - k(x_{LT} - \underline{x})) - k(\hat{x}_2 - x_{LT}) \qquad \hat{x}_2 \in (\underline{x}, x_{LT})
\end{aligned}$$

which exist and are unique by continuity, strict monotonicity, and the Intermediate Value Theorem. Because we are taking convex combinations of y_{LT} and $\Phi - k|x - x_{LT}|$, where the weights depend on x as well, we will only want to place less weight on y_{LT} if $\Phi - k|x - x_{LT}| > y_{LT}$ (precisely on $[\hat{x}_2, x_{LT}]$ and $[x_{LT}, \hat{x}_1]$).

The derivative of $U(x)$ for $x > x_{LT}$ is:

$$\frac{2\sigma(1-\delta)\delta\Delta'(x)(\Phi - k(x-x_{LT}) - y_{LT}) - 2\sigma(1-\delta)\delta\Delta(x)k - \delta^2\Delta(x)^2k}{(2\sigma(1-\delta)+\delta\Delta(x))^2}$$

A nearly identical derivative for $x < x_{LT}$ is omitted. At $x = x_{LT}$, the right derivative is $\frac{\delta\Delta'(x)(\Phi - y_{LT})}{2\sigma(1-\delta)} > 0$. At $\hat{x}_1$, when $\Phi - k(\hat{x}_1 - x_{LT}) - y_{LT} = 0$, the derivative is $-\frac{\delta\Delta(x)k}{2\sigma(1-\delta)+\delta\Delta(x)} < 0$. The derivative is ≥ 0 when:

$$\begin{aligned}
2\sigma(1-\delta)\delta\Delta'(x)(\Phi - k(x-x_{LT}) - y_{LT}) &\geq \delta\Delta(x)(2\sigma(1-\delta)+\delta\Delta(x))k \\
\Phi - k(x-x_{LT}) - y_{LT} &\geq \frac{\delta\Delta(x)(2\sigma(1-\delta)+\delta\Delta(x))k}{2\sigma(1-\delta)\delta\Delta'(x)} \\
\Phi - k(x-x_{LT}) &\geq y_{LT} + \frac{\Delta(x)(2\sigma(1-\delta)+\delta\Delta(x))}{2\sigma(1-\delta)\Delta'(x)}k \\
&= y_{LT} + k(x-x_{LT})\left(1 + \frac{\delta\Delta(x)}{2\sigma(1-\delta)}\right)
\end{aligned}$$

The final line is strictly increasing in x with a zero at $x_{LT} < \hat{x_1}$. $\Phi - k(x - x_{LT})$ is strictly decreasing in x with a zero at $\hat{x_1}$. Hence there is a *unique* maximizer $\overline{x}^\star$ to this problem in $(x_{LT}, \hat{x}_1)$. Define c implicitly as $\Phi - k(\overline{x}^\star - x_{LT}) = y_{LT} + c$. Defining c using the derivative to the left of x_{LT} gives the same c, since Φ, $\Delta(x)$, and $k|x - x_{LT}|$ are symmetric about x_{LT}. Then:

$$\begin{aligned}
W(1/2, x_{LT}) \leq \max_x U(x) &= y_{LT} + \frac{\delta\Delta(\overline{x}^\star)}{2\sigma(1-\delta)+\delta\Delta(\overline{x}^\star)}\left(k(\overline{x}^\star - x_{LT})\left(1 + \frac{\delta\Delta(\overline{x}^\star)}{2\sigma(1-\delta)}\right)\right) \\
&= y_{LT} + \frac{\delta\Delta(\overline{x}^\star)}{2\sigma(1-\delta)+\delta\Delta(\overline{x}^\star)}c < y_{LT} + c.
\end{aligned}$$

Since $x^\star > x_{LT}$, we also have $\Phi(1/2, x_{LT}) - k(x^\star - x_{LT}) = y_{LT} + c \implies \Phi(1/2, x_{LT}) > y_{LT} + c$. That c is not *so* large that the people are discouraged from overthrowing at $q = 0, 1$ follows from the fact that $\Phi(1/2, x_{LT}) = \frac{1}{2}W(1, x_{LT}) + \frac{1}{2}W(0, x_{LT})$ and that $|\beta_g - \beta_b|$ is small. Since $W(q, x')$ is continuous, the inequality $W(1/2, x_{LT}) < y_{LT} + c$ also holds in a *neighborhood* of $(1/2, x_{LT})$. This means that $LT : \{(q, x') : W(q, x') \leq y_{LT} + c\}$ has nonempty interior. □

Proof of Solution to Leader's Problem

Theorem 1. *Suppose $\sigma(1-\delta^\ell)$ is small or $\underline{u}(x')$ is sufficiently low. Then, there exist thresholds $\underline{q} < \overline{q} \in (0,1)$ such that the following hold:*

1. *If $q \leq \underline{q}$, then $x_\ell(q,x') = \underline{x}_{LT}(q,x')$ or $\overline{x}_{LT}(q,x')$. The leader plays the* closest policy *to the learning trap.*

2. *If $q \in [\underline{q},\overline{q}]$, then $x_\ell(q,x') = \underline{x}_{LT}(q,x')$ or $\overline{x}(q,x')$. The leader plays either the lowest policy below the learning trap or her highest feasible policy.*

3. *If $q > \overline{q}$, = then $x_\ell(q,x') = \overline{x}(q,x')$ or $\underline{x}(q,x')$ (and it is necessary that $\Delta(\underline{x}(q,x')) > \Delta(\overline{x}(q,x'))$). The leader either plays her highest feasible policy or policies maximizing information revelation.*

Lemma 6. *Suppose that the leader can implement any $x \in [x_\ell(0,x'), x_\ell(1,x')]$. Then, there exist thresholds $\underline{q} < \overline{q}$ such that: for $q \leq \underline{q}$, the derivative of the objective in the maximization problem is positive for $x < x_{LT}$ and negative for $x > x_{LT}$; for $q \in [\underline{q},\overline{q}]$, the derivative of the objective is positive for $x < x_{LT}$ and positive for $x > x_{LT}$; and for $q \geq \overline{q}$, the derivative of the objective is negative for $x < x_{LT}$ and positive for $x > x_{LT}$.*

Proof. Let $X_{unc} = [x_\ell(0,x'), x_\ell(1,x')]$. The leader's Bellman in the unconstrained case can be written as:

$$\tilde{V}(q) = \max_{x \in X_{unc}} (1-\delta)u_\ell(x) + \delta\frac{\Delta(x)}{2\sigma}(q\overline{u}(x') + (1-q)\underline{u}(x')) + \delta\left(1 - \frac{\Delta(x)}{2\sigma}\right)\tilde{V}(q)$$

Let $\Psi(q) = q\overline{u}(x') + (1-q)\underline{u}(x')$. Replicating an earlier argument implies

$$\tilde{V}(q) = \max_{x \in X_{unc}} \frac{2\sigma(1-\delta)u_\ell(x) + \delta\Delta(x)\Psi(q)}{2\sigma(1-\delta) + \delta\Delta(x)}.$$

The numerator of the derivative with respect to x for $x \neq x_{LT}$ is:

$$= 4\sigma^2(1-\delta)^2 u_\ell'(x) + 2\sigma\delta(1-\delta)(\Delta(x)u_\ell'(x) + \Delta'(x)\Psi(q) - \Delta'(x)u_\ell(x)).$$

Observe that $\Delta(x)$ is the absolute value of a linear function with zero at x_{LT}.

$$\Delta(x) = |\beta_g - (x - \tilde{x})^2 - (\beta_b - x^2)| = |\beta_g - \beta_b - \tilde{x}^2 + 2\tilde{x}x| \equiv \xi|x - x_{LT}|$$

This means that for $x > x_{LT}$, the derivative of the objective is negative if and only if

$$2\sigma(1-\delta)u'_\ell(x) \leq \delta\big(\Delta'(x)u_\ell(x) - \Delta(x)u'_\ell(x) - \Delta'(x)\Psi(q)\big)$$
$$2\sigma(1-\delta)u'_\ell(x) \leq \delta\xi\big(u_\ell(x) - u'_\ell(x)(x - x_{LT}) - \Psi(q)\big)$$

$u_\ell(x) - u'_\ell(x)(x - x_{LT})$ is equal to u_{LT} at $x = x_{LT}$ and is weakly increasing for $x > x_{LT}$.[20] By weak concavity, the LHS below is decreasing and the RHS increasing:

$$2\sigma(1-\delta)u'_\ell(x) \leq \delta\xi\big(u_\ell(x) - u'_\ell(x)(x - x_{LT}) - \Psi(q)\big) \tag{1.1}$$

If $\sigma(1-\delta)$ is small, since $\underline{u}(x') < u_{LT} \leq u_\ell(x) - u'_\ell(x)(x - x_{LT}) < \overline{u}(x')$, there is a threshold $\underline{q}$ such that the inequality satisfies strictly if and only if $q < \underline{q}$. Moreover, for any $\sigma(1-\delta)$, as $\underline{u}(x')$ grows more negative, $-\Psi(q)$ grows large, which can also allow the inequality to satisfy. A similar argument shows that for $x < x_{LT}$, the derivative is positive when:

$$2\sigma(1-\delta)u'_\ell(x) \geq \delta\xi\big(\Psi(q) - \big(u_\ell(x) - u'_\ell(x)(x - x_{LT})\big)\big) \tag{1.2}$$

When (1) holds with equality, its RHS is strictly positive, meaning the RHS of (2) is strictly negative, showing $\underline{q} < \overline{q}$. Hence: for $x > x_{LT}$, the derivative of the objective is strictly negative when $q < \underline{q}$ and strictly positive when $q > \underline{q}$, with equality at $q = \underline{q}$; for $x < x_{LT}$, the derivative of the objective is strictly positive when $q < \overline{q}$ and strictly negative when $q > \overline{q}$, with equality at $q = \overline{q}$. This completes the lemma. □

Lemma 7. *Let $x_\ell(q, x')$ denote a policy for the leader such that $x_\ell(q, x_\ell(q, x')) = x_\ell(q, x')$. Then, the*

[20] To see this, note that the derivative of this expression with respect to x is $u'_\ell(x) - u''_\ell(x)(x - x_{LT}) - u'_\ell(x) = -u''_\ell(x)(x - x_{LT}) \geq 0$.

people overthrow if and only if $\tilde{W}(x_\ell, q, x') \geq W(q, x') - c$*, where*

$$\tilde{W}(x_\ell, q, x') = \frac{2\sigma(1-\delta)\mathbb{E}[y|x_\ell(q,x')] + \delta\Delta(x_\ell(q,x'))\tilde{\Phi}(q, x_\ell(q,x'))}{2\sigma(1-\delta) + \delta\Delta(x)} - k|x_\ell(q,x') - x'|$$

where $\tilde{\Phi}(q,x) = q\big(f_g(x_\ell(1,x) - k|x_\ell(1,x) - x'|\big) + (1-q)\big(f_b(x_\ell(0,x) - k|x_\ell(0,x) - x|\big)$.

Proof. Let $\tilde{W}(x_\ell, q, x')$ be the value for the people of accepting for all time a policy $x_\ell(q,x')$ satisfying the hypothesis. Let

$$\begin{aligned}
\tilde{\Phi}(q,x) &= q\tilde{W}(x_\ell, 1, x_\ell(1, x_\ell(1,x'))) + (1-q)\tilde{W}(x_\ell, 1, x_\ell(0, x_\ell(0,x'))) \\
\tilde{W}(x_\ell, 1, x') &= \Big(\sum_{t=0}^{\infty}(1-\delta) f_g(x_\ell(1,x'))\Big) - k|x_\ell(1,x') - x'| = f_g(x_\ell(1,x') - k|x_\ell(1,x') - x'| \\
\tilde{W}(x_\ell, 0, x') &= \Big(\sum_{t=0}^{\infty}(1-\delta) f_b(x_\ell(0,x'))\Big) - k|x_\ell(0,x') - x'| = f_b(x_\ell(0,x') - k|x_\ell(0,x') - x'|.
\end{aligned}$$

Then:

$$\begin{aligned}
\tilde{W}(x_\ell, q, x_\ell(q,x')) &= (1-\delta)\mathbb{E}[y|x_\ell(q,x')] + \delta\frac{\Delta(x_\ell(q,x'))}{2\sigma}\Phi(q, x_\ell(q,x')) \\
&\quad + \delta\big(1 - \frac{\Delta(x)}{2\sigma}\big)\tilde{W}(q, x_\ell(q,x')) \\
&= \frac{2\sigma(1-\delta)\mathbb{E}[y|x_\ell(q,x')] + \delta\Delta(x_\ell(q,x'))\Phi(q, x_\ell(q,x'))}{2\sigma(1-\delta) + \delta\Delta(x)}
\end{aligned}$$

Replicating arguments from earlier lemmata gives the result:

$$\tilde{W}(x_\ell, q, x') = \frac{2\sigma(1-\delta)\mathbb{E}[y|x_\ell(q,x')] + \delta\Delta(x_\ell(q,x'))\Phi(q, x_\ell(q,x'))}{2\sigma(1-\delta) + \delta\Delta(x)} - k|x_\ell(q,x') - x'|$$

□

Lemma 8. $x_\ell(q,x') = x_\ell\big(q, x_\ell(q,x')\big)$*, meaning the previous lemma characterizes the equilibrium overthrow decision.*

Proof. If $\mathrm{NR}(q,x')$ does not bind, the leader implements her preferred policy by the previous lemma, which satisfies the property. So, suppose $\mathrm{NR}(q,x')$ binds and that by contradiction there exists x' such that $x_1 \equiv x_\ell(q,x') \neq x_\ell\big(x_\ell(q,x')\big) \equiv x_2$. Let $h(x) = (1-\delta)\mathbb{E}[y|x] + \frac{\delta\Delta(x)}{2\sigma}\tilde{\Phi}(q,x)$. Then,

we must have

$$\begin{aligned} W(q,x') - c &= h(x_1) - k|x_1 - x'| \\ &\quad + \delta\big(1 - \frac{\Delta(x_1)}{2\sigma}\big)\big(h(x_2) + \delta\big(1 - \frac{\Delta(x_2)}{2\sigma}\big)\tilde{W}_{x_\ell}(x_2) - k|x_2 - x_1|\big) \\ W(q,x_1) - c &= h(x_2) + \delta\big(1 - \frac{\Delta(x_2)}{2\sigma}\big)\tilde{W}_{x_\ell}(x_2) - k|x_2 - x_1| \end{aligned}$$

Let $x^* = x_p(q,x')$. Since $\mathrm{NR}(q,x')$ is binding, it must be that:

$$\begin{aligned} \frac{2\sigma(1-\delta)\mathbb{E}[y|x^*] + \delta\Delta(x^*)\Phi(q,x^*)}{2\sigma(1-\delta) + \delta\Delta(x^*)} - k|x^* - x'| - c > \\ h(x_2) + \delta\big(1 - \frac{\Delta(x_2)}{2\sigma}\big)\tilde{W}_{x_\ell}(x_2) - k|x_2 - x'| \end{aligned}$$

where the RHS is equal to $W(q,x')$. Since either $x^* < x_1 < x_2$ or $x^* > x_1 > x_2$ by a previous lemma (policy is attracted in a single direction; implementing x^* and then x_2 is dominated by a policy between x^* and x_2), as $x' \to x_1$, both sides increase/decrease by the same amount or the right side increases by more, meaning:

$$\frac{2\sigma(1-\delta)\mathbb{E}[y|x^*] + \delta\Delta(x^*)\Phi(q,x^*)}{2\sigma(1-\delta) + \delta\Delta(x^*)} - k|x^* - x_1| > h(x_2) + \delta\big(1 - \frac{\Delta(x_2)}{2\sigma}\big)\tilde{W}_{x_\ell}(x_2) - k|x_2 - x_1|$$

But then:

$$\begin{aligned} W(q,x_1) &= \max_{x\in[0,1]} \frac{2\sigma(1-\delta)\mathbb{E}[y|x] + \delta\Delta(x)\Phi(q,x)}{2\sigma(1-\delta) + \delta\Delta(x)} - k|x - x_1| \\ &\geq \frac{2\sigma(1-\delta)\mathbb{E}[y|x^*] + \delta\Delta(x^*)\Phi(q,x^*)}{2\sigma(1-\delta) + \delta\Delta(x^*)} - k|x^* - x_1| \end{aligned}$$

a contradiction.

□

Proof. We use the lemmata above to prove the original theorem; the solution can be written as:

$$V(q,x') = \max_{x\in\mathrm{NR}(q,x')} \frac{2\sigma(1-\delta)u)\ell(x) + \delta\Delta(x)\Psi(q)}{2\sigma(1-\delta) + \delta\Delta(x)}$$

where, replicating the previous lemma's argument, we have $\Psi(q) = q\overline{u}(x') - (1-q)\underline{u}(x')$. If $q \leq \underline{q}$,

the derivative for $x > x_{LT}$ is negative and the derivative for $x < x_{LT}$ is positive, meaning the leader is attracted to the learning trap. Hence, $x_\ell(q, x') = \underline{x}_{LT}(q, x')$ or $\overline{x}_{LT}(q, x')$. If $q \in [\underline{q}, \overline{q}]$, the solution is either $\underline{x}_{LT}(q, x')$ or some $x \geq x_{LT}$. If $q \geq \overline{q}$. the derivative for $x > x_{LT}$ is positive and the derivative for $x < x_{LT}$ is negative. Hence the optimal solution is either $\overline{x}(q, x')$ or $\underline{x}(q, x')$. □

1.8 Additional Results

Preference Microfoundation: Land Reform

Overview This subsection provides a microfoundation for the preferences of the "people" and "leader" with an eye to the fiscal setting of Russian land reform. Peasants produce output, which is extracted by the leader via taxation that is siphoned away by elites. We represent the people's utility as the sum of post-tax surplus for the gentry and peasantry. The main mechanism driving the leader's preference for disempowering the gentry — beyond effects on aggregate output by changing peasant production incentives — is that disenfranchising the gentry (high x in the main model) allows them to siphon less tax revenue.

We allow the preference of the leader, in this case, to depend on the state of the world, and relax the assumption of strict monotonicity; however, this does not radically alter the proofs in the previous section.

Setup Consider an economy made up of peasants, gentry, and leader. The union of the peasants and gentry compose the "people." Each period, peasants produce surplus $i(x) \geq 0$, depending on $x \in [0,1]$ and the state of the world (pro or anti-leader). $x = 0$ represents a system of overseen serf labor. $x = 1$ represents a system where peasant households are given private property and labor mobility. x_{LT} represents an arrangement where the commune remains and the peasantry have some economic freedom.

The leader collects revenue from the peasantry in a process intermediated by the gentry by setting a tax $t \in [0,1]$ on $i(x)$. Assume there is some maximal-level of taxation $\bar{t} < 1$ that the peasantry will permit (reflecting e.g. subsistence or overthrow threats) so that the leader sets $t = \bar{t}$ in equilibrium.

Since x reflects not only peasants' economic incentives but also gentry involvement, suppose that when the government taxes at t, the gentry siphons away a share $s(x) \in [0,1]$ of revenue. We assume s is strictly decreasing for $x < x_{LT}$ and constant for $x \geq x_{LT}$. For $x < x_{LT}$, the leader gradually expropriates gentry land, meaning they can siphon less-and-less. For $x \geq x_{LT}$, the leader only has to worry about compensating the gentry for their land (so $s(x)$ is constant but is likely > 0).

Writing y as Post-Tax Surplus Let $u_p(x)$ denote the sum of both peasants' and gentry utility:

$$u_p(x) = (1 - \bar{t} + \bar{t}s(x))i(x)$$

Suppose that for each state of the world, $i(x) =:$

$$\frac{f_g(x)}{1 - \bar{t} + \bar{t}s(x)} \qquad \text{Pro-Leader}$$
$$\frac{f_b(x)}{1 - \bar{t} + \bar{t}s(x)} \qquad \text{Anti-Leader}$$

Assume that $f_g(x), f_b(x) \geq 0$ on $[0,1]$ since $i(x)$ represents output. With this interpretation, y_t in each stage of the main game represent the *post-tax* output left to the "people" (gentry plus peasantry); substituting these expressions for $i(x)$ back into $u_p(x)$ gives $u_p(x) = f_g(x)$ in the pro-leader state and $f_b(x)$ in the anti-leader state of the world. Note that since $f_g(x_{LT}) = f_b(x_{LT})$, $i(x_{LT})$ is the same in both states of the world; hence, both aggregate output and post-tax output are the same in each state of the world, and inference made with either quantity yields similar results. Next, since $1 - \bar{t} + \bar{t}s(x)$ is weakly decreasing in x and positive, $i(x)$ is strictly increasing in x in the pro-leader state of the world, just as $f_g(x)$. The derivative with respect to x in the anti-leader state is

$$\frac{f_b'(x)}{1 - \bar{t} + \bar{t}s(x)} - \bar{t}s'(x)\frac{f_b(x)}{\left(1 - \bar{t} + \bar{t}s(x)\right)^2}$$

which is strictly decreasing for $x > x_{LT}$ and for some region of $x < x_{LT}$ as well. This means that *aggregate* output may not be higher under full-serfdom, but that departing from x_{LT} results in some gains, and that *post-tax surplus* for the sum of the elite and peasant surpluses is higher in this state of the world.

Establishing Leaders' Preferences The leader's revenue is given by $\bar{t}(1-s(x))i(x)$. Using the mapping above, utility $u_\ell(x)$ in each state of the world is

$$\bar{t}(1-s(x))\frac{f_g(x)}{1-\bar{t}+\bar{t}s(x)} \qquad \text{Pro-Leader}$$
$$\bar{t}(1-s(x))\frac{f_b(x)}{1-\bar{t}+\bar{t}s(x)} \qquad \text{Anti-Leader}$$

In the pro-leader state of the world, $u_\ell(x)$ is clearly strictly increasing in x. In the anti-leader state of the world, for the main results of the main model to go through, we simply require that

$$\bar{t}(1-s(x_\ell(0,x')))\frac{f_b(x_\ell(0,x'))}{1-\bar{t}+\bar{t}s(x_\ell(0,x'))} < \bar{t}(1-s(x_{LT}))\frac{y_{LT}}{1-\bar{t}+\bar{t}s(x_{LT})}$$

recalling that $x_\ell(0,x')$ is the solution to the leader's problem in the main model when the anti-leader state of the world. This holds, for example, if $s(x_\ell(0,x'))$ is large relative to $s(x_{LT})$, and would seem to hold in the Russian case, where the gentry have a strong hold in the peasantry during serfdom but have their property expropriated after the reforms (at x_{LT}).

These pieces exemplify precisely how a "divide-the-dollar" land-reform arrangement can serve as a foundation for the preferences of the leader in the model and how these preferences align with features of the 1861 Russian reforms.

External Information Revelation

Proposition 5. *Suppose η increases. Then, $\underline{q}$ decreases and $\overline{q}$ increases. Fixing x', the leader plays the learning trap for fewer values of q.*

Proof. The objective of an unconstrained problem here here can be written as:

$$\max_{x\in[0,1]} \frac{2\sigma(1-\delta)u_\ell(x)+\delta((1-\eta)\Delta(x)+2\sigma\eta)\Psi(q)}{2\sigma(1-\delta)+\delta((1-\eta)\Delta(x)+2\sigma\eta)}$$

The numerator of the derivative can be expressed as:

$$
\begin{aligned}
=&\big(2\sigma(1-\delta)+\delta(1-\eta)\Delta(x)+2\sigma\delta\eta\big)\big(2\sigma(1-\delta)u_\ell'(x)+\delta(1-\eta)\Delta'(x)\Psi(q)\big)\\
&-\delta(1-\eta)\Delta'(x)\big(2\sigma(1-\delta)u_\ell(x)+\delta\big((1-\eta)\Delta(x)+2\sigma\eta\big)\Psi(q)\big)\\
=&4\sigma^2(1-\delta)^2u_\ell'(x)+2\sigma\delta(1-\delta)(1-\eta)\big(\Delta(x)u_\ell'(x)+\Delta'(x)\Psi(q)-\Delta'(x)u_\ell(x)\big)\big)\\
&+2\sigma\delta\eta\big(2\sigma(1-\delta)u_\ell'(x)+\delta(1-\eta)\Delta'(x)\Psi(q)-(1-\eta)\Delta'(x)\Psi(q)\big)\\
=&4\sigma^2(1-\delta)^2u_\ell'(x)+2\sigma\delta(1-\delta)(1-\eta)\big(\Delta(x)u_\ell'(x)+\Delta'(x)\Psi(q)-\Delta'(x)u_\ell(x)\big)\big)\\
&+4\sigma\delta(1-\delta)\eta u_\ell'(x)
\end{aligned}
$$

For $x > x_{LT}$, the derivative is negative, replicating a previous lemma, if:

$$2\sigma(1-\delta)u_\ell'(x)+2\sigma\delta\eta u_\ell'(x)\leq\delta(1-\eta)\xi\big(u_\ell(x)-u_\ell'(x)(x-x_{LT})-\Psi(q)\big)$$

where $\underline{q}$ is the unique point that solves this equation with equality. For each x, the LHS is increasing in η and the RHS is decreasing in η. This means that as η increases from 0, $\underline{q}$ shifts to the left. The equation defining $\overline{q}$ can be written as:

$$2\sigma(1-\delta)u_\ell'(x)+2\sigma\delta\eta u_\ell'(x)=\delta(1-\eta)\xi\big(\Psi(\overline{q})-\big(u_\ell(x)-u_\ell'(x)(x-x_{LT})\big)\big)$$

An increase in η likewise causes the LHS to increase and RHS to decrease, causing $\overline{q}$ to increase. □

Three States of the World

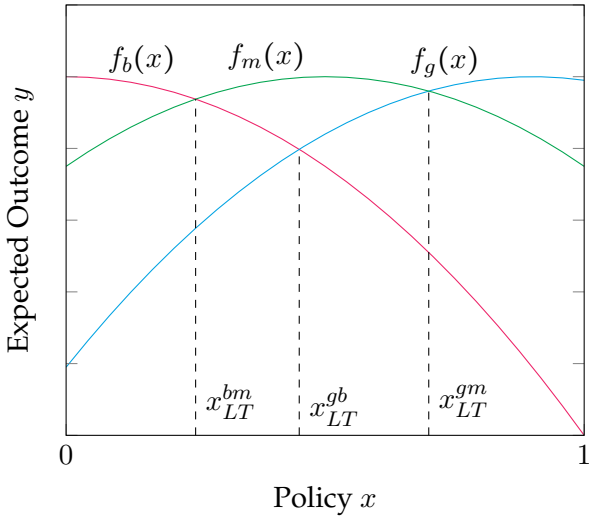
This section shows robustness of this paper's fundamental insight on leaders' policies to three states of the world by establishing a version of Lemma 6. Analogues of Proposition 1 or Theorem 1 are difficult to work with because $\mathrm{NR}(q,x')$ is hard to pin down.

The three states of the world are described as follows

$$
\begin{aligned}
f_g(x) \qquad & y_t=\beta_g-(x_t-\tilde{x}_g)^2+\epsilon_t\\
f_m(x) \qquad & y_t=\beta_m-(x_t-\tilde{x}_m)^2+\epsilon_t\\
f_b(x) \qquad & y_t=\beta_b-x_t^2+\epsilon_t
\end{aligned}
$$

for $\tilde{x}_g > \tilde{x}_m \in (0,1]$ and $\epsilon_t \sim \mathcal{U}[-\sigma,\sigma]$ in all cases, as earlier. We assume $f_i(x) - f_j(x)$ each single-cross at some x_{LT}^{ij} for all $i \neq j$. Examples are graphed below.

Figure 1.5: $f_g(x)$, $f_m(x)$, and $f_b(x)$ graphed



Our state variables are q_g (belief that g holds) and q_m (belief that m holds) and we fix $q_g, q_m > 0$ and $q_g + q_m < 1$ for the remainder of the analysis. Beginning at a vector of states (q_g, q_m), we can write the following expected values of the problem, as well as the probabilities with which they occur (which are all functions of x).

- V^{gmb}: initial value of the problem at (q_g, q_m).
- V^{g}: value of problem if g is true; $(q_g, q_m) \to (1,0)$; occurs with prob. p_g. Solution is denoted x_ℓ^g.
- V^{m}: value of problem if m is true; $(q_g, q_m) \to (0,1)$; occurs with prob. p_m Solution is denoted x_ℓ^m.
- V^{b} value of problem if b is true; $(q_g, q_m) \to (0,0)$; occurs with prob. p_b. Solution is denoted x_ℓ^b.
- V^{gb} value of problem if only m is not true; $(q_g, q_m) \to (\frac{q_g}{1-q_m}, 0)$; occurs with prob. p_{gb}
- V^{gm} value of problem if only b is not true; $(q_g, q_m) \to (\frac{q_g}{q_g+q_m}, \frac{q_m}{q_g+q_m})$; occurs with prob. p_{gm}
- V^{mb} value of problem if only g is not true; $(q_g, q_m) \to (0, \frac{q_m}{1-q_g})$; occurs with prob. p_{mb}

Each of these probabilities can be computed explicitly in terms of the beliefs and x; for example, $p_g(x)$ is given by $q_g \cdot \frac{|f_g(x)-f_b(x)|}{2\sigma} \cdot \frac{|f_g(x)-f_m(x)|}{2\sigma}$.

We assume that $x_\ell^b < x_{LT}^{gb} < x_\ell^g$ and $x_\ell^b < x_{LT}^{mb} < x_\ell^m$. Fixing $1 - q_m - q_b$, suppose the difference between $V^b = u_\ell(x_\ell^b)$ and $V^m = u_\ell(x_\ell^m)$ is sufficiently large. Applying Lemma 6 shows:

- $V^{gb} = x_{LT}^{gb}$: suppressing the risk of information revelation that could reveal b as true, the leader will choose a "pairwise" learning trap x_{LT}^{gb}
- $V^{mb} = x_{LT}^{mb}$: suppressing the risk of information revelation that could reveal b as true, the leader will choose a "pairwise" learning trap x_{LT}^{mb}

This gives us the following result:

Proposition 6. *Suppose the following: $q_g, q_m > 0$ and $q_g + q_m < 1$; the leader's policies are constrained if g, m, or b are revealed as true but can otherwise play any policy in $[x_\ell^b, x_\ell^g]$; and V^b is sufficiently small. Then, the leader either implements a learning trap policy x_{LT}^{gb} or x_{LT}^{mb}.*

Proof. The proof for this result is intuitive. Replicating earlier arguments, the leader's problem can be written as a convex combination of flow and continuation values:

$$V^{gmb} = \max_x \frac{(1-\delta)u_\ell(x) + \delta\big(p_g V^g + p_m V^m + p_b V^b + p_{gb} V^{gb} + p_{gm} V^{gm} + p_{mb} V^{mb}\big)}{(1-\delta) + \delta\big(p_g + p_m + p_b + p_{gb} + p_{gm} + p_{mb}\big)}$$

$$V^{gmb} = \max_x \frac{(1-\delta)u_\ell(x) + \delta\big(p_g V^g + p_m V^m + p_b V^b + p_{gb} x_{LT}^{gb} + p_{gm} V^{gm} + p_{mb} x_{LT}^{mb}\big)}{(1-\delta) + \delta\big(p_g + p_m + p_b + p_{gb} + p_{gm} + p_{mb}\big)}$$

Since we assume V^b is sufficiently small, the leader will do anything in her power to suppress revealing that b is true, i.e. will try its best to set $p_b = 0$. p_b is expressed as:

$$p_b = (1 - q_g - q_m)\frac{|f_g(x) - f_b(x)|}{2\sigma} \cdot \frac{|f_m(x) - f_b(x)|}{2\sigma}$$

With probability $1 - q_g - q_m$, b is true. Then, with probability $\frac{|f_g(x)-f_b(x)|}{2\sigma}$, state b is separable from state g and with probability $\frac{|f_g(x)-f_b(x)|}{2\sigma}$ it is separable from state m. Because the left and right derivatives of p_b are always nonzero, the only way to set p_b equal to zero is either to set $f_g(x) = f_b(x)$ by playing x_{LT}^{gb}, or to set $f_m(x) = f_b(x)$ by implementing x_{LT}^{mb}. The option that is chosen will depend on the parameters of the model (precise beliefs and values of x_{LT}^{gb} and V^m). □

The fundamental insight is that a leader, when faced with a threat of future overthrow, will *always* be attracted to a moderate policy that shuts down information revelation about b. More subtly, the terminal history of the model always reduces to one or two states of the world . Suppose the leader implements x_{LT}^{gb}, allowing information revelation only about whether $\{g, b\}$ or $\{m\}$ is true. In the former case, we arrive at the baseline two-state model. On the equilibrium path, uncertainty is resolved only when upon revelation of states to which the leader is not highly averse.

Model with Non-Uniform Distribution of ϵ_t

Finally, we return to the model with myopia and solve a version of the model assuming that $\epsilon_t \sim p$, where $p(\cdot)$ is continuous, has mean 0, is symmetric around 0, and single-peaked. Denote the distribution of posteriors q_t given a prior q_{t-1} and today's policy x_t as $\tau(q_t|q_{t-1}, x_t)$. Note that, if $|x_t' - x_{LT}| > |x_t - x_{LT}|$, for $x \leq \tilde{x}$, because the expected value of the posterior is the prior, $\tau(q_t|q_{t-1}, x_t')$ is a mean preserving spread of $\tau(q_t|q_{t-1}, x_t)$; that is, policies farther from x_{LT} reveal more information.

Next, note that in the model with myopia, $\mathrm{NR}(q) = \mathrm{NR}(q, x') = [\tilde{x}q - \sqrt{c}, \min\{\tilde{x}q + \sqrt{c}, 1\}]$ is convex in (q, x). If the people are not myopic and have a large preference for learning, this assumption may not necessarily hold, since the people simply prefer any extreme policy that reveals information.

The leader's problem can now be expressed as:

$$V(q) = \max_{x_\ell(q) \in \mathrm{NR}(q)} \underbrace{(1-\delta)u_\ell(x)}_{\text{Flow utility}} + \underbrace{\delta \int V(q')\tau(q'|q, x)}_{\text{Value of information}},$$

When the leader implements a policy x, this policy potentially generates information, in addition to flow utility. Note that, because the expected value of the posterior is the prior, the (continuation) value of information is simply $V(q)$ integrated over the distribution $\tau(q'|q, x)$. When $x = x_{LT}$, $\tau(q'|q, x)$ places mass 1 on the prior q; as x grows more extreme, the variance of posteriors increases. If V is concave in q, there is hence an aversion to information revelation.

The following proposition characterizes $x_\ell(q)$; the proof is identical for arbitrary $\mathrm{NR}(q, x')$ as

long as this set is convex.

Proposition 7. *There exists a threshold $\underline{q} \in (0,1]$ such that:*

1. *if $q \leq \underline{q}$, $x_\ell(q) = \overline{x}(q)$, where $\underline{q}$ solves $\tilde{x}\underline{q} + \sqrt{c} = x_{LT}$. The leader plays the most extreme policy preventing overthrow.*

2. *If $q \geq \underline{q}$, the leader implements $x_\ell(q) \in [\min\{x_{LT}, \underline{x}(q)\}, \overline{x}(q)]$. Moreover, for δ high, $x_\ell(q) < \overline{x}(q)$; i.e., she moderates her actions to pursue the learning trap.*

Proof. Suppose q is such that $\overline{x}(q) \leq 1$. Because $\mathrm{NR}(q)$ is convex in (q, x), by Theorem 9.8 in Stokey, Lucas, and Prescott (1989), $V(q)$ is concave in q. In particular, V's concavity in the belief q suggests a general *aversion* to information revelation, and hence a tendency to pursue policies closer to x_{LT} whenever a leader values this information sufficiently, i.e. when she is patient.

Let $\underline{q} \in (0,1)$ solve $\tilde{x}\underline{q} + \sqrt{c} = x_{LT}$. Suppose $q \leq \underline{q}$. Note that, on this range, an increase from x to x' has two effects. First, it increases flow utility $u_\ell(q)$. Second, because $q \leq \underline{q}$, we must have $x_{LT} \geq x' > x$, so $|x - x_{LT}| > |x' - x_{LT}|$, meaning $\tau(q'|q, x')$ is a mean-preserving contraction of $\tau(q'|q, x)$. This means that, because $V(q)$ is concave in q, $\int V(q')\tau(q'|q, x') > \int V(q')\tau(q'|q, x)$, i.e. the value of information increases. Hence, the leader will pursue $\overline{x}(q)$, which is both the highest possible policy and the closest policy to the learning trap. This shows the first part of the proposition.

Next, suppose $q > \underline{q}$. By the previous argument, all feasible $x < x_{LT}$ are dominated by x_{LT}. Now, above x_{LT}, an increase from x to x' has two effects. First, it increases flow utility: $u_\ell(x') > u_\ell(x)$. Second, because V is concave, it *decreases* the value of continuation utility, i.e. $\int V(q')\tau(q'|q, x') < \int V(q')\tau(q'|q, x)$. For δ large, this second informational effect will tend to outweigh the former flow effect, suggesting that $\min\{\underline{x}(q), x_{LT}\} \leq x_\ell(q) < \overline{x}(q)$. □

Chapter 2

Persuasion with Partisanship

2.1 Introduction

Scholars have documented a rise in partisanship in the U.S. Congress since the 1990s that is apparent in political speeches, voting behavior, and policy proposals.[1] However, the role of the executive branch in shaping partisanship has been relatively understudied. Executives have played an increasingly prominent role in policymaking, so much so that they must often successfully pass and implement their policy agendas to win reelection.[2] Since passing partisan policies often requires more political skill than policies appealing to many parties, voters may assess the success or failure of a partisan policy agenda differently than a bipartisan one. Changes in what executives need to communicate about their skill to win reelection may hence generate variations in partisanship differing from those in legislative settings.

In this paper, I use a theoretical model to study electoral incentives for political executives to pursue partisan policies. I then apply its insights to explain empirical variations in the partisanship of U.S. governors' policy agendas. I develop a model where an incumbent's policy agenda generates an *information structure* mapping her political ability to a likelihood of policy success.

[*]Chas Brown and Kevin Guan provided excellent research assistance in hand-coding speech snippets used in the data for this chapter.

[1]See McCarty (2019) for an overview of the forces driving trends in partisanship.

[2]Andeweg, Elgie, Helms, Kaarbo, and Muller-Rommel (2020) discuss the reemergence of executive power since the mid-20th century, as well as the expanding role of the U.S. Presidency and its administrative apparatus in shaping legislative discourse. Goldgeier and Saunders (2018) note a rise in the executive branch's control over foreign policy, which has become increasingly left unchecked by Congress. Reynolds (2024) and Peterson (2024) provide stylized overviews of the eroding power of Congress over the executive branch. The burgeoning power of U.S. state governors, especially after the 1990s, is discussed by Heidbreder (2012).

Partisan policies that appeal only to the incumbent's party require more skill to pass, meaning their failure can be attributable to legislative opposition, while bipartisan policies that appeal to the opposition require less skill to pass, meaning their failure signals low ability. The model predicts a U-shaped relationship between an incumbent's popularity and the partisanship of her agenda. Unpopular incumbents pursue partisan agendas to salvage reelection. Popular incumbents propose partisan policies whose failure can be blamed on legislative antagonism, allowing them to maintain their popularity. Incumbents in the middle pursue safe bipartisan policies to secure reelection. The U-shape weakens as the incumbent's alignment with the legislature increases.

I then study the partisanship of U.S. gubernatorial policy agendas and test the predictions of the model. I compile a panel of governors' annual "State of the State" addresses to the legislature. The presence of 50 governors at any given time across the states offers a rich laboratory for uncovering the executive branch's role in generating variation in partisanship. I utilize large-language modeling to isolate text related to policy proposals. I measure the partisanship of these proposals via divergence in phrases used by Democrats and Republicans. I show that while aggregate gubernatorial partisanship has steadily increased since the early 2000s, it is lower than that estimated from the Congressional Record and that there are substantial deviations from these trends at the governor-level. [3]

With this heterogeneity in mind, I turn to the panel data to test the model's predictions. I compare the partisanship of reelectable governors' proposals to those of lame duck governors ineligible for election. This controls for forces that influence partisanship of both lame duck and reelectable governors — such as preferences for partisan policies — and thereby isolates variations in partisanship emerging from electoral concerns. I show a nonmonotonic relationship between gubernatorial approval and partisanship of policy proposals for reelectable governors, in line with the model's prediction. This relationship weakens as legislative alignment increases. Finally, I argue that the model's qualitative insights can be applied to study the policy choices of executives like the U.S. President.

[3]Specifically, the level of gubernatorial partisanship is below that of the Congressional level estimated by Gentzkow, Shapiro, and Taddy (2019).

Theoretical Model The baseline model considers an incumbent executive of unknown ability who chooses a policy agenda of varying partisanship. I interpret ability as an executive's managerial skill, capacity to follow through on promises, or valence, and refer to expected ability as *reputation*. An agenda's probability of passing is decreasing in its partisanship but increasing in the incumbent's ability. This captures the notion that partisan policies appealing to a single party are harder to garner support for than bipartisan compromises amenable to most of the legislature, which even low-skilled politicians may pass. A representative voter observes the incumbent's choice of agenda, whether it passes or fails, updates beliefs over the incumbent's ability, and reelects her if and only if her reputation is sufficiently high. Because incumbents are office-motivated, the predictions of the model capture the behavior of reelectable incumbents relative to lame ducks.

The key idea of the model is that an incumbent's choice of policy is, in fact, also a choice of information structure, as in models of persuasion. Given a prior over the incumbent's ability, the success or failure of an agenda generates a distribution of posteriors over ability whose skewness is determined by the agenda's partisanship. The incumbent chooses this distribution — the partisanship of her policy agenda — to maximize chances of reelection. The informativeness of success or failure is disciplined by the incumbent's alignment with the legislature. Intuitively, an incumbent aligned with her legislature has power over the government and receives credit or blame for policy successes or failures. An incumbent unaligned with her legislature can often blame policy failures on an uncooperative legislature.

The baseline model highlights a fundamental nonmonotonicity in how partisan the incumbent's policy agenda is as her reputation increases. This is the core theoretical prediction of the model. Partisan policies are likely to fail, but reveal a high ability incumbent if they succeed. Bipartisan policies are likely to succeed, but reveal a low ability incumbent if they fail. Hence, low reputation incumbents likely to lose reelection pursue partisan policies to save their chances of retaining office. Middling reputation incumbents pursue low-risk bipartisan policies to secure reelection. High reputation incumbents pursue partisan policies whose failure can be blamed on legislative antagonism, meaning the incumbent's reputation falls by relatively less. Therefore, the model delivers a U-shaped relationship between an incumbent's reputation and the partisanship of her policy agenda.

This U-shape flattens out if legislative alignment increases. Decreasing electoral competition also makes the U-shape less salient — for example, in constituencies where an incumbent's party always wins reelection. I show that voter utility — measured as the value of learning about an incumbent's ability or flow utility provided by the incumbent's policy — is generally nonmonotonic in reputation, legislative alignment, and electoral competition.

After presenting this baseline model, I extend it by allowing endogenous differentiation of policy passage by ability. This generalization renders the model similar to a Bayesian Persuasion problem where a sender cannot totally obfuscate information; in this case, the incumbent cannot pursue a policy agenda whose success or failure is totally uninformative of her ability. [4] The mathematical flexibility of the general approach allows me to add a preference shock and write the model's predictions in terms of ex-ante *winning probability*.

I show that the partisanship of the incumbent's policy agenda indeed exhibits a U-shape as a function of ex-ante winning probability, which flattens when legislative alignment increases or competition weakens. Since proxies for politicians' winning probability — approval ratings — are readily available, the generalization delivers testable predictions of the theory: the partisanship of an executive's policy agenda should follow a U-shape as a function of her approval rating and flatten under high legislative alignment or low electoral competition.

The more general model shows that the U-shape relationship between reputation and partisanship may in fact be asymmetric. Popular incumbents may pursue *more partisan* agendas than unpopular incumbents, since low reputation incumbents prefer policies that differentiate high and low-ability incumbents, while high reputation incumbents prefer policies whose failure is as uninformative as possible. I additionally show, using the technology of the generalization, that the model's key predictions are robust to allowing multiple periods of agenda passage or allowing incumbents to choose policies from the opposing party.

Documenting Gubernatorial Partisanship I apply the insights of the model to explain variation in the state-level partisanship of U.S. governors' policy agendas. I begin by assembling a cor-

[4] In particular, the model can be seen as a Bayesian Persuasion problem where a receiver (voter) is uncertain whether he observes a signal because it was endogenously chosen by a sender (incumbent) or whether it was generated by an outside source (i.e. whether a policy outcome is a consequence of the incumbent's inherent skill or the legislature's involvement).

pus of U.S. governors' "State of the State" speeches. These addresses to the state legislature are given in the first quarter of each year and lay out the governor's policy proposals for the coming year. I collect over 800 speeches and merge them with existing data to assemble an annual panel of over $1,300$ speeches from 1990-2020. Because these speeches also contain reflections on past policies and rhetoric, I fine-tune a large-language-model called "BERT" (Devlin, Chang, Lee, and Toutanova, 2018) to isolate speech related to policy proposals. I construct an aggregate measure of partisanship of policy proposals for each year and a panel measure for each governor-year by measuring divergence of phrase-use between Republican and Democratic governors using Gentzkow et al. (2019).

My aggregate series shows a steady rise in partisanship of policy proposals beginning in the early 2000s that finishes with a spike in 2017.[5] Gubernatorial partisanship is consistently lower than Congressional partisanship, and the timing of changes is different. However, I show that these trends disguise sizable variation in partisanship at the governor level, such as spikes in the late 1990s surpassing the levels of the late 2010s and dips in the mid-2010s lower than those in the 1990s, raising questions about the sources of this heterogeneity.

Testing Model and Implications I then use the gubernatorial panel to explain variation in the partisanship of policy proposals through the lens of my model. To capture the effects of winning probability and legislative alignment, I interact deciles of gubernatorial approval ratings with an indicator for whether the governor's party controls the state legislature. I study how the governor-year measure of partisanship moves with changes in these variables. To highlight how electoral incentives drive variation in partisanship, I compare the partisanship of reelectable and lame duck governors, thereby controlling for governor-level policy preferences or time trends that drive partisanship but impact both groups.

I show that when comparing reelectable governors to lame ducks, partisanship of proposals exhibits a U-shape as a function of approval decile for governors unaligned with their legislatures. The speech of high or low approval governors is over one standard deviation higher relative to middle approval governors. The U-shape is weaker for governors aligned with their legislatures and states where Democratic governors consistently win elections.

[5] These trends are identical when estimated on the full corpus of governor speech.

These insights suggest that partisanship should be viewed not as a slow-moving aggregate but as a dynamic force shaped by electoral and strategic concerns, popularity, and legislative alignment. They offer a reputational explanation for why executives compromise with the opposition, assign credit or blame for policy outcomes, or pursue radical partisan reforms.

Literature This paper contributes to a theoretical literature on political reputation and career concerns by developing a model that exhibits a nonmonotonic relationship between policy extremity and reputation. By contrast, most other settings establish a negative monotonicity known as "gambling for resurrection" (Dur, 2001; Fu and Li, 2014; Izzo, 2024; Majumdar and Mukand, 2004), which would not explain the nonmonotonicities evident in the gubernatorial data on partisan policy proposals. Additionally, I develop a model where partisanship endogenously emerges from an incumbent's desire to win reelection in a Bayesian setting, distinguishing the present paper from models where partisanship emerges as a byproduct of voters' behavioral updating methods (Izzo et al., 2021; Levy and Razin, 2021b). The paper's welfare insight on the ambiguous effects of electoral competition is also related to Dewan and Hortala-Vallve (2019), is studied in the settings of electoral accountability by Ashworth et al. (2017) and Bils and Izzo (2022), and is applied to executives' willingness to pursue unilateral actions in Judd (2017). These papers show how equilibrium outcomes are sensitive to parameters of competition, office rents, and voter beliefs. My model predicts a singular U-shaped relationship between partisanship and popularity that shifts or flattens with such changes. A somewhat similar nonomonotonicity has been established in settings of political pandering, most notably Canes-Wrone, Herron, and Shotts (2001) and Maskin and Tirole (2004), which establish that low and high popularity incumbents pursue policies that reveal their competence, while those in the middle pursue uninformative policies that "pander" to voters' priors over optimal policy. This insight is driven by an assumption of voter inattention: voters may not observe the outcomes of executives' policy choices, allowing middling-popularity incumbents to pursue policies that generate little accountability. Relaxing inattention erodes any incentive to obfuscate information and results in all politicians pursuing the same policy.[6]

[6]Canes-Wrone et al. (2001) assume executives want to both win reelection and implement the optimal policy. The latter preference drives non-pandering in edge cases of their model, which is absent if politicians only want to win reelection.

The present paper provides insights into solving a constrained Bayesian Persuasion problem, following Kamenica and Gentzkow (2011), where a receiver is uncertain about the origin of a sender's signal. The key constraint on the sender (incumbent) in my environment is that she is unable to implement totally uninformative information structures (policy agendas). A lemma in the appendix shows how to solve this problem when senders possess concave, strictly increasing value functions, which is the primary setting in which inability to shut down information binds. To this end, the model is related to a literature on Bayesian Persuasion with mediation, multiple senders, and cheap talk (Alonso and Câmara, 2016; Arieli, Babichenko, and Sandomirskiy, 2022; Ichihashi, 2019; Lipnowski, Ravid, and Shishkin, 2022).

I contribute to the empirical literature on political accountability by comprehensively linking ideological partisanship to legislative alignment, popularity, and reelection eligibility. Prior papers study such topics as the effects of term limits and career-concerns on partisanship (Besley and Case, 1995; Iaryczower, Lopez-Moctezuma, and Meirowitz, 2024); the effect of reelection on welfare reform (Bernecker, Boyer, and Gathmann, 2021) or COVID-19 stringencies (Pulejo and Querubín, 2021); or the economic returns to holding a Congressional seat (Diermeier, Keane, and Merlo, 2005). I address the paper's connection to a literature on voter responsiveness to executives' policy choices in the data section.

I add to research on partisanship of political speech by providing a novel, unidimensional series on partisanship of political executives' speech and partisanship at the state level. Series for partisanship of speech have been studied in the U.S. Congressional Record (Gentzkow et al., 2019; Jensen, Naidu, Kaplan, Wilse-Samson, Gergen, Zuckerman, and Spirling, 2012) and other countres like the UK (Peterson and Spirling, 2018), but the few studies at the state level have been limited to counts of pre-selected words or nationalization trends (Butler and Sutherland, 2023; Hopkins, Schickler, and Azizi, 2022). I also contribute to a literature analyzing State of the State speeches, which I discuss in the data section.

My utilization of advances in large-language-modeling techniques adds to a growing literature in Economics utilizing LLMs, including applications to labor contracts and the FOMC (Arold, Ash, MacLeod, and Naidu, 2024; Hansen, McMahon, and Prat, 2018), which are reviewed by Gentzkow, Kelly, and Taddy (2019) and Ash and Hansen (2023). I utilize the same methods as Card, Chang, Becker, Mendelsohn, Voigt, Boustan, Abramitzky, and Jurafsky (2022), who fine-tune two models

to isolate speeches related to immigration in the Congressional Record and then categorize those speeches by sentiment.

Finally, I complement a literature on the mixed (and heterogeneous) behavior of partisanship in state and local politics utilizing roll-call data, close elections, and other non-text measures. Some studies argue that governor and mayoral party identity has little impact on partisanship of most political outcomes (Ferreira and Gyourko, 2009; Leigh, 2008), while others argue for modest or larger effects of party control on partisanship after the 1990s, measured using liberalism scores; law passage by policy domain; taxation, debt, housing stock, and public expenditures; or policy diffusion across states (Carlino, Drautzburg, Inman, and Zarra, 2023; Caughey, Xu, and Warshaw, 2017; de Benedictis-Kessner, Jones, and Warshaw, 2024; de Benedictis-Kessner and Warshaw, 2020; DellaVigna and Kim, 2022; Grumbach, 2018). Increases in partisanship of legislative voting behavior at the state level have been documented by Shor and McCarty (2011) and updated by DellaVigna and Kim (2022), although evidence is again heterogenous by state. By viewing partisan variation as products of incumbents' electoral incentives, my paper provides structure to these more heterogeneous outcomes from the standpoint of executives. To this end, the present paper differs from attempts to measure partisanship of inherent gubernatorial ideology, such as the campaign finance DIME measure constructed by Bonica (2014) and evaluated for governors in Warner (2023).

Plan of the Paper I introduce the theoretical model in the next section. I solve the baseline model, investigate welfare implications of the equilibrium result, solve the general model with a valence shock, detail data predictions, and address extensions. I then introduce the data, explain the relevance of the model to the gubernatorial setting, and detail estimation of partisanship. Next, I study how aggregate trends in partisanship mask substantial heterogeneity in the panel data. With this heterogeneity in mind, I test the predictions of the model. I conclude with a qualitative analysis of U.S. Presidents through the lens of the model and comment on the broader implications of the research.

2.2 Theoretical Model

I begin by introducing a benchmark model highlighting the key intuitions of the theory. After showing the central result and assessing welfare, I move to an extended model that maintains the key intuitions of the benchmark while permitting mathematical flexibility to solve the model with a valence shock. This allows me to write the model's comparative statics in terms of the incumbent's win probability and thereby utilize approval ratings as an independent variable when testing the model's comparative statics in the empirical section. I finish by addressing two extensions to the model.

2.2.1 Preliminaries

Setup There are two time periods. At $t = 1$, an incumbent politician R (she) holds office. R chooses a policy agenda $\pi \in [0, 1]$, which may pass or fail. Higher π represents a more partisan agenda that favors R's party. The probability of passage is increasing in R's ability $a_R \in \{0, 1\}$, which is unknown to all agents.[7] Higher ability politicians are more effectively able to pass and/or implement legislation.

A representative voter V (he) observes R's choice of agenda, whether it passes or fails, and updates his beliefs about R's ability. Then, at $t = 2$, V chooses to retain R or replace her with a challenger L of unknown ability $a_L \in \{0, 1\}$. q_i^t is the belief politician i is high ability ($a_i = 1$) at time t, which we refer to as i's "reputation." V's utility is the ability a_i of politician i in office. Politicians are office-motivated, receiving a payoff of 1 upon reelection and 0 otherwise.

Policies The success of R's policy agenda is a function of its partisanship $\pi \in [0, 1]$; R's ability a_R; and the legislature's alignment with R, $\lambda \in [0, 1]$, which represents the relative strength of R's party in the legislature. The probability of passage is:

$$\lambda a_R + (1 - \lambda)(1 - \pi).$$

Legislative alignment λ dictates the relative control of the incumbent over the legislature and, relatedly, how much blame can be placed on the incumbent for a success or failure. When $\lambda = 1$,

[7] The model's equilibrium is robust to an incumbent privately knowing her type.

the legislature is totally aligned with the incumbent. Success or failure are entirely driven by R's ability. When $\lambda < 1$, the legislature involves itself in the passage of policies so that passage is a weighted combination of R's ability and her agenda's partisanship. Pursuing the least partisan agenda $\pi = 0$ always leads to success when $a_R = 1$ and may still succeed when $a_R = 0$. Pursuing the most partisan agenda $\pi = 1$ succeeds only if $a_R = 1$, and even then may still fail. For $\lambda = 0$, only partisanship affects whether an agenda passes.

The choice of $\pi \in [0, 1]$ then amounts to the choice of an *information structure*. That is, the agenda maps a state (the incumbent's underlying ability) to a distribution of signals (success or failure of that agenda). The voter infers the incumbent's type from these signals and uses this to update his beliefs over how competent the incumbent is.

Comments Because incumbents are office-motivated, the model should be thought of as isolating variation in partisanship due to reelection incentives. Adding flow utility over partisan policies (e.g. due to an incumbent's preferences or external constraints) would add a unidirectional preference for more partisan policies on top of that of the model. Because such forces would also operate if an incumbent were term-limited, the predictions of the model should emerge when comparing reelection eligible incumbents to lame ducks.[8] This comparison controls not only for executives' preferences but also contemporaneous shocks that affect the partisanship of both reelectable and lame duck executives equally, and will serve as a cornerstone of the empirical tests of the model.

The model interprets ability as a politician's managerial skill, which may include efficacy in crafting and implementing laws, managing bills and the budget, and maintaining a smooth-running bureaucracy. Governors themselves often use the analogy of a "CEO" to describe the duties and skills ascribed to their job (Behn, 1991). Ability can also represent a capacity to commit to following through on a policy agenda; or the incumbent's valence, which is higher if she is perceived to be better at passing legislation. The data section provides examples of these skills from the National Governors Association, noting that the success or failure of a policy agenda is one of the most crucial inputs into voters' perceptions of governors.

[8]Relatedly, because partisan policies are identified by their informational feature that they are hard to pass, an incumbent facing an aligned legislature may counterintuitively attempt to pass policies from the opposing party to signal her skill. I show in an extension at the end of the section that allowing for this accentuates the model's result.

I assume that the representative voter V prefers a high ability incumbent to a challenger of unknown ability to a low ability incumbent. This assumption may capture two sorts of phenomena. First, V may be a member of R's party and, all else equal, prefers that the executive of her party is high ability rather than low ability. The second interpretation is that V is a median voter who may have positive or negative ideological preferences over R's partisan policies but, regardless of these preferences, values R's managerial ability to address nonpartisan tasks of governing. As long as V values this managerial ability sufficiently than these ideological preferences, which is highly plausible in a setting like that of U.S. governors, the findings of the model should hold. I provide a microfoundation for this phenomenon in Section 2.7.2. [9]

I assume that all else equal, bipartisan policies are more likely to pass than partisan policies. In many polarized electoral environments, passing even bipartisan legislation may be seen as a success. What matters for the analysis is that even if passing bipartisan policies is a strong signal of political skill, passing even more partisan policies is an even stronger signal of political skill. To this end, bipartisan policies should be thought of as *ideologically* bipartisan (or even nonpartisan) policies, as opposed to bills requiring bipartisan *support* to pass. For U.S. governors, bipartisan or nonpartisan policies may include such topics as infrastructure investing or education spending. [10]

Notice that in the model, independently of the incumbent's policy choice, some information is bound to be revealed about her type with positive probability. This feature is motivated by the fact that, in high-stakes political settings, incumbents oversee programs run by the executive branch, basic functioning of government, and passage of "business as usual" legislation.

The model defines partisanship as *relative* to a legislature's composition. Suppose λ increases to λ'. A bipartisan policy for the λ' legislature will likely be more partisan than a bipartisan policy

[9] In particular, I assume there are two dimensions of policy that are functions of ability: managerial and ideological. Voter utility in the managerial dimension is always increasing in ability. Voter utility in the ideological dimension is given by an absolute loss function in π with a weight δ. I assume that the representative voter's preferences align with the legislature, writing her bliss point as λ; and assume that this loss is experienced only if the incumbent's agenda π is passed. I show that as long $\delta \leq 4$, the insights of the model hold.

[10] Curry and Lee (2020) provide a comprehensive picture of the landscape of (bi)partisan bill passage in the U.S. Congress. One Congressional staffer notes that "[a] bill still has a lot greater chance to make it into law if it's bipartisan" (p. 44), while another states that "[g]etting anything done is hard, but it's even harder on a partisan basis" (p. 45). Another staffer points out that while some bipartisan legislation may be passed with easy support from both parties, more ambitious legislation is only able to pass with negotiation between party leaders and veto-holders, suggesting a certain level of persuasion skill or political acumen — such as that at the level of a political executive — to secure passage of more partisan initiatives. The authors additionally point out that, despite increasing party polarization, the sorts of bipartisan support and tactics required to secure legislation is empirically no different than the 1970s and 1980s.

for the λ legislature. π, to this end, identifies partisanship relative to a base level for the legislature.

Voters in the model can learn about an incumbent but not the challenger, meaning the expected ability of a challenger is effectively an outside option for voters. I take the expected ability of the challenger to generally describe how electorally competitive the incumbent's jurisdiction is. Voters learn about the incumbent's ability precisely because she is in office, and her ability can be assessed through performance of regular duties.

2.2.2 Equilibrium

The voter elects whichever politician has higher expected ability. R wins reelection if and only if $q_R^2 \geq q_L^2$. Hence, the equilibrium utility of R as a function of q_R^2 is $u_R(q_R^2) \equiv \mathbf{1}[q_R^2 \geq q_L^2]$.

Information Given a prior q_R^1, the partisanship of each policy agenda generates posteriors q_R^2 about R's ability, given by $F(q_R^2|q_R^1, \pi)$. In the setting with success/failure, we end up with either a posterior higher than q_R^1 (denoted $\overline{q}_R^2$) or lower than q_R^1 (denoted $\underline{q}_R^2$). The incumbent's general problem is then given by:

$$V_R(q_R^1) = \max_{\pi \in [0,1]} \int_{q_R^2} u_R(q_R^2) dF(q_R^2|q_R^1, \pi). \tag{2.1}$$

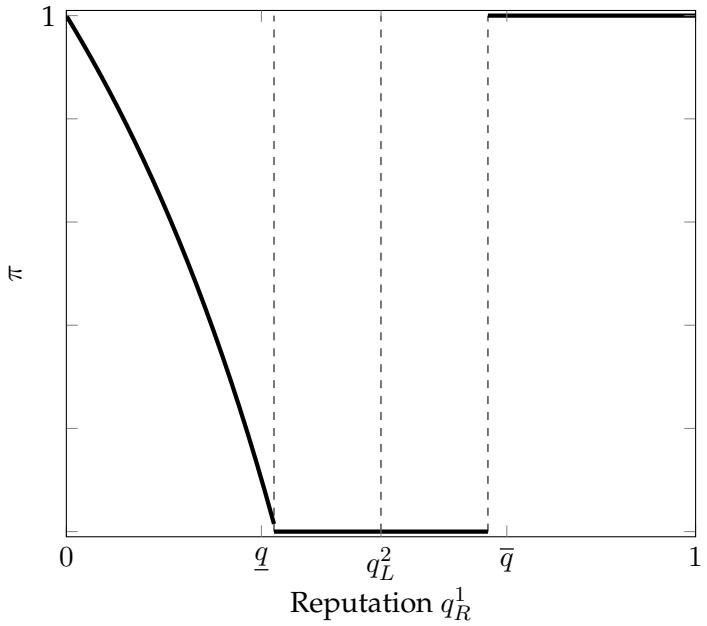
Let $\pi^*(q_R^1)$ represent the solution of the program above. The following proposition describes π^*.

Proposition 1. *There exist thresholds* $0 < \underline{q} < q_L^2 < \overline{q} < 1$ *such that:*

- *for* $q_R^1 \in [0, \underline{q})$, $\pi^*(q_R^1)$ *is strictly decreasing, with* $\pi^*(0) = 1$ *and* $\pi^*(\underline{q}) = 0$*;*
- *for* $q_R^1 \in [\underline{q}, \overline{q})$, $\pi^*(q_R^1) = 0$*;*
- *for* $q_R^1 \in [\overline{q}, 1]$, $1 \in \pi^*(q_R^1)$ *with equality at* $\overline{q}$*.*

The solution is graphed below:

Figure 2.1: Partisanship of Optimal Policy π^*



All proofs are contained in Section 2.7. The intuition for this result is best viewed through comparing the payoffs from $\pi = 0$ and $\pi = 1$. When $\pi = 0$, observing a success generates the posterior over ability $\overline{q}_R^2 = \frac{q_R^1}{q_R^1+(1-\lambda)(1-q_R^1)}$. Observing a failure generates the posterior $\underline{q}_R^2 = 0$, since failure to pass bipartisan policies is damning evidence that $a_R = 0$.

When $\pi = 1$, observing a success is confirmatory evidence that $a_R = 1$, so $\overline{q}_R^2 = 1$. Because high ability incumbents may fail to pass partisan policies, the posterior upon observing failure is $\underline{q}_R^2 = \frac{(1-\lambda)q_R^1}{1-\lambda q_R^1}$. Hence, the expected winning probabilities from pursuing each policy are:

$$\pi = 1: \qquad (q_R^1 + (1-\lambda)(1-q_R^1))\mathbf{1}\left[\frac{q_R^1}{q_R^1+(1-\lambda)(1-q_R^1)} \geq q_L^2\right],$$
$$\pi = 0: \qquad \lambda q_R^1 + (1-\lambda q_R^1)\mathbf{1}\left[\frac{(1-\lambda)q_R^1}{1-\lambda q_R^1} \geq q_L^2\right].$$

Consider the following cases:

- Suppose q_R^1 is so low that $\frac{q_R^1}{q_R^1+(1-\lambda)(1-q_R^1)} < q_L^2$. Then R never wins if she pursues $\pi = 0$. By contrast, if she pursues $\pi = 1$, she wins with positive probability, since successes are confirmatory news of high ability. Hence, $\pi = 1$ dominates $\pi = 0$.

- Suppose q_R^1 is larger so that $\frac{q_R^1}{q_R^1+(1-\lambda)(1-q_R^1)} \geq q_L^2$ but such that $\frac{(1-\lambda)q_R^1}{1-\lambda q_R^2} < q_L^2$. The expected win

probability from pursuing $\pi = 0$ is $q_R^1 + (1-\lambda)(1-q_R^1)$, whereas the expected win probability from pursuing $\pi = 1$ is only λq_R^1. Hence $\pi = 0$ dominates $\pi = 1$.

- Finally, suppose q_R^1 is high, so that $\frac{(1-\lambda)q_R^1}{1-\lambda q_R^1} \geq q_L^2$. Pursuing $\pi = 1$ then leads to a win with probability 1, which dominates $\pi = 0$.

Note that for $q_R^1 \geq \overline{q}$, π^* may contain other values besides $\pi = 1$ because win probability is flat above q_L^2. Any indifference can be broken in favor of $\pi = 1$ by assuming that, all else equal, R prefers more partisan policies in her direction.

The proof of the baseline result yields the following comparative statics.

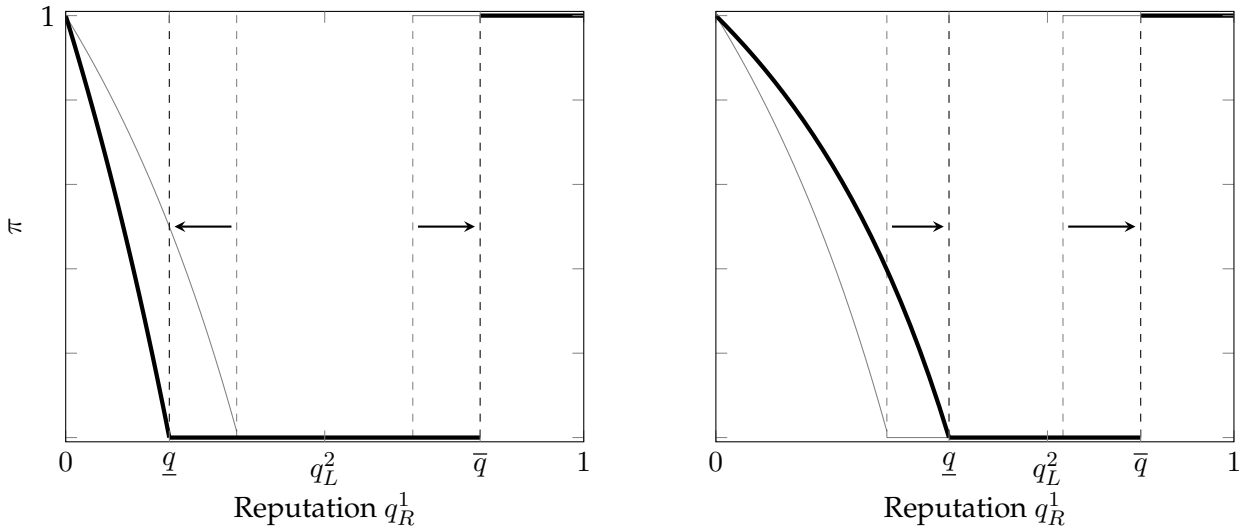
Corollary 1. *$\overline{q}$ and $\underline{q}$ vary as follows with λ:*

- *$\underline{q}$ is decreasing in λ; as $\lambda \to 0$, $\underline{q} \to 0$; as $\lambda \to 1$, $\underline{q} \to q_L^2$.*
- *$\overline{q}$ is increasing in λ; as $\lambda \to 1$, $\underline{q} \to 1$; as $\lambda \to 0$, $\overline{q} \to q_L^2$.*

$\overline{q}$ and $\underline{q}$ are both increasing in q_L^2. As $q_L^2 \to 0$, both terms $\to 0$. As $q_L^2 \to 1$, both terms $\to 1$.

Recall that $1 - \lambda$ also measures R's capacity to blame outcomes on the legislature. As λ increases, this obfuscatory power decreases, meaning in particular that the failure of partisan policies is a more negative signal of ability. Bipartisan policies $\pi = 0$ are the "safest" in the sense that they have a high probability of success; and so, incumbents lean more on bipartisan policies for their relative safety. As q_L^2 increases, the electoral environment becomes more competitive, i.e. R needs to achieve a higher threshold to win reelection. This ultimately causes a "right shift" in the incentives of the model, as R requires more partisan policies at the low end of q_R^1 to signal sufficiently high ability. Flipping the result suggests that in *less* competitive environments — e.g. those where an incumbent's party always tends to win reelection — we should observe a "left shift" in the baseline predictions.

Figure 2.2: Variation of Optimal Policy π^* with λ, q_L^2



(a) Increase in Leg. Alignment λ

(b) Increase in Competitition q_L^2

2.2.3 Welfare

I show that voter welfare — measured as the additional value of learning or flow utility — may be nonmonotonic with respect to reputation; and that changes in legislative alignment and competition may have ambiguous effects on welfare.

Learning

The first benchmark for welfare will be voter learning. The voter's value function at the beginning of $t = 2$, as a function of his beliefs, is $\max\{q_2^R, q_2^L\}$. With some abuse of notation, the value of learning for the voter given $\pi^*(q_R^1)$ can be written as:

$$V_V(\pi^*) = \int_{q_R^2} \max\{q_2^R, q_2^L\} dF(q_R^2 | q_R^1, p_1^*, p_0^*),$$

which can be summarized as follows.

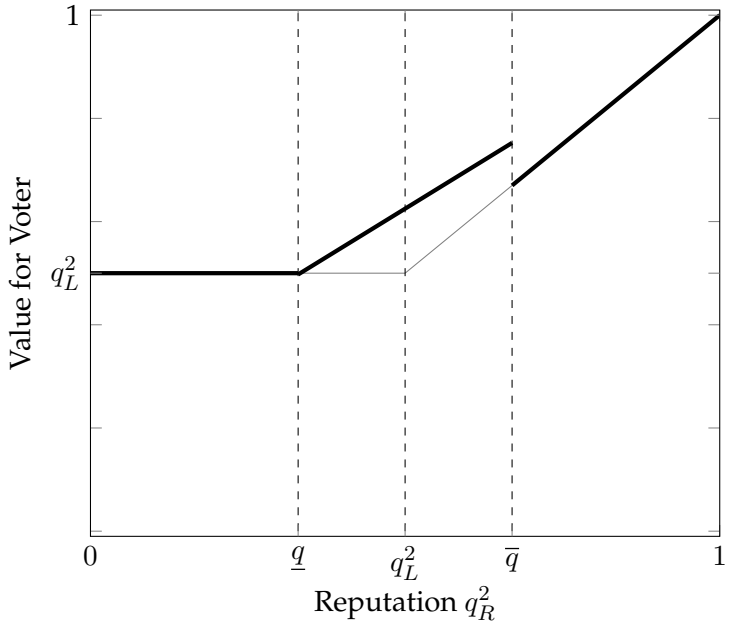
Proposition 2. *The equilibrium value of learning for the voter, $V_V(\pi^*)$, is as follows.*

- *For $q_R^1 \in [0, \underline{q}]$, $V_V(\pi^*) = q_L^2$;*
- *for $q_R^1 \in [\underline{q}, \overline{q})$, $V_V(\pi^*) = q_R^1 + \lambda(1 - q_R^1)q_L^2$;*

- *for* $q_R^1 \in [\overline{q}, 1]$, $V_R(\pi^*) = q_R^1$.

The value of learning is graphed below. The thick lines indicate the value of the solution, while the light gray lines represent the baseline value of the problem for the voter.

Figure 2.3: Learning Value of π^* for Voter



The only region where R generates a strictly positive effect on learning is on $[\overline{q}, \underline{q})$. The intuition is that when R succeeds in passing a bipartisan agenda here, q_R^2 strictly increases. If R fails, q_R^2 goes to 0, but this allows V to replace R with L. In the other two regions, V's *strict* preference for retaining or replacing R does not change after the realization of information. Increasing λ always widens this range; hence, increasing λ always improves agents' relative learning. However, changing q_L^2 results in shifts to this region; I show in Section 2.7 that there is an intermediate value of competition q_L^2 at which the size of the interval over which R adds any value to voter learning is maximal.

Flow Utility

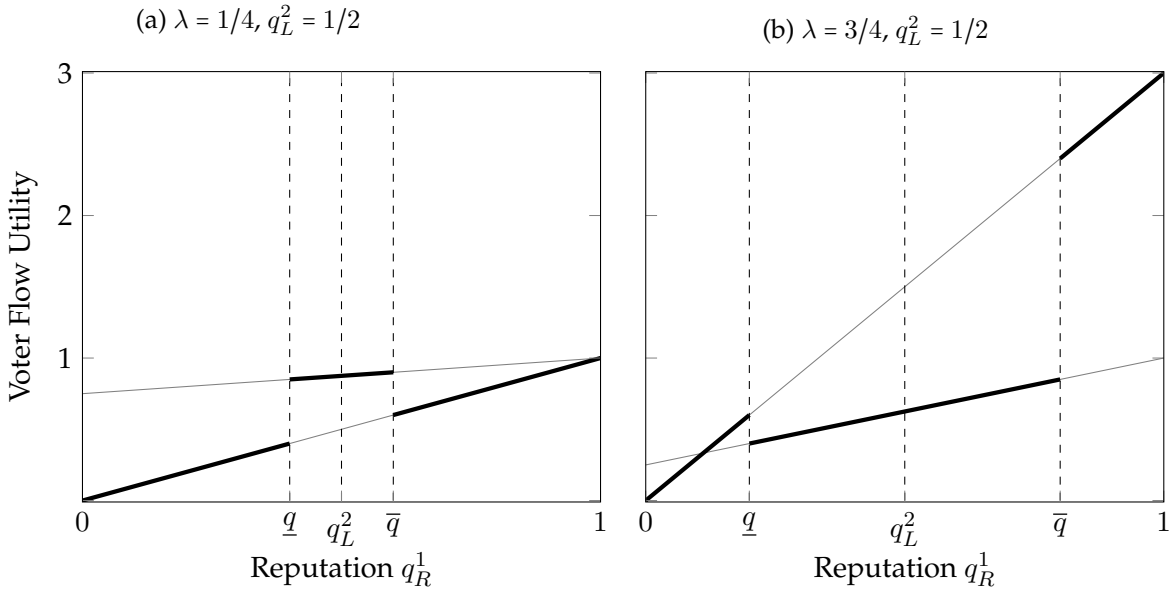
So far, the model is silent about the concurrent effects of pursuing different policy agendas on the representative voter's welfare.[11] Partisan policies, when manifest as policy reforms, are often

[11]While more partisan policies may please members of R's party, they may displease members of L's party. I hence take these results as an upper bound on the potential welfare gains of different policy agendas.

assumed to be valuable when implemented by a skilled incumbent but worse than a (bipartisan) status quo when pursued by an unskilled incumbent.

To illustrate the nonmonotonic variation of welfare with reputation and legislative alignment, I allow two policies, $\pi \in 0, 1$, as in the proof intuition for Proposition 1. I assume voter utility from $\pi = 0$ is always equal to 1 upon passage. I assume utility from $\pi = 1$ is equal to 0 if passed by a low ability incumbent and 4 if passed by a high ability incumbent. Utility from non-passage is always 0. The two panels of the figure below illustrate the tradeoffs of increasing or decreasing λ when $q_L^2 = 1/2$. The left panel shows the result under low legislative alignment ($\lambda = 1/4$) and the latter under high alignment ($\lambda = 3/4$).

Figure 2.4: Variation of Voter Flow Utility with λ



The pursuit of partisan policies results in less than ideal flow utility at low values of reputation, since incumbents' abilities are low in expectation. At high values of partisanship, pursuing partisan policies is optimal in expectation when legislative alignment is high but suboptimal when legislative alignment is low. At middling values of reputation, pursuit of bipartisan policies is optimal at low values of legislative alignment but suboptimal at high values of alignment, since the incumbent has tighter control over the legislature and could use this control to pass more ambitious policies.

2.2.4 Persuasion Model and Data Predictions

While the benchmark is useful for conveying the core intuitions of the theory, the information structures it captures are quite stark. The model does not directly capture, for example, that high ability incumbents may be marginally better at passing partisan policies. Additionally, the incumbent in the model, immediately prior to an election, always wins with probability 0 or 1. This latter point is particularly limiting given my empirical application, where proxies for winning probability (such as approval ratings) are readily available, while proxies for reputation are hard to uncover.

To improve mathematical tractability and address these points, I now describe an extended model utilizing a constrained persuasion structure. This approach gives the incumbent flexibility in directly choosing success probabilities for high and low ability incumbents. The extended model not only shows that the insights of the benchmark are more robust to a general setting, but also allows me to solve the model with a valence shock and thereby write the theoretical predictions in terms of an incumbent's ex-ante winning probability. Additionally, it allows me to easily address two extensions — described in the next section — for a multi-period version of the model as well as an extension regarding politician flow utility.

Persuasion Model I now characterize R's choice of policy agenda as a signal structure (p_1, p_0) for $p_1 \geq p_0$. These respectively represent the conditional probabilities an $a_R = 1$ and 0 incumbent would be able to earnestly pass an agenda if she were totally unaligned with the legislature. Total passage probabilities are:

$$\begin{aligned} a_r = 1: &\qquad \lambda + (1-\lambda)p_1, \\ a_r = 0: &\qquad (1-\lambda)p_0. \end{aligned}$$

The partisanship π of an agenda (p_1, p_0) is defined as $\pi(p_1, p_0) = 1 - \frac{p_1+p_0}{2}$.

This model is a form of a constrained persuasion problem. When $\lambda = 0$, the problem boils down to a standard Bayesian persuasion problem with two states (high/low ability) and two signals (success/failure). (p_1, p_0) represent the conditional probabilities of observing a success conditional on each state. For $\lambda > 0$, R chooses a signal structure (p_1, p_0), but the voter V only

receives a signal from this structure with probability $1-\lambda$. With probability λ, he receives a signal from a structure $T = (1,0)$ (success $\iff a_R = 1$), and cannot differentiate whether the signal they received came from P or T. In this sense, the set of posteriors over ability the policy agenda can induce is constrained.

Proving an analogue of Proposition 1 for this case is nearly identical. I hence use the flexibility of this more general model to prove the baseline result with a valence shock. Let $\epsilon \sim N$ be a single-peaked preference shock favoring L, which has support on $\mathbb{R}$, has mean 0, is symmetric around 0, is twice continuously differentiable, and is strictly increasing (decreasing) above (below) 0. V retains L if and only if $q_R^2 \geq q_L^2 + \epsilon$, meaning she wins the election with ex-ante probability $N(q_R^2 - q_L^2)$. Her expected value from an agenda (p_1, p_0) is then

$$\max_{p_1 \geq p_0} \int_{q_R^2} N(q_R^2 - q_L^2) dF(q_R^2 | q_R^1, p_1, p_0).$$

where $F(q_R^2|q_R^1, p_1, p_0)$ now represents the distribution of posteriors given (p_1, p_0). Let $p^*(q_R^1) = (p_1^*, p_0^*)$ represent the agenda that solves the equation above and π^* the partisanship of this agenda. The generalization of the first proposition is as follows.

Theorem 1. *There exist thresholds $\underline{q} < \overline{q}$ such that:*

- *for $q_R^1 \leq \underline{q}$, $\pi^*(q_R^1)$ is strictly decreasing. $p^*(0) = (1,0)$ and $\pi^*(0) = 1/2$, while $\pi^*(\underline{q}) = 0$ and $\pi^*(\underline{q}) = 0$.*
- *For $q_R^1 \in [\underline{q}, \overline{q})$, $p^* = (1,1)$ and $\pi^* = 0$;*
- *for $q_R^1 \geq \overline{q}$, $p^* = (0,0)$ and $\pi^* = 1$.*

Note in particular that $p^* = (1,1)$ corresponds to the $\pi = 0$ policy in the initial model. The $p^* = (0,0)$ policy corresponds to the $\pi = 1$ policy in the initial model.

As a (constrained) persuasion problem, I utilize insights from the Bayesian Persuasion literature to address parts of the proof. The optimal solution in an unconstrained persuasion problem is given by the concavification of the sender's value function (in this case, the incumbent's value function). If a policy agenda induces posteriors that achieve the concavification of $N(q_R^2 - q_L^2)$, that agenda is optimal. I show that for $q_R^1 < \underline{q}$, we can indeed use (p_1, p_0) to achieve that concavifica-

tion, which is a lottery between some posterior $\overline{q}_R^2 = \tilde{q} > q_L^2$ and $\underline{q}_R^2 = 0$.[12] However, because the persuasion problem is *constrained*, the concavification is no longer achievable after some $\underline{q}$, where an incumbent would ideally implement a less informative or uninformative policy agenda, but cannot due to the constraints of the problem. At $\underline{q}$, the solution itself is given by $(p_1, p_0) = (1, 1)$. After $\underline{q}$, the incumbent pursues a boundary solution. I show that the solution is either given by $(p_1, p_0) = (1, 1)$ or $(p_1, p_0) = (0, 0)$, just as in the baseline model. In particular, I show that if V is an everywhere concave function, then $(p_1, p_0) = (0, 0)$ is the optimal solution. Because when q_R^2 is sufficiently high, $N(q_R^2 - q_L^2)$ is locally concave (and equal to its concavification), $(p_1, p_0) = (0, 0)$ is the optimal solution for sufficiently high q_R^2. Otherwise, between $\underline{q}$ and $\overline{q}$, R continues to implement $(p_1, p_0) = (1, 1)$.

The persuasion extension suggests a potential *asymmetry* in nonmonotonicity relative to the baseline model. When $q_R^1 = 0$, $\pi = 1/2$, while when q_R^1 is high, $\pi = 1$. The baseline comparative statics also emerge from the proof of the result. Increasing λ causes $\underline{q}$ to decrease and $\overline{q}$ to increase. Increasing q_L^2 causes $\underline{q}$ and $\overline{q}$ to shift to the right.

Data Mapping The expression $N(q_R^2 - q_L^2)$ can be inverted to write the predictions of Theorem 1, as well as the comparative statics, in terms of win probabilities. Let P represent the incumbent's ex-ante probability of reelection.

Corollary 2. *There exist thresholds $\underline{P} < \overline{P}$ such that:*

- *when $P \leq \underline{P}$, partisanship π^* is strictly decreasing in P until $\pi^* = 0$ at $\underline{P}$;*
- *when $P \in (\underline{P}, \overline{P})$, partisanship $\pi^* = 0$;*
- *when $P \geq \overline{P}$, partisanship $\pi^* = 1$.*

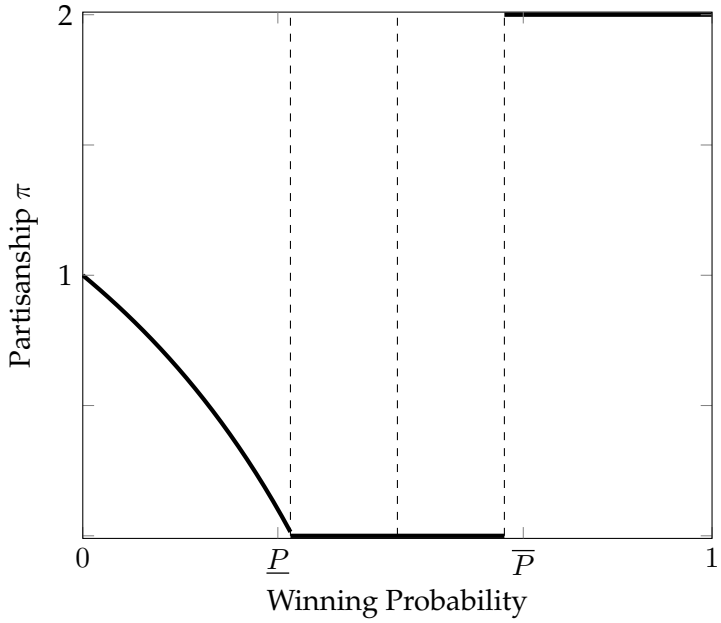
Moreover,

- *When legislative alignment λ increases, $\underline{P}$ decreases and $\overline{P}$ increases.*
- *When competition q_L^2 increases, $\underline{P}$ and $\overline{P}$ both increase.*

[12] A picture of the value function and its concavification is displayed in Section 2.7.1.

This corollary summarizes the key predictions I will take to test the partisanship of gubernatorial policy agendas in the model. The basic graph of this pattern is shown below.

Figure 2.5: Partisanship of Optimal Policy $\mathbf{P}^*$



The data predictions are as follows:

- **P1**: There is a nonmonotonic relationship between the partisanship of a reelectable incumbent's policy agenda and her probability of reelection, when compared to a reelection-ineligible incumbent. These incumbents should pursue partisan agendas when their probability of reelection is low or high, and relatively bipartisan agendas when their probability of reelection is moderate.

- **P2**: When increasing legislative alignment, the nonmonotonicity of **P1** should flatten out. However, the minimum level of partisanship of **P1** at moderate levels of win probability should also increase.

- **P3**: At high levels of party entrenchment, the nonmonotonicity of **P1** should flatten out; the nonmonotonicity should be strongest in states that lack party entrenchment.

P1 emerges because the model isolates informational incentives for partisan policies that drive reelection, meaning its predictions apply when comparing reelection eligible to ineligible execu-

tives. **P2** utilizes the fact that the definition of a bipartisan policy is relative to the legislature's composition, measured using λ. This means that raising λ may also raise the base level of partisanship in the data. **P3** emerges because increasing competition q_L^2 can be thought of as increasing party entrenchment.

Finally, although direct measures of winning probability are hard to track historically for executives — particularly in non-election years — approval data are readily available. These data are often used by politicians themselves to assess their popularity and chances at reelection; as such, I utilize approval data to proxy for ex-ante win probability in the data section.

2.2.5 Extensions

I utilize the persuasion model without a valence shock to address two extensions to the model.

Politician Utility and L Policies First, I show that the model is robust to allowing R to choose policies from the opposing party. In particular, a more partisan policy's informational utility is that it is "hard to pass." However, if R is incredibly aligned with her legislature, then passing policies favored by the opposing parties may in fact be harder to pass. Paradoxically, R may then attempt to go against her party's preferences and pursue an agenda of the opposition to "prove her worth."

I address this concern by allowing R to choose from a spectrum of partisan policies. She can choose either an R-partisan agenda $p^R \in (p_1^R, p_0^R)$ or an L-partisan agenda $p^L = (p_1^L, p_0^L)$. I endow R with a preference for R policies over L policies. I show that the model's central nonmonotonicity is left untouched. If flow utility concerns are weak, the region in $[\underline{q}, \overline{q}]$ may exhibit additional nonmonotonicities in the direction of L policies.

Suppose that R chooses a policy agenda p^R or p^L. Success probabilities are:

$$\text{Choose } \pi_R : \lambda a_R + (1-\lambda) p_{a_R}^R$$
$$\text{Choose } \pi_L : (1-\lambda) a_R + \lambda p_{a_R}^L$$

Intuitively, if R chooses π^L, she is adopting the policies an L politician would implement, meaning her de facto legislative alignment flips. I assume R receives a disutility $-c$ from pursuing

π^L.

Let $F^i(q_R^2|q_R^1, p_1^i, p_0^i)$ be the distribution of posteriors over R's ability for $i \in \{R, L\}$. Then, R implements an agenda from π^R if and only if

$$\max_{p_1^R \geq p_0^R} \int_{q_R^2} u_R(q_R^2) dF^R(q_R^2|q_R^1, p_1^R, p_0^R) \geq -c + \max_{p_1^R \geq p_0^R} \int_{q_R^2} u_R(q_R^2) dF^L(q_R^2|q_R^1, p_1^L, p_0^L).$$

The left expression is the value from the solution to the baseline problem. The right expression is the value from choosing an L policy agenda less the cost $-c$ from choosing policies preferred by the opposing party. We can write the optimal solution of each program as p^{i*}, which is characterized by thresholds $\underline{q}^i, \overline{q}^i$. While c sufficiently large may mechanically shut down any preference for L policies, I show that even for c small, the model's nonmonotonicity still holds.

Proposition 3. *There exists $\overline{\lambda} > 1/2$ such that for $\lambda \leq \overline{\lambda}$, R implements p^{R*}. For $\lambda \geq \overline{\lambda}$ and c sufficiently low, there exist thresholds $\underline{q}^f, \overline{q}^f$, with $\underline{q}^R < \underline{q}^f < \overline{q}^f < \overline{q}^R$such that R implements*

- *p^{R*} for $q_R^1 \leq \underline{q}^f$;*
- *p^{L*} or p^{R*} for $q_R^1 \in (\underline{q}^f, \overline{q}^f)$;*
- *p^{L*} for $q_R^1 \geq \overline{q}^f$;*

The intuition is as follows. Below $\underline{q}^R$ and above $\overline{q}^R$, R is able to achieve the concavification of $u_R(\cdot)$ utilizing p^R. The integrand term on the left hand side is maximized, suggesting a strict preference for p^{R*}. Note that R's maximal utility from pursuing p^{R*} is strictly decreasing in λ — since lower λ allows greater informational flexibility — while the maximal utility from pursuing p^{L*} is strictly increasing in λ. The utility from pursuing either policies is always equal at $\lambda = 1/2$, modulo the $-c$ term. This means that p^{R*} is still strictly preferred at $\lambda = 1/2$. Preference for p^{L*} can override if and only if c is sufficiently small; and only then when the value from pursuing π^{L*} is sufficiently higher than that of p^{R*}. This can occur only within a strict subset of $[\underline{q}^R, \overline{q}^R]$.

Extending the space of partisanship downwards — so that L policies are even *less* partisan than $\pi = 0$ (for an R incumbent) — suggests that partisanship weakly decreases from $\pi = 1/2$ to 0 at $\underline{q}^f$; potentially plummets further between $\underline{q}^f$ and $\overline{q}^f$, either with the implementation of a moderate or extremely partisan L policy; and then shoots back up to $\pi = 1$ above $\overline{q}^f$. That is, within

$(\underline{q}^f, \overline{q}^f)$, R implements either $\pi^R = 0$ or $\pi^L \geq 0$, potentially adding additional nonmonotonicities to the pattern, but ensuring that partisanship (relative to R's party) is always lower in this middle region than when reputation is low or high.

Multi-Period Dynamics I show that allowing for an additional period of policy passage preserves the fundamental insights of the model. While the baseline model assumes that politicians' policy agendas represent the totality of their term in office, we can consider a three period model where politicians pursue agendas in periods $t = 0$ and $t = 1$ before an election at $t = 2$. In between $t = 0$ and $t = 1$, agents update beliefs over a_R as before.

Following notation from earlier, let the belief that R is high ability at the beginning of $t = 0$ be q_R^0. The twoperiodproof is characterized as follows.

Proposition 4. *Given $\underline{q}, \overline{q}$, there exist thresholds $\underline{\underline{q}} < \underline{q} < \overline{q} < \overline{\overline{q}}$ such that*

- *For $q_R^0 \leq \underline{\underline{q}}$, partisanship π^* is decreasing from $1/2$ at q_R^0 to 0 at $\underline{\underline{q}}$.*
- *There exists $q_\dagger \in [\underline{\underline{q}}, \overline{q} < \overline{\overline{q}}]$ such that for all $q_R^0 \geq q_\dagger$, $\pi^* < 1$.*
- *For $q_R^0 \geq \overline{\overline{q}}$, $\pi^* = 1$.*

2.3 Data

This section begins with information about U.S. gubernatorial State of the State addresses, including their relevance to my setting. I then discuss how the data is sourced and pre-processed before analysis. I finish by detailing how I use large-language modeling to isolate policy proposals from speeches.

2.3.1 U.S. Governors and State of the State Addresses

Reputational Priorities of U.S. governors The policy agendas of U.S. governors provide an ideal setting to test the predictions of the model. Governors are high-profile political executives who, since the end of the 20th century, have played an outsized role in setting, pursuing, and implementing state policy agendas.[13] Compared to legislators, governors "enjoy organizational, insti-

[13]The "devolution revolution" of the 1990s marks the period where the federal government heavily devolved authority over many public policies to state governments

tutional, and popular advantages similar to and arguably even greater than the president [over Congress]" (Heidbreder, 2012). In any given year there are 50 U.S. governors in office, all of whom vary in their popularity, eligibility for reelection, party entrenchment, and alignment with state legislatures.

There is substantial evidence that voters evaluate governors based on their productivity in passing their policy agendas. The National Governors Association (NGA) provides a series of guides to incoming U.S. governors, which provide rich insight into how governors themselves view incentives to pass policies while in office. One excerpt from a guide remarks that "[t]he media and public will judge the governor's leadership ability and success... [by] whether the administration's legislative program succeeds... The governor's ability to manage and secure legislation also affects his or her ability to serve as a strong leader of the party... If the legislation fails, it will be considered a political defeat... [T]he passage of priority legislation usually will signal a political success." (National Governors Association, 2018). An executive director of the NGA has gone as far as to say that "the ultimate measure of success [is] the ability of... governors to get his or her initiatives enacted" (Scheppach, 2005).

The NGA guides also provide insights on the dimensions of skill by which a governor is evaluated, often using the language of a managerial "CEO" to describe these responsibilities. "As chief executive officers (CEOs), governors are responsible for the leadership and management of their states. As leaders, they set priorities for their administration and enact new policies and programs designed to achieve those priorities" (National Governors Association, 2019). Moreover, an "effective process to craft and implement a legislative program and strategy, as well as to cultivate and maintain working relationships with legislative leaders and members, is critical to ensuring the success of a governor's legislative program." (National Governors Association, 2018)[14]

There is evidence that voters pay attention to and respond to governors' actions, and that this response may be mediated by alignment with legislatures. Wolak and Parinandi (2022) find a positive response of gubernatorial approval to both ideological alignment and substantive mea-

[14] Anecdotal evidence supports the idea that executives strengthen their reelection chances by successfully pursuing ambitious policies in the face of partisan adversity. Iowa Governor Terry Branstad, for instance, describes how his ability to massively restructure state government aided in a competitive reelection to his second term. Branstad writes that he managed to secure "90 percent of the reorganization package" despite a "legislature overwhelmingly controlled by the opposition party" (Behn, 1991), going on to describe the techniques of solid management, legislative persuasion, and drive necessary to achieve those goals.

sures of performance, such as economic performance and policy outputs. Across economic and policy outcomes, Brown (2010) shows that voters are more likely to blame opposing parties for problems if the opposition holds an executive position and their own party has a hold on the legislature. Leyden and Borrelli (1995) find that voters are much more responsive to changes in state economic outcomes when state government is unified (rather than divided, with different parties holding the executive and legislative branches). Jacobson (2006) shows that governors who are not aligned with their state legislatures enjoy slightly higher approval than those with aligned legislatures, arguably in part due to the fact that misalignment reduces blame. Larimer (2015) studies the limits of gubernatorial party control and argues that "[u]nified control provides the public an easy target to blame when things go bad" (p. 96), going on to note that divided governments may permit flexibility in allowing voters to learn about governors' managerial strengths.

Policy Agendas The model's predictions apply to executives' policy agendas. I utilize the text of each governor's annual "State of the State" address as a baseline corpus of policy agendas; utilizing text data also allows us to use methods from the literature to analyze partisanship of policy proposals.

At the beginning of each legislative session, the governor of each state in the United States is required to address a joint session of the legislature to deliver a "State of the State" (SotS) speech. [15] The speech is delivered annually (in some states biennially) in the first quarter of the year, and is the US States' analogue of the presidential "State of the Union." Since these data are given around the same time every year in (almost) every US state, these data allow an annual panel of policy agendas for each governor, state, and year.

According to a National Governors Association guide on legislative relations, the "inaugural address, State of the State address and budget message are all excellent forums to communicate and build momentum for the executive branch's legislative agenda" (National Governors Association, 2018). A large literature in political science has documented the importance of these speeches as vehicles for the governor reflect on her administration's past accomplishments and lay out her policy priorities for the coming year. Coffey (2005) and Heidbreder (2012) argue that these

[15] In some states, the speech is called the "State of the Commonwealth"; in others, the governor's budget or inaugural address take the same role.

speeches accurately represent the incumbent's current policy priorities rather than pure policy preferences. Governors themselves also view these addresses as a highly salient, public platform for signaling policy agendas.

There is also evidence that governors are indeed able to pass many of the agenda items laid out in their SotS addresses. Kousser and Phillips (2012) investigate over $1,000$ proposals laid out in a set of State of the States speeches in the mid-2000s, showing that 41% passed in some form similar to what the governor proposed and 18% with compromises. They show that passage is more likely when a governor is aligned with the legislature or holds more political capital, with the notable exception of budgetary items — where these variables exhibit no apparent effect on passage probability. One implication of this finding is that, regardless of partisan affiliation, there are basic duties (such as budget management) that a skilled executive should be able to address. Indeed, an ex-governor of Maryland noted that "the realities of running a state do not allow for ideological rigidity. Governors must ensure that the budget is balanced, that the state can adequately respond to its day-to-day challenges, and must be able to work with lawmakers from both parties and across the ideological spectrum" (Kousser and Phillips, 2012, p. 97).

In stark contrast to the Congressional Record (Gentzkow et al., 2019; Jensen et al., 2012), comprehensive analysis of partisanship in the State of the State speeches has been relatively scarce. Many articles — e.g. DiLeo (1997), Coffey (2005), Weinberg (2010), Heidbreder (2012), Warner (2023) — have been able to analyze a few years worth of SotS data to document cursory facts about gubernatorial ideology. For example, DiLeo (1997) shows governors in Democratic states are more likely to pursue redistribution. Ferguson (2003) shows that a governor's priorities are disciplined by the composition of her legislature and economic conditions in states. Coffey (2005), Weinberg (2010), and Kousser and Phillips (2012) attempt to measure ideology in these speeches by either manually coding sentences or relying on dictionaries of partisan words. Heidbreder (2012) shows that Democratic governors are more attentive to healthcare and social policies.

One of the key issues with analysis of SotS addresses is the lack of a systematic, centralized data source of speeches, with the exception of two potential databases. Lushkov (2019) collects hundreds of SotS addresses back to the 1800s to look at the frequency with which governors discuss education. Butler and Sutherland (2023) have digitized almost all SotS addresses from 1960 onwards, documenting in increase in nationalization of speech. However, this result does not nec-

essarily imply any distinct trend in gubernatorial partisanship. Relatedly, Hopkins et al. (2022) document a divergence in the speech of 1,783 state party platforms from 1918-2017 beginning in the 1990s, but restrict their analysis to the frequencies of certain topics and phrases as opposed to en-masse text analysis of their corpus that delivers a unidimensional measure for partisanship.

2.3.2 Sourcing and Pre-Processing

I compile a digitized text library of the State of the State speeches from 1990-2020, allowing me to employ methods from the literature on partisanship of speech and large-language-modeling to measure partisanship, study its heterogeneity, and test the model's predictions.

I began with the database of Lushkov (2019), which contained 330 speeches post-1995 due to data loss. Next, I scoured state news sites, educational resources, archived governor websites, and online state libraries to assemble SotS speeches for an initial dataset of $1,144$ usable speeches. I later gained access to data from Butler and Sutherland (2023), allowing me to assemble a final dataset of $1,345$ speeches for 1990-2020. Figure 2.8.1 graphs the set of usable speeches for analysis.

Finally, I broke these speeches into snippets of thematically contiguous thoughts using the NLTK (Natural Language Toolkit) Text-Tiling tokenizer. Each snippet was about 7 sentences and can be thought of either as a paragraph or sets of small paragraphs. I end up with $78,702$ snippets for the period $1990 - 2000$.

2.3.3 Speech Processing with BERT

To separate portions of SotS addresses discussing the governor's policy agenda, I fine-tuned a large-language-model called "BERT" from the huggingface transformers library to identify relevant snippets of text. In particular, I asked it to classify whether a given snippet of text discussed a concrete policy (yes/no) and, if so, whether it was a discussion of a policy proposal or made mention of a past policy (yes/no).

BERT is a pretrained model that learns the structure of provided text examples, allowing it to be "fine-tuned" to classify text (Devlin et al., 2018). The present paper's process utilizes the same methodology as Card et al. (2022), which identifies whether Congressional speeches were a) about immigration; b) if they were about immigration, whether the tone was positive, negative or

neutral. I repurpose much of these authors' github code to my setting.

Hand-Coding I randomly selected around 9500 snippets from the $70,368$ snippets of the initial $1,144$ speech 1995-2020 dataset to be hand-labelled and provided to BERT as fine tuning data. Two research assistants were given the following instructions to code these snippets.

1. "Policy." Coded as "1" if the snippet discusses the enactment of a state-level policy (either passed by the governor, state government, or referendum) and "0" if it does not. A policy discussion is a reference to a specific act of legislation or law, a concrete proposal to increase or decrease funding to a certain cause, other legal orders proposed by the government to take certain concrete actions, and discussions of details of any of the above.

2. "Proposal/Past." Only applies if "policy" coded as "1." Coded as "1" if the snippet refers to a policy that has just put into place or will be put into place in the future. Coded as "0" otherwise — in particular, if a governor is reflecting on the effects of a policy in the past.

The full coding guide is contained in Section 2.9. Of these hand-coded data, we identified $4,144$ snippets as discussing policy and, of those, $2,993$ as discussing proposals.

Detailing BERT Model I utilized the "BERT-Base-Uncased" model, which contains approximately 110 million parameters which are adjusted through the process of fine-tuning, to analyze the data. I make use of a two-layered model approach. I first train a model to classify whether a snippet is a policy discussion, which takes in as inputs all the snippets in our hand-coded data. The second model is trained to classify whether a snippet is about a policy proposal or some reflection on a past policy. The latter takes as inputs only the hand-coded data corresponding to policies.[16]

Once these models were fine-tuned, I ran the full dataset of snippets through each model. I used the policy classification to identify policies, and then the proposal classification model to identify proposals, allowing me to classify each snippet as desired: snippets about policy proposals, snippets about policy but not proposals, and other snippets. The policy classification model

[16]The input (hand-coded) data for each model is divided into seven stages called "epochs." Each epoch takes the training data and partitions it into three groups. The first set, the "train set," is the main dataset used to train the model. The "dev set," comprised of 400 snippets, is used to adjust the model after its initial training. Finally, the "test set," comprised of 300 snippets, is used to calculate the models' accuracy. Before feeding the data into the model, we also tokenize the data using spaCy, which takes each snippet and breaks it into its component words.

ended with a low cross-entropy loss of 0.0111 nits and the proposal classification 0.0058 nits. Of the $78,702$ snippets, I identify $35,636$ (45%) as corresponding to policy and $26,720$ (34%) as corresponding to policy proposals.

2.3.4 Additional Data Sources

The following additional data sources are used in the analysis:

- Governor names/dates in office: National Governors Association (2024)
- Gubernatorial win margins and seat status: Algara and Amlani (2021)
- Quarterly governor approval ratings: Singer (2023)
- Legislative composition: Klarner (2013) and National Conference of State Legislatures (2024)
- Term limit rules and missing election dates: Ballotpedia (2024a) and Ballotpedia (2024b)
- Additional covariates: Grossmann, Jordan, and McCrain (2021)

2.4 Documenting Partisanship

The section begins by detailing measurement of aggregate partisanship. I then document aggregate changes in partisanship from 1990-2000 before describing how aggregate trends disguise substantive variation at the level of governors, states, and regions. This variation is interpretable within my model, which is explored in Section 5.

2.4.1 Calculating Partisanship

With the corpus of governor speeches, partisanship of gubernatorial speech can be measured for both full speeches and just policy proposals using techniques from the literature on political speech. I calculate a measure of partisanship π_{it} for each governor i and year t. I then take a weighted sum of π_{it} to construct the familiar measure of annual aggregate partisanship Π_t used in the literature.

The metric for defining and calculating partisanship is that detailed in Gentzkow et al. (2019). Aggregate partisanship for year t, Π_t, is measured as the expected *informativeness* of a randomly

selected phrase in inferring a governor's party. Specifically, suppose we randomly select a governor — selecting from one of Democrats or Republicans with (prior) probability $1/2$. Partisanship is measured as the expected posterior of guessing that governor's party correctly. If partisanship is less than or equal to $1/2$, speech is — on average — uninformative of governor party. If Republican and Democratic governors use dissimilar phrases, partisanship is greater than $1/2$ since, in expectation, language is informative of party.

I use "bigrams," i.e. two-word phrases, as my measure of phrase. I use the NLTK Porter-Stemmer to reduce words to their base form. I then use the Gentzkow et al. (2019) "leave-out-estimator," for calculating aggregate partisanship Π_t. For a sample period of interest, let R be a set of Republican governors and D Democratic governors.[17] Let c_{ij} be the count of phrase j used by governor i. Let $C_{ij} = \frac{c_i j}{\sum_{j \in J} c_{ij}}$ be the normalized count of phrase j used by governor i, where J is the set of all phrases used in the time period of interest. Let C_j^P be the normalized count of phrase j used by party P: $C_j^P = \frac{\sum_{i \in P} c_{ij}}{\sum_{i \in P} \sum_{j \in J} c_{ij}}$. Finally, let $T(t)$ be a five-year window around time t: $T(t) = \{t-2, t-1 \ldots, t+2\}$.[18] A subscript t represents the value of the variable at time t, while $t \in T(t)$ represents its value within the five-year window. For a phrase j and speaker i, define ρ_{-ijt} as the ratio of j's use by Republican vs. Democratic governors, modulo governor $i \in R_t \cup D_t$, within the five-year window acting as the reference group:

$$\rho_{-ijt} = \frac{C_{jt \in T(t)}^{R-\{i\}}}{C_{jt \in T(t)}^{R-\{i\}} + C_{jt \in T(t)}^{D-\{i\}}}.$$

Concretely, given a flat prior, ρ_{-ijt} is the posterior probability assigned to a speaker being a Republican governor upon observing phrase j.

Aggregate partisanship in the US States at time t, Π_t, can then be calculated as:

$$\Pi_t = \frac{1}{2}\frac{1}{|R_t|} \sum_{i \in R_t} \sum_j C_{ijt} \cdot \rho_{-ijt} + \frac{1}{2}\frac{1}{|D_t|} \sum_{i \in D_t} \sum_j C_{ijt} \cdot (1 - \rho_{-ijt}). \tag{2.2}$$

The interpretation of Π_t is precisely as above. With probability $1/2$, we randomly select a party

[17] The few independent governors in our sample lean Democrat, so I classify them as Democratic for all intents and purposes.

[18] I make use of a five-year window since each year only has at most 50 speeches. Gentzkow et al. (2019) use Congressional Record Data, where the number of speakers and quantity of text is much larger, and the length of a Congressional session is two years.

(R or D), and from there randomly draw a governor. With probability C_{ijt}, the chosen governor i uses phrase j. Then, conditional on phrase j, the posterior over governor party moves to ρ_{ijt}. Π_t averages this posterior across governors and phrases.

I measure the partisanship of the speech of governor i in state s at time t as:

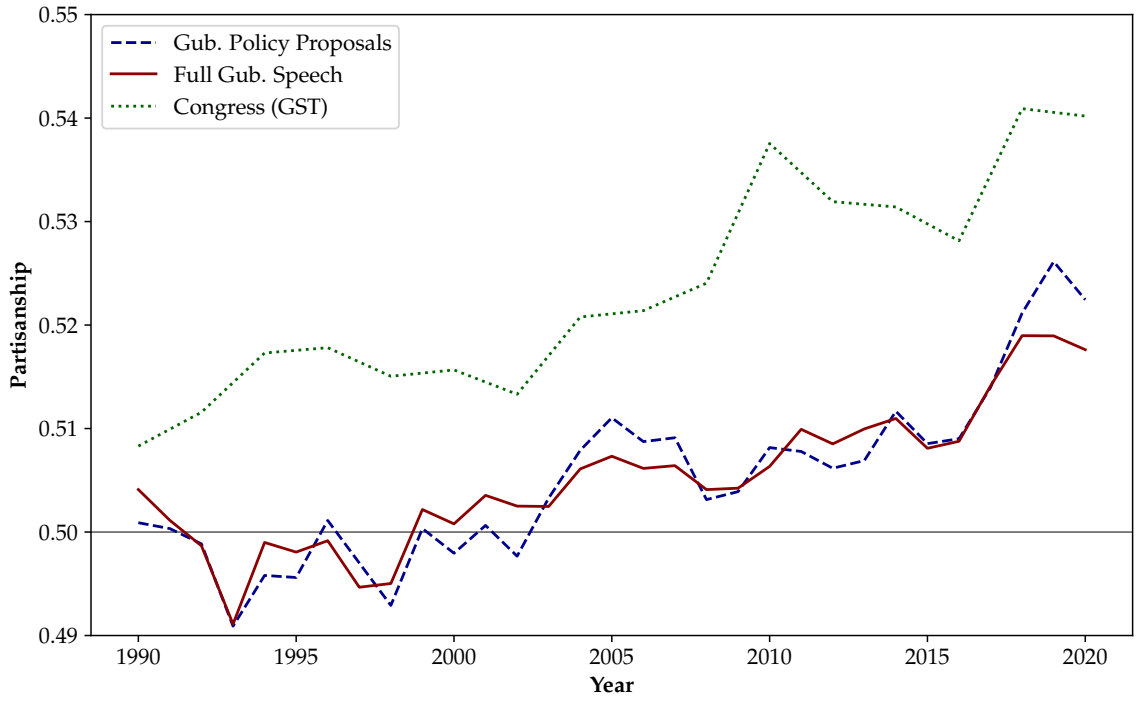
$$\begin{aligned} \pi_{it} &= \sum_j C_{ijt} \cdot \rho_{-ijt} && i \in R_t \\ \pi_{it} &= \sum_j C_{ijt} \cdot (1 - \rho_{-ijt}) && i \in D_t, \end{aligned} \tag{2.3}$$

which measures how partisan governor i's speech is relative to her party. Governors with $\pi_{it} \geq 1/2$ use language mostly in line with their party, while those with $\pi_{it} \leq 1/2$ use language in line with that of the other party.

2.4.2 Aggregate Behavior of Partisanship

The figure below shows the evolution of partisanship over the sample period for three series. Partisanship calculated using the entire corpus of speeches is shown in the solid red line, while partisanship calculated using only the policy proposal component of speeches is shown in the dashed blue line. Finally, the dotted green line plots the series of Congressional partisanship from Figure 2A of Gentzkow et al. (2019), which also uses the leave-out estimator. In Figure 2.8.2, I also compare these measures of partisanship to those estimated from past policy discussions and rhetoric.

Figure 2.6: Partisanship of U.S. Gubernatorial Speeches, Full Speech vs. Policy Proposals vs. Congressional Record: 1990-2020



Partisanship of gubernatorial speech calculated using Gentzkow et al. (2019) leave-out estimator. "Gub. Policy Proposals" computes estimator for gubernatorial speech snippets coded as discussing policy proposals on corpus of U.S. governors' speech from 1990-2020. "Full Gub. Speech" computes estimator for all gubernatorial speech in given year. "Congress (GST)" series is partisanship of Congressional speech using leave-out estimator, replicating Figure 2A of Gentzkow et al. (2019).

I begin by discussing the measures of state partisanship using both full speeches and the policy proposal data. Partisanship for both series remains low in the early 1990s. The policy proposals data do not experience their first prominent increase until a hump lasting from 2001-2008, with the end of the hump coinciding with the Great Recession. Partisanship dips during the recessionary period before resuming its slow upward creep. Partisanship then spikes in 2017, coinciding with the arrival of governors elected in the 2016 electoral cycle. Both series track each other, a fact borne out in the governor-level panel data.[19] The stages in which we see these increases is likely correlated with the electoral cycle; while Congressional terms are only two years, governors mostly serve four-year terms, and many serve two consecutive terms.

[19]The correlation between the full speech and proposal measures is 0.92 in the panel data.

Recall that partisanship measures the posterior probability that an observer with a flat prior can accurately guess a governor's party after randomly drawing a phrase. When partisanship is less than or equal to $1/2$, speech is on average so similar between parties that observing a randomly drawn phrase does not improve inference about the correct governor party. By this metric, partisanship is negligible at the level of state governors until around 2000. At the end of the sample, however, observing just one phrase leads to, in expectation, a two percentage point movement in the prior towards guessing the governor's party.

The measure of gubernatorial partisanship appears to track many features of the Congressional partisanship series from Gentzkow et al. (2019), with a few notable exceptions. First, the Congressional series exhibits a spike from 2008-2010 that is absent from the gubernatorial series. Second, the timing of increases in the gubernatorial series appears to be more gradual, with the magnitude of the change in the 2010s in the Congressional series being double that in the gubernatorial series. Third, the level of gubernatorial speech appears to be lower than that of the Congressional record. In fact, it is not until 2015 that the *level* of gubernatorial partisanship even reaches that of Congressional Partisanship in 1990. This finding also contrasts with that of Hopkins et al. (2022), who find that divergence in speech of state party platforms began in the mid-1990s. Nevertheless, both series experience increases after 2000 that finish with spikes in 2017.

As a validation exercise, Table 2.8.1. summarizes the most partisan Republican and Democratic phrases utilizing the policy proposal text data. I again utilize a method from Gentzkow et al. (2019), where the partisanship of a phrase is determined as measuring the informational loss in inferring governor's party when removing phrase j from governors' vocabulary. This formula for phrase j is given by

$$1/2 - 1/2 \sum_{j \neq k} \left(\frac{C_k^R}{1 - C_j^R} + \frac{C_k^D}{1 - C_j^D} \right) \frac{C_k^R}{C_k^R + C_k^D}, \tag{2.4}$$

where more positive numbers correspond to more Republican phrases and more negative numbers to more Democratic phrases.[20] I calculate this metric for each bigram in the six epochs of 1990-1994, 1995-1999, 2000-2004, 2005-2010, 2010-2014, and 2015-2020. The table lists the ten "most Republican" and "most Democratic" phrases. Republicans are likely to mention taxes — includ-

[20]Following the norm established by their paper, I code Republican governors' party as $+1$ and Democratic governors' as -1.

ing phrases like tax reduction, tax relief, and cutting taxes. They also emphasize the budget and, later in the sample, discuss crime. Democrats are more likely to mention topics related to the environment, welfare and childcare, the minimum wage, and affordable housing.

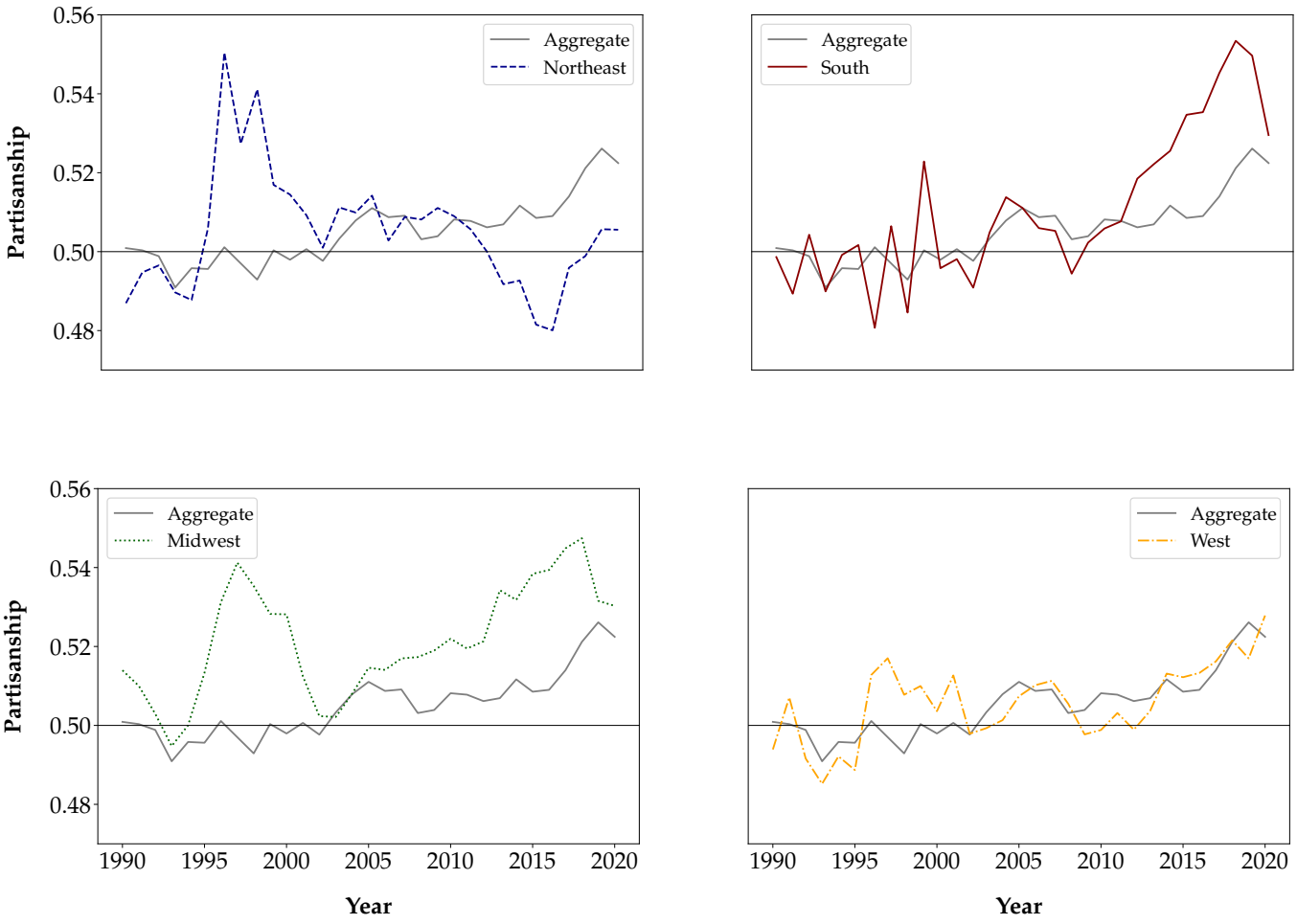
2.4.3 Heterogeneity

The aggregate series in Figure 2.6 mask substantial heterogeneity in partisanship at the state level. Taking the policy proposal data, the figure below shows this heterogeneity by displaying the average of governor-level partisanship π_{it} for each of the four U.S. Census regions.

Figure 2.7: Partisanship of U.S. Gubernatorial Policy Proposals, by U.S. Region: 1990-2020

(a) Northeast

(b) South



(c) Midwest

(d) West

Partisanship of gubernatorial speech calculated using Gentzkow et al. (2019) leave-out estimator on U.S. gubernatorial speech snippets coded as discussing policy proposals. Northeastern states are CT, ME, MA, NH, NJ, NY, PA, RI, VT. Midwestern states are IL, IN, IA, KS, MI, MN, MO, NE, ND, OH, SD, WI. Southern states are AL, AR, DE, FL, GA, KY, LA, MD, MS, NC, OK, SC, TN, TX, VA, WV. Western states are AK, AZ, CA, CO, HI, ID, MT, NM, NV, OR, UT, WA, WY.

The Northeast and Midwest both experience spikes in partisanship at the end of the 1990s, well-surpassing the apex of partisanship achieved at the end of the aggregate series. Partisanship in the Northeastern states then experiences a trough around 2015 and ends at the same level it began in the 1990s The Southern states in panel (b) exhibit significant variation in partisanship before 2005. The average level of partisanship for these states accelerates throughout the 2010s

but then dips after 2017. The Western States in panel (d) follow the aggregate trend most closely, but even these states see regular dips and troughs.

This regional heterogeneity is indicative of deeper variations in partisanship at the state and even governor level. Table 1 below summarizes standard deviations in partisanship by each of the three decades for the aggregate sample, within states, and within governors. It also replicates corresponding measures of Congressman-level partisanship from Gentzkow et al. (2019), which is displayed in their Figure 2.8B.

Table 2.1: Standard Deviations of Partisanship, Governors vs. Congress, 1990-2020

Standard Deviation	1990-1999	2000-2009	2010-2020
Overall	0.053	0.031	0.041
Within-State	0.041	0.027	0.029
Within Governor	0.028	0.016	0.017
N	375	482	479
Num. Governors	80	92	104
Overall	0.015	0.019	0.028
Within-State	0.014	0.018	0.027
Within-Congressman	0.008	0.012	0.017
N	2,701	2,686	3,261
Num. Congressmen	906	779	1,030

Partisanship of gubernatorial speech calculated using Gentzkow et al. (2019) leave-out estimator on U.S. gubernatorial speech snippets coded as discussing policy proposals. Partisanship of Congressional speech replicated from Gentzkow et al. (2019) using leave-out estimator. 1990-1999 utilizes Congresses 101-105, 2000-2009 Congresses 106-110, 2010-2020 Congresses 111-116.

The top panel of the table shows that variation in partisanship for governors within states is relatively on par with variation in the entire panel. There is also substantive variation in partisanship within governors, around half that in the full sample, suggesting that these effects are not driven by conditions unique to certain governors. [21] Notably, the magnitude of these standard de-

[21] Differences in partisanship between governors in the same state may be driven not by differences in their styles of governing or policy preferences, but by differences in governors' approval ratings, legislative alignment, or other changes in the governors' environment.

viations is nearly equal to the average increase in partisanship of about 0.02 seen in the governor series in Figure 2.6.

The bottom panel shows that variation in partisanship within state and Congressman in the Congressional record increases over time, and is systematically lower than that for governors prior to the 2010s. There is no obvious pattern for governors, suggesting that this heterogeneity has consistently been a part of gubernatorial politics. These data suggest that trending partisanship at the aggregate level disguises variations at the state and year level, as well as countervailing dynamic fluctuations — such as in the late 1990s — that are smoothed out in the aggregate series. Because much of this variation occurs within the same states and governors, I turn to test the predictions of the model in explaining fluctuations in partisanship in the next section.

2.5 Testing the Model

I begin by overviewing the estimation methods and variables I will use to capture the predictions of the model. I then present evidence supporting the model's predictions in aggregate. Afterwards, I disaggregate by competition and state type to further probe these predictions. Finally, I use the model's lens to qualitatively assesspartisanship in the U.S. presidency.

2.5.1 Estimation

Recall the central predictions of the model.

- **P1**: There is a nonmonotonic relationship between the partisanship of a reelectable incumbent's policy agenda and her probability of reelection, when compared to a reelection-ineligible incumbent. These incumbents should pursue partisan agendas when their probability of reelection is low or high, and relatively bipartisan agendas when their probability of reelection is moderate.

- **P2**: When increasing legislative alignment, the nonmonotonicity of **P1** should flatten out. However, the minimum level of partisanship of **P1** at moderate levels of win probability should also increase.

- **P3**: At high levels of party entrenchment, the nonmonotonicity of **P1** should flatten out; the nonmonotonicity should be strongest in states that lack party entrenchment.

To test these predictions, I use the following variables for governor i in state s at time t.

- Partisanship, π_{ist}. I obtain this measure using the partisanship formula in equation (2.3) and normalize it by the sample mean and standard deviation.[22]

- Win probability, q_{ist}. I utilize quarterly gubernatorial approval data compiled by Singer (2023).[23] Since State of the State addresses are given in Q1 of each year, I use approval data from Q4 of the previous year. For governors' inaugural years, I use the approval data from Q1.[24]

- Legislative alignment, λ_{ist}. I combine the data on legislative composition from Klarner (2013) and National Conference of State Legislatures (2024). I choose a binary representation of λ_{ist}. The variable is equal to 1 if more than 50% are members of the governor's party and 0 otherwise.

- r_{ist}, reelection eligibility. This is equal to 1 if a governor is eligible to run another term and 0 otherwise.[25]

To pinpoint the model's nonmonotonicity, I break approval ratings q_{ist} into deciles by state.[26] Let q_{kist} correspond to the k^{th} decile of approval for a governor in state s. Finally, let ξ_s be a state fixed effect and χ_t a year fixed effect. I estimate the following regression equation.

$$\begin{aligned}\pi_{ist} =&\alpha_0 + \sum_{k\neq 4}^{10} \alpha_k q_{kist} + \gamma_0 \cdot \lambda_{ist} + \sum_{k\neq 4}^{10} \gamma_k q_{kist} \cdot \lambda_{ist} \\ &+ r_{ist} \cdot \Big(\beta_0 + \sum_{k\neq 4}^{10} \beta_k q_{kist} + \delta_0 \cdot \lambda_{ist} + \sum_{k\neq 4}^{10} \delta_k q_{kist} \cdot \lambda_{ist}\Big) + \xi_s + \chi_t + \epsilon_{ist} \end{aligned} \tag{2.5}$$

[22] The sample mean is 0.511 and the standard deviation 0.042.

[23] I assume win probability is a strictly increasing function of gubernatorial approval within each state.

[24] For the sample period, these data are missing approval data from Idaho. There are also substantial gaps for Hawaii, Louisiana, and North Dakota.

[25] I count those governors who are eligible for reelection but *choose* not to run again as still being reelection eligible, as choosing not to run again is an endogenous choice. The central results are robust to counting only governors who actually run for their seats as reelection eligible.

[26] This accounts for persistent differences in the means of state gubernatorial approval. This also ensures state-level balance in the approval data, allowing the regression to capture departures from average approval within each state.

The baseline expression in the first line summarize describes how the level of partisanship varies with approval decile interacted with legislative alignment. The estimands of interest are the β_k and δ_k coefficients, which capture the difference in partisanship for reelection eligible governors relative to lame ducks ineligible for reelection. β_k represents partisanship for the k^{th} decile of approval, relative to the fourth decile, for a reelection eligible governor with an unaligned legislature, adjusting for lame ducks' behavior.[27] $\beta_k + \delta_k$ represents the level of partisanship for the k^{th} decile of approval, relative to the fourth decile, for a reelection eligible governor with an *aligned* legislature, adjusting for lame duck behavior.[28]

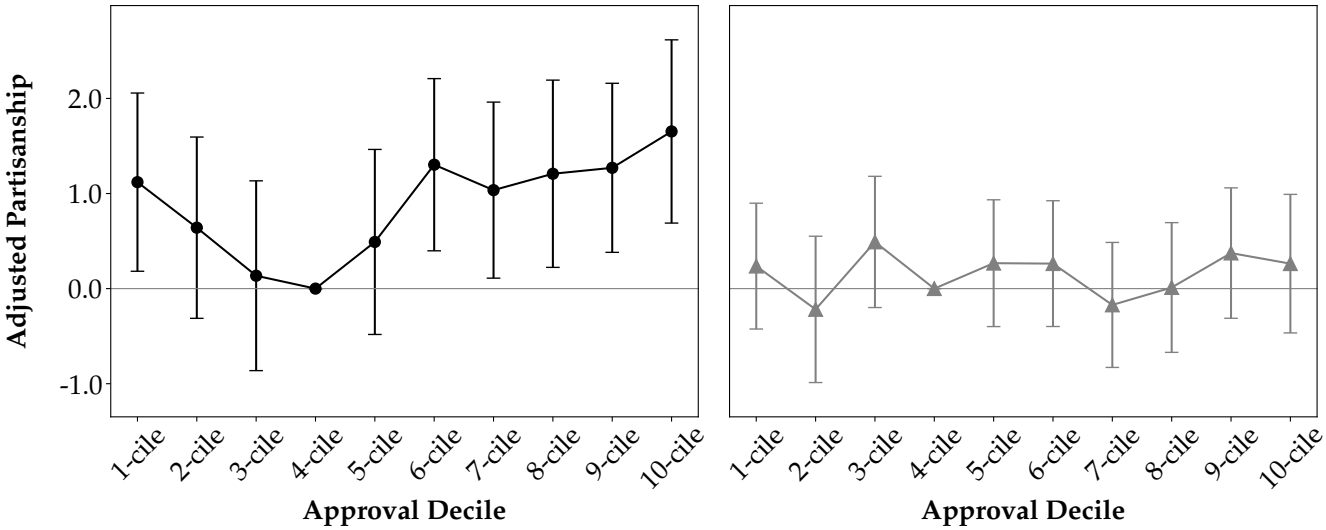
2.5.2 Estimating Model's Predictions

The figure below plots the estimates of the β_k coefficients using black circles in panel (a) on the left. The sums of the $\beta_k + \delta_k$ coefficients are plotted using gray triangles in panel (b) on the right.

[27] I choose the fourth decile as the base category because, as we will see, partisanship experiences a trough in the fourth decile.

[28] The α_k and γ_k coefficients have an analogous interpretation for reelection ineligible (lame duck) governors.

Figure 2.8: Partisanship of Reelectable Governors by Approval Decile, Adjusting for Lame Duck Behavior, Unaligned vs. Aligned Legislatures, 1990-2020



(a) Unaligned Legislatures (b) Aligned Legislatures

Dependent variable measures partisanship of U.S. governor State of the State speeches using leave-out estimator of Gentzkow et al. (2019), as calculated in equation (2.3), normalized by sample mean and standard deviation. Horizontal axis plots approval decile of gubernatorial approval by state, calculated using previous year's fourth quarter approval from Singer (2023) after first year in office, with first quarter approval utilized for first year in office. Black circles plots β_k coefficients from equation (2.5), measuring level of partisanship for reelectable governors with unaligned legislatures in decile k, adjusting for behavior of lame ducks, relative to fourth decile. Gray triangles plot $\beta_k + \delta_k$, measuring level of partisanship for reelectable governors with aligned legislatures in decile k, adjusting for behavior of lame ducks, relative to fourth decile. Bands around coefficient estimates display 95% confidence intervals. Legislative alignment measured as whether more than half of state legislators match governor's party. N: 1108; num. states: 48; within R^2: 0.06. Regression contains year and state fixed effects.

The left panel showcases a nonmonotonic relationship between approval decile and level of partisanship for unaligned legislatures. Being in the first decile of approval means governors' speeches are, on average, 1.1 standard deviations more partisan than in the fourth decile, falling gradually in the second and third deciles. Partisanship relative to the fourth decile begins rising in the fifth decile before remaining at between 1 to 1.5 standard deviations higher in the sixth through tenth deciles. This nonmonotonicity is consistent with prediction **P1** of our model.[29] For

[29]I also look at how approval deciles map onto winning gaps in gubernatorial elections. For governors eligible for reelection in the first approval decile, the average win gap is about six percentage points, i.e. a governor on average wins a race with about 53% of votes. This drops to less than four percentage points for swing states, which I define in the next section, with a large left tail. The average gap rises steadily until about 10 percentage points in the fifth decile and then all the way to 17 percentage points in the top decile.

reference, the overall standard deviation of partisanship in the data is 0.042 percentage points, roughly twice the increase we see in Figure 6 from 2000 to 2020. Figure 2.8.3. plots the consequent *levels* of partisanship for reelectable governors and lame ducks — i.e. the α_k and $\alpha + \gamma_k$ coefficients and the $\alpha_k + \beta_k$ and $\alpha_k + \beta_k + \gamma_k + \delta_k$ from equation (2.5). It shows that the nonmonotone pattern in Figure 2.8(a) emerges as the difference between a shallower (but statistically significant) U pattern for reelectable governors and an upside-down U pattern for lame duck governors.

P2 suggests that when legislative alignment increases, the nonmonotonicity of **P1** flattens out. The lack of any evident relationship between approval and partisanship for aligned legislatures in the right panel is consistent with prediction **P2**. Figure 2.8.4 shows that this finding is robust to alternative thresholds of legislative alignment. [30]

Table 2.8.2. lists the main regression coefficients for equation (2.5) from which Figure 2.8 is constructed. It shows that the coefficient on legislative alignment for reelectable governors is positive and statistically significant at the 10% level with a magnitude of 0.367 standard deviations. This is in line with the prediction of **P2** that increases in legislative alignment also cause increases in the baseline level of partisanship.

Figure B.5. additionally shows that the baseline result is robust to excluding the year reelectable governors are up for reelection; including *only* the year reelectable governors are eligible for reelection; and dropping observations for lame duck governors in their last year in office.[31] The lack of any obvious dynamic effects is in line with the theoretical model's robustness to multiple periods of passage.

Disaggregation by Competition The model also predicts, via **P3**, that competitive "swing states" should exhibit more prominent nonmonotonicities than states where a single party always wins the governorship. To test this prediction of the model, I break states into three groups based on the frequency of governor party over the sample period. This allows states to be categorized via a time-invariant definition.

[30] The results are strongest when defining legislative alignment as equal to 1 if more than 50 or 55% of legislators as matching the governor's party. The results are weaker using 45% or 60%. Although the former suffers from power issues, both findings are consistent with the comparative statics in Corollary 1. The median number of legislators of the governor's party is 55%. The 25^{th} and 75^{th} percentiles are respectively 43% and 65%.

[31] The nonmonotonicity is starkest when dropping lame ducks in their last year in office since, presumably, these governors have a substantively minimal policy agenda relative to prior years in office.

1. Republican States: states with a Republican governor more than 60% of the sample period (1990-2020).

2. Democratic States: states with a Democratic governor more than 60% of the sample period.

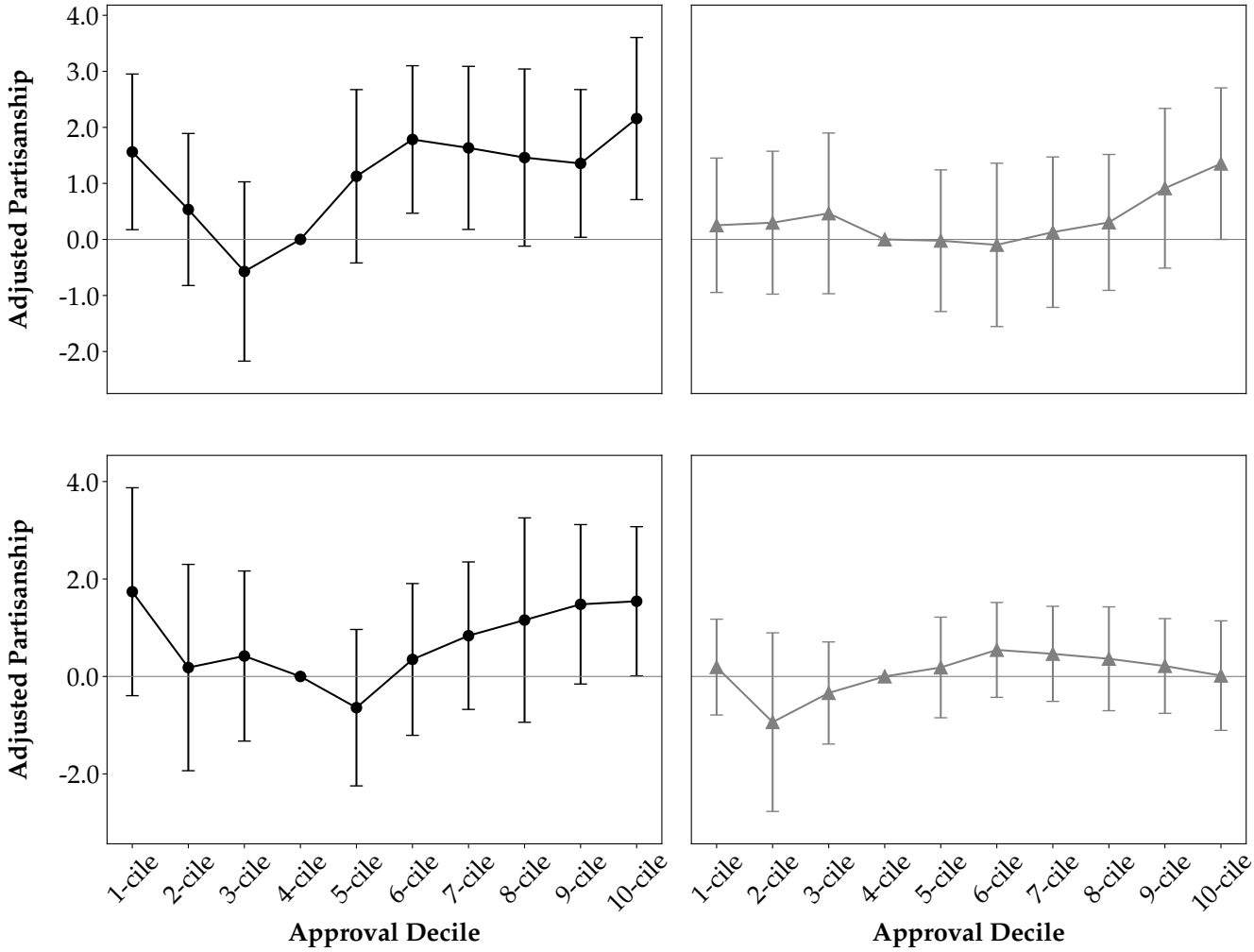
3. Swing States: remaining states, where the State governorship experiences fluctuations in party over the sample period.

A list of all states belonging to each category is displayed in Table 2.8.3. The six panels of the figure below showcase estimations of equation (2.4) for the three categories of state competition: Swing, Republican, and Democratic.

Figure 2.9: Partisanship of Reelectable Governors by Approval Decile, Adjusting for Lame Duck Behavior, Unaligned vs. Aligned Legislatures, Disaggregated by Competition, 1990-2020

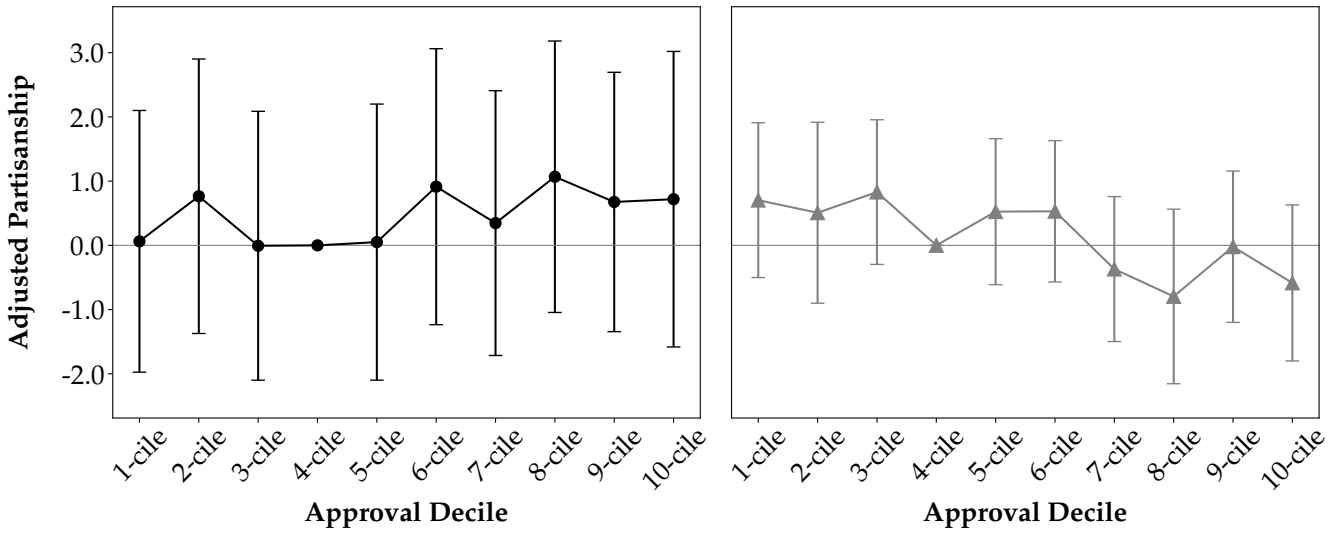
(a) Swing States, Unaligned Legislatures

(b) Swing States, Aligned Legislatures



(c) Rep. States, Unaligned Legislatures

(d) Rep. States, Aligned Legislatures



(e) Dem. States, Unaligned Legislatures (f) Dem. States, Aligned Legislatures

Dependent variable measures partisanship of U.S. governor State of the State speeches using leave-out estimator of Gentzkow et al. (2019), as calculated in equation (2.3), normalized by sample mean and standard deviation. Horizontal axis plots approval decile of gubernatorial approval by state, calculated using previous year's fourth quarter approval from Singer (2023) after first year in office, with first quarter approval utilized for first year in office. Black series plots β_k coefficients from equation (2.4), measuring level of partisanship for reelectable governors with unaligned legislatures in decile k, adjusting for behavior of lame ducks, relative to fourth decile. Gray circles plot $\beta_k + \delta_k$, measuring level of partisanship for reelectable governors with aligned legislatures in decile k, adjusting for behavior of lame ducks, relative to fourth decile. Bands around coefficient estimates display 95% confidence intervals. Legislative alignment measured as whether more than half of state legislators match governor's party. Republican states defined as those states with Republican governor more than 60% of sample period. Democratic states defined as those states with Democratic governor more than 60% of sample period. Swing states defined as remaining states. For panels (a) and (b), N: 421; num. states: 18; within R^2: 0.16. For panels (c) and (d), N: 369; num. states: 16; within R^2: 0.19. For panels (e) and (f), N: 318; num. states: 14; within R^2: 0.22.

A nonmonotonicity resembling that in Figure 2.7(a) emerges most prominently in panels 8(a) and 8(c), respectively corresponding to swing and Republican states. Partisanship for these states in the first decile of approval is approximately two standard deviations higher than a trough in the third or fifth decile. For swing states, partisanship in the highest decile of approval is about 2.5 standard deviations than the trough, while for Republican states, partisanship is about 2 standard deviations higher.[32] Consistent with the model, these results vanish for governors aligned with their legislatures.

Democratic states in panels (e) and (f) exhibit no relationship between partisanship and approval ratings except for a slightly negative relationship for aligned legislatures in panel (f). To

[32]The asymmetry in partisanship, where the level of partisanship for the lowest approval governors is lower than that of the highest approval governors, is consistent with the general version of the model presented in Theorem 1.

this end, Democratic states provide strong evidence of prediction **P3**: the U-shape of the model should be weaker in states with entrenched parties. While the nonmonotonicity of Republican states in panel (c) is ever so slightly attenuated relative to the swing states in panel (a), one explanation for the more drastic effect for Democrats is that the governors sample tends slightly Republican. States with Democratic governors for more than 60% of the sample comprise 25% of the data, while 36% have Republican governors for more than 60% of the sample. [33]

2.5.3 Presidential Speech

Finally, although the existence of only one president at any given time limits empirical analysis of presidential speech, the framework of the paper can also provide qualitative insight into patterns in partisanship of presidential policies, as well as their effects on reelection.

For example, the resounding success of Franklin D. Roosevelt's first two New Deal policies amidst high approval ratings contributed to a landslide reelection in 1936, illustrating how the success of a partisan agenda can broadly secure reelection. At the same time, Roosevelt continued to pursue more radical reform in his second term in what scholars sometimes call a "Third New Deal" — "far reaching" efforts to greatly expand a social welfare state, establish fiscal Keynesianism, and address the plight of the "one-third of a nation ill-housed, ill-clad, ill-nourished" — which met significant opposition from many Republicans and centrist Democrats that, along with other failures, stymied his approval later in his second term (Jeffries, 1996, p. 396). FDR's consequent shift to bipartisan policies towards the end of his second term can be seen as a response to this negative change in approval, a decision likely helping him secure reelection. [34]

Similarly, Lyndon B. Johnson enjoyed high popularity ratings leading up to his reelection. [35] His consequent exhibition of political acumen — a keen handling of the Kennedy assassination and passage of landmark civil rights legislation through a "congressional logjam" in less than a month (Caro, 2012) — contributed to a landslide reelection.[36] Johnson's soaring popularity may

[33] While a more strict criterion for Republican states is appealing — such as, e.g., defining them as states with Republican governors for more than 70% of the sample — a definition like suffers from missing values for some lame duck categories.

[34] "The reinforcing impacts of the court-packing bill, the 1937-38 recession, labor activism, and the 1938 purge attempt both diminished popular support for FDR and increased congressional opposition to the New Deal. By shifting attention to foreign and military affairs, restoring prosperity, and enhancing the prestige and political clout of business, World War II contributed to the ebbing of reform energies and to stalemate and consolidation" (Jeffries, 1996, p. 398).

[35] References for presidential approval come from The American Presidency Project (2024).

[36] Caro writes: "To watch him deal with Congress, deal with the Kennedys, confront a dozen other challenges for

have further influenced his pursuit of the ambitious Great Society reforms and the War on Poverty, despite the contentions of many conservative politicians, in line with the sort of phenomena predicted by our model.

Conversely, the passage or failure of more bipartisan or nonpartisan legislation may not generate a strong enough signal to win politicians reelection when the odds are stacked against them. While Jimmy Carter, for example, "did have some successes with Congress", it was "often because he backed existing Democratic programs, such as raising the minimum wage" while his party held a bicameral supermajority; Carter gained a reputation for "political ineptitude" compounded by a hostage crisis, communication blunders, and contentious relations with legislators amidst plummeting approval ratings (Strong, 2024). To this end, his success in achieving bipartisan goals was not informativeness enough of his skill, given his context, to win him reelection . Similarly, George H.W. Bush's bipartisan accomplishments — the Americans with Disabilities Ac and Clean Air Act — could not save him from a perceived negligence of domestic affairs and an ongoing recession that led to his loss to Bill Clinton (Knott, 2024). Clinton performed well in the election — corresponding to an incumbent with high initial reputation in the model — setting him up to pursue an ambitious partisan healthcare reform within his first 100 days in office. However, after the bill's eventual failure and Clinton's subsequently middling approval ratings, the president shifted to a decidedly more bipartisan program that made significant concessions to the Republicans — manifest in his plan ending "welfare as we know it," ratification of NAFTA, and the 1996 Crime Bill (Riley, 2024). Clinton's initial pursuit of an ambitious partisan policy, the failure of that partisan policy hurting (although not dooming) his reputation, and his subsequent success in pursuing concessionary politics to secure reelection is in line with the story of this paper's model.

2.6 Conclusion

This paper explores how political executives may utilize partisan policies as a means of winning reelection and, thereby, how changes in executives' electoral environments may generate variation in partisanship. I develop a theory that interprets an incumbent executive's choice of policy

which there was no precedent — for which he had to create his own precedents — is to watch a President, in very difficult circumstances, triumph over them, and it is therefore a means of gaining new insight into some fundamental realities about the pragmatic potential in the American presidency."

agenda as an information structure over incumbent ability, where less partisan policies generate left skewness in the distribution of posteriors over incumbent ability and more partisan policies generate right skewness. Incumbents who face threshold retention rules as a function of their ability, as in much of the voting literature, then exhibit a nonmonotonic relationship between the partisanship of their policies and their reputation. As their reputation increases, they first pursue partisan policy agendas, then bipartisan agendas, and then partisan agendas again. Utilizing an extension akin to a constrained persuasion model, I show that high reputation incumbents may even pursue more partisan policies than their low reputation counterparts. I show that these insights are robust to uncertainty in elections, allowing incumbents to choose opposing parties' platforms, and multiple periods of agenda setting.

I then apply these insights to explain how the partisanship of U.S. governors' policy proposals varies with changes in their electoral environments. I first document that gubernatorial partisanship —measured using governors' annual State of the State speeches — only becomes substantive after the early 2000s, creeps up slowly for the next decade, spikes after 2017, and consistently has a lower level than the comparable series for Congress. I then move to panel data to study partisanship of governors' policy proposals in these speeches, using the model to explore differences in the partisanship of reelectable and lame duck governors. I show that a nonmonotonic variation of gubernatorial partisanship with approval decile in unaligned legislatures matches that of the model, especially in swing states. I show that, in line with the model, increasing legislative alignment flattens this nonmonotonicity; that increasing legislative alignment also shifts up the baseline level of partisanship; and that states with entrenched parties are also less likely to showcase this nonmonotonicity. The panel analysis coupled with the theoretical analysis permits a view that changes in partisanship as not solely driven by latent trends in political landscapes but also shaped by elected officials' reelection incentives.

The paper broadly advances three research agendas. First, the model links the documentary literature on partisanship directly to the theoretical literature on political accountability. It provides a reputational and informational mechanism through which partisanship may manifest, grow, and fluctuate as a function of an incumbent executive's electoral environment.

Second, the documentation of trends in gubernatorial partisanship provides a novel series measuring partisanship of political executives by utilizing governors' State of the State speeches.

It also provides a novel series measuring partisanship of political speech at the state level, and utilizes advances in large-language modeling to specifically probe the partisanship of policy proposals.

Finally, the paper links a theoretical framework to aggregate partisanship with specific attention to explaining *variations* in partisanship, contrasting with a large literature which has largely documented *trends* in partisanship. Crucially, the model does not interpret departures from trends or widely discussed state-level heterogeneity in partisanship as noise, but as structured deviations generated by incumbents' incentives to win reelection. Regression analysis of panel data on governor-level partisanship provides evidence of the model's mechanism, manifest in a non-monotonic relationship between partisanship and popularity strongest for governors unaligned with their legislatures.

Executives at large, especially governors, exercise close control over public policies. In the dataset at hand, partisan attitudes manifest themselves in efforts to create affordable housing, combat climate change, restructure law enforcement, fund charter schools, extend healthcare coverage, raise minimum wages, create hospitable business environments, or cut taxes.[37] These policies have large effects on the shape of economic inequality, social stratification, and human capital acquisition within the States. Providing insight on when, how, and why politicians may pursue more or less partisan agendas permits richer insights into the timing, nature, and potential effects of these sorts of policies on the landscape of partisanship in American politics.

2.7 Proofs and Theoretical Extensions

2.7.1 Main Proofs

Proposition 1. *There exist thresholds* $0 < \underline{q} < q_L^2 < \overline{q} < 1$ *such that:*

- *for* $q_R^1 \in [0, \underline{q})$, $\pi^*(q_R^1)$ *is strictly decreasing, with* $\pi^*(0) = 1$ *and* $\pi^*(\underline{q}) = 0$*;*
- *for* $q_R^1 \in [\underline{q}, \overline{q})$, $\pi^*(q_R^1) = 0$*;*
- *for* $q_R^1 \in [\overline{q}, 1]$, $1 \in \pi^*(q_R^1)$ *with equality at* $\overline{q}$*.*

[37] These insights are gleaned from Table 2.8.1.

Proof. Note in general the following formulae for posteriors as a function of π, fixing q_R^1:

$$\overline{q}_R^2(\pi) = \frac{(\lambda + (1-\lambda)(1-\pi))q_R^1}{\lambda q_R^1 + (1-\lambda)(1-\pi)}$$
$$\underline{q}_R^2(\pi) = \frac{(1-\lambda)\pi q_R^1}{\lambda(1-q_R^1) + (1-\lambda)\pi}.$$

Note that $\overline{q}_R^2$ is decreasing in π while $\underline{q}_R^2$ is increasing in π.

Next, fix $q_R^1 \in (0,1)$. Consider the following sets $\overline{P}$ and $\underline{P}$:

$$\underline{P}(q_R^1) := \{\pi : \overline{q}_R^2(\pi) \geq q_L^2, \underline{q}_R^2(\pi) < q_L^2\}$$
$$\overline{P}(q_R^1) := \{\pi : \underline{q}_R^2(\pi) \geq q_L^2\}.$$

Note that $\underline{P}$ is always nonempty ($\ni \pi = 1$) but $\overline{P}$ is empty for q_R^1 low. I claim that $\pi^*(q_R^1) = \min\{\underline{P}(q_R^1)\}$ when $\overline{P} = \varnothing$ and $\overline{P}$ otherwise.

First, any $\pi \notin \overline{P}, \underline{P}$ always leads to an expected win probability of 0; these are dominated by any $\pi \in \underline{P}$. Next, for any $\pi \in \underline{P}$, the expected win probability is $\lambda q_R^1 + (1-\lambda)(1-\pi)$, which is the likelihood of seeing $\overline{q}_R^2$. This probability is maximized when π is minimized, i.e. at $\min\{\underline{P}(q_R^1)\}$. If $\overline{P}$ is empty, then $\pi^*(q_R^1) = \min\{\underline{P}(q_R^1)\}$. Finally, if $\overline{P}$ is nonempty, any $\pi \in \overline{P}$ leads to a win with probability 1, meaning $\pi^*(q_R^1) = \overline{P}(q_R^1)$.

Next, we show $\min\{\underline{P}(q_R^1)\}$ is strictly decreasing in π until a point $\underline{q}$, whereafter it is equal to 0. In general, $\min\{\underline{P}(q_R^1)\}$ is the solution to $\frac{(\lambda+(1-\lambda)(1-\pi))q_R^1}{\lambda q_R^1+(1-\lambda)(1-\pi)} = q_L^2$. As q_R^1 rises, the left-hand-side increases for each π, meaning the π solving this equation decreases. The solution to this equation exists up until some $\underline{q} < q_L^2$, defined by $\frac{(\lambda+(1-\lambda))\underline{q}}{\lambda\underline{q}+(1-\lambda)} = q_L^2$, where $\pi = 0$ and can no longer decrease. For $q_R^{1'} > q_R^1$, if $\pi \in \underline{P}(q_R^1)$, then $\pi \in \underline{P}(q_R^{1'})$, meaning $\min\{\underline{P}(q_R^1)\} = 0$ for $q_R^1 \geq \overline{q}$. Finally, as $q_R^1 \to 0$, $\pi \to 1$.

Finally, we show that there exists $\overline{q} > q_L^2$ such that $\overline{P}(q_R^1) \neq \varnothing$ if and only if $q_R^1 \geq \overline{q}$. Note that, fixing q_R^1, the expression for $\underline{q}_R^2$ is maximized when $\pi = 1$. This means that $\overline{P}$ is nonempty if and only if $\frac{(1-\lambda)q_R^1}{1-\lambda q_R^1} \geq q_L^2$, which occurs if and only if $q_R^1 \geq \overline{q} > q_L^2$ defined as the implicit solution to $\frac{(1-\lambda)\overline{q}}{1-\lambda\overline{q}} = q_L^2$.

The comparative statics with respect to q_L^2 and λ emerge directly from the equations defining $\overline{q}$ and $\underline{q}$. □

Proposition 2. *The equilibrium value of learning for the voter, $V_V(\pi^*)$, is as follows.*

- *For $q_R^1 \in [0, \underline{q}]$, $V_V(\pi^*) = q_L^2$;*
- *for $q_R^1 \in [\underline{q}, \overline{q})$, $V_V(\pi^*) = q_R^1 + \lambda(1 - q_R^1)q_L^2$;*
- *for $q_R^1 \in [\overline{q}, 1]$, $V_R(\pi^*) = q_R^1$.*

Proof. Below $\underline{q}$, we either have $q_R^1 \to \underline{q}_R^2 < q_L^2$ or $q_R^1 \to \overline{q}_R^2 = q_L^2$. In the former case, V elects L and receives q_L^2; in the latter, she is indifferent between R and L but receives q_L^2 either way. In the region $(\underline{q}, \overline{q})$, $q_R^1 \to 0$ or some $\overline{q}_R^2 > q_L^2$. In the former case, V replaces R but in the latter she is retained, and her expected ability is greater than q_L^2. Finally, for $q_R^1 \geq \overline{q}$, R is always retained. The expected value of the posterior q_2^R is, as a result, simply the prior q_R^1. □

Proposition 3. *As $q_L^2 \to 1$, $\underline{q} \to \overline{q}$. As $q_L^2 \to 0$, $\overline{q} \to \underline{q}$. There exists q_L^{2*} that maximizes the size of the interval $[\underline{q}, \overline{q})$.*

Proof. The closed forms for $\underline{q}, \overline{q}$ are respectively, based on the previous proposition:

$$\frac{(1-\lambda)q_L^2}{1-\lambda q_L^2}, \frac{q_L^2}{1-\lambda(1-q_L^2)}.$$

Both terms are increasing in q_L^2. The derivative of the difference of these two terms with respect to q_L^2 is given by:

$$\frac{1-\lambda}{(1-\lambda(1-q_L^2))^2} - \frac{1-\lambda}{(1-\lambda q_L^2)^2}$$

which is positive for $q_L^2 \leq 1/2$, negative when $q_L^2 \geq 1/2$, and $= 0$ at $q_L^{2*} = 1/2$. □

The following two lemmata are used to prove Theorem 1.

Lemma 1. *Suppose w is a differentiable, strictly concave, and strictly increasing function defined over q_R^2. Then, there exists $p \in [0, 1]$ such that the expected value of w over the lottery of posteriors $\overline{q}_R^2, \underline{q}R^2$ is maximized when $(p_1, p_0) = (p, p)$.*

Proof. We make a slight change of notation, $q_R^1 = q$. Our problem describing the maximum of w over the lotteries over q_R^2 is given by:

$$\max_{p_1 \geq p_0} \int_{q_R^2} w(q_R^2) dF(q_R^2 | p_1, p_0, q).$$

Let $g(p_1)$ be the probability of a success conditional on $a_R = 1$; and $h(p_0)$ the probability of success conditional on $a_R = 0$. Note that the distribution of posteriors $F(q_R^2 | p_1, p_0, q_R^2)$ is a two-point mean preserving spread of the prior q.

Suppose by contradiction that $p_1 > p_0$. This means that there would exist p_1, p_0 such that

$$\begin{aligned} \overline{q}_R^2(p_1, p_0) &= \frac{g(p_1)q}{g(p_1)q + h(p_0)(1-q)} \\ \underline{q}_R^2(p_1, p_0) &= \frac{q - qg(p_1)}{1 - qg(p_1) - (1-q)h(p_0)}. \end{aligned}$$

The partial derivatives of $\overline{q}_R^2$ are:

$$\begin{aligned} p_1 \qquad & q(1-q)g'(p_1)\frac{h(p_0)}{[g(p_1)q + h(p_0)(1-q)]^2} > 0 \\ p_0 \qquad & -q(1-q)h'(p_0)\frac{g(p_1)}{[g(p_1)q + h(p_0)(1-q)]^2} < 0. \end{aligned}$$

The partial derivatives of $\underline{q}_R^2$ are:

$$\begin{aligned} p_1 \qquad & -q(1-q)g'(p_1)\frac{(1-h(p_0))}{[1 - qg(p_1) - (1-q)h(p_0)]^2} < 0 \\ p_0 \qquad & q(1-q)h'(p_0)\frac{(1-g(p_0))}{[1 - qg(p_1) - (1-q)h(p_0)]^2} > 0 \end{aligned}$$

Suppose we marginally decrease p_1 by Δ_1 and increase p_0 by Δ_0 so that $\overline{q}_R^2$ remains the same, i.e.

$$\begin{aligned} q(1-q)g'(p_1)\frac{h(p_0)}{[g(p_1)q + h(p_0)(1-q)]^2}\Delta_1 &= q(1-q)h'(p_0)\frac{g(p_1)}{[g(p_1)q + h(p_0)(1-q)]^2}\Delta_0 \\ \implies g'(p_1)h(p_0)\Delta_1 - g(p_1)h'(p_0)\Delta_0 &= 0 \end{aligned}$$

The sign of the change in $\underline{q}^2_R$ is then the sign of:

$$\begin{aligned} &-g'(p_1)(1-h(p_0))\Delta_1 + h'(p_0)(1-g(p_1)\Delta_0 \\ &= g'(p_1)h(p_0))\Delta_1 - g(p_1)h'(p_0)\Delta_0 - g'(p_1)\Delta_1 + h'(p_0)\Delta_0 \\ &= -g'(p_1)\Delta_1 + h'(p_0)\Delta_0 \end{aligned}$$

$g(p_1)$ is given by $\lambda+(1-\lambda)p_1$, so its derivative is $(1-\lambda)$. $h(p_0)$ is given by $(1-\lambda)p_0$, so its derivative is also $(1-\lambda)$. Hence, the sign of the change in $\underline{q}^2_R$ is the sign of $\Delta_0 - \Delta_1$.

I claim that $\Delta_0 > \Delta_1$. The expressions for these are

$$\Delta_1 = q(1-q)g'(p_1)h(p_0)$$
$$\Delta_0 = q(1-q)g(p_1)h'(p_0).$$

The latter is larger than the former if and only if

$$g(p_1)h'(p_0) > g'(p_1)h(p_0).$$

Since $h'(p_0) = g'(p_1) = 1-\lambda$, this is true if and only if $g(p_1) > h(p_0)$, which is always true by the constraint that $p_1 \geq p_0$ (i.e. the conditional likelihood of a success is higher for $a_R = 1$ than $a_R = 0$). Hence, decreasing p_1 by Δ_1 and raising p_0 by Δ_0 keeps $\overline{q}^2_R$ fixed while raising $\underline{q}^2_R$, generating a mean preserving contraction of the original lottery over posteriors. Because w is strictly increasing and concave, this new lottery over posteriors is preferred to the original generated by (p_1, p_0), a contradiction. Hence, we always have that w is maximized when $p_1 = p_0 = p$. □

Lemma 2. *Suppose w is a differentiable, strictly concave, and strictly increasing function defined over q^2_R. Then, the lottery generated by $(p,p) = (0,0)$ over q^2_R dominates all other (p_1, p_0).*

Proof. Note first that $\overline{q}^2_R$ is strictly decreasing as a function of p, meaning $\overline{q}^2_R$ is maximized at $p = 0$. Note that any lottery over posteriors can be represented with a line segment connecting $(\underline{q}^2_R, w(\underline{q}^2_R))$ and $(\overline{q}^2_R, w(\overline{q}^2_R))$, with the expected value of the lottery given by the point on the segment corresponding to the prior q.

Note that, fixing $\overline{q}^2_R$, flatter line segments correspond to higher expected values. It then suffices

to show that the line segment connecting $(\underline{q}^2_R, w(\underline{q}^2_R))$ and $(\overline{q}^2_R, w(\overline{q}^2_R))$ is shallowest at $p = 0$ — where $w(\overline{q}^2_R)$ is additionally maximal. I.e., its slope, given by

$$\frac{w(\overline{q}^2_R(p)) - w(\underline{q}^2_R(p))}{\overline{q^2_R} - \underline{q}^2_R},$$

achieves a minimum at $p = 0$. The numerator of the derivative of this expression is given by:

$$[\overline{q^2_R} - \underline{q}^2_R][w'(\overline{q}^2_R)\overline{q^2_R}' - w'(\underline{q}^2_R)\underline{q^2_R}'] - [w(\overline{q}^2_R) - w(\overline{q}^2_R)][\overline{q}^{2'}_R - \underline{q}^{2'}_R]$$

Note that $h'(p) = g'(p) = (1 - \lambda)$, and that $g(p) - h(p) = \lambda$, so that:

$$\begin{aligned}\overline{q}^{2'}_R - \underline{q}^{2'}_R &= \frac{q(1-q)[g'(p)h(p) - h'(p)g(p)]}{[qg(p) + (1-q)h(p)]^2} - \frac{q(1-q)[g'(p)(1-h(p)) - h'(p)(1-g(p))}{[1 - qg(p) - (1-q)h(p)]^2} \\ &= \frac{q(1-q)\lambda(1-\lambda)}{[1 - qg(p) - (1-q)h(p)]^2} - \frac{q(1-q)\lambda(1-\lambda)}{[qg(p) + (1-q)h(p)]^2}\end{aligned}$$

Hence the derivative above is ≥ 0 if and only if

$$\begin{aligned}&[\overline{q^2_R} - \underline{q}^2_R][-w'(\overline{q}^2_R)\frac{q(1-q)\lambda(1-\lambda)}{[qg(p) + (1-q)h(p)]^2} + w'(\underline{q}^2_R)\frac{q(1-q)\lambda(1-\lambda)}{[1 - qg(p) - (1-q)h(p)]^2}] \\ &-[w(\overline{q}^2_R) - w(\overline{q}^2_R)]\left[\frac{q(1-q)\lambda(1-\lambda)}{[1 - qg(p) - (1-q)h(p)]^2} - \frac{q(1-q)\lambda(1-\lambda)}{[qg(p) + (1-q)h(p)]^2}\right] \geq 0 \\ &\frac{w'(\underline{q}^2_R)}{[1 - qg(p) - (1-q)h(p)]^2} - \frac{w'(\overline{q}^2_R)}{[qg(p) + (1-q)h(p)]^2} \\ &\geq \frac{\frac{w(\overline{q}^2_R) - w(\overline{q}^2_R)}{\overline{q^2_R} - \underline{q}^2_R}}{[1 - qg(p) - (1-q)h(p)]^2} - \frac{\frac{w(\overline{q}^2_R) - w(\overline{q}^2_R)}{\overline{q^2_R} - \underline{q}^2_R}}{[qg(p) + (1-q)h(p)]^2}\end{aligned}$$

Note in particular that by concavity, $w'(\underline{q}^2_R) > \frac{w(\overline{q}^2_R) - w(\overline{q}^2_R)}{\overline{q^2_R} - \underline{q}^2_R} > w'(\underline{q}^2_R)$. This means that the first term on the left is always strictly larger than the first time on the right; and that the magnitude of the second term on the left is smaller than the magnitude of the second term on the right, so that this expression is always true, i.e. the derivative is positive for all p. Hence, the slope of the segment achieves a minimum at $p = 0$, showing the result. □

Theorem 1. *There exist thresholds $\underline{q} < \overline{q}$ such that:*

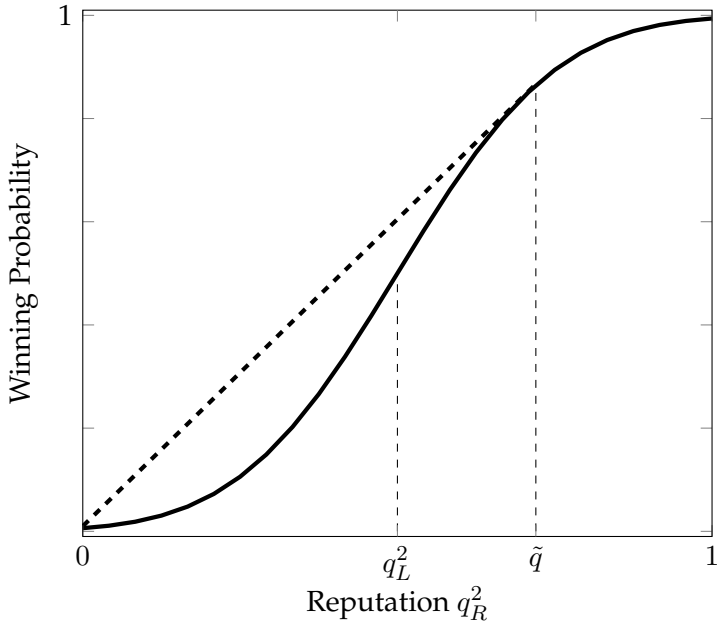
- *for $q^1_R \leq \underline{q}$, $\pi^*(q^1_R)$ is strictly decreasing. $p^*(0) = (1, 0)$ and $\pi^*(0) = 1/2$, while $\pi^*(\underline{q}) = 0$ and*

$\pi^*(\underline{q}) = 0$.

- *For* $q_R^1 \in [\underline{q}, \overline{q})$, $p^* = (1,1)$ *and* $\pi^* = 0$;
- *for* $q_R^1 \geq \overline{q}$, $p^* = (0,0)$ *and* $\pi^* = 1$.

Proof. The proof of this result makes use of the previous two lemmata. First, note that the concavification of $N(q_R^2 - q_L^2)$ is characterized by a point $\tilde{q} > q_L^2$. Specifically, the concavification is given by a line segment connection the points $(0, N(-q_L^2))$ and $(\tilde{q}, N(\tilde{q} - q_L^2))$, where $\tilde{q}$ solves $N'(\tilde{q} - q_L^2)\tilde{q} = N(\tilde{q} - q_L^2) - N(-q_L^2)$, followed by the curve N itself for $q \geq \tilde{q}$. Note also that N is strictly concave for $q_R^1 > q_L^2$ and strictly convex for $q_R^1 < q_L^2$. These are shown in the figure below, where the solid line is $N(q_R^2 - q_L^2)$ and the dashed line the concavification.

Figure 2.10: Concavification of $N(q_R^2 - q_L^2)$



Next, notice that there exists $0 < \underline{q} < \tilde{q}$ such that for all $q_R^1 \leq \underline{q}$, we can choose p_1 and p_0 to achieve the concavification. Since $\underline{q}_R^2$ must $= 0$ for any point $\leq \tilde{q}$ to achieve the concavification, we must have $p_1 = 1$ and vary p_0. p_0 must then solve, for each q_R^1,

$$\frac{q_R^1}{q_R^1 + (1-\lambda)(1-q_R^1)p_0} = \tilde{q}.$$

Notice that this expression is minimized when $p_0 = 1$, when it is equal to $\frac{q_R^1}{q_R^1+(1-\lambda)(1-q_R^1)}$. Let $\underline{q}$ solve $\frac{\underline{q}}{\underline{q}+(1-\lambda)(1-\underline{q})} = \tilde{q}$. Note that for all $q_R^1 > \underline{q}$, there does not exist $p_0 \in [0,1]$ such that the posteriors $(\overline{q}_R^2, \underline{q}_R^2) = (\tilde{q}, 0)$ are achievable. It is easy to see that for all $q \leq \underline{q}$, there exists $p_0(q_R^1)$ that achieves these posteriors, with p_0 given directly by

$$p_0 = \frac{q_R^1(1-\tilde{q})}{(1-\lambda)(1-q_R^1)\tilde{q}},$$

which is strictly increasing from 0 at $q_R^1 = 0$ to 1 at $\underline{q}$. This argument uses a similar approach to Proposition 1.[38]

Next, we show that for all $q_R^1 \geq \underline{q}$, $(p_1, p_0) = (1,1)$ dominates any (p_1, p_0) such that $\underline{q}_R^2 < q_L^2$. Suppose by contradiction that there exists $(p_1, p_0) \neq (1,1)$ such that $\underline{q}_R^2 < q_L^2$, which dominates $(1,1)$. Because $\overline{q}_R^2$ is minimized at $(p_1, p_0) = (1,1)$, we necessarily have that $\overline{q}_R^2 > q_L^2$, so that $\overline{q}_R^2$ is on the concave portion of N. Suppose first that $p_1 = 1$ so that $\underline{q}_R^2 = 0$. The expected value of this lottery is the line segment from $(0, N(-q_L^2))$ to $(\overline{q}_R^2, N(\overline{q}_R^2 - q_L^2))$. Since $p_0 < 1$, we can generate a strict improvement by increasing p_0, which slightly lowers $\overline{q}_R^2$. However, because $\overline{q}_R^2$ is on the concave portion of N, the line segment from $(0, N(-q_L^2))$ to $(\overline{q}_R^2, N(\overline{q}_R^2 - q_L^2))$ becomes steeper while originating from the same point $(0, N(-q_L^2))$. Because the expected value from (p_1, p_0) lies on the point on the segment corresponding to q_1^2, this steepening generates an improvement on the original (p_1, p_0), a contradiction.

Hence, suppose $p_1 < 1$ so that $\underline{q}_R^2 > 0$. If $p_0 > 0$, we can reverse the argument in Lemma 1 and slightly raise p_1 and lower p_0 so that $\overline{q}_R^2$ remains the same but $\underline{q}_R^2$ decreases. The line segment connecting $(\underline{q}_R^2, N(\underline{q}_R^2 - q_L^2))$ to $(\overline{q}_R^2, N(\overline{q}_R^2 - q_L^2))$ becomes shallower, while its upper point $(\overline{q}_R^2, N(\overline{q}_R^2 - q_L^2))$ remains the same, meaning the line rotates upwards, again generating an improvement. If $p_0 = 0$, increasing p_0 causes both $\underline{q}_R^2$ and $\overline{q}_R^2$ to decrease. However, because $\overline{q}_R^2$ is on the concave portion of N and $\underline{q}_R^2$ on the convex portion, this generates a left/upward shift in the line segment connecting $(\underline{q}_R^2, N(\underline{q}_R^2 - q_L^2))$ to $(\overline{q}_R^2, N(\overline{q}_R^2 - q_L^2))$, again generating an improvement.

Next, we show that there exists $\overline{q}$ such that for all $q_R^1 \geq \overline{q}$, $(p_1, p_0) = (0,0)$ dominates $(1,1)$. $(0,0)$ generates posteriors $\overline{q}_R^2 = 1$ and $\underline{q}_R^2 = \frac{(1-\lambda)q_R^1}{1-\lambda q_R^1}$. Note that because $1 > \frac{q_R^1}{q_R^1+(1-\lambda)(1-q_R^1)}$, a sufficient

[38] In particular, $\frac{\underline{q}}{\underline{q}+(1-\lambda)(1-\underline{q})} = \tilde{q}$ gives the comparative static of $\underline{q}$ with respect to λ and q_L^2. Increasing λ decreases $\underline{q}$. Increasing q_L^2 increases $\tilde{q}$ and hence increases $\underline{q}$.

condition for this to hold is that the line segment connecting and $(\frac{(1-\lambda)q_R^1}{1-\lambda q_R^1}, N(\frac{(1-\lambda)q_R^1}{1-\lambda q_R^1} - q_L^2)$ and $(1, N(1-q_L^2))$ is shallower than that connecting $(0, N(-q_L^2))$ to $(\frac{q_R^1}{q_R^1+(1-\lambda)(1-q_R^1)}, N(\frac{q_R^1}{q_R^1+(1-\lambda)(1-q_R^1)} - q_L^2))$. As $q_R^1 \to 1$, the slope of the former line segment approaches 0; while the slope of the latter approaches $N(1-q_L^2) - N(-q_L^2)$. Because the change in these slopes is monotone as long as q_R^1 is sufficiently high (i.e. as long as $\frac{(1-\lambda)q_R^1}{1-\lambda q_R^1} \geq q_L^2$) , by continuity, there exists $\overline{q}$ such that for $q_R^1 \geq \overline{q}$, $(0,0)$ dominates $(1,1)$.

Finally, using Lemma 2, we know that $(0,0)$ dominates any other (p_1, p_0) such that $\overline{q}_R^2, \underline{q}_R^2 \geq q_L^2$. Because $(1,1)$ dominates any (p_1, p_0) with $\overline{q}_R^2 \geq q_L^2 \geq \underline{q}_R^2$, $(0,0)$ also dominates all these points, showing the result.

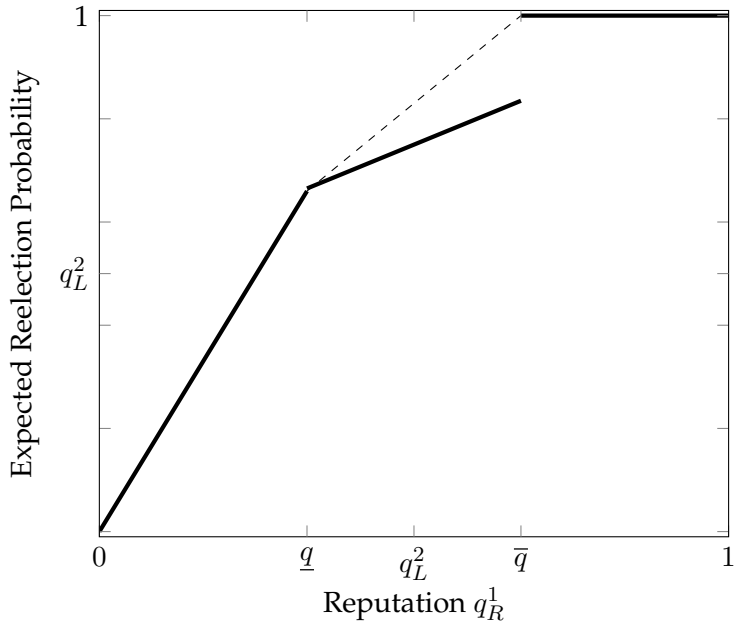
□

Proposition 4. *Given $\underline{q}, \overline{q}$, there exist thresholds $\underline{\underline{q}} < \underline{q} < \overline{q} < \overline{\overline{q}}$ such that*

- *For $q_R^0 \leq \underline{\underline{q}}$, partisanship π^* is decreasing from $1/2$ at q_R^0 to 0 at $\underline{\underline{q}}$.*
- *There exists $q_\dagger \in [\underline{\underline{q}}, \overline{q} < \overline{\overline{q}}]$ such that for all $q_R^0 \geq q_\dagger$, $\pi^* < 1$.*
- *For $q_R^0 \geq \overline{\overline{q}}$, $\pi^* = 1$.*

Proof. We graph the value function of R, $V_R(q_R^1)$, as a function of q_R^1 in the figure below.

Figure 2.11: Value Function of R As Function of q_R^1



The concavification of the value function is achievable for low q_R^1 as a lottery between the beliefs 0 and $\underline{q}$, achieved via $(p_1, p_0) = (1, p_0^*)$ for p_0^* solving $\frac{q_R^0}{q_R^0 + (1-q_R^0)(1-\lambda)p_0^*} = \underline{q}$. p_0^* increases in q_R^0 until some $\underline{\underline{q}}$, when it is equal to 1. By a similar argument from before, partisanship $\pi^* = 1 - \frac{1+p_0^*}{2}$ is then decreasing from $1/2$ at 0 to 0 at $\underline{\underline{q}}$. An identical argument to proposition 1 shows that above some threshold $\overline{\overline{q}}$, the optimal agenda is given by $(p_1, p_0) = (0, 0)$.

Let $q_\dagger$ solve $\underline{q} = \frac{(1-\lambda)q_R^0}{1-\lambda q_R^0}$. Note that for each (p_1, p_0), we have a line segment connecting $(\underline{q}_R^1, V_R(\underline{q}_R^1))$ to $(\overline{q}_R^1, V_R(\overline{q}_R^1))$. Moreover, for $q_R^0 \in [q_\dagger, \overline{\overline{q}}]$, the expected value of $(p_1, p_0) = (0, 0)$ is the segment connecting the value function at $\underline{q}$ to the value function at 1 (i.e. an extension of the value function on $[\underline{q}, \overline{q}]$).

I claim that for all $q_R^0 \in (q_\dagger, \overline{\overline{q}})$, $\pi^* < 1$, i.e. $(p_1, p_0) \neq (0, 0)$. To see this, note that by slightly increasing both p_1 and p_0, $(\underline{q}_R^1, V_R(\underline{q}_R^1))$ decreases linearly. However, $(\overline{q}_R^1, V_R(\overline{q}_R^1))$ slides to the left without decreasing, meaning the line segment steepens and, at the prior q_R^0, generates an improvement on $(p_1, p_0) = (0, 0)$. The expressions for points in between are generally dependent on λ and q_L^2.

□

2.7.2 Voter Preference Microfoundation

I address a microfoundation for second period voter utility, providing conditions under which, regardless of a voter's ideological preferences she retains an incumbent if and only if the incumbent's ability is sufficiently high.

Consider two dimensions of utility which are functions of ability. The first is a managerial component; for politician R, the voter's utility from this component simply a_R. The second dimension is an ideological component, weighted by some $\delta > 0$. Given that the incumbent pursues a partisan policy π, the voter has preferences given by $-|\lambda - \pi|$, i.e. the voter's ideal point is given by alignment with the legislature. This loss is experienced only if π actually passes.[39] Voter utility is given by the sum of these two components, as follows:

$$a_R - \delta(\lambda a_R + (1-\lambda)(1-\pi))|\lambda - \pi|$$

[39] Any outside option whose utility is ≤ 0 upon failure to pass the policy will accentuate the result.

I assume that, in the second period, R pursues $\pi = 1$. The voter's utility is given then by:

$$a_R - \delta(\lambda a_R + (1-\lambda)(1-\pi))(1-\lambda),$$

which is increasing in a_R if and only if

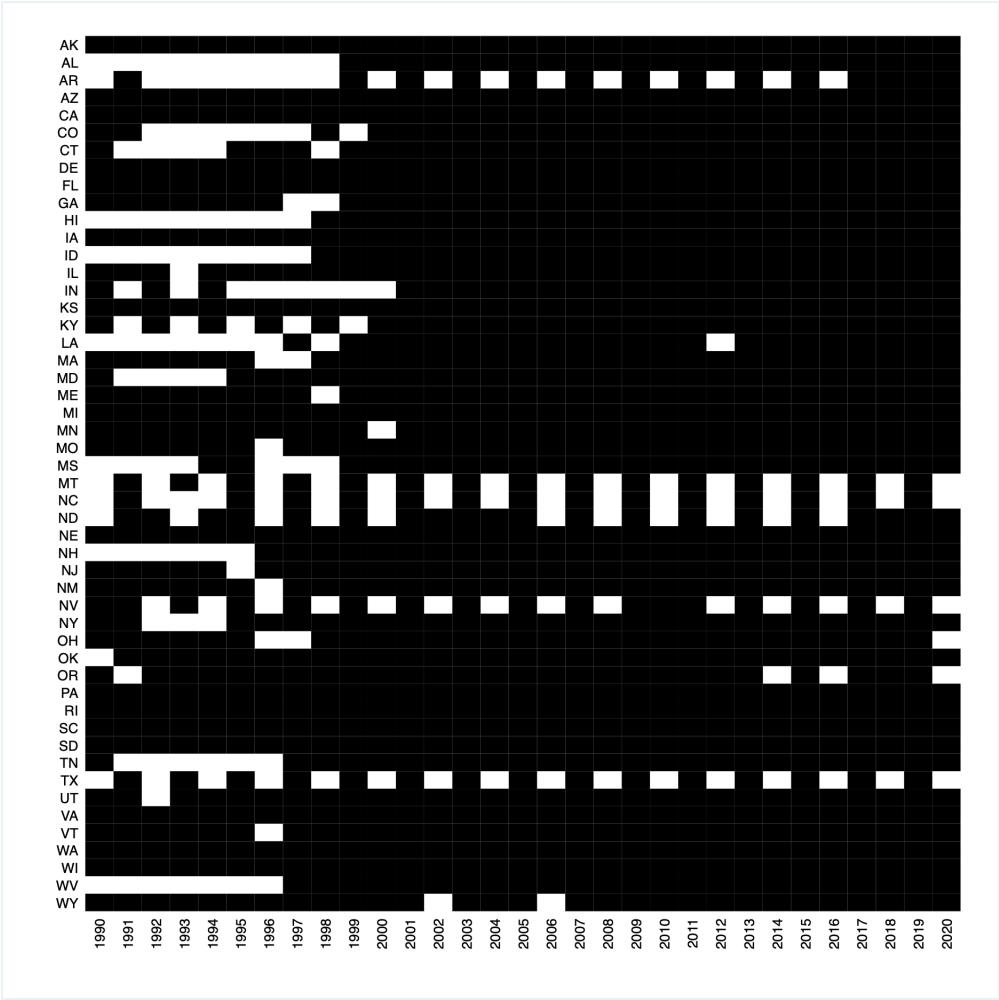
$$1 - \delta\lambda(1-\lambda) \geq 0 \iff 1 \geq \delta\lambda(1-\lambda).$$

The right hand side is maximized at $\lambda = 1/2$ with value $\frac{\delta}{4}$. This inequality hence holds everywhere for $\delta \leq 4$. If $\delta = 1$ (managerial and ideological components of utility are weighted equally), this always holds. The inequality also always holds in neighborhoods of $\lambda = 0$ and $\lambda = 1$. The intuition is that when λ is low, partisan legislation passes with low probability. Hence, even if the incumbent is skilled (including if she is skilled at passing legislation V would heavily dislike), this unlikeable legislation never passes.

Note that voter utility is a linear function of a_R; hence, when compared to a constant outside option at $t = 2$ (i.e. the utility from replacing R with L, plus any additional valence terms for L), V will choose to re-elect R if and only if q_R^2 exceeds a threshold.

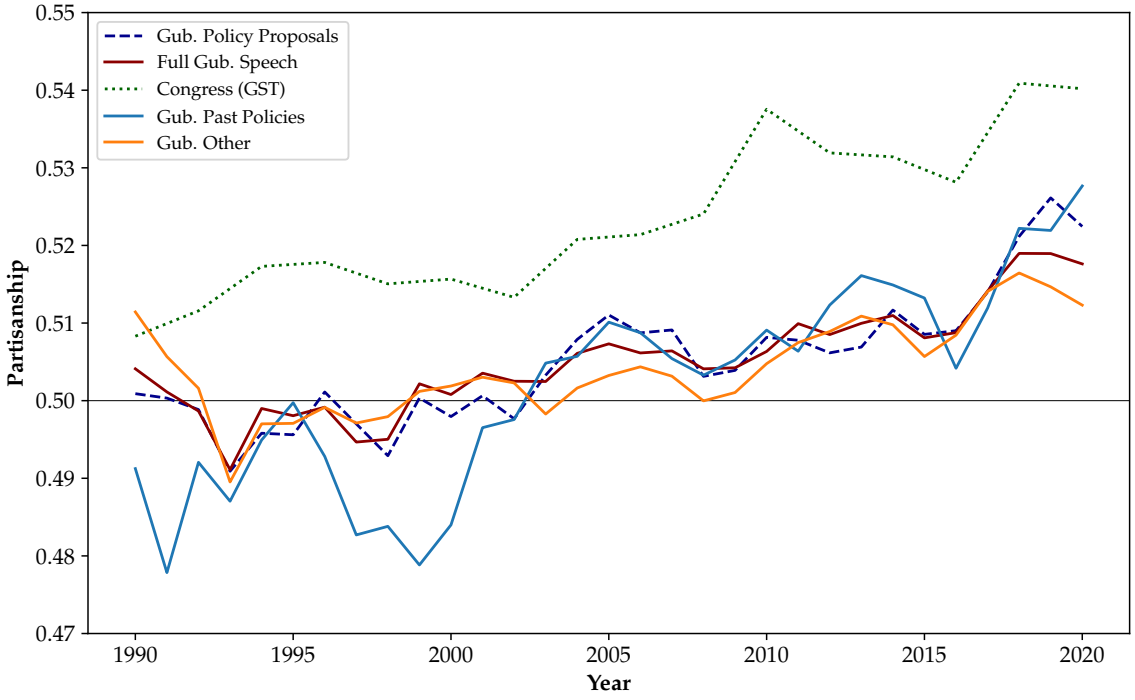
2.8 Auxiliary Tables and Figures

Figure 2.8.1: Usable Data Coverage by State and Year



Solid black squares indicate availability of usable speech for given state and year. Empty squares indicate lack of speech data for that year. Some states experience periodicity in missing speeches due to biannual delivery of addresses, such as Texas.

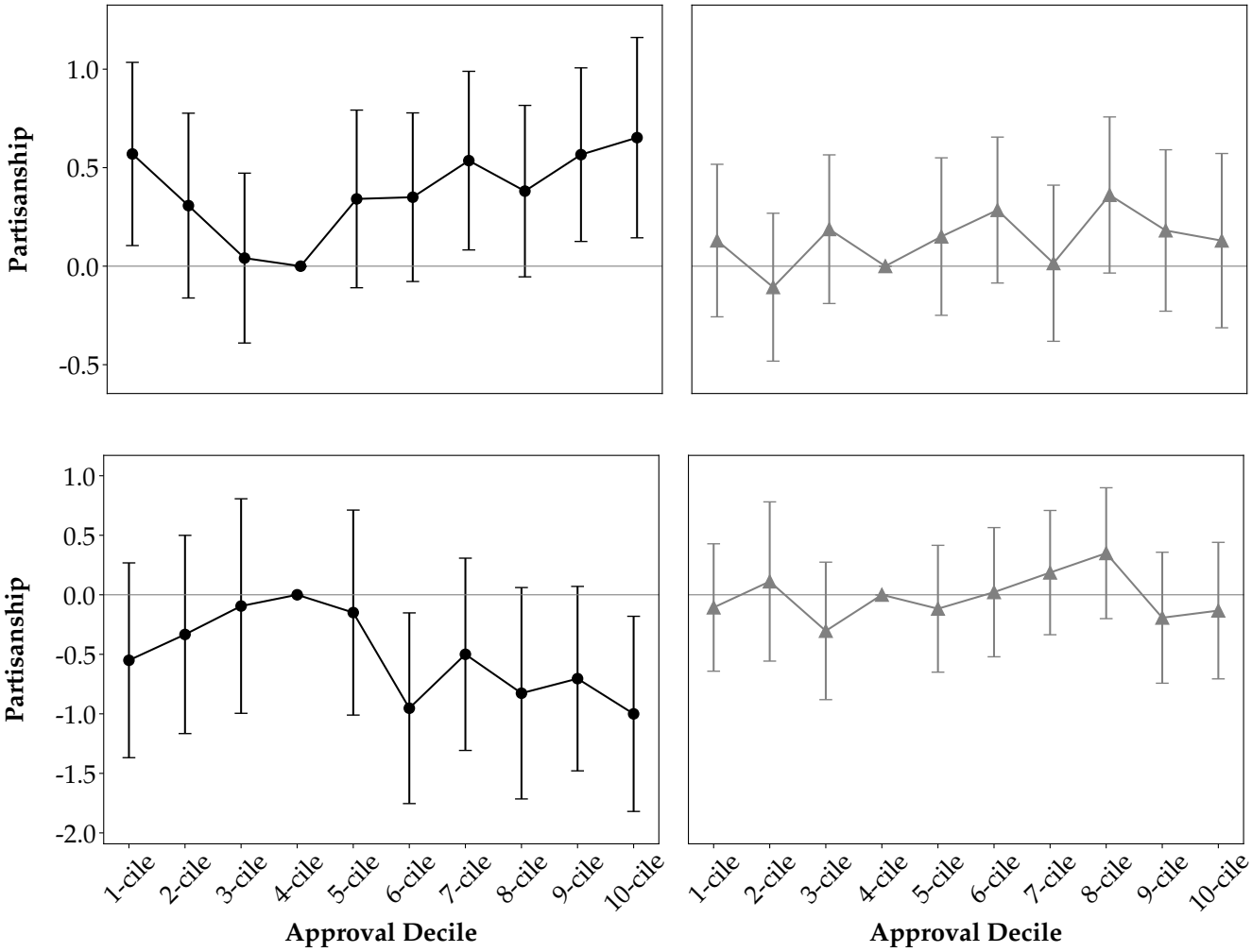
Figure 2.8.2: Partisanship of U.S. Gubernatorial Speeches, Full Speech, Policy Proposals, Past Policies, and Other Speech, vs. Congressional Record: 1990-2020



Partisanship of gubernatorial speech calculated using Gentzkow et al. (2019) leave-out estimator. "Gub. Policy Proposals" computes estimator for gubernatorial speech snippets coded as discussing policy proposals on corpus of U.S. governors' speech from 1990-2020. "Full Gub. Speech" computes estimator for all gubernatorial speech in given year. "Gub. Past Policies" computes estimator for snippets coded as discussing policies but not policy proposals. "Gub. Other" computes estimator for snippets coded as not discussing policies. "Congress (GST)" series is partisanship of Congressional speech using leave-out estimator, replicating Figure 2A of Gentzkow et al. (2019).

Figure 2.8.3: Unadjusted Levels of Partisanship for Reelectable and Lame Duck Governors by Approval Decile, Aligned vs. Unaligned Legislatures, 1990-2020

(a) Reelectable Govs., Unaligned Legislatures (b) Reelectable Govs., Aligned Legislatures



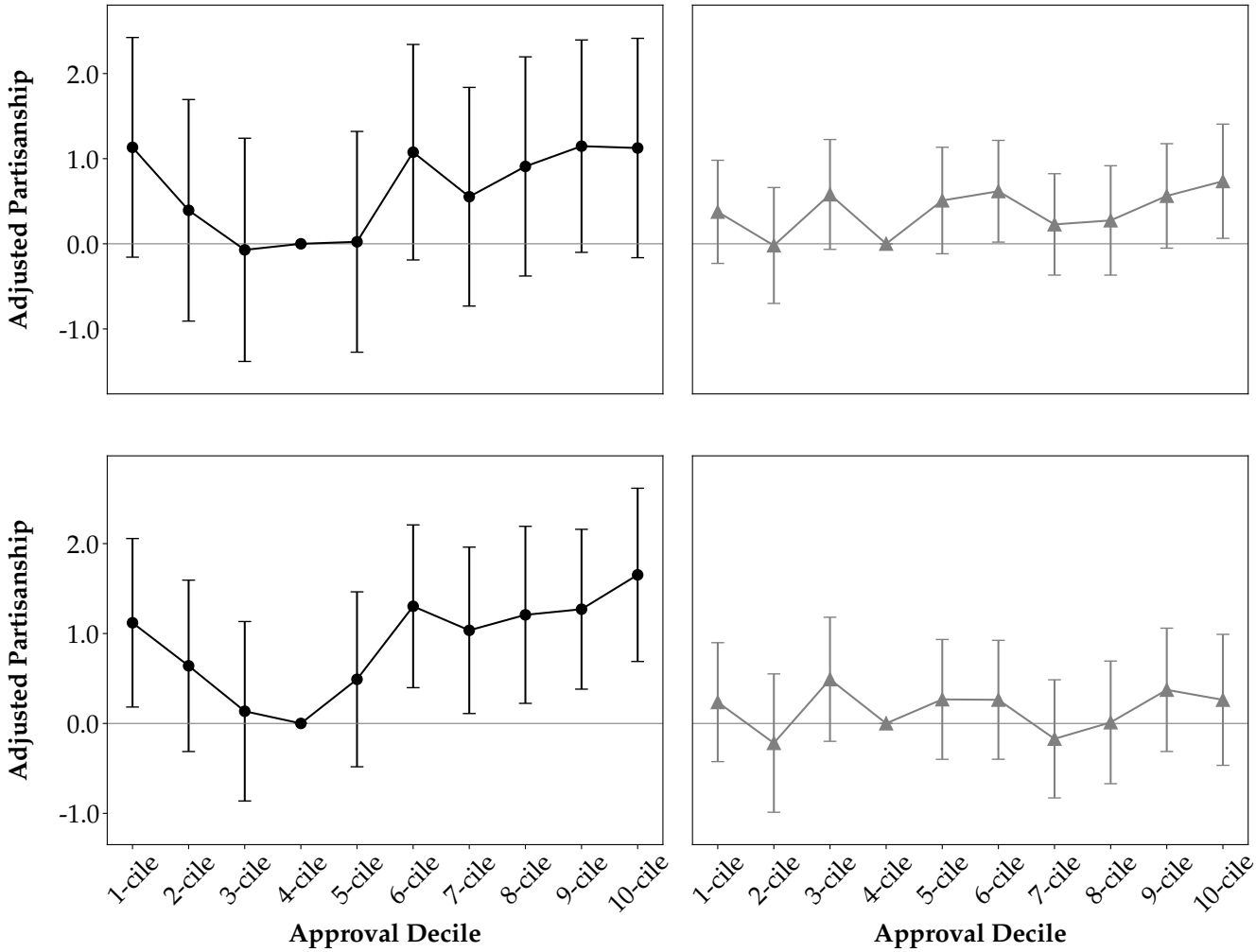
(c) Lame Ducks, Unaligned Legislatures (d) Lame Ducks, Aligned Legislatures

Dependent variable measures partisanship of U.S. governor State of the State speeches using leave-out estimator of Gentzkow et al. (2019), as calculated in equation (2.3), normalized by sample mean and standard deviation. Horizontal axis plots approval decile of gubernatorial approval by state, calculated using previous year's fourth quarter approval from Singer (2023) after first year in office, with first quarter approval utilized for first year in office. Panels (a) and (b) plot levels of partisanship for reelectable governors. Panel (a) plots $\alpha_k + \beta_k$ coefficients from equation (2.5), measuring level of partisanship for reelectable governors with unaligned legislatures in decile k relative to fourth decile. Panel (b) plots $\alpha_k + \beta_k + \gamma_k + \delta_k$ coefficients from equation (2.5), measuring level of partisanship for reelectable governors with unaligned legislatures in decile k relative to fourth decile. Panels (c) and (d) plot levels of partisanship for lame duck governors. Panel (c) plots α_k coefficients from equation (2.5), measuring level of partisanship for lame ducks with unaligned legislatures in decile k relative to fourth decile. Panel (d) plots $\alpha_k + \gamma_k$ coefficients from equation (2.5), measuring level of partisanship for lame ducks with unaligned legislatures in decile k relative to fourth decile.

Figure 2.8.4: Partisanship of Reelectable Governors by Approval Decile, Adjusting for Lame Duck Behavior, Varying Definitions of Legislative Alignment, 1990-2020

(a) < 45% Legislators of Gov. Party

(b) ≥ 45% Legislators of Gov. Party

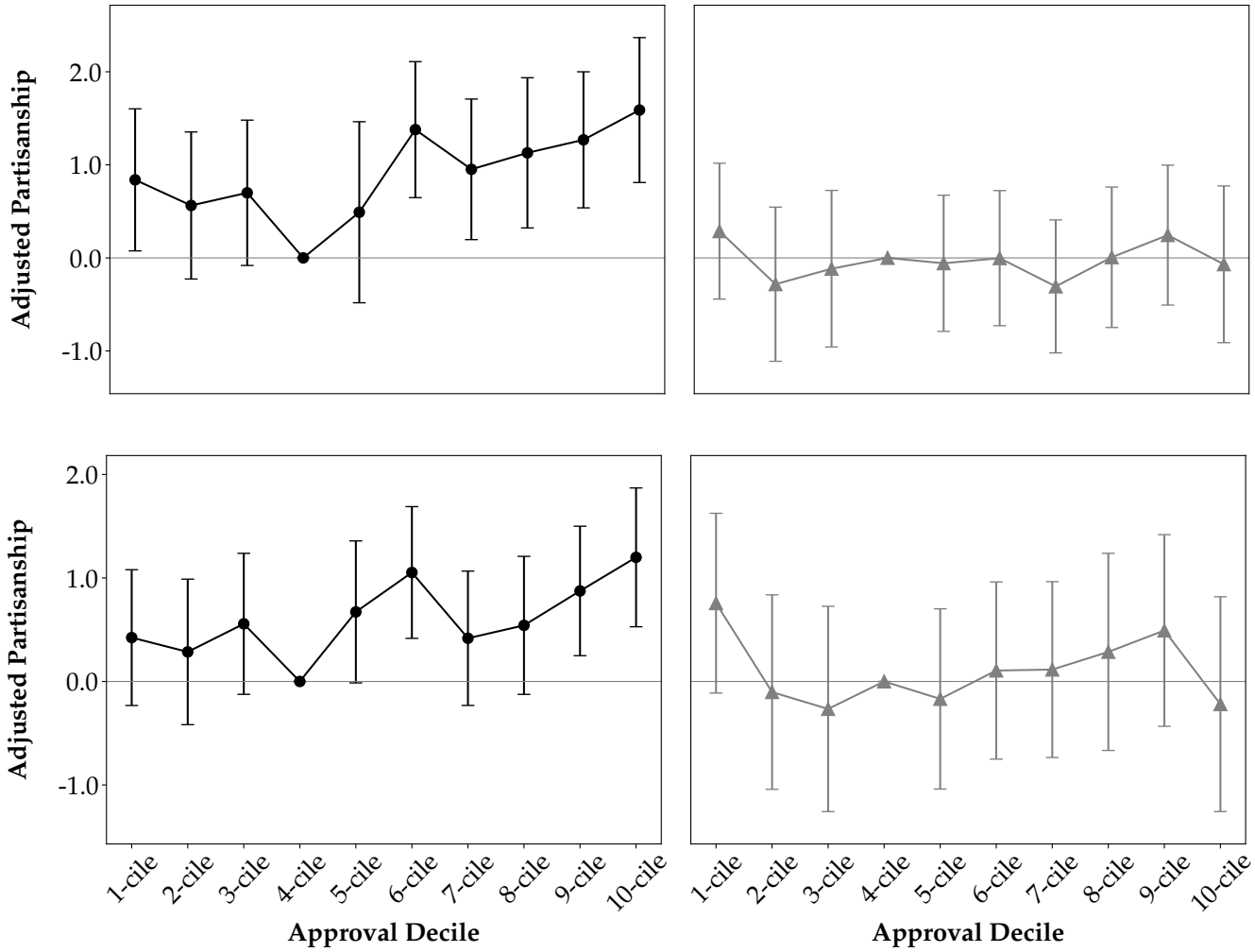


(c) < 50% Legislators of Gov. Party

(d) ≥ 50% Legislators of Gov. Party

(e) < 55% Legislators of Gov. Party

(f) ≥ 55% Legislators of Gov. Party



(g) < 60% Legislators of Gov. Party

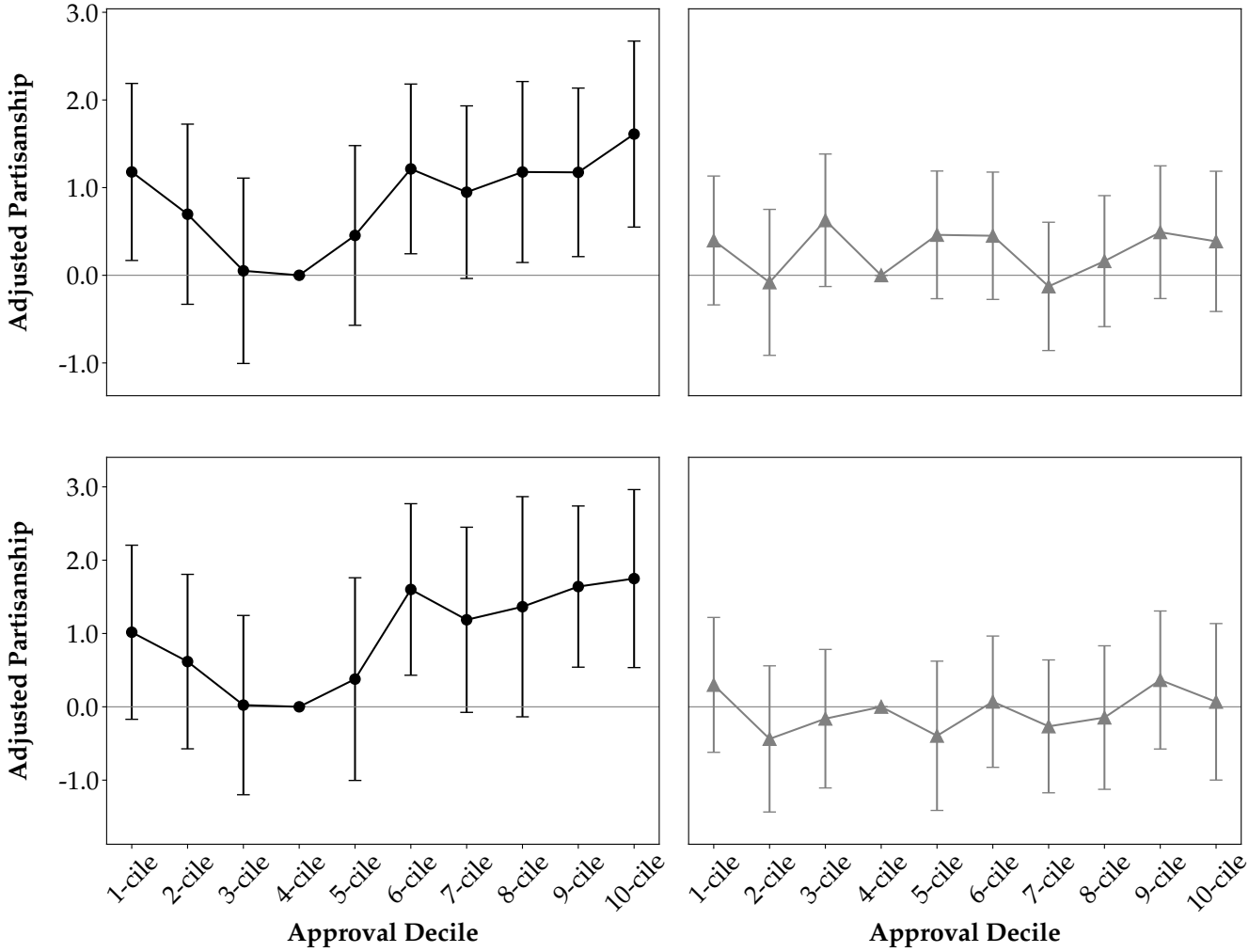
(h) ≥ 60% Legislators of Gov. Party

Dependent variable measures partisanship of U.S. governor State of the State speeches using leave-out estimator of Gentzkow et al. (2019), as calculated in equation (2.3), normalized by sample mean and standard deviation. Horizontal axis plots approval decile of gubernatorial approval by state, calculated using previous year's fourth quarter approval from Singer (2023) after first year in office, with first quarter approval utilized for first year in office. Black circles plots β_k coefficients from equation (2.5), measuring level of partisanship for reelectable governors with unaligned legislatures in decile k, adjusting for behavior of lame ducks, relative to fourth decile. Gray triangles plot $\beta_k + \delta_k$, measuring level of partisanship for reelectable governors with aligned legislatures in decile k, adjusting for behavior of lame ducks, relative to fourth decile. Bands around coefficient estimates display 95% confidence intervals. Legislative alignment defined as whether more than x% of legislators match governor's party. In panels (a) and (b), x = 45%; in (c) and (d), x = 50%; in (e) and (f), x = 55%, in (g) and (h), x = 60%. For panels (a) and (b), N: 1108; num. states: 48; within R^2: 0.06. For panels (c) and (d), N: 1108; num. states: 48; within R^2: 0.06. For panels (e) and (f), N: 1108; num. states: 48; within R^2: 0.07. For panels (e) and (f), N: 1108; num. states: 48; within R^2: 0.07.

Figure 2.8.5: Partisanship of Reelectable Governors by Approval Decile, Adjusting for Lame Duck Behavior, Variation with Electoral Cycle, 1990-2020

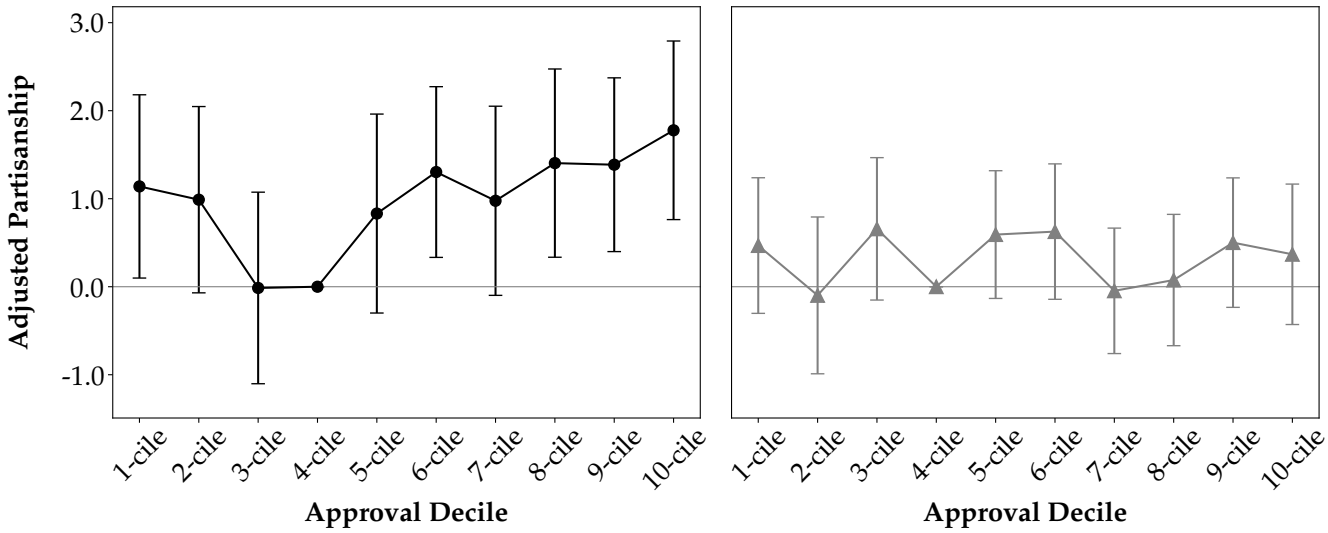
(a) Non-Election Years, Unaligned Leg.

(b) Non-Election Years, Aligned Leg.



(c) Election Years Only, Unaligned Leg.

(d) Election Years Only, Aligned Leg.



(e) No Lame Duck Last Year, Unaligned Leg. (f) No Lame Duck Last Year, Aligned Leg.

Dependent variable measures partisanship of U.S. governor State of the State speeches using leave-out estimator of Gentzkow et al. (2019), as calculated in equation (2.3), normalized by sample mean and standard deviation. Horizontal axis plots approval decile of gubernatorial approval by state, calculated using previous year's fourth quarter approval from Singer (2023) after first year in office, with first quarter approval utilized for first year in office. Black circles plots β_k coefficients from equation (2.5), measuring level of partisanship for reelectable governors with unaligned legislatures in decile k, adjusting for behavior of lame ducks, relative to fourth decile. Gray triangles plot $\beta_k + \delta_k$, measuring level of partisanship for reelectable governors with aligned legislatures in decile k, adjusting for behavior of lame ducks, relative to fourth decile. Bands around coefficient estimates display 95% confidence intervals. Legislative alignment defined as whether more than half of state legislators. Panels (a) and (b) drop datapoints where reelectable governors face gubernatorial elections later that year. Panels (c) and (d) drops all datapoints for reelectable governors except governors facing gubernatorial elections later that year. Panels (e) and (f) drop datapoints for lame ducks in their last full year in office. For panels (a) and (b), N: 910. Num. States: 48. Within R^2: 0.07. For panels (c) and (d), N: 517; num. states: 46; within R^2: 0.11. For panels (e) and (f), N: 1028; num. states: 48; within R^2: 0.06.

Table 2.8.2: Most Republican and Most Democratic Phrases, 1990-2020

(a) 1990-1994

Republican	Democratic
properti tax	state govern
gener assembl	long term
sale tax	commun colleg
incom tax	health insur
amend section	human resourc
public school	welfar recipi
act appropri	clean air
tax relief	child care
school district	econom develop
tax reduct	state employe

(b) 1995-1999

Republican	Democratic
properti tax	class size
school district	high school
tax relief	child care
million dollar	public safeti
cut tax	perman fund
charter school	tax credit
budget recommend	transport system
econom develop	privat sector
look forward	year thi
thi budget	econom growth

(c) 2000-2004

Republican	Democratic
charter school	health care
incom tax	prescript drug
high tech	state agenc
mental health	properti tax
center excel	minimum wage
tax cut	domest violenc
million state	school construct
long term	billion dollar
tax relief	health insur
low incom	educ lotteri

(d) 2005-2009

Republican	Democratic
incom tax	health care
tax relief	thi budget
charter school	health insur
million dollar	new job
math scienc	feder govern
tax rate	clean energi
gener fund	energi effici
qualiti life	pre k
properti tax	creat job
budget provid	afford health

(e) 2010-2014

Republican	Democratic
incom tax	health care
econom develop	sale tax
charter school	tax credit
state govern	creat job
budget recommend	minimum wage
school district	earli childhood
feder govern	mental health
job creator	21st centuri
high school	gener assembl
gener fund	thi budget

(f) 2015-2020

Republican	Democratic
incom tax	clean energi
tax relief	afford hous
budget recommend	health care
high school	minimum wage
law enforc	middl class
tax cut	renew energi
properti tax	climat chang
task forc	child care
pay rais	let pass
depart correct	work togeth

Table lists top 10 most Democratic and top 10 most Republican bigrams, in descending order, for each time period. Partisanship of phrase measured using equation (2.4), which describes informational loss in inferring governor's party based upon removal of bigram, as in Gentzkow et al. (2019) Table 1. Table removes certain procedural phrases in calculation of partisanship of phrase use.

Table 2.8.3: Partisanship by Approval Decile, Legislative Alignment, and Reelection Eligibility, 1990-2020

		Reelect ×	Leg. Align ×	Reelect × Leg. Align ×
Appr. 1-cile	-0.550	1.120**	0.443	-0.439
	(0.417)	(0.478)	(0.497)	(0.310)
Appr. 2-cile	-0.333	0.641	0.445	-0.414
	(0.424)	(0.486)	(0.545)	(0.309)
Appr. 3-cile	-0.095	0.136	-0.209	0.147
	(0.459)	(0.509)	(0.543)	(0.294)
Appr. 5-cile	-0.149	0.491	0.032	-0.191
	(0.439)	(0.496)	(0.516)	(0.309)
Appr. 6-cile	-0.953**	1.303***	0.975**	-0.065
	(0.408)	(0.461)	(0.496)	(0.291)
Appr. 7-cile	-0.500	1.036**	0.686	-0.521*
	(0.412)	(0.472)	(0.495)	(0.310)
Appr. 8-cile	-0.827*	1.208**	1.176**	-0.020
	(0.452)	(0.502)	(0.538)	(0.305)
Appr. 9-cile	-0.704*	1.270***	0.512	-0.385
	(0.395)	(0.453)	(0.490)	(0.309)
Appr. 10-cile	-1.000**	1.653***	0.867	-0.523*
	(0.418)	(0.491)	(0.513)	(0.345)
Constant	0.325	-0.832**	-0.216	0.367*
	(0.218)	(0.378)	(0.391)	(0.218)
N		1108		
Num. States		48		
Year F.E.		Yes		
State F.E.		Yes		
R^2 Within		0.06		

Dependent variable measures partisanship of U.S. governor State of the State speeches using leave-out estimator of Gentzkow et al. (2019), as calculated in equation (2.3), normalized by sample mean and standard deviation. Approval decile represents decile of approval rating by state, calculated using previous year's fourth quarter approval from Singer (2023) after first year in office, with first quarter approval utilized for first year in office. First column of coefficients measures baseline levels of partisanship by approval decile for lame duck governors, i.e. α_k terms in equation (2.5), with constant corresponding to omitted category of fourth decile. Names of second through fourth columns represent variables being interacted with approval decile. Second column of coefficients measures baseline partisanship of reelectable governors relative to lame ducks by approval decile, i.e. β_k coefficients. Third column measures level of partisanship for legislatively aligned lame duck governors relative to unaligned, i.e. γ_k coefficients. Fourth column measures level of partisanship for reelectable governors with aligned legislatures relative to unaligned legislatures, i.e. δ_k coefficients. Third column compares partisanship of reelectable governors in gubernatorial election years to all lame duck governors. Fourth column compares partisanship of reelectable governors to all lame duck governors except those in their last year in office. Data for Idaho (no approval data) and Nebraska (does not recognize political parties in legislature) omitted. Standard errors in parentheses. $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Table 2.8.4: Swing States, Republican States, and Democratic States

Swing States	**Republican States**	**Democratic States**
AK	AL	CO
AR	AZ	DE
CA	FL	HI
CT	IA	KY
GA	ID	MD
IL	MA	MO
IN	MI	NC
KS	MS	NY
LA	ND	OR
ME	NE	PA
MN	NM	VA
MT	NV	VT
NH	OH	WA
NJ	SC	WV
OK	SD	
RI	TX	
TN	UT	
WY	WI	

2.9 Hand-Coding Guide

Below is the guide provided to research assistants for the hand-coding task.

General remark. Remember that the goal of this task is to identify (past or current) policy proposals among gubernatorial speech snippets — as well as risky policy proposals — that will be used to train a large-language-model. As a guiding principle when coding, it may be useful to ask yourself: is the language in this snippet relevant to identifying policy proposals, past/current proposals, or risky policy proposals?

For example, if a governor spends a lot of time in a snippet on rhetoric, but then at the very end mentions a policy she passed, we wouldn't code that as "yes, this is about as policy" even though a policy may be mentioned by name at the end. This is because the language of that snippet, by and large, does not talk about past policy proposals.

If a governor is clearly reflecting on the content of a policy proposal — but the policy is not mentioned by name — this would also be coded as "referring to a policy proposal." The reason is that we are trying to figure out how much time governors spend in their speeches discussing policy proposals (as opposed to other things). So the relevancy of snippets to this category — or any of the other categories — should be assessed using these sorts of heuristics.

Coding Guidelines. The outline below details each of the main categories to be coded, as well as examples ("easy" and "hard") of each of the codings.

1. "Policy." Coded as "1" if the snippet discusses the enactment of a state-level policy (either passed by the governor, state government, or referendum) and "0" if it does not. A policy discussion is a reference to a specific act of legislation or law, a concrete proposal to increase or decrease funding to a certain cause, other legal orders proposed by the government to take certain concrete actions, and discussions of details of any of the above.

    - Example of 0 (easy): "I will continue to speak out against those who promote prejudice. I know you will too. And I will tell you this: a handful of people who may want to

burn a cross are no match for 10,000 Idahoans who marched to support the Table Rock Cross." (ID, 2000).

- Example of 0 (hard): "It is simply not pono for our families to be living in cars, people to be sleeping in the doorways of businesses downtown or on picnic tables in our parks. There is no one silver bullet to solve the problems of homelessness and affordable housing, but there are many good ideas that can and should be enacted." (HI, 2006). Discusses an issue and hints at the concept of a solution, but does not concretely address a policy.
- Example of 0 (hard): "The budget is balanced but great risks and uncertainties lie ahead. The federal government, the courts or changes in the economy all could cost us billions and drive a hole in the budget. The ultimate costs of expanding our health care system under the Affordable Care Act are unknown. Ignoring such known unknowns would be folly, just as it would be to not pay down our wall of debt. That is how we plunged into a decade of deficits." (CA, 2013). Does not actually discuss a governor or state-led policy initiative, despite referring to the ACA (a federal initiative).
- Example of 1 (easy): "The Reform Albany Act will have as its centerpiece an independent ethics commission that will have jurisdiction over State government. This commission will have the power to enforce campaign finance and end pay-to-play and bring jurisdiction and oversight to so-called good government groups, who hide their donors behind walls of sanctimony." (NY, 2010)
- Example of 1 (hard): "We were asked to meet yet another list of requirements. The federal government objects not on a scientific basis, but upon a vaguely defined legal risk analysis. This is not just about semantics. It is about achieving wolf delisting on rational terms that work for Wyoming. I do not care what we call them as long as we can manage them. The new demands from the federal government go far beyond the word predator and include changing how a pack is defined and even questioning whether the national parks will assume responsibility for half of the 15 packs it plans for Wyoming." (WY, 2004). Refers to federal policy/definitions, which is confusing, but also details of how the state will implement such a policy out in the context of their

own state.

2. "Proposal/Past." Only applies if "policy" coded as "1." Coded as "1" if the snippet refers to a policy that has just put into place or will be put into place in the future. Coded as "0" otherwise — in particular, if a governor is reflecting on the effects of a policy in the past. Coded as "0.5" if it contains substantive elements of both. Continuing a preexisting policy implemented from the past without any substantive changes also does not constitute a (future) policy proposal — this would be coded as a "0."

    - Example of 0 (easy): "Clearly we are doing things right. We are making progress. But the job numbers are only part of the story. In addition to making it easier for businesses to create jobs, we have also invested in public works projects. In doing so, we improved our public infrastructure, made it more attractive for businesses to relocate or stay here, and directly created even more jobs." (OR, 2006). Clearly refers to a past policy action but not a future action.
    - Example of 0 (hard): "...We have budgeted more than 260 million for higher ed capital. That funds new science facilities at Jackson State Community College and the University of Tennessee. It also includes nearly 25 million for improvements to our colleges of applied technology all across the state, and it includes the funds to complete the long awaited fine arts building at East Tennessee State University. The reason we continue to make these investments in education is we want Tennesseans to have the education, training and skills necessary to have a good paying, high-quality job..." (TN, 2015). Suggests that we "have budgeted" (not will budget) money, and that this process will continue, but does not propose something inherently novel.
    - Example of 1 (easy): "By reducing our dependence on foreign energy sources we can not only stop sending our energy dollars to unstable parts of the world, but we can become a world leader in clean energy technologies, from wind and solar power to geothermal and fuel cells...." (NY, 2004). References clear future policy action (reducing dependence on foreign energy source.)
    - Example of 1 (hard): "We need a new, more dynamic, economic development strategy. One that can leverage the resources of our business sector, as well as higher education

and not for profits. The Delaware Economic Development Office needs to be at the forefront of moving Delaware into the 21st century economy. So my first act as governor was to find a way to energize our economic development efforts. We are going to do that by bringing private sector involvement into DEDO." (DE, 2017). Reflects on a past action, but lays out a policy for the coming year (private sector involvement).

- Example of 0.5: "Some said that Louisiana could not change its stripes and make a new start…but we did.And now…after our recent successes in ethics reform and tax reform…we must take the next step forward…an overhaul of our workforce development system." (LA, 2008). Second sentence reflects on past risky policies, and then talks about workforce development.

Chapter 3

Transgressions in Organizations*

3.1 Introduction

Organizations often uphold standards that prevent misconduct and abuse of employees at the hands of authority figures like managers. "Transgressions" of workplace norms may include verbal or sexual harassment, forced overtime work, or abuses of power that surpass what employees or the organization find acceptable. Accountability mechanisms for managers who commit transgressions often operate as follows: employees file a report, triggering an investigation into the incident. If a manager is found guilty, employees receive a payout for their grievance, while managers are punished — through suspension, termination of employment, or legal prosecution. However, while many organizations have rules against committing transgressions, managers may not always have a perfect understanding of how egregious their actions must be to constitute a transgression. For example, while firms may have policies banning sexual harassment, managers may not understand exactly what sorts of comments employees consider sexual harassment. Or, while some level of overtime work may be necessary to achieve a firm's objectives, managers may not understand where employees draw the line and consider work excessive.

This paper studies an organizational environment where "managers" have the capacity to commit potentially harmful actions against "employees" but, crucially, are uncertain about which of these actions constitutes a "transgression." In the event that an action crosses a (private) threshold for employees, the action is considered a *transgression* and is thus reportable to and punishable

*This chapter is coauthored with Kim Sarnoff.

by the organization. We show that policies that discourage managers from committing transgressions may in fact backfire and make employees worse off, depending on the severity of harm that transgressions generate.

We focus on two sets of mechanisms that generate these welfare effects. First, discouraging managers from committing transgressions leads them to not interact with employees in the first place. Interacting with a manager and producing some match value at the risk of experiencing a transgression may be preferred by an employee to no interaction at all *if* the relative harm risked by a transgression is small. Second, managers may moderate the severity of their actions to avoid being punished. On the one hand, these actions are less likely to constitute transgressions, making employees better off. However, in the event they are harmful, it is harder for employees to prove that these actions are transgressions, disincentivizing reporting and making them worse off. When the size of harm is relatively small, the latter channel outweighs the former, and the effect on employee welfare is negative. We show that these dueling welfare effects run through changes to the value of a employee-manager match, the cost of reporting, and the degree to which managers are punished for deviant behavior. In particular, we show that it may not be optimal for employers to punish managers as much as possible, and that optimal punishment may be interior or at the lower bound for possible punishments.

We view the dependence of these welfare effects on the severity of harm as generating insights for different forms of organizational transgressions. "Microaggressions" — subtle but uncomfortable verbal abuse — may be seen as less harmful, and may be particularly privy to the backfiring our model speaks to. More egregious actions that result in physical safety risks for employees, the generation of dangerous precedents for manager behavior, or large and persistent effects on future career prospects may be those that should discourage abuse at all costs.

We finally consider a dynamic extension, where managers and the organization learn about what employees consider transgressions through the verification and adjudication of reports. We show that when only information about verified reports is available publicly, the model converges to a steady state equilibrium where no further organizational learning occurs and transgressions, if they do occur, are never reported or punished. Moreover, we show that employees' expected utility may decrease as learning occurs and we converge towards the steady state.

Model Overview Our model considers a set of managers and employees in an infinite time-horizon environment. Each period, a manager and employee pair are born and active for only that period. In that time, managers can choose whether to interact with an employee — generating a match value V — and commit an action a_t. No interaction yields an outside option for both the employee and manager. Employees are homogeneous, and are characterized by a common threshold a^*. If a manager's action a_t surpasses this threshold, it is considered a *transgression*, causing employees to incur a disutility h. Managers and the organization do not know the value of a^*, but know that it is below some upper bound A_t.[1] Managers are heterogenous, and are characterized by a bliss point or "type" b, characterizing their affinity for committing transgressions. Managers would like to take actions a_t closer to their bliss point.

Upon experiencing a transgression, employees choose whether to report the action at cost c, triggering an investigation that results in the collection of evidence. Under the interpretation of verifiable evidence, truthful reporting is incentive-compatible for employees. We assume that the probability that the investigation turns up evidence favoring an employee is decreasing in the distance between a_t and A_t, reflecting the idea that A_t reflects an organization's general understanding of what constitutes a transgression, and departures from that understanding may require interpretations or collection of novel evidence. Finally, if an investigation verifies a transgression, managers are punished with a disutility $\gamma \geq V$, while employees receive a payout that compensates for the reporting cost.

We first show that employees report a transgressive action if and only if it exceeds a reporting threshold $\overline{r}_t$, which depends on the organization's understanding of transgressions A_t. Then, we show that manager behavior follows one of two equilibrium cases. In both, managers with blisspoints below the reporting threshold $\overline{r}_t$ simply play their preferred action and remain unpunished. Managers with blisspoints slightly above $\overline{r}_t$ play actions right at the reporting threshold $\overline{r}_t$, keeping employees indifferent to reporting. From here, behavior bifurcates. In one equilibrium, managers with high blisspoints opt out of interaction entirely. In the second, managers with moderately high blisspoints play "interior actions" that are slightly above the reporting threshold which are reportable as transgressions with positive probability, trading off the value of playing

[1] Formally, we assume that at time $t = 0$, managers' beliefs over $a^* \sim \mathcal{U}[0, A_0]$. If managers observe that some action a' is a transgression, by Bayes' Rule, their posterior over a^* is $\mathcal{U}[0, a']$.

their blisspoints with the risk of potential punishment. We denote these "interior actions" $a_t^\dagger$ Finally, managers in the second equilibrium with very high blisspoints opt out of interaction.

We then study how comparative statics of the model affect employee welfare. We broadly highlight three mechanisms that influence employees' expected utility, depending on the severity of harm h that employees face upon experiencing a transgression. The first mechanism relates to switches between participation and no participation. If h is relatively small, employees may prefer interacting with a manager — receiving a match value V and risking harm h — to the outside option of no interaction. We show that this mechanism is in effect when the match value of interaction V increases; if h is relatively small, increasing V encourages manager participation from a larger range of types b. If h is large, the effects on welfare are potentially ambiguous, since the increase in V for employees is counterbalanced by the harmful participation of managers with high values of b.

The second mechanism considers switches from interior actions $a_t^\dagger$ to the reporting threshold $\overline{r}_t$. If a type b manager switches from playing an interior action to the (lower) reporting threshold $\overline{r}_t$, this generates two effects. On the one hand, $\overline{r}_t$ is ex-ante less likely to constitute a transgression. On the other hand, conditional on being a transgression, employees are strictly worse off; they go from experiencing harm and having a strict incentive to report it (and receive a payout) to being indifferent to reporting and not reporting the transgression. If the size of harm h is small, the latter effect dominates and employees are worse off; otherwise, employees are better off.

We show that an increase in the reporting cost c captures the effects of both the first and second mechanisms. Increasing the cost of reporting first encourages the participation of a broader range of managers; the utility from matching with these managers is higher if h is small. However, even if h is large, interacting with certain types of managers could improve employee utility. An increase in c encourages some managers to switch from interior actions $a_t^\dagger$ to the reporting threshold $\overline{r}_t$, which improves expected utility precisely if h is large. This creates a tension between the negative effect of the first mechanism and positive effect of the second. These behavioral effects operate in tandem with the negative, mechanical effects of costlier reporting.

Our third mechanism is related to the second, and considers decreases in the intensity of interior actions $a_t^\dagger$. If a manager goes from playing one interior action $a_t^\dagger$ to a less intense interior action $a_t^{\dagger\prime}$ that is *still* above the reporting threshold, employees again deal with the two effects

above. If h is small, this switch leaves employees worse off, and if h is large, they are better off.

We use this mechanism, in tandem with the first two, to analyze how increasing the magnitude of manager punishment γ affects employee welfare. This change has three effects. First, certain managers switch from playing an interior action $a_t^\dagger$ to the reporting threshold, in line with the second mechanism. Second, managers who play $a_t^\dagger$ moderate their actions in line with the third mechanism. Third, managers with high blisspoints b opt out of interaction altogether. These effects are welfare-decreasing for employees if h is small, welfare-improving for h large, and ambiguous if h is intermediate. We use the combination of our three mechanisms to explore the optimal punishment γ for transgressive managers, i.e. that which maximizes employees' expected utility. When the size of harm h is small, we show that the optimal value of γ is at its lower bound, when it is large, the organization should make γ as large as possible. When h is intermediate, an intermediate level of punishment may maximize employees' expected utility.

Finally, we analyze the dynamics of learning on welfare by exploring how decreases in A_t — the organization's understanding of transgressions — affect employee expected utility. We operationalize learning by assuming that the organization keeps a public record of all past *successfully verified* transgressions Thus, when an action a_t is committed and verified as a transgression, the organization and managers know with certainty that the transgression threshold a^* is less than a_t. The upper bound on organizational beliefs about what constitutes a transgression then moves to $A_{t+1} = a_t$. We show that the model always converges to a steady state, in teh sense that the value of A_t remains stationary for the rest of time. This implies that there is no equilibrium learning about the value of a^* upon convergence to the steady state. We then use the tessellation of our three mechanisms to show how learning may negatively affect manager participation and the expected impact of participating managers' actions. We characterize conditions under which the model may move to a steady state where employee welfare is worse off than prior to convergence.

The remainder of this section reviews literature relevant to our setting. We then lay out the basic structure of the model — as well as its applications to other organizational, political, or behavioral settings — and characterize its static equilibrium. We then analyze how changes in model parameters affect employees' expected utility. We finish by looking at the dynamics of learning and reporting on welfare.

Literature Our primary motivation for studying "transgressions" in organizations is driven by a growing interest in workplace harassment and abuse — particularly sexual harassment — in both the mainstream press and scholarly literature. Harassment may include persistent subjugation of employees to unsavory behavior that impacts their emotional wellbeing or productivity, abuse of authority and privilege, or broader violations of workplace *norms* of respect and dignity (Aquino and Thau, 2009). Sexual harassment can have negative effects on victims' willingness to be hired at or remain in firms (Adams-Prassl, Huttunen, Nix, and Zhang, 2024) and pursuit of leadership positions (Folke, Rickne, Tanaka, and Tateishi, 2020). Harassment increases abseentism, productivity, job satisfaction, and labor turnover (Hersch, 2015). Studies such as (Hersch, 2018) have calculated statistical values of sexual harassment for women, arguing that its estimates are above the maximal payouts available for victims under American federal law. This research is complemented by an organizational literature that lays out the startling frequency of harassment across a wide array of firms, and how both toxic organizational climates and vertical relationships where harassers have managerial power over victims exacerbate the potential for abuse. Cortina and Areguin (2021) reviews this literature and argues that formal complaint processes often make matters worse for victims, whether by failing to end harassment, triggering retaliation, or putting victims through further psychological stress. To this end, our paper aims to provide a deeper understanding of how policy changes that may be traditionally thought to improve prospective victims' welfare — more effective reporting, lower reporting costs, or increased punishment for managers — may backfire and hurt them.

These empirical insights have paved the way for a theoretical literature on incentives for committing and reporting harassment (as well as other behaviors like corruption) in organizational settings. Our model is most closely related to those of Lee and Suen (2020) and Cheng and Hsiaw (2022), where employees report private experiences with managers who are heterogenous in their propensity for harassment. The latter's model — like ours — argues that disincentivizing harassment may make employees worse off by discouraging manager participation or "mentoring." However, both of these papers focus on the role of coordination problems and verifiability of information in shaping victim reporting in settings where serial harassers are more likely to be punished if more victims come forward. By contrast, our model assumes away these reporting frictions to focus on how managers' *uncertainty* about what actions actually constitute transgres-

sions affect their willingness to commit harm and, subsequently, their effects on victim welfare. To this end, our paper differs from an emerging literature on how informational frictions — such as coordination problems, false reports, or unverifiable evidence — affect incentives for reporting bad behavior (Bac, 2018; Chassang and Miquel, 2019; Siggelkow, Trockel, and Dieterle, 2018; Zhu, 2024). In particular, we show that when there is uncertainty over what employees consider harassment, traditional policy levers that discourage managers from committing transgressions can backfire by not only discouraging participation but also encouraging the pursuit of slightly less egregious actions that are harder for employees to report.

Our theoretical environment also connects to a long-standing literature on the economics of crime deterrence, first pioneered by Becker (1968) and reviewed by Chalfin and McCrary (2017). This literature often suggests that more stringent punishment reduces crime, which is balanced against the costs of law enforcement or costs that result from punishing innocents. However, viewing crime prevention through the lens of our model suggests that increased punishment may have more subtle effects — namely, by discouraging manager participation and encouraging the pursuit of actions that are not worth reporting. This means that a moderate level of punishment for crimes may be socially optimal due to its, even when law enforcement itself is costless and there is no risk of type-I error.

Finally, our model can also be thought of as a model of political accountability, where an executive who can abuse her authority is uncertain about the degree to which she will be held accountable by the judicial branch. To this end, our paper is related to formal models of the judiciary such as Beim, Hirsch, and Kastellec (2014) and Patty and Turner (2021), which study how the review of evidence, the threat of whistleblowing, and the degree of preference alignment affect the judiciary's efficacy in holding others accountable.

3.2 Setup

Agents and Actions Consider an organization composed of managers (M) and employees (E). Time is discrete beginning at $t = 1$. Each period, a new manager-employee pair (m, e) is born and is active only for that period. During that period, the manager m (she) chooses whether to interact with an employee e (he), where interaction generates a symmetric benefit $V > 0$ for both parties.

This could represent, for example, the value of a project that the manager and employee decide to carry out together. If m decides not to interact, both agents receive 0 and we move to the next period.

Managers are heterogenous; each m is characterized by a bliss point $b \sim F(b)$ with support on $[0, A_0]$. If m decides to interact, m takes some action $a \in [0, A_0]$ for $A_0 > 0$, where a represents m's behavior towards e. We denote by a_t the action of a manager m at time t. We refer to b as a manager's *type*. Upon taking action $a_t = a$, a type b manager receives a quadratic loss $(a - b)^2$. Employees, on the other hand, are homogeneous. Each employee e shares a common cutoff $a^* \in [0, A_0]$ such that $a > a^*$ imposes a utility cost of $h > 0$ to them, while $a \leq a^*$ induces no cost.[2] a^* represents the threshold for the manager's treatment of the employee to constitute overwork, harassment, or other harmful behavior. We refer to values of a that exceed a^* as *transgressions*. We provide further interpretations of managers, employees, and their actions at the end of the section.

The employee e can choose to report a transgression to the organization, triggering an investigation and possible punishment of the manager. Reporting incurs a cost $c > 0$ to e and either results in the report being *verified* or *unverified*. If the report is *verified*, m is found guilty of a transgression and is punished with a loss γ, while the employee receives a payout normalized to 1. We assume $c < 1$ so that the payout from a successful report is worth the cost. We also assume $V < \gamma$ so that no interaction is always better for a manager than an interaction that is punished with certainty.

Thus, the baseline utilities for each agent, conditional on interacting, are:

$$\begin{aligned} &\text{Manager } m: && V - (a-b)^2 - \gamma \mathbf{1}[m \text{ punished}] \\ &\text{Employee } e: && V - h\mathbf{1}[a \geq a^*] - c\mathbf{1}[m \text{ reported}] + \mathbf{1}[m \text{ punished}]. \end{aligned}$$

Updating and Reporting While neither managers nor the organization know a^*, they may learn about its value through reports and their adjudication. We assume that the organization keeps a record of all past, verified reports, which are public knowledge to both employees and managers. We assume that at $t = 1$, the prior over a^* is uniform on $[0, A_0]$.

At time t, let $R(t) \subseteq \{1, 2, \ldots, t\}$ denote the set of time periods t where a transgression was

[2] The insights of the model are identical if employees are heterogeneous in their thresholds.

reported and verified. Let $A_t = \min\{a_t : t \in R(t)\}$ be the minimum of all past verified reports. This means, by Bayes' rule, that at time t, the posterior belief over the value of a^* is given by $\mathcal{U}[0, A_t]$.

Given a reported action a_t and A_t, we denote the probability with which a transgression is verified as $p(a_t, A_t)$. We assume the functional form:

$$p(a_t, A_t) = \begin{cases} 1 & a_t > A_t \\ \frac{a_t}{A_t} & a_t \leq A_t \end{cases}$$

This assumption means that an employee's incentive to report depends on the current standard for what actions constitute a transgression. Specifically, for $a_t < A_t$, there is only a probability of verification, which decreases linearly in the action. We view this as according with the intuition that it is harder to find or interpret evidence when the reported action is further from what is already established as a transgression. Under this interpretation of verification, truth-telling on the part of employees is incentive-compatible. If one reports an action which is not a transgression, the probability of a report being verified is 0, since no evidence would turn up in support of an action that constitutes a transgression, meaning reporting simply entails a cost.

Comments and Interpretations First, note that while managers are heterogeneous in their affinity for committing transgressions, employees are homogeneous and share a common threshold a^*. Since managers are uncertain about the value of a^*, the key welfare insights of the model were employees to have different thresholds would be identical. However, in such a model, there would be no capacity for reports to generate information about employees' thresholds.

Second, we assume the linear functional form of the verification probability $p(\cdot, \cdot)$ for tractability purposes. What matters for the analysis is that the probability of verification decreases as managers commit actions farther from A_t. To this end, we may interpret A_t as reflecting an organization's relative understanding of what constitutes a transgression; actions farther from that standard lack precedent, and may require learning how to interpret evidence in new ways. This implicitly means that the probability of verification depends on a^* insofar as A_t reflects the organization's understanding of the value of a^*. Similarly, the uniform prior over employees' threshold a^* simplifies calculations of managers' expected utility that aid in our later welfare analysis.

Third, note that neither the cost of reporting nor the manager punishment/employee reward for a verified transgression depend on the magnitude of the manager's action a_t or the transgression threshold a^*. However, what matters for the analysis is, that conditional on reporting, a manager's (employees') expected payout is decreasing (increasing) in a_t and increasing (decreasing) in A_t , which is reflected in $p(\cdot, \cdot)$. Hence, allowing for these effects would deliver similar insights as the existing model. Similarly, note that neither the value of a match V nor the disutility of transgressions h not depend on a_t. If the value of a match were to depend on a_t and, in particular, were concave, the model's predictions would be similar, since managers would trade off the benefits of matching and interacting with the risks of punishment and value of non-interaction. Allowing the disutility of transgressions to vary with h would introduce a negative distortionary effect for employees' utility but would leave our basic equilibria unchanged, as well as our basic welfare predictions.

Additionally, notice that employees themselves face no incentive-compatibility constraints. However, as we will see, many of the inefficiencies that arise in our setting occur when the disutility from a transgression h is relatively small, i.e. precisely when an incentive-compatibility constraint would not bind. Thus, providing employees an outside option would not eliminate the central mechanisms driving our results.

Finally, we provide several interpretations of the model through the lens of organizations, political institutions, and consumer behavior. Our motivating interpretation is that transgressions can be seen as "harassment" in an organizational setting. Firms may have policies that protect employees from sexual or racial harassment at the hands of supervisors; however, supervisors may not have a good understanding of what constitutes "harassment" in the first place. Costs of reporting may represent the psychic burden of coming forward with an accusation. The verification process involves the interviewing of witnesses and other parties privy to an incident, and punishment for transgressive managers may range from suspension to termination of employment. Relatedly, the model can be used to analyze overtime work and burnout in firms. Managers can force employees to work overtime on projects, but excessive overtime work may cause employee burnout, which constitutes workplace abuse. Managers are uncertain about employees' tolerance for overtime.

The model can also describe a setting of political accountability. Consider a political exec-

utive (manager) who can use her authority as leader to achieve policy objectives — such as by issuing executive orders that override the oversight of the legislature. The executive faces uncertainty about the extent to which she will be held accountable for potential abuses of authority. Employees in this setting can represent political constituents or the judiciary, who can choose to investigate the executive's actions and put a stop to them, acknowledging that their ability to hold the executive accountable is a function of the egregiousness of the executive's actions relative to legal precedent, as well as their ideological tolerance for an executive's actions or an understanding that executive authority may be beneficial in times of crisis. While executives themselves understand precedents for behavior, they may face uncertainty over the degree to which departures from those precedents may be enforced.

Finally, the model can also describe brand affinity and consumers' tolerance for price increases. Consider a firm (manager) selling a product to a set of consumers (employees). The firm has the option to mark up the price of its product. While consumers may stomach small price hikes, they may have a distaste for excessive price hikes and boycott the company's product if markups exceed a threshold. The firm faces uncertainty over the degree to which they can hike prices, while consumers may face uncertainty about the value of competitors' products if they boycott the original firm.

3.3 Static Equilibrium

Since managers and employees are short-lived, we consider static equilibria of the game and the implications this generates for the organization. A pure strategy static equilibrium for an interaction between a manager and employee at time t is given by a profile $\{(i_t^*(b), \alpha_t(b)), r_t^*(a_t)\}$, where $i_t^*(b) \in \{I, NI\}$ is the decision of a bliss point b manager to interact/not interact, $\alpha_t(b)$ is the action taken by a type b manager, conditional on interacting, and $r_t^*(a_t) \in \{R, NR\}$ is the decision of the employee to report (R) or not report (NR), conditional on experiencing a_t. Because agents live only a single period and employees make choices after managers, managers' and employees' equilibrium strategies are identical across time.

Employee's Problem First, we write the employee's maximization problem, conditional on experiencing an action a_t:

$$U^E(a_t) = \max \begin{cases} V - c + \mathbf{1}[a_t \geq a^*](-h + \min\{\frac{a_t}{A_t}, 1\}) & R \\ V + \mathbf{1}[a_t \geq a^*](-h) & NR \end{cases}$$

The first line represents the employee's utility if he chooses to report at cost c. If the manager's action is a transgression, he incurs disutility h but a potential reward of 1 with probability $\frac{a_t}{A_t}$. Not reporting simply involves the match utility with a loss of h if $a_t \geq a^*$.

Lemma 1 shows that reporting follows a threshold rule: the employee has a strict incentive to report a transgression if and only if $a_t > \overline{r}_t \equiv cA_t$ and $a_t \geq a^*$.

Lemma 1. *Reporting follows a threshold rule: the employee has a strict incentive to report a transgression if and only if $a_t > \overline{r}_t \equiv cA_t$ and $a_t \geq a^*$.*

We refer to $\overline{r}_t$ as the *reporting threshold* at time t. Note that $\overline{r}_t$ is *not* a function of a^*, meaning it is known to managers.

Given that the employee reports if and only if $a_t > \overline{r}_t$, a type b manager's maximization problem, conditional on interacting, is:

$$U^M(b) = \max_{a_t} \begin{cases} V - (a_t - b)^2 & a_t \leq \overline{r}_t \\ V - (a_t - b)^2 - \gamma(\frac{a_t}{A_t})^2 & \overline{r}_t < a_t \leq A_t \\ V - (a_t - b)^2 - \gamma & a_t > A_t \end{cases}$$

Taking action a_t below the threshold $\overline{r}_t$ will never be reported, and so utility is simply a function of V and the disutility of deviating from the bliss point. Taking an action greater than A_t additionally entails the cost of sure punishment. For intermediate actions $a_t \in (\overline{r}_t, A_t]$, γ is weighted by the joint probability that the action is a transgression and is reported.

The following theorem characterizes employee and manager equilibrium behavior in the static game.

Theorem 1. *Equilibrium behavior in the static game at time t is characterized as follows. The employee*

reports an action a_t if and only if $a_t > \overline{r}_t = cA_t$. There exist thresholds $\underline{a}_t$, $\underline{i}_t$, and $\overline{a}_t$ such that a type b manager's behavior, in equilibrium, is characterized as follows:

- *If $b \in [0, \overline{r}_t]$, then the manager interacts and $\alpha_t(b) = b$.*
- *If $b \in (\overline{r}_t, \min\{\underline{i}_t, \underline{a}_t, A_0\}]$, the manager interacts and plays $\alpha_t(b) = \overline{r}_t$.*

From here, equilibrium behavior bifurcates into one of two cases.

1. ***Equililbrium 1**: if $\underline{i}_t < b \leq \min\{\underline{a}_t, A_0\}$, the manager does not interact for all $b \geq \overline{i}_t$. We will refer to this as E^{NL}.*
2. ***Equililbrium 2**: if $\underline{a}_t < \min\{\underline{i}_t, A_0\}$, the manager interacts and plays $\alpha_t(b) = a_t^\dagger(b) = \frac{bA_t^2}{A_t^2+\gamma} > \overline{r}_t$ for all $b \in (\underline{a}_t, \min\{\overline{a}_t, A_0\}]$. If $b > \overline{a}_t$, the manager never interacts. We will refer to this as E^L.*

Here, $\underline{a}_t$ is given by $cA_t + c\frac{\gamma+\sqrt{\gamma(A_t^2+\gamma)}}{A_t}$, $\underline{i}_t$ is given by $cA_t + \sqrt{V}$, and $\overline{a}_t$ is given by $\sqrt{(\frac{A_t^2}{\gamma}+1)V}$.

On the end of the employee, an action a_t is only worth reporting if it is worse than a^* and the probability of verification is sufficiently high. His decision to report is hence given by a simple threshold rule.

A manager's optimal behavior follows a richer cutoff structure. We begin by considering action choices, conditional on interacting. Those with type $b < \overline{r}_t$ simply play their bliss point, since there is no cost to doing so: the action will go unreported, and thus unpunished, with certainty. This delivers the first cutoff for manager actions: the reporting threshold itself.

A manager with $b \geq \overline{r}_t$ has two options. She can play the reporting threshold, which is the non-reportable action that minimizes deviation from her bliss point. Or, she can play some $a_t > \overline{r}_t$, which is an action that would be closer to her bliss point, but would entail a risk of punishment. We refer to the optimal action that takes into account the risk of punsihment as an "interior action", and we denote by $a_t^\dagger(b) = \frac{bA_t^2}{A_t^2+\gamma}$.

For managers with b sufficiently close to $\overline{r}_t$, avoiding punishment dominates the cost of an action further from the bliss point. Thus, there exists an interval of b's, with lower bound $\overline{r}_t$, who bunch at the reporting threshold. For those with b sufficiently far from $\overline{r}_t$, the cost of deviating to $\overline{r}_t$ dominates, so they play an interior action that internalizes the risk of punishment. This gives the second threshold for manager actions: the type that is indifferent between the reporting threshold and an interior action, denoted $\underline{a}_t$.

We then characterize a manager's choice to interact altogether. This decision depends on two additional thresholds: the type that is indifferent between participating and not participating (denoted $\underline{i}_t$), and the type that is indifferent between participating and playing an interior action (denoted $\overline{a}_t$). The ordering of $\underline{i}_t$ with respect to $\underline{a}_t$ and $\overline{a}_t$ plays an important role, as it determines whether learning can occur in equilibrium.

In particular, two kinds of static equilibria may realize in the model. We will refer to these as a non-learning equilibrium (E^{NL}) and a learning equilibrium (E^{L}). The non-learning equilibrium E^{NL} is illustrated in Figure 3.3.1(a). The key feature of this equilibrium is that the participation constraint binds for all managers who would play an interior action, i.e. $\underline{i}_t < \underline{a}_t$. So, managers either participate and play an action that is not incentive compatible to report (the red line), or they do not participate at all. Because manager behavior eliminates reporting, no learning can occur.

The learning equilibrium E^{L} is illustrated in Figure 3.3.1(b). In contrast to E^{NL}, the participation constraint does not bind for some interval of b who would play an interior action (the blue line), i.e. $\underline{i}_t \in (\underline{a}_t, \overline{a}_t)$. As a result, actions that are incentive compatible to report (if transgressions) are played with positive probability. This, in turn, makes learning possible.

Finally, we can establish conditions on parameters that characterize whether the equilibrium is E^{L} or E^{NL}.

Corollary 1. *The static equilibrium is E^{L} if and only if*

$$c\frac{\gamma + \sqrt{\gamma(A_t^2 + \gamma)}}{A_t} < \sqrt{V}$$

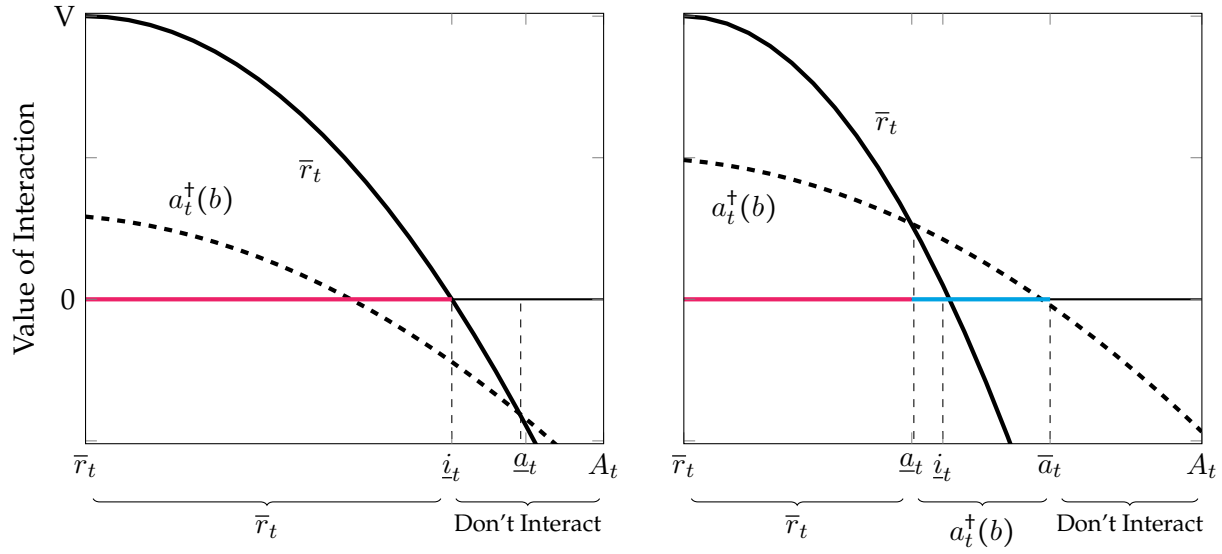
In particular, there exists $\overline{\gamma}$ such that if $\gamma \geq \overline{\gamma}$, the equilibrium is E^{NL}.

Intuitively, when γ is sufficiently large, the manager's expected payoff from an interior action is negative, and so she will not play one. In fact, she will not participate at all: since the type that is indifferent between the reporting threshold and an interior action receives negative payoff from both, so must all b more extreme.

Comments Note that the participation constraint binds for some $b \in (\overline{r}_t, A_0]$, or binds for no one.[3] If it binds for no one, either equilibrium is possible, depending on the values of V and γ. To highlight the main forces of the model, we will focus on the case where it binds for some $b \in (\overline{r}_t + \epsilon, A_0]$.

Additionally, note that equilibrium learning may *still* halt in E^L if $a^* > a_t^\dagger(\overline{a}_t)$. In this case, managers of type $b \in [\underline{a}_t, \overline{a}_t]$ commit actions above the reporting threshold, but the maximal action played is below the true cutoff for a transgression. So, none of these actions would ever be reported. Conditional on being in E^L, this case occurs with probability $1 - \frac{\overline{a}_t}{A_t}$.

Figure 3.3.1: Value of Manager Interactions: Equilibrium 1 (E^{NL}) vs. Equilbrium 2 (E^L)



(a) E^{NL}: $\underline{i}_t < \underline{a}_t$

(b) E^L: $\underline{a}_t < \underline{i}_t < \overline{a}_t$

Threshold Comparative Statics The following proposition summarizes comparative statics of the main thresholds and actions.

Proposition 1. *Comparative statics of the main equilibrium actions for the employee and manager are as follows.*

- *An increase in V generates an increase in $\underline{i}_t$ and $\overline{a}_t$.*

[3] It binds for $b \in (\overline{r}_t + \epsilon, A_0]$, where $\epsilon > 0$, when either $\underline{a}_t < \underline{i}_t$ and $\overline{a}_t < A_0$ or $\underline{i}_t < A_0, \underline{a}_t$. Note that when $V > 0$, there will always exist some interval of b (with lower bound $\overline{r}_t$) who receive positive utility from playing the reporting threshold. This ensures that the participation constraint can start to bind only for some $b > \overline{r}_t$.

- *An increase in c generates an increase in $\overline{r}_t$, $\underline{a}_t$, and $\underline{i}_t$.*
- *An increase in γ generates an increase in $\underline{a}_t$ and a decrease in $\overline{a}_t$. Moreover, for each b, $a_t^\dagger(b) = \frac{bA_t^2}{A_t^2+\gamma}$ decreases.*
- *A* decrease *in A_t generates a decrease in $\overline{r}_t$, $\underline{i}_t$, and $\overline{a}_t$. For each b, $a_t^\dagger(b) = \frac{bA_t^2}{A_t^2+\gamma}$ decreases. There exists, $\tilde{A} > 0$ such that the sign of $\underline{a}_t$ is negative for $A_t \leq \tilde{A}$ and is positive above.*

Notably, by shifting the ordering of thresholds, changes in manager behavior due to changes in model parameters can change the equilibrium that holds, i.e. induce a switch from E^{NL} to E^L or E^L to E^{NL}.

3.4 Welfare

With comparative statics of the major thresholds characterizing behavior in hand, we now move to the welfare effects of policy changes. Our analysis will consider welfare as calculated from the perspective of the organization. The following corollary follows from Theorem 1, and characterizes the effects of parameter changes on managers' welfare.

Corollary 2. *Suppose V increases, c increases, γ decreases, or A_t increases. Then, managers' expected utility increases.*

Whilethese parameter changes have straightforward effects on manager welfare, the effects on employees' welfare are more nuanced. Increasing in the value of an interaction V or punishment γ may actually reduce employees utility in expectation, while increases in the reporting cost c may improve their utility conditional on interacting with certain managers.

We will say that employees' welfare decreases if, for every distribution $F(b)$ with support on $[0, A_0]$, employees' expected utility decreases, and if there exists some F such that employees' expected utility *strictly* decreases. This is equivalent to saying that for each $b \in [0, A_0]$, employees' expected utility from interacting with that type either stays the same or *strictly* decreases. We define an increase in welfare analogously. If welfare neither increases nor decreases in the aforementioned sense, we will say that the effect is *ambiguous*. In particular, this means that there exist b and $b' \in [0, A_0]$ such that interacting with a type b manager increases employee utility, while

interacting with a type b' decreases utility. Hence, from the perspective of the organization, the overall effect on *expected* utility depends on the distribution $F(b)$.[4]

First, we write out the indirect utilities of employees in each of the two equilibria, conditional on experiencing a transgression.

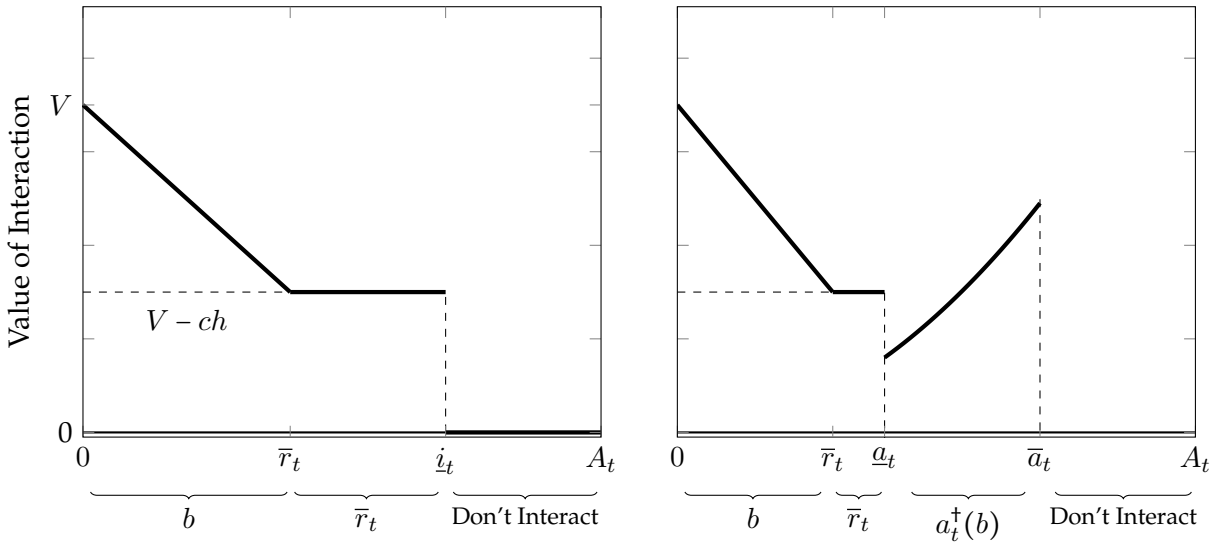
$$U_{NL}^{E}(b) = \begin{cases} V - h\frac{b}{A_t} & b \leq \overline{r}_t \\ V - ch & b \in (\overline{r}_t, \underline{i}_t] \\ 0 & b > \underline{i}_t \end{cases} \tag{3.1}$$

$$U_{L}^{E}(b) = \begin{cases} V - h\frac{b}{A_t} & b \leq \overline{r}_t \\ V - ch & b \in (\overline{r}_t, \underline{a}_t) \\ V + (\frac{bA_t}{A_t^2+\gamma} - c - h)(\frac{bA_t}{A_t^2+\gamma}) & b \in [\underline{a}_t, \overline{a}_t] \\ 0 & b > \overline{a}_t \end{cases} \tag{3.2}$$

Note that from the perspective of the organization, employees' expected utility from experiencing $\overline{r}_t$ is $V - h\frac{\overline{r}_t}{A_t} = V - h\frac{cA_t}{A_t} = V - ch$, whose value does not depend on b or A_t directly. The indirect utilities for the two equilibria are graphed in Figure 3.4.1. Both panels are calibrated with $V - ch > 0$, meaning that matching with a manager and experiencing a transgression is better for employees in expectation than no interaction at all. Calibrating with $V - ch < 0$ does not change the equilibrium thresholds in the figure, since the value of h does not affect managers' optimal strategy.

[4]That is, there may exist a distribution $F(b)$ such that employees' expected utility increases and $G(b)$ where it decreases.

Figure 3.4.1: Value of Employee Interactions: Non-Learning Equilbrium (E^{NL}) vs. Learning Equilibrium (E^L)



(a) E^{NL}: $\underline{i}_t < \underline{a}_t$ with $V - h > 0$

(b) E^L: $\underline{a}_t < \underline{i}_t < \overline{a}_t$ with $V - h < 0$

We highlight three mechanisms through which parameter changes affect employee welfare. The first emerges from changes in manager participation. The second and third emerge from changes in whether managers play interior actions and the intensity of these actions.

1. **Mechanism 1: switching between participation and no participation**. An increase in $\overline{a}_t$ or $\underline{i}_t$ induces some previously non-participating managers to participate. This increases welfare if and only if an employee prefers interacting with any newly participating b to not interacting, given the equilibrium (E^{NL} or E^L) after the parameter change. Formally, from (3.1) and (3.2), welfare increases if an employee's utility from interacting with type b satisfies:

$$V - ch \geq 0 \qquad E^{NL} \tag{P1}$$

$$V + \left(\frac{bA_t}{A_t^2 + \gamma} - c - h\right)\left(\frac{bA_t}{A_t^2 + \gamma}\right) \geq 0 \qquad E^L \tag{P2}$$

   Decreases in $\overline{a}_t$ or $\underline{i}_t$ result in the opposite effects.

2. **Mechanism 2: switching from interior actions to the reporting threshold**. If $\underline{a}_t$ decreases, or

if the initial equilibrium is E^L, managers who previously interacted and played actions $a_t^\dagger(\cdot)$ above the reporting threshold now play the reporting threshold $\overline{r}_t$. From the perspective of the organization, a type b lowering their action to $\overline{r}_t$ is better for employees if and only if

$$V - ch \geq V + (\frac{bA_t}{A_t^2+\gamma} - c - h)(\frac{bA_t}{A_t^2+\gamma})$$
$$\iff h \geq \frac{bA_t}{A_t^2+\gamma} \tag{S}$$

These expressions follow from (3.2).

The reporting threshold is less likely to be a transgression in the first place. But in the event that the original action and the reporting threshold are both transgressions, employees are worse off. The two actions incur the same harm, but switching to the reporting threshold eliminates reporting, and thus the possibility of receiving compensation. For a given b, the organization weighs the expected gain from a lower action against the expected loss from disincentivizing reports of actual transgressions.

3. **Mechanism 3: decrease in intensity of interior actions**. Suppose A_t falls to A_t' or γ increases to γ'. Consider a type b manager who played an interior action $a_t^\dagger(b)$ and continues to play an interior action $a_t^\dagger(b)'$. Since $a_t^\dagger(b) = \frac{bA_t^2}{A_t^2+\gamma} > \frac{bA_t'^2}{A_t'^2+\gamma'} = a_t^\dagger(b)'$, the intensity of the manager's action decreases.

   The tradeoff in terms of welfare is analogous to Mechanism 2. This lower action is less likely to be a transgression in the first place. But in the event that the original action and new action are both transgressions, employees are worse off: both actions induce a report, but at the new action, the probability of verification is lower. From the organization's perspective, a type b manager switching from $a_t^\dagger(b)$ to $a_t^\dagger(b)'$ is better for employees if and only if

$$(\frac{a_t^\dagger(b)}{A_t} - c - h)\frac{a_t^\dagger(b)}{A_t} \geq (\frac{a_t^\dagger(b)'}{A_t} - c - h)\frac{a_t^\dagger(b)'}{A_t}$$
$$c + h \geq a_t^\dagger(b) + a_t^\dagger(b)' \tag{I}$$

   These expressions also follow from (3.2).

These three mechanisms generate subtleties in how equilibrium welfare responds to changes in

the match value of an interaction V, the reporting cost c, and the size of punishment for managers γ.

Changes in Value of Interaction

Proposition 2. *Suppose V increases to V'. If h is sufficiently small, employee welfare increases. In particular:*

- *if we move from E^{NL} to E^{NL} and (P1) holds at V', welfare increases.*
- *If we move from E^{NL} or E^{L}, welfare increases if (P1) holds at V' and (P2) holds at V' for $b \in [\max\{\overline{a}_t, \underline{a}_t\}, \overline{a}'_t]$.*
- *If we move from E^{L} to E^{L}, welfare increases if (P2) holds at V' for $b \in [\max\{\overline{a}_t, \underline{a}_t\}, \overline{a}'_t]$.*

Otherwise, welfare changes are ambiguous.

An increase in V does not affect a manager's action, conditional on participating, but does shift the thresholds for participation to the right. This extends the upper bound on which types interact. In regions where managers were already participating, welfare increases: they play the same action, and V is higher. Newly participating managers, however, play actions that are weakly higher than the most extreme action played prior to their entry. So, if h is sufficiently large, these types may generate a negative or ambiguous change in welfare overall.

The equilibrium may stay in E^{NL} if the increase in V is small, or switch to E^{L} if it is large. However, the equilibrium can never switch from E^{L} to E^{NL}.

Changes in Reporting Costs

Proposition 3. *Suppose c increases marginally to c'.*

- *If we move from E^{NL} to E^{NL}, employee welfare decreases if h is large, i.e. if P1 holds at c'.*
- *If we move from E^{L} to E^{L}, welfare decreases if h is small, i.e. if $h \leq \frac{\frac{bA_t}{A_t^2+\gamma} - c}{\frac{bA_t}{A_t^2+\gamma} - c'} \frac{bA_t}{A_t^2+\gamma}$ for $b \in [\underline{a}'_t, \overline{a}_t]$.*
- *If we move from E^{L} to E^{NL}, welfare decreases if h is small, i.e. if $h \leq \frac{\frac{bA_t}{A_t^2+\gamma} - c}{\frac{bA_t}{A_t^2+\gamma} - c'} \frac{bA_t}{A_t^2+\gamma}$ for $b \in [\underline{a}_t, \underline{i}'_t]$ and (P2) holds for $b \in [\underline{i}'_t, \overline{a}_t]$.*

Otherwise, changes in welfare are ambiguous.

An increase in c increases the reporting threshold, which has three different effects on welfare. Whether the interaction of these effects increases or decreases welfare depends on the value of h.

The first, present in all cases in Proposition 7, is mechanical: increasing reporting costs directly decreases employee welfare in the event of a report.

The second emerges from Mechanism 1: increasing c may crowd in manager participation. If h is sufficiently low, increasing the probability of matching can increase welfare; otherwise, it may not. This is relevant in the first and third cases, as participation cannot change at all when starting and ending in E^L.

The third effect emerges from Mechanism 2. If we start in E^L, so that there are managers who play an interior action, then increasing c will induce some of them to play the *new* reporting threshold instead. It is possible that for some of these managers, the new reporting threshold is less intense than their original action, while for others it is more intense. This is because the reporting threshold is higher at $c' > c$. If h is sufficiently large, the former dominates the latter in terms of its welfare effect, so welfare increases. The constraint on h in the second bullet point is similar to (S) but additionally internalizes the increase in the reporting threshold.

Mechanism 2 governs whether the equilibrium may change from E^L to E^{NL}. Namely, the new reporting threshold may be sufficiently high that any participating b would rather deviate to it than play a higher, interior action that risks punishment.

Changes with Respect to γ

Proposition 4. *An increase in γ does not change employee welfare if we begin in E^{NL}. If we start in E^L, an increase in γ increases welfare if h is sufficiently large. If h is sufficiently small, welfare decreases. In particular, welfare increases (decreases)*

- *if (S) holds (does not hold) for $b \in [\underline{a}_t, \min\{\underline{a}'_t, \underline{i}_t\}]$,*
- *if (I) holds (does not hold) on $[\min\{\underline{a}'_t, \underline{i}_t\}, \max\{\underline{i}_t, \overline{a}'_t\}]$,*
- *and if (P2) holds (does not hold) on $[\max\{\underline{i}_t, \overline{a}'_t\}, \overline{a}_t]$.*

Otherwise, welfare changes are ambiguous.

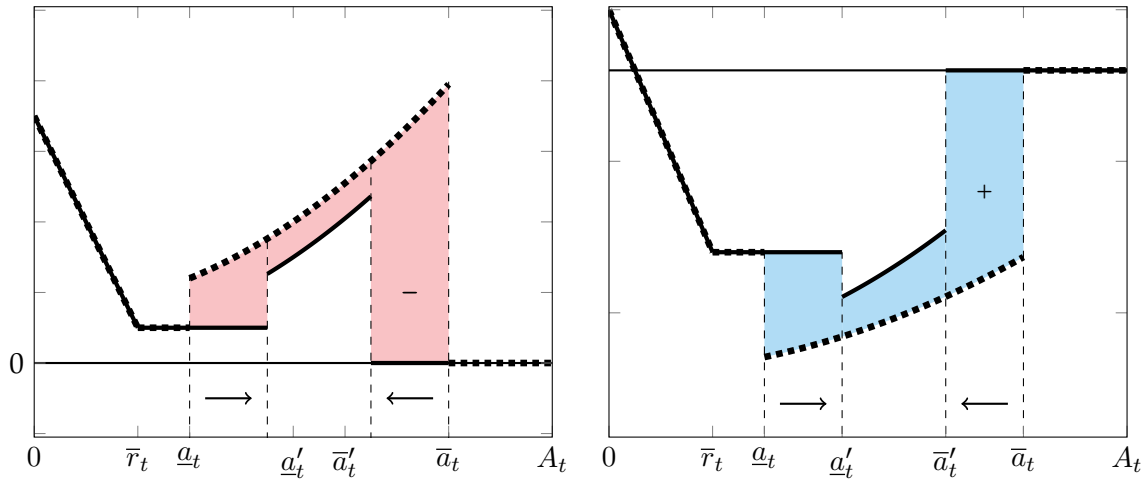
If we begin in E^{NL}, punishment is already large enough that no manager wants to play above the reporting threshold. A fortiori, this will be the case at higher γ as well. So, it is only possible for increasing γ to have an effect if we begin in E^{L}. When γ increases in E^{L}, all three mechanisms discussed above are operative, and their overall effect again depends on the value of h.

Figure 3.4.2 illustrates welfare changes as a function of h. When γ increases, $[\underline{a}_t, \overline{a}_t]$ shrinks to $[\underline{a}'_t, \overline{a}'_t]$. This triggers all three mechanisms: a range of b to the right select out of participation (Mechanism 1), a range of b to the left now take action $\overline{r}_t$ (Mechanism 2), and the intensity of actions in $[\underline{a}'_t, \overline{a}'_t]$ shifts downwards (Mechanism 3). This generates a benefit in the form of reducing action intensity. But, each mechanism has its concomitant cost: reducing the probability of interaction (Mechanism 1), or reducing the chance of compensation — either through reducing reports (Mechanism 2) or decreasing the probability of report verification (Mechanism 3).

Figure 3.4.2(a) illustrates the welfare impact of these changes when h is small. The dotted curves indicate the initial equilibrium and the solid the new equilibrium. In this case, there is a strict loss of welfare, shaded in red. Intuitively, since h is small, the match value V is large relative to the cost of a transgression h, as is the value of preserving the chance of compensation. So, for all three mechanisms, the benefit of reduced action intensity is dominated by the cost.

In contrast, when h is high, the cost of each mechanism is dominated by the benefit of reduced action intensity. This is illustrated in Figure3.4.2(b), with the gain shaded in blue.

Figure 3.4.2: Effects of Increase in γ in E^L on Equilibrium Utility: h Small vs. h Large

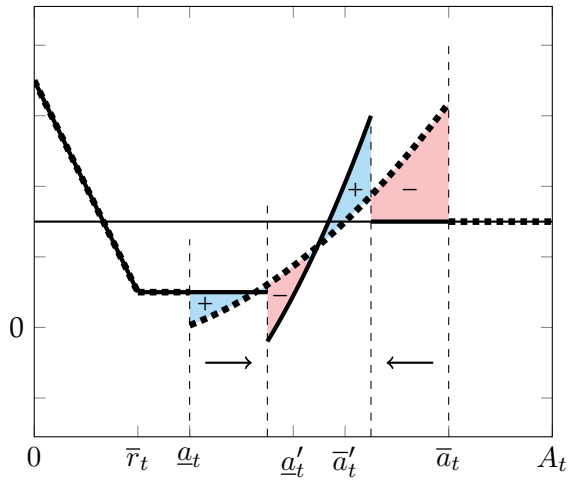


Bliss Point b

(a) h Small: Welfare Decreases

Bliss Point b

(b) h Large: Welfare Increases



Bliss Point b

(c) h Intermediate: Welfare Ambiguous

Finally, 3.4.2(c) illustrates the case of intermediate h. In such cases, on some interval(s), h may be sufficiently large to generate a welfare gain, while on others, h is sufficiently small to generate a loss. In the calibration in 3.4.2(c), h is large enough that the loss from eliminating reports on $[\underline{a}_t, \overline{a}'_t]$ is better than the interior action $a_t^\dagger(\cdot)$ played prior. In the next interval, $[\underline{a}'_t, \overline{a}'_t]$, h is small enough that decreasing action intensity only improves welfare when b (and so the resulting action)

is sufficiently high. In the last interval, $[\overline{a}'_t, \overline{a}_t]$, h is sufficiently small that the loss from reduced participation outweighs the benefit of cutting extreme actions.

The following Theorem synthesizes the patterns in Figure 3.4.2, showing that the size of the optimal punishment is increasing in h. When transgressions themselves only have a small effect on employee welfare, managers should be punished as little as possible. When they incur great harm, punishment should be maximal. And in between, the value of optimal punishment can take an interior value.

Theorem 2. *Let $\gamma^*(h)$ be the value of $\gamma \geq V$ that maximizes expected employee utility, as a function of h. $\gamma^*(h)$ is nondecreasing in h. For h sufficiently small, $\gamma^*(h) = V$, and for h sufficiently large, $\gamma^*(h) = \overline{\gamma} > V$.*

3.5 Dynamics and Non-Learning Steady State

Within the organizational model of the paper, verified reports update the organizational standard for what actions constitute transgressions. As this process continues, the model may eventually converge to a steady state equilibrium. This section details conditions for the steady state equilibrium to occur, how employee welfare changes as we converge to this steady state, and assesses the qualitative effects of this process on organizational learning.

Definition A^* corresponds to a *steady state equilibrium* if when $A_t = A^*$, $A_{t+j} = A^*$ with probability 1 for all $j > 0$.

E^{NL} is always a steady state, since no learning can occur. Managers either never interact, or interact and keep employees precisely indifferent to reporting. E^L may or may not be a steady state, depending on whether the interior actions being played in fact constitute a transgression. If for some $b \in [\underline{a}_t, \overline{a}_t]$, $a_t^\dagger(b) > a^*$, a report — and therefore learning — can occur. So, this is not a steady state. If $a_t^\dagger(b = \overline{a}_t) \leq a^*$, i.e. the maximal action played in t with positive probability is below the true cutoff for transgressions, then reporting will stop. This is a steady state according to the definition above. However, employees and the organization differ in their perception of whether a steady state has been reached. In particular, the organization still believes that reports will occur with positive probability, since interior actions are being played.

The following proposition establishes that the model converges to a steady state.

Proposition 5. *We converge to a steady state.*

Dynamic Welfare We analyze the welfare effects of *decreases* in A_t. Since a decrease in A_t corresponds to learning, this allows us to study how refinement of standards for transgression can generate inefficiences in welfare. The following proposition summarizes comparative statics when $A_{t+1} < A_t$.

Proposition 6. *Suppose A_t decreases marginally to A_{t+1}. Conditional on matching with $b \in [0, \overline{r}_{t+1}]$, employee welfare decreases. Moreover,*

- *If we begin and stay in E^{NL}, welfare decreases if h is small, that is, if (P1) holds.*
- *If we begin in E^{NL} and move to E^{L}, welfare decreases if h is sufficiently large. That is,*
  - *welfare decreases if (S) holds for $b \in [\overline{r}_t, \underline{a}_{t+1}]$*
  - *and (P2) for $b \in [\underline{i}_t, \overline{a}_{t+1}]$.*
- *If we begin in E^{L} and stay in E^{L}, welfare decreases if h is sufficiently small and $\underline{a}_t < \underline{a}_{t+1}$. In particular, it decreases if*
  - *$\underline{a}_{t+1} < \underline{a}_t$ and (S) holds;*
  - *$\underline{a}_t < \underline{a}_{t+1}$ and (S) does not hold,*
  - *(I) holds for $b \in [\max\{\underline{a}_t, \underline{a}_{t+1}\}, \overline{a}_{t+1}]$, and (P2) holds for $b \in [\overline{a}_{t+1}, \overline{a}_t]$.*
- *If we begin in E^{L} and move to E^{NL}, welfare decreases if h is sufficiently small. That is, welfare decreases if*
  - *$\underline{i}_{t+1} < \underline{a}_t$ and (P1) holds,*
  - *$\underline{a}_t < \underline{i}_{t+1}$ and (S) does not hold for $b \in [\underline{a}_t, \underline{i}_{t+1}]$, and*
  - *and (P2) holds for $b \in [\max\{\underline{i}_{t+1}, \underline{a}_t\}, \overline{a}_t]$.*

Otherwise, welfare changes are ambiguous.

A decrease in A_t to A_{t+1} has several effects, depending on the value of b. These are illustrated in Figure 3.5.1, for the case where we begin and end in E^L.[5]

We start by considering managers with $b < r_{t+1}$, who play their bliss point regardless. While their action does not change, the organization's belief that these actions are transgressions increases from $\frac{b}{A_t}$ to $\frac{b}{A_{t+1}}$. This makes employees who match with a manager with low b worse off *in expectation*.

Those $b \in [\overline{r}_{t+1}, \overline{r}_t]$ go from playing their bliss point to the new reporting threshold. This, too, results in a strict decrease in welfare, despite being a lower action. Recall that an employee's expected utility from experiencing the reporting threshold is independent of A_t. Since, for b in this range, $\overline{r}_t$ is worse than if they played their bliss point, switching to $\overline{r}_{t+1}$ must necessarily be worse as well. Those who play the reporting threshold in either case, as discussed, incur no change in expected utility. In the figure, this is illustrated for $b \in [\overline{r}_t, \underline{a}_{t+1}]$[6]

The remaining interval of b's are impacted by the mechanisms discussed earlier. Mechanism 2 operates for $b \in [\underline{a}_{t+1}, \underline{a}_t]$: some managers will shift from an action $a_t^\dagger$ above the old reporting threshold to the (new) reporting threshold $\overline{r}_{t+1}$. As in the static case, this decreases welfare if h is small (panel a), but can increase welfare if h is large (panel b).[7]

Mechanism 3 operates for $b \in [\underline{a}_t, \overline{a}_{t+1}]$. Suppose a type b manager takes action $a_t^\dagger(b)$ at A_t and $a_{t+1}^\dagger(b)$ at A_{t+1}. Then, $a_t^\dagger(b) > a_{t+1}^\dagger(b)$. On the one hand, these actions are less intense, and hence less likely to constitute a transgression. On the other hand, conditional on a transgression occurring, it is harder to verify. As in the static case, for h small, welfare decreases (panel a), while it potentially increases for h large (panel b).

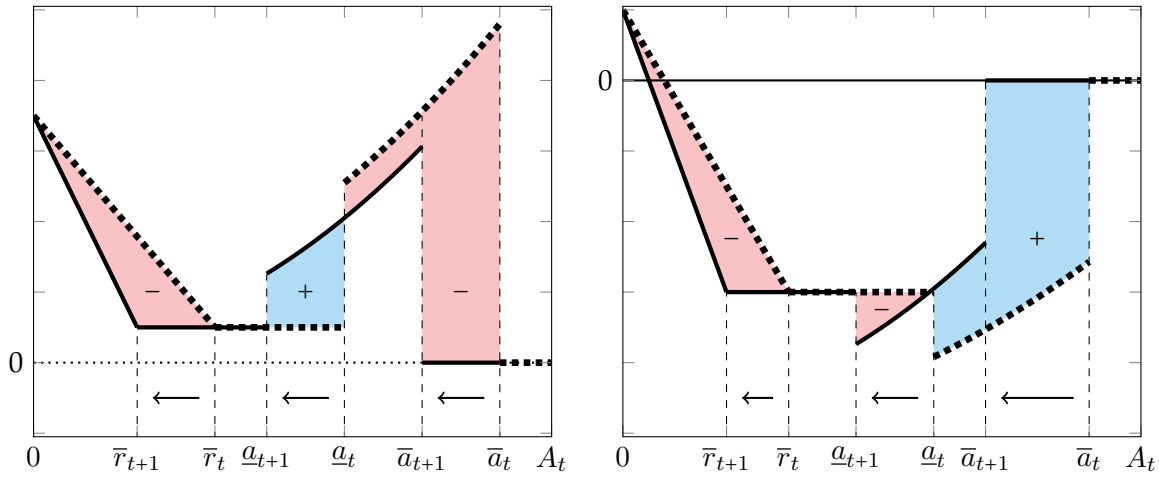
Finally, Mechanism 1 operates for $[\overline{a}_{t+1}, A_t]$: managers with high b who previously interacted now no longer interact. As in the static case, this decreases utility if h is small (panel a), and increases if h is large (panel b).

[5] We choose this case because it illustrates all of the effects generated by changes in A_t, whereas other cases do not.

[6] Note that when A_t changes, it can be that $\underline{a}_{t+1} < \underline{a}_t$, in which case the range would be $b \in [\overline{r}_t, \underline{a}_t]$.

[7] If $\underline{a}_{t+1} < \underline{a}_t$, some managers will shift from the reporting threshold to an interior action. This has the opposite interaction with h; welfare decreases for h large and increases for h small.

Figure 3.5.1: Effects of Decrease in A_t to A_{t+1} on Equilibrium Utility Beginning and Ending in E^L, $\underline{a}_{t+1} < \underline{a}_t$, h Small vs. h Large



Bliss Point b

(a) h Small: Welfare Decreases

Bliss Point b

(b) h Large: Welfare Increases

Convergence to Suboptimal Steady State The proposition above further implies that, when h is sufficiently small, a decrease in A_t that generates a switch from E^L to E^{NL} results in a decrease in welfare.

Corollary 3. *Fixing all other parameters, suppose A_t falls to A_{t+1}, such that A_t corresponds to E^L and A_{t+1} to E^{NL}. If h is sufficiently small, welfare decreases.*

Intuitively, the welfare effects of a switch from E^L to E^{NL} are determined by Mechanisms 1 and 2. Some managers switch from interaction to no interaction, while others switch from playing an action $a_t^\dagger$ above the reporting threshold to the new reporting threshold. Both effects decrease welfare when h is small.

The consequence of this result is that, if E^L is *not* a steady state, learning may result in convergence to a steady state E^{NL} where welfare is *strictly worse* than prior to being in a steady state.[8]

[8]Recall that, from the perspective of the organization, E^L is *never* a steady state. If A_t decreases but we remain in E^L at $t+1$, a necessary condition for a decrease in utility given low h is that $\underline{a}_t < \underline{a}_{t+1}$. A sufficient condition for this, by Proposition 3.3, is that A_t small.

3.6 Conclusion

We study an organizational model where "managers" commit transgressions — such as harassment, overwork, or other firms of workplace abuse — against "employees." Managers consider first whether to interact with an employee. Managers then take an action of varying intensity. While more intense actions are more likely to constitute a transgression, they are uncertain about what employees actually consider transgressions. In the event a manager does commit a transgression, employees incur a disutility and face the decision to report this action to the organization at a cost. Verifying a transgression is easier for the organization when it is closer to what has previously been established as a transgression. Verification of a transgression leads to a payout for the victim and a punishment for the manager. Managers are heterogenous, and have different propensities for committing more intense actions.

We show that policies that lower managers' expected utility from committing transgressions may have ambiguous or negative effects on employees' expected utility. We highlight three mechanisms that generate these ambiguous effects. First, managers' expected utility may decrease when policy changes discourage participation. For example, if managers receive a match value upon interacting with an employee, and that match value decreases, managers may be less likely to interact with an employee ex-ante. In the event the disutility of experiencing a transgression is large, employees may be better off, no interaction may be preferred to being interacted with and experiencing abuse. However, if the disutility is small, employees' expected utility may decrease.

Second, managers may switch from committing actions that employees have a strict incentive to report to the organization to actions that keep them indifferent. Policy changes that encourage these changes — such as increasing the punishment for committing a transgression — have two effects. On the one hand, because managers commit less intense actions, these are ex-ante less likely to constitute transgressions. On the other hand, in the event an action is actually a transgression, employees no have no incentive to report it, and are worse off. If the disutility from a transgression is small, the latter channel dominates, and employees' expected utility may decrease. Third, managers may continue committing actions that employees have a strict incentive to report, but again decrease their intensity or opt out altogether. These can decrease employees' welfare if the disutility from experiencing a transgression is large.

We use the tessellation of these three mechanisms to derive the optimal punishment for transgressive managers that maximizes employees' expected utility. We show that optimal punishment depends on the magnitude of harm incurred by a transgression, and is non-decreasing in this magnitude. When the size of harm is low, optimal punishment may involve minimal punishment of abusers. When it is large, maximal punishment is optimal. Otherwise, it may lie in between these two extremes.

Since managers' propensity to commit transgressions is heterogeneous within the organization, the precise welfare effects of these three mechanisms depend on the distribution of managers. However, the general dependence of these different effects allows us to think about policy changes in the setting of different forms of transgressions. One-off microaggressions that make employees feel uncomfortable but may not result in the sabotage of may correpond to transgressiosn that result in minimal harm. Physical forms of abuse and sabotage that cause long-lasting and persistently negative effects on employees' reputations may correpond to higher levels of disutility.

Finally, we consider a learning extension to the model where reports generate information for the organization and managers. Specifically, we assume that a record of *only* past, successfully-verified reports of transgressions are public information. This means that, as transgressions occur and are reported over time, more precise information about what constitutes a transgression is revealed, which may be thought to improve employees' welfare. We show, however, that learning of this sort encompasses all three of the mechanisms above, and may actually decrease employees' welfare. Moreover, we show that this dynamic model always converges to a steady-state equilibrium. In this steady state equilibrium, no learning about transgressions occurs, and if transgressions do occur, they are never punished.

Future directions for this line of research include richer connections to empirical data on harassment and workplace abuse using micro-data on labor turnover, reporting of abuse, and billed hours in organizations. Our model can be applied to settings beyond harassment and workplace of abuse. We have commented on its applications to studying consumer boycotts of excessive price hikes or how judiciaries may hold political executive accountable for abuses of power, but they may also apply to macroeconomic policymaking. For example, consider a government that taxes its citizens. Citizens tolerate taxes up to a certain threshold, but beyond that threshold, start

evading payment. Or, consider a Central Bank that sets interest rates that affects financial markets; markets again tolerate these hikes up to a certain point, but if they are too excessive, this triggers a sell-off. This framework provides a variety of avenues for further theoretical research on the externalities of harmful actions.

3.7 Proofs

Lemma 1. *Reporting follows a threshold rule: the employee has a strict incentive to report a transgression if and only if* $a_t > \overline{r}_t \equiv cA_t$ *and* $a_t \geq a^*$.

Proof. For $a_t \geq a^*$, the employee has a strict incentive to report if and only if $-c + \frac{a_t}{A_t} > 0$, which occurs if and only if $a_t > cA_t$. □

Lemma 2. *Conditional on interacting, the optimal action* $a^*(t)$ *for a type* b *manager is characterized by the thresholds*

- $\underline{a}_t \equiv cA_t + c\frac{\gamma+\sqrt{\gamma(A_t^2+\gamma)}}{A_t^2}$, *and*
- $\overline{b}_t \equiv A_t + \frac{\gamma}{A_t}$

such that:

- *if* $b \in [0, \overline{r}_t]$, $\alpha_t(b) = b$ *(The manager plays her bliss point.)*
- *if* $b \in (\overline{r}_t, \min\{\underline{a}_t, \overline{b}_t, A_0\}]$, *then* $\alpha_t(b) = \overline{r}_t$ *(The manager plays the reporting threshold.)*
- *if* $b \in (\underline{a}_t \min\{\overline{b}_t, A_0\})$, *then* $\alpha_t(b) = \frac{bA_t^2}{A_t^2+\gamma} \equiv a_t^\dagger(b)$ *(The manager plays an interior action above the reporting threshold but below her preferred action.)*
- *and if* $b > \min\{\overline{b}_t, A_0\}$, *then* $\alpha_t(b) = b$ *(The manager plays her bliss point.)*

Proof. Suppose $a_t \leq \overline{r}_t$. In this case, if $b \leq \overline{r}_t$, the manager simply plays her preferred action, $a_t = b$. If $a_t \geq \overline{r}_t$, the action achieving her maximal payoff is defined by:

$$-2(a_t - b) - 2\frac{a_t}{A_t^2}\gamma = 0 \implies a_t = \frac{bA_t^2}{A_t^2 + \gamma} \equiv a_t^\dagger(b)$$

This interior action is better than $\overline{r}_t$ for $A_t \geq a_t \geq \overline{r}_t$ and $A_t \geq \frac{bA_t^2}{A_t^2+\gamma}$ if and only if

$$\begin{aligned}
V - (\frac{bA_t^2}{A_t^2+\gamma} - b)^2 - \gamma\left(\frac{bA_t}{A_t^2+\gamma}\right)^2 &\geq V - (cA_t - b)^2 \\
-\left(\frac{b\gamma}{A_t^2+\gamma}\right)^2 - \gamma\left(\frac{bA_t}{A_t^2+\gamma}\right)^2 &\geq -(cA_t - b)^2 \\
-b^2\frac{\gamma^2}{(A_t^2+\gamma)^2} - b^2\frac{\gamma A_t^2}{(A_t^2+\gamma)^2} &\geq -b^2 + 2bcA_t - c^2A_t \\
\left[1 - \frac{\gamma(A_t^2+\gamma)}{(A_t^2+\gamma)^2}\right]b^2 - 2bcA_t + c^2A_t &\geq 0 \\
\left[\frac{A_t^2}{(A_t^2+\gamma)}\right]b^2 - 2bcA_t + c^2A_t &\geq 0 \\
A_tb^2 - 2bc(A_t^2+\gamma) + c^2(A_t^2+\gamma) &\geq 0 \\
b \geq \frac{2c(A_t^2+\gamma) + \sqrt{4c^2(A_t^2+\gamma)^2 - 4A_tc^2k^2(A_t^2+\gamma)}}{2A_t} & \\
b \geq \frac{c(A_t^2+\gamma) + \sqrt{c^2(A_t^2+\gamma)^2 - A_tc^2(A_t^2+\gamma)}}{A_t} & \\
b \geq \frac{c(A_t^2+\gamma) + \sqrt{c^2(A_t^2+\gamma)(A_t+\gamma-A_t)}}{A_t} & \\
b \geq \left(\frac{(A_t^2+\gamma) + \sqrt{\gamma(A_t^2+\gamma)}}{A_t}\right)\left(c\right) & \\
= cA_t + c\frac{\gamma + \sqrt{\gamma(A_t^2+\gamma)}}{A_t} \equiv \underline{a}_t &
\end{aligned}$$

If $b \leq \min\{\underline{a}_t, A_t\}$, then a type b manager will commit action $\overline{r}_t$. If $A_t > b > \underline{a}_t$, a type b manager will commit action $\alpha_t(b) = \frac{bA_t^2}{A_t^2+\gamma}$. Finally, if $\frac{bA_t^2}{A_t^2+\gamma} > A_t$, the optimal action (conditional on interaction) is simply b, since transgressions will always be reported and punished with probability 1, and changing a_t has no marginal impact on punishment probability. This switch is given by the indifference point $\overline{b}_t$ defined by:

$$\frac{\overline{b}_tA_t^2}{A_t^2+\gamma} = A_t \implies \overline{b}_t = A_t + \frac{\gamma}{A_t}$$

Hence, for $b > \overline{b}_t$, the manager simply plays her preferred action b.

□

Lemma 3. *The manager's decision to interact, conditional on playing $\alpha_t(b)$ if she does, is characterized by*

two additional thresholds,

- $\underline{i}_t \equiv cA_t + \sqrt{V}$ *and*
- $\overline{a}_t \equiv \sqrt{\left(\frac{A_t^2}{\gamma} + 1\right)V}$

such that

- *if* $b \in [0, \min\{\underline{a}_t, \underline{i}_t, A_0\}]$, $i_t^*(b) = I$ *(The manager interacts.)*
- *if* $\underline{a}_t < \underline{i}_t$ *and* $b \in (\underline{a}_t, \min\{\overline{a}_t, A_0\}]$, $i_t^*(b) = I$ *(The manager interacts.)*
- *if* $\underline{i}_t \leq \underline{a}_t$ *and* $b > \underline{i}_t$, $i_t^*(b) = NI$ *(The manager never interacts.)*
- *if* $\underline{a}_t < \underline{i}_t$ *and* $b > \overline{a}_t$, $i_t^*(b) = NI$ *(The manager never interacts.)*

Proof. For $b \in [0, \overline{r}_t]$, the manager is playing her optimal action and is not punished, so she always interacts. For $b \in (\overline{r}_t, \min\{\underline{a}_t, A_t\}]$, the manager plays at the reporting threshold. Interaction is optimal if and only if

$$V - (\overline{r}_t - b)^2 \geq 0$$
$$\overline{r}_t + \sqrt{V} \equiv \underline{i}_t \geq b.$$

Next, if $\underline{a}_t < A_t$ and $b \in (\underline{a}_t, \min\{\overline{b}_t, A_0\}]$, the manager plays an interior action $\frac{bA_t^2}{A_t^2+\gamma}$. Interaction is optimal if and only if

$$\begin{aligned} V - (\frac{bA_t^2}{A_t^2+\gamma} - b)^2 - \gamma\frac{b^2A_t^2}{(A_t^2+\gamma)^2} &\geq 0 \\ V - \gamma^2\frac{b^2}{(A_t^2+\gamma)^2} - \gamma\frac{b^2A_t^2}{(A_t^2+\gamma)^2} &\geq 0 \\ V - \frac{\gamma^2b^2 + \gamma b^2A_t^2}{(A_t^2+\gamma)^2} &\geq 0 \\ V - \frac{\gamma}{A_t^2+\gamma}b^2 &\geq 0 \end{aligned}$$

Note that the expression above is decreasing in b. Moreover, at $b = \overline{b}_t$, the expression is equal to $V - \gamma < 0$. Hence, there exists $\overline{a}_t < \overline{b}_t$ such that there is an incentive to interact if and only if $b \leq \overline{a}_t$.

This is i is given by solving the expression above at equality, which yields

$$\overline{a}_t = \sqrt{\left(\frac{A_t^2}{\gamma} + 1\right)V}$$

□

Proposition 1. *Comparative statics of the main equilibrium actions for the employee and manager are as follows.*

- *An increase in V generates an increase in $\underline{i}_t$ and $\overline{a}_t$.*
- *An increase in c generates an increase in $\overline{r}_t$, $\underline{a}_t$, and $\underline{i}_t$.*
- *An increase in γ generates an increase in $\underline{a}_t$ and a decrease in $\overline{a}_t$. Moreover, for each b, $a_t^\dagger(b) = \frac{bA_t^2}{A_t^2+\gamma}$ decreases.*
- *A* decrease *in A_t generates a decrease in $\overline{r}_t$, $\underline{i}_t$, and $\overline{a}_t$. For each b, $a_t^\dagger(b) = \frac{bA_t^2}{A_t^2+\gamma}$ decreases. There exists, $\tilde{A} > 0$ such that the sign of $\underline{a}_t$ is negative for $A_t \leq \tilde{A}$ and is positive above.*

Proof. All comparative statics are immediate, with the exception of $\underline{a}_t$ with respect to A_t, which is proportional to

$$\frac{(A_t^2 + \gamma) + \sqrt{\gamma(A_t^2 + \gamma)}}{A_t}$$

The sign of its derivative is positive if and only if:

$$\begin{aligned}
A_t[2A_t + \frac{\gamma A_t}{\sqrt{\gamma(A_t^2+\gamma)}}] - [(A_t^2+\gamma) + \sqrt{\gamma(A_t^2+\gamma)}] &\geq 0 \\
A_t^2 + \frac{\gamma A_t^2}{\sqrt{\gamma(A_t^2+\gamma)}} - \gamma - \sqrt{\gamma(A_t^2+\gamma)}] &\geq 0 \\
(A_t^2-\gamma)\sqrt{\gamma(A_t^2+\gamma)} + \gamma A_t^2 - \gamma(A_t^2+\gamma) &\geq 0 \\
(A_t^2-\gamma)\sqrt{\gamma(A_t^2+\gamma)} - \gamma^2 &\geq 0 \\
(A_t^2-\gamma)^2\gamma(A_t^2+\gamma) &\geq \gamma^4 \\
(A_t^2-\gamma)(A_t^2+\gamma)\gamma(A_t^2-\gamma) &\geq \gamma^4 \\
(A_t^4-\gamma^2)(A_t^2+\gamma) &\geq \gamma^3 \\
A_t^6 + \gamma A_t^4 - \gamma^2 A_t^2 &\geq 0 \\
A_t^2(A_t^2-\gamma) &\geq \gamma^2
\end{aligned}$$

Note that the above is quadratic in A_t^2, is negative below some $\tilde{A}$ which solves the equation above with equality, and is positive above $\tilde{A}$. Moreover, as A_t^2 approaches 0, $\underline{i}_t$ appraoches ∞. □

Proposition 2. *Suppose V increases to V'. If h is sufficiently small, employee welfare increases. In particular:*

- *if we move from E^{NL} to E^{NL} and (P1) holds at V', welfare increases.*
- *If we move from E^{NL} or E^L, welfare increases if (P1) holds at V' and (P2) holds at V' for $b \in [\max\{\overline{a}_t, \underline{a}_t\}, \overline{a}'_t]$.*
- *If we move from E^L to E^L, welfare increases if (P2) holds at V' for $b \in [\max\{\overline{a}_t, \underline{a}_t\}, \overline{a}'_t]$.*

Otherwise, welfare changes are ambiguous.

Proof. First, conditional on matching with a manager who has $b \leq \min\{\underline{i}_t, \underline{a}_t\}$, managers' actions stay the same, but employees derive a strictly higher match value V', so their welfare strictly increases.

Next, consider E^{NL}. If we remain in E^{NL} — so $\underline{a}'_t > \underline{i}'_t > \underline{i}_t$, managers who previously did not interact on $[\underline{i}_t, \underline{i}'_t]$ now interact and keep employees indifferent to reporting a transgression; welfare on $[\underline{i}_t, \underline{i}'_t]$ increases if and only if $V' - ch \geq 0$.

If we switch from E^{NL} to E^L — so $\underline{i}'_t > \underline{a}_t$ — then managers on $[\underline{i}_t, \underline{a}_t]$ switch from not interacting to interacting at $\overline{r}_t$. Welfare hence increases for employees if and only if $V' - ch \geq 0$. Manager behavior changes to the interior action on $[\underline{a}_t, \overline{a}'_t]$, meaning welfare increases if and only if $V - h - c + \frac{bA_t}{A_t^2+\gamma} \geq 0$.

If we begin in E^L, note first that we never switch to E^{NL}, since $\underline{a}_t$ does not change and $\underline{i}_t$ only increases. Here, welfare strictly increases below $\overline{a}_t$. Managers on $[\overline{a}_t, \overline{a}'_t]$ who did not interact previously now interact and play an action that an employee reports. Welfare on this range increases if and only if $V' + (\frac{bA_t}{A_t^2+\gamma} - c - h)(\frac{bA_t}{A_t^2+\gamma}) \geq 0$. □

Proposition 7. *Suppose c increases marginally to c'.*

- *If we move from E^{NL} to E^{NL}, employee welfare decreases if h is large, i.e. if P1 holds at c'.*
- *If we move from E^L to E^L, welfare decreases if h is small, i.e. if $h \leq \frac{\frac{bA_t}{A_t^2+\gamma} - c}{\frac{bA_t}{A_t^2+\gamma} - c'} \frac{bA_t}{A_t^2+\gamma}$ for $b \in [\underline{a}'_t, \overline{a}_t]$.*
- *If we move from E^L to E^{NL}, welfare decreases if h is small, i.e. if $h \leq \frac{\frac{bA_t}{A_t^2+\gamma} - c}{\frac{bA_t}{A_t^2+\gamma} - c'} \frac{bA_t}{A_t^2+\gamma}$ for $b \in [\underline{a}_t, \underline{i}'_t]$ and (P2) holds for $b \in [\underline{i}'_t, \overline{a}_t]$.*

Otherwise, changes in welfare are ambiguous.

Proof. Both $\overline{r}_t$ and $\underline{i}_t$ increase marginally by a factor of A_t. $\underline{a}_t$ increases by a factor larger than A_t; hence, an equilibrium can move from E^L to E^{NL}, but never from E^{NL} to E^L.

Next, for managers with $b \leq \overline{r}_t$, their actions do not change. However, managers with $b \in (\overline{r}_t, \overline{r}'_t]$ now play their bliss point, which is strictly higher than $\overline{r}_t$, generating a strictly negative shift in employee utility. Suppose we are E^{NL} and remain in E^{NL}. Welfare on $[\overline{r}'_t, \underline{i}_t]$ strictly decreases, since managers on this range are playing at the new reporting threshold $\overline{r}'_t > \overline{r}_t$. Finally, on $[\underline{i}_t, \underline{i}'_t]$, welfare increases if and only if $V - c'h \geq 0$.

Next, suppose we are in E^L. Due to a marginal increase, we have either $\overline{r}'_t < \underline{a}_t < \underline{a}'_t < \underline{i}'_t$, in which case we remain in E^L, or $\overline{r}'_t < \underline{a}_t < \underline{i}'_t < \underline{a}'_t$, in which case we move to E^{NL}. As above, welfare decreases on $[\overline{r}_t, \overline{r}'_t]$. Welfare on $[\overline{r}'_t, \underline{a}_t]$ decreases; managers are playing at a strictly higher reporting threshold.

If we remain in E^L, welfare on $[\underline{a}_t, \underline{a}'_t]$ welfare increases if and only if the new reporting threshold is worse than experiencing $a_t^\dagger$ under the old c, i.e. if and only if

$$\begin{aligned} -c'h &\geq (-h - c + \frac{bA_t}{A_t^2+\gamma})\frac{bA_t}{A_t^2+\gamma} \\ (-c' + \frac{bA_t}{A_t^2+\gamma})h &\geq (-c + \frac{bA_t}{A_t^2+\gamma})\frac{bA_t}{A_t^2+\gamma} \\ h &\geq \frac{\frac{bA_t}{A_t^2+\gamma} - c}{\frac{bA_t}{A_t^2+\gamma} - c'}\frac{bA_t}{A_t^2+\gamma} \end{aligned}$$

Welfare on $[\underline{a}'_t, \overline{a}_t]$ strictly decreases; employees experience the same interior action $a_t^\dagger$ but pay a higher cost to report it.

Finally, in the case where we move to E^{NL}, welfare on $[\underline{i}'_t, \overline{a}_t]$ increases if and only if no interaction is better than the old $a_t^\dagger$, i.e. if and only if $V + (-h - c - + \frac{bA_t}{A_t^2+\gamma})(\frac{bA_t}{A_t^2+\gamma}) \leq 0$.

□

Proposition 4. *An increase in γ does not change employee welfare if we begin in E^{NL}. If we start in E^L, an increase in γ increases welfare if h is sufficiently large. If h is sufficiently small, welfare decreases. In particular, welfare increases (decreases)*

- *if (S) holds (does not hold) for $b \in [\underline{a}_t, \min\{\underline{a}'_t, \underline{i}_t\}]$,*
- *if (I) holds (does not hold) on $[\min\{\underline{a}'_t, \underline{i}_t\}, \max\{\underline{i}_t, \overline{a}'_t\}]$,*
- *and if (P2) holds (does not hold) on $[\max\{\underline{i}_t, \overline{a}'_t\}, \overline{a}_t]$.*

Otherwise, welfare changes are ambiguous.

Proof. Consider then E^L, where an increase in γ leads to an increase in $\underline{a}_t$ to $\underline{a}'_t$ and a decrease in $\overline{a}_t$ to $\overline{a}'_t$. If we remain in E^L, on $[\underline{a}_t, \underline{a}'_t]$, managers who previously interacted above the reporting threshold and risked potential punishment now interact right at the reporting threshold. This is

better for employees if and only if

$$\begin{aligned} V + (\frac{bA_t}{A_t^2+\gamma} - c - h)(\frac{bA_t}{A_t^2+\gamma}) &\leq V - h\frac{\overline{r}_t}{A_t} \\ (\frac{bA_t}{A_t^2+\gamma} - c - h)(\frac{bA_t}{A_t^2+\gamma}) &\leq -h \cdot c \\ \big(-c + \frac{bA_t}{A_t^2+\gamma}\big)(\frac{bA_t}{A_t^2+\gamma}) &\leq h\big(\frac{bA_t}{A_t^2+\gamma} - c\big) \\ h &\geq \frac{bA_t}{A_t^2+\gamma} \end{aligned}$$

On $[\underline{a}'_t, \overline{a}'_t]$, managers continue to interact at an action above the reporting threshold, but that action becomes less intense. This results in a potential decrease to expected utility (harder to report a transgression conditional on it occurring) but also a potential increase (an action for a given type b manager in this range is less likely to be a transgression). An increase is better if and only if

$$\begin{aligned} (\frac{bA_t}{A_t^2+\gamma} - c - h)(\frac{bA_t}{A_t^2+\gamma}) &\leq (-h - c + \frac{bA_t}{A_t^2+\gamma'})(\frac{bA_t}{A_t^2+\gamma'}) \\ \frac{-h-c}{A_t^2+\gamma} + \frac{bA_t}{(A_t^2+\gamma)^2} &\leq \frac{-h-c}{A_t^2+\gamma'} + \frac{bA_t}{(A_t^2+\gamma')^2} \\ bA_t\bigg(\frac{1}{(A_t^2+\gamma)^2} - \frac{1}{(A_t^2+\gamma')^2}\bigg) &\leq (h+c)\big(\frac{1}{A_t^2+\gamma} - \frac{1}{A_t^2+\gamma'}\big) \\ bA_t\bigg(\frac{1}{A_t^2+\gamma} + \frac{1}{A_t^2+\gamma'}\bigg) &\leq (h+c) \\ h - \frac{bA_t}{A_t^2+\gamma'} &\geq \frac{bA_t}{A_t^2+\gamma} - c \\ h &\geq \frac{bA_t}{A_t^2+\gamma'} + \frac{bA_t}{A_t^2+\gamma} - c \end{aligned}$$

On $[\overline{a}'_t, \overline{a}_t]$, managers who previously interacted now do not interact at all. Because employees on this range had a strict incentive to report, this is an improvement if and only if $0 \geq V + (\frac{bA_t}{A_t^2+\gamma} - c - h)(\frac{bA_t}{A_t^2+\gamma})$.

Finally, if we move from E^L to E^{NL}, we are simply in an equilibrium equivalent to having $\underline{a}'_t = \underline{i}_t = \overline{a}'_t$, and can apply the insights from above.

□

Theorem 2. *Let $\gamma^*(h)$ be the value of $\gamma \geq V$ that maximizes expected employee utility, as a function of h. $\gamma^*(h)$ is nondecreasing in h. For h sufficiently small, $\gamma^*(h) = V$, and for h sufficiently large,*

$\gamma^*(h) = \overline{\gamma} > V$.

Proof. Note that conditional on being in E^{NL}, welfare does not change with γ, so suppose we are in E^L. An employee's expected utility conditional on γ is given by:

$$F(\overline{a}_t)V + \int_0^{\overline{r}_t} -h\frac{b}{A_t}f(b)db + \int_{\overline{r}_t}^{\underline{a}_t} -h\cdot cf(b)db + \int_{\underline{a}_t}^{\overline{a}_t}(-h-c+\frac{bA_t}{A_t^2+\gamma})\frac{bA_t}{A_t^2+\gamma}f(b)db$$

Noting that $\overline{a}_t$ and $\underline{a}_t$ are both functions of γ, the derivative of this expression with respect to γ is

$$f(\overline{a}_t)\overline{a}_t'V - hcf(\underline{a}_t)\underline{a}_t' + (-h-c+\frac{\overline{a}_tA_t}{A_t^2+\gamma})\frac{\overline{a}_tA_t}{A_t^2+\gamma}f(\overline{a}_t)\overline{a}_t' - (-h-c+\frac{\underline{a}_tA_t}{A_t^2+\gamma})\frac{\underline{a}_tA_t}{A_t^2+\gamma}f(\underline{a}_t)\underline{a}_t' + \int_{\underline{a}_t}^{\overline{a}_t} -2\frac{(bA_t)^2}{(A_t^2+\gamma)^3} + (h+c)\frac{bA_t}{(A_t^2+\gamma)^2}f(b)db.$$

Note that $\overline{a}_t'(\gamma) < 0$ and $\underline{a}_t'(\gamma) > 0$. Hence, after reorganizing terms, the derivate's sign can be characterized via the following pieces:

$$\underbrace{f(\overline{a}_t)\overline{a}_t'V + (-c+\frac{\overline{a}_tA_t}{A_t^2+\gamma})\frac{\overline{a}_tA_t}{A_t^2+\gamma}f(\overline{a}_t)\overline{a}_t' - (-c+\frac{\underline{a}_tA_t}{A_t^2+\gamma})\frac{\underline{a}_tA_t}{A_t^2+\gamma}f(\underline{a}_t)\underline{a}_t'}_{\text{Negative}}$$
$$+\underbrace{h\left[\frac{\underline{a}_tA_t}{A_t^2+\gamma}f(\underline{a}_t)\underline{a}_t' - \frac{\overline{a}_tA_t}{A_t^2+\gamma}f(\overline{a}_t)\overline{a}_t' - cf(\underline{a}_t)\underline{a}_t' + \int_{\underline{a}_t}^{\overline{a}_t}\frac{bA_t}{(A_t^2+\gamma)^2}f(b)db\right]}_{\text{Positive}}$$
$$\underbrace{-\int_{\underline{a}_t}^{\overline{a}_t}\frac{bA_t}{(A_t^2+\gamma)^2}\left[\frac{2bA_t}{A_t^2+\gamma}-c\right]f(b)db}_{\text{Positive}} \geq 0$$

To see that the middle term is positive, note that

$$\overline{r}_t = cA_t < \frac{\underline{a}_tA_t^2}{A_t^2+\gamma} \implies c < \frac{\underline{a}_tA_t}{A_t^2+\gamma} \implies cf(\underline{a}_t)\underline{a}_t' < \frac{\underline{a}_tA_t}{A_t^2+\gamma}f(\underline{a}_t)\underline{a}_t' < \frac{\underline{a}_tA_t}{A_t^2+\gamma}[f(\underline{a}_t)\underline{a}_t' - f(\overline{a}_t)\overline{a}_t']$$

Since $\overline{a}_t' < 0 < \underline{a}_t'$, this gives the result. Moreover, differentiating the expression above with respect to h yields

$$\left[\frac{\underline{a}_tA_t}{A_t^2+\gamma}f(\underline{a}_t)\underline{a}_t' - \frac{\overline{a}_tA_t}{A_t^2+\gamma}f(\overline{a}_t)\overline{a}_t' - cf(\underline{a}_t)\underline{a}_t' + \int_{\underline{a}_t}^{\overline{a}_t}\frac{bA_t}{(A_t^2+\gamma)^2}f(b)db\right] > 0$$

suggesting, by Topkis' Monotonicity Theorem, that $\gamma^*(h)$ is nondecreasing in h. Finally, note that

at $h = 0$, the sign of the derivative is negative everywhere, suggesting that the optimal γ is a corner solution at its lower bound, i.e. $\gamma = V$. As h grows arbitrarily large, for all γ, the middle positive term grows arbitrarily large, suggesting that the derivative is everywhere positive and we end up at a corner solution with γ as high as possible. This upper bound is defined by the point where $\underline{a}_t = \underline{i}_t$, which gives the second part of the result; for γ larger than this value, welfare does not change, since we remain in E^{NL}, where utility does not change.

□

Proposition 5. *We converge to a steady state.*

Proof. First, if A_t is such that we begin in E^{NL}, we are already in a steady state. Hence, suppose we begin in E^L.

Next, if $A_{t+1} \neq A_t$, then it must be that $A_{t+1} < A_t$; to see this, note that $A_{t+1} \neq A_t$ only if an employee matches with a manager of type $bin[\underline{a}_t, \overline{a}_t]$, which results in an action $\frac{bA_t^2}{A_t^2+\gamma}$ being played; in this case, $A_{t+1} = \frac{bA_t^2}{A_t^2+\gamma} < A_t$. Hence, by the monotone convergence theorem, each sequence of A_ts converges.

Finally, fixing stuff, suppose A^* is not a steady state. Can happen if and only if $\overline{a}_t^* > a^*$. But then with positive probability, it changes.

Fix a^*. Suppose A_t does not converge to a steady state. This can happen only if, as $j \to \infty$, A_{t+j} is such that we remain in E^L. Additionally, this can happen only if, as $j \to \infty$, $\overline{a}_{t+j} > a^*$

$A_{t+1} \neq A_t$ with probability 1 if and only if $a^* \leq \overline{a}_t$. Pick the sequence of A_ts s.t. doesn't change.

Suppose that as $t \to \infty$, A_t does converge to a steady state. This implies that

With probability $1 - \frac{\overline{a}_t}{A_t}$, $a^* > \overline{a}_t$, and we are already in a steady state. □

Proposition 6. *Suppose A_t decreases marginally to A_{t+1}. Conditional on matching with $b \in [0, \overline{r}_{t+1}]$, employee welfare decreases. Moreover,*

- *If we begin and stay in E^{NL}, welfare decreases if h is small, that is, if (P1) holds.*
- *If we begin in E^{NL} and move to E^L, welfare decreases if h is sufficiently large. That is,*
  - *welfare decreases if (S) holds for $b \in [\overline{r}_t, \underline{a}_{t+1}]$*
  - *and (P2) for $b \in [\underline{i}_t, \overline{a}_{t+1}]$.*

- *If we begin in E^L and stay in E^L, welfare decreases if h is sufficiently small and $\underline{a}_t < \underline{a}_{t+1}$. In particular, it decreases if*
  - *$\underline{a}_{t+1} < \underline{a}_t$ and (S) holds;*
  - *$\underline{a}_t < \underline{a}_{t+1}$ and (S) does not hold,*
  - *(I) holds for $b \in [\max\{\underline{a}_t, \underline{a}_{t+1}\}, \overline{a}_{t+1}]$, and (P2) holds for $b \in [\overline{a}_{t+1}, \overline{a}_t]$.*
- *If we begin in E^L and move to E^{NL}, welfare decreases if h is sufficiently small. That is, welfare decreases if*
  - *$\underline{i}_{t+1} < \underline{a}_t$ and (P1) holds,*
  - *$\underline{a}_t < \underline{i}_{t+1}$ and (S) does not hold for $b \in [\underline{a}_t, \underline{i}_{t+1}]$, and*
  - *and (P2) holds for $b \in [\max\{\underline{i}_{t+1}, \underline{a}_t\}, \overline{a}_t]$.*

Otherwise, welfare changes are ambiguous.

Proof. A decrease in A_t causes a downward shift in all the major thresholds of the model with the exception of $\underline{a}_t$, whose shift is ambiguous.

Since we are analyzing a marginal decrease, we have $\overline{r}_{t+1} < \overline{r}_t < \min\{\underline{a}_{t+1}, \underline{i}_{t+1}\}$. Welfare conditional on interacting with $b \leq \overline{r}_{t+1}$ decreases, since $-\frac{b}{A_t} > -\frac{b}{A_{t+1}}$. Additionally, welfare on $[\overline{r}_{t+1}, \overline{r}_t]$ also decreases, since, on this range, $-\frac{b}{A_t} > -c$.

$E^{NL} \to E^{NL}$ Next, suppose we begin in E^{NL} and stay in E^{NL}. Welfare conditional on $b \in [\overline{r}_t, \underline{i}_{t+1}]$ remains the same; it is $V - ch$ in both cases. Welfare on $[\underline{i}_{t+1}, \underline{i}_t]$ decreases if and only if $V - ch \geq 0$; employees go from being interacted with and receiving $V - ch$ to no interaction.

$E^{NL} \to E^L$ Suppose we begin in E^{NL} and move to E^L. By a similar argument as above, welfare conditional on $b \in [\overline{r}_t, \underline{a}_{t+1}]$ remains the same. Managers on $[\underline{a}_{t+1}, \underline{i}_t]$ previously kept employees

indifferent but now interact; welfare from employeees decreases here if and only if

$$V + (-h - c + \frac{bA_{t+1}}{A_{t+1}^2 + \gamma})(\frac{bA_{t+1}}{A_{t+1}^2 + \gamma}) \leq V - ch$$
$$(-c + \frac{bA_{t+1}}{A_{t+1}^2 + \gamma})(\frac{bA_{t+1}}{A_{t+1}^2 + \gamma}) \leq (-c + \frac{bA_{t+1}}{A_{t+1}^2 + \gamma})h$$
$$h \geq (\frac{bA_{t+1}}{A_{t+1}^2 + \gamma})$$

Finally, employees who match with a manager with $b \in [\underline{i}_t, \overline{a}_{t+1}]$ go from no interaction to interaction at $a_t^\dagger$. Their welfare decreases if and only if $V + (-h - c + \frac{bA_{t+1}}{A_{t+1}^2+\gamma})(\frac{bA_{t+1}}{A_{t+1}^2+\gamma}) \leq 0$ on this range.

$E^L \to E^L$ Now, suppose we start in E^L and stay in E^L. We either have $\overline{r}_t < \underline{a}_{t+1} < \underline{a}_t < \overline{a}_{t+1} < \overline{a}_t$, or $\overline{r}_t < \underline{a}_t < \underline{a}_{t+1} < \overline{a}_{t+1} < \overline{a}_t$. Utility on $[\overline{r}_t, \min\{\underline{a}_t, \underline{a}_{t+1}\}]$ does not change, following the argument above.

If $\underline{a}_{t+1} < \underline{a}_t$, on $[\underline{a}_{t+1}, \underline{a}_t]$, employees go from experiencing the reporting threshold to the interior action $a_t^\dagger$. Welfare decreases conditional on matching with $b \in [\underline{a}_{t+1}, \underline{a}_t]$ if and only if

$$V + (-h - c + \frac{bA_{t+1}}{A_{t+1}^2 + \gamma})(\frac{bA_{t+1}}{A_{t+1}^2 + \gamma}) \leq V - ch$$
$$h \geq (\frac{bA_{t+1}}{A_{t+1}^2 + \gamma})$$

If $\underline{a}_t < \underline{a}_{t+1}$, the opposite holds; welfare decreases conditional on matching with a manager of type $b \in [\underline{a}_t, \underline{a}_{t+1}]$ if and only if $h \leq (\frac{bA_{t+1}}{A_{t+1}^2+\gamma})$.

Next, welfare on $[\max\{\underline{a}_t, \underline{a}_{t+1}\}, \overline{a}_{t+1}]$ decreases if and only if $(-h - c + \frac{bA_{t+1}}{A_{t+1}^2+\gamma})(\frac{bA_{t+1}}{A_{t+1}^2+\gamma}) \leq (-h - c + \frac{bA_t}{A_t^2+\gamma})(\frac{bA_t}{A_t^2+\gamma})$; in both cases, employees experience interaction above the reporting threshold,

but with the drop to A_{t+1}, they are less intense. This condition can be simplified as follows:

$$\begin{aligned}
(-h-c+\frac{bA_{t+1}}{A_{t+1}^2+\gamma})(\frac{bA_{t+1}}{A_{t+1}^2+\gamma}) &\leq (\frac{bA_t}{A_t^2+\gamma}-c-h)(\frac{bA_t}{A_t^2+\gamma}) \\
h\big(\frac{bA_t}{A_t^2+\gamma}-\frac{bA_{t+1}}{A_{t+1}^2+\gamma}\big) &\leq (\frac{bA_t}{A_t^2+\gamma}-c)(\frac{bA_t}{A_t^2+\gamma})-(\frac{bA_{t+1}}{A_{t+1}^2+\gamma}-c)(\frac{bA_{t+1}}{A_{t+1}^2+\gamma}) \\
h &\leq \frac{(\frac{bA_t}{A_t^2+\gamma}-c)\frac{bA_t}{A_t^2+\gamma}-(\frac{bA_{t+1}}{A_{t+1}^2+\gamma}-c)\frac{bA_{t+1}}{A_{t+1}^2+\gamma}}{\frac{bA_t}{A_t^2+\gamma}-\frac{bA_{t+1}}{A_{t+1}^2+\gamma}} \\
h &\leq \frac{(\frac{bA_t}{A_t^2+\gamma})^2-(\frac{bA_{t+1}}{A_{t+1}^2+\gamma})^2+c(\frac{bA_t}{A_t^2+\gamma}-\frac{bA_{t+1}}{A_{t+1}^2+\gamma})}{\frac{bA_t}{A_t^2+\gamma}-\frac{bA_{t+1}}{A_{t+1}^2+\gamma}} \\
h &\leq \frac{bA_t}{A_t^2+\gamma}+\frac{bA_{t+1}}{A_{t+1}^2+\gamma}-c
\end{aligned}$$

On $[\overline{a}_{t+1}, \overline{a}_t]$, welfare decreases if and only if only if $0 \leq V + (-h-c+\frac{bA_t}{A_t^2+\gamma})(\frac{bA_t}{A_t^2+\gamma})$; managers here used to play $a_t^\dagger$ but now no longer interact.

$E^L \to E^{NL}$ Finally, suppose we start in E^L and move to E^{NL}. This means $\overline{r}_{t+1} < \underline{i}_{t+1} < \underline{a}_t < \overline{a}_t$ or $\overline{r}_{t+1} < \underline{a}_t < \underline{i}_{t+1} < \underline{a}_{t+1} < \overline{a}_t$.

If $\underline{i}_{t+1} < \underline{a}_t$, welfare decreases on $[\underline{i}_{t+1}, \underline{a}_t]$ if and only if no interaction is worse than experiencing an action at the reporting threshold, i.e. if and only if $V - ch \geq 0$. If $\underline{a}_t < \underline{i}_{t+1}$, welfare decreases if and only if now experiencing actions at the reporting threshold is worse than experiencing an action above the reporting threshold, i.e. if and only if

$$\begin{aligned}
V - ch &\leq V + (-h-c+\frac{bA_{t+1}}{A_{t+1}^2+\gamma})(\frac{bA_{t+1}}{A_{t+1}^2+\gamma}) \\
(-c+\frac{bA_{t+1}}{A_{t+1}^2+\gamma})h &\leq (-c+\frac{bA_{t+1}}{A_{t+1}^2+\gamma})(\frac{bA_{t+1}}{A_{t+1}^2+\gamma}) \\
h &\leq (\frac{bA_{t+1}}{A_{t+1}^2+\gamma})
\end{aligned}$$

Finally, on $[\max\{\underline{i}_{t+1}, \underline{a}_t\}, \overline{a}_t]$, welfare decreases if and only if no interaction is worse than the previous interior actions, i.e. if and only if $V + (-h-c+\frac{bA_t}{A_t^2+\gamma})(\frac{bA_t}{A_t^2+\gamma}) \geq 0$, as above.

□

Corollary 3. *Fixing all other parameters, suppose A_t falls to A_{t+1}, such that A_t corresponds to E^L and A_{t+1} to E^{NL}. If h is sufficiently small, welfare decreases.*

Proof. Note that the previous proposition only considers marginal decreases, i.e. assumes $\underline{i}_{t+1}$ does not fall below $\overline{r}_t$, so we briefly address these cases. If $\overline{r}_{t+1} < \overline{r}_t < \underline{i}_{t+1} < \underline{a}_t$ or $\overline{r}_{t+1} < \overline{r}_t < \underline{a}_t < \underline{i}_{t+1}$, the previous proposition applies. If $\overline{r}_{t+1} < \underline{i}_{t+1} < \overline{r}_t < \underline{a}_t$, welfare on $[\overline{r}_{t+1}, \underline{i}_{t+1}]$ decreases. On $[\underline{i}_{t+1}, \overline{r}_t]$, it decreases if $V - \frac{b}{A_t}h \geq 0$. On $[\overline{r}_t, \underline{a}_t]$, it decreases if $V - ch \geq 0$. Note that if $V - ch \geq 0$, this is a sufficient condition for $V - \frac{b}{A_t}h \geq 0$

Next, since we are switching from E^L to E^{NL}, we necessarily have $\underline{a}_t < \underline{i}_t$ but $\underline{i}_{t+1} < \underline{a}_{t+1}$, which means that

$$c\frac{\gamma + \sqrt{\gamma(A_t^2 + \gamma)}}{A_t} < \sqrt{V} < c\frac{\gamma + \sqrt{\gamma(A_{t+1}^2 + \gamma)}}{A_{t+1}}$$

Note that since, as $A_{t+1} \to 0$, $\frac{\gamma+\sqrt{\gamma(A_{t+1}^2+\gamma)}}{A_{t+1}} \to \infty$, that there always exists A_{t+1} small such that this expression holds.

Next, following from the previous proposition, we have that welfare decreases in the following situations:

- $\underline{a}_t < b < \underline{i}_{t+1}$ and $h \leq \frac{bA_t}{A_t^2+\gamma}$ OR $\underline{i}_{t+1} < b < \underline{a}_t$ and $V - ch \geq 0$
- $\max\{\underline{i}_{t+1}, \underline{a}_t\} < b < \overline{a}_t$ and $V + (-h - c + \frac{bA_t}{A_t^2+\gamma})(\frac{bA_t}{A_t^2+\gamma}) \geq 0$.

These conditions hold by choosing h sufficiently small without disturbing the initial condition that $c\frac{\gamma+\sqrt{\gamma(A_t^2+\gamma)}}{A_t} < \sqrt{V} < c\frac{\gamma+\sqrt{\gamma(A_{t+1}^2+\gamma)}}{A_{t+1}}$. □

References

Acemoglu, Daron and James A Robinson (2000). Why Did the West Extend the Franchise? Democracy, Inequality, and Growth in Historical Perspective. *The Quarterly Journal of Economics 115*(4), 1167–1199.

Acemoglu, Daron and James A Robinson (2001). A Theory of Political Transitions. *American Economic Review 91*(4), 938–963.

Acemoglu, Daron and James A Robinson (2008). Persistence of Power, Elites, and Institutions. *American Economic Review 98*(1), 267–93.

Adams-Prassl, Abi, Kristiina Huttunen, Emily Nix, and Ning Zhang (2024). Violence against women at work. *The Quarterly Journal of Economics 139*(2), 937–991.

Aghion, Philippe and Matthew O Jackson (2016). Inducing Leaders to Take Risky Decisions: Dismissal, Tenure, and Term Limits. *American Economic Journal: Microeconomics 8*(3), 1–38.

Algara, Carlos and Sharif Amlani (2021). Replication Data for: Partisanship and Nationalization in American Elections: Evidence from Presidential, Senatorial, & Gubernatorial Elections in the U.S. Counties, 1872–2020.

Alonso, Ricardo and Odilon Câmara (2016). Persuading voters. *American Economic Review 106*(11), 3590–3605.

Andeweg, Rudy B., Robert Elgie, Ludger Helms, Juliet Kaarbo, and Ferdinand Muller-Rommel (2020, 07). The Political Executive Returns: Re-empowerment and Rediscovery. In *The Oxford Handbook of Political Executives*. Oxford University Press.

Aquino, Karl and Stefan Thau (2009). Workplace victimization: Aggression from the target's perspective. *Annual review of psychology 60*(1), 717–741.

Arieli, Itai, Yakov Babichenko, and Fedor Sandomirskiy (2022). Bayesian persuasion with media-

tors. *arXiv preprint arXiv:2203.04285*.

Arold, Benjamin W, Elliott Ash, W Bentley MacLeod, and Suresh Naidu (2024). Do Words Matter?

Ash, Elliott and Stephen Hansen (2023). Text algorithms in economics. *Annual Review of Economics 15*(1), 659–688.

Ashworth, Scott, Ethan Bueno De Mesquita, and Amanda Friedenberg (2017). Accountability and Information in Elections. *American Economic Journal: Microeconomics 9*(2), 95–138.

Bac, Mehmet (2018). Wages, performance and harassment. *Journal of Economic Behavior & Organization 145*, 232–248.

Ballotpedia (2024a). List of Current Governors in the United States. Accessed: 2024-09-18.

Ballotpedia (2024b). States with Gubernatorial Term Limits. Accessed: 2024-09-18.

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

Becker, Gary S (1968). Crime and punishment: An economic approach. *Journal of political economy 76*(2), 169–217.

Behn, Robert D (1991). *Governors on governing*. University Press of America.

Beim, Deborah, Alexander V Hirsch, and Jonathan P Kastellec (2014). Whistleblowing and compliance in the judicial hierarchy. *American Journal of Political Science 58*(4), 904–918.

Bernecker, Andreas, Pierre C Boyer, and Christina Gathmann (2021). The role of electoral incentives for policy innovation: Evidence from the us welfare reform. *American Economic Journal: Economic Policy 13*(2), 26–57.

Besley, Timothy and Anne Case (1995). Does Electoral Accountability Affect Economic Policy Choices? Evidence from Gubernatorial Term Limits. *The Quarterly Journal of Economics 110*(3), 769–798.

Bils, Peter and Federica Izzo (2022). Policymaking in Times of Crisis.

Boix, Carles and Milan W Svolik (2013). The Foundations of Limited Authoritarian Government: Institutions, Commitment, and Power-sharing in Dictatorships. *the Journal of Politics 75*(2), 300–316.

Bonica, Adam (2014). Mapping the ideological marketplace. *American Journal of Political Science 58*(2), 367–386.

Brown, Adam R (2010). Are governors responsible for the state economy? Partisanship, blame,

and divided federalism. *The Journal of Politics 72*(3), 605–615.

Buera, Francisco J, Alexander Monge-Naranjo, and Giorgio E Primiceri (2011). Learning the Wealth of Nations. *Econometrica 79*(1), 1–45.

Butler, Daniel M and Joseph L Sutherland (2023). Have State Policy Agendas Become More Nationalized? *The Journal of Politics 85*(1), 351–355.

Callander, Steven (2011). Searching for Good Policies. *American Political Science Review 105*(4), 643–662.

Canes-Wrone, Brandice, Michael C Herron, and Kenneth W Shotts (2001). Leadership and pandering: A theory of executive policymaking. *American Journal of Political Science*, 532–550.

Card, Dallas, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky (2022). Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences 119*(31), e2120510119.

Carlino, Gerald, Thorsten Drautzburg, Robert Inman, and Nicholas Zarra (2023). Partisanship and fiscal policy in economic unions: Evidence from us states. *American Economic Review 113*(3), 701–737.

Caro, Robert A (2012). *The passage of power: The years of Lyndon Johnson*, Volume 4. Knopf.

Caughey, Devin, Yiqing Xu, and Christopher Warshaw (2017). Incremental democracy: The policy effects of partisan control of state government. *The Journal of Politics 79*(4), 1342–1358.

Chalfin, Aaron and Justin McCrary (2017). Criminal deterrence: A review of the literature. *Journal of Economic Literature 55*(1), 5–48.

Chassang, Sylvain and Gerard Padró I Miquel (2019). Crime, intimidation, and whistleblowing: A theory of inference from unverifiable reports. *The Review of Economic Studies 86*(6), 2530–2553.

Chen, Yuyu and David Y Yang (2019). Historical Traumas and the Roots of Political Distrust: Political Inference from the Great Chinese Famine.

Cheng, Ing-Haw and Alice Hsiaw (2022). Reporting Sexual Misconduct in the# MeToo Era. *American Economic Journal: Microeconomics 14*(4), 761–803.

Coffey, Daniel (2005). Measuring gubernatorial ideology: A content analysis of state of the state speeches. *State Politics & Policy Quarterly 5*(1), 88–103.

Cortina, Lilia M and Maira A Areguin (2021). Putting people down and pushing them out: Sexual

harassment in the workplace. *Annual Review of Organizational Psychology and Organizational Behavior 8*(1), 285–309.

Cox, Gary W (2017). Political Institutions, Economic Liberty, and the Great Divergence. *The Journal of Economic History 77*(3), 724–755.

Curry, James M. and Frances E. Lee (2020). *The Limits of Party*. Chicago: University of Chicago Press.

de Benedictis-Kessner, Justin, Daniel Jones, and Christopher Warshaw (2024). How partisanship in cities influences housing policy. *American Journal of Political Science*.

de Benedictis-Kessner, Justin and Christopher Warshaw (2020). Politics in forgotten governments: the partisan composition of county legislatures and county fiscal policies. *The Journal of Politics 82*(2), 460–475.

De Mesquita, Bruce Bueno and Alastair Smith (2010). Leader Survival, Revolutions, and the Nature of Government Finance. *American Journal of Political Science 54*(4), 936–950.

de Mesquita, Ethan Bueno and Wioletta Dziuda (2024). Partisan Traps.

DellaVigna, Stefano and Woojin Kim (2022). Policy diffusion and polarization across US states. Technical report, National Bureau of Economic Research.

Dennison, Tracy (2006). Did Serfdom Matter? Russian Rural Society, 1750–1860. *Historical Research 79*(203), 74–89.

Dennison, Tracy (2011). *The Institutional Framework of Russian Serfdom*. Cambridge University Press.

Dennison, Tracy (2014). *The Institutional Context of Serfdom in England and Russia* (NED - New edition ed.)., pp. 249–268. Boydell & Brewer.

Dennison, Tracy (2020). Overcoming Institutional Inertia: Serfdom, the State and Agrarian Reform in Prussia and Russia. In *A History of the European Restorations*, Volume 2, pp. 188–201. Bloomsbury Academic.

Dennison, Tracy (2023). Weak State, Strong Commune: Rural Authority in Imperial Russia. *Rivista Storica Italiana 135*(2), 595–622.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR abs/1810.04805*.

Dewan, Torun and Rafael Hortala-Vallve (2019). Electoral Competition, Control and Learning.

British Journal of Political Science 49(3), 923–939.

Diamond, Jared M (2005). *Guns, Germs, and Steel: The Fates of Human Societies*. Norton.

Diermeier, Daniel, Michael Keane, and Antonio Merlo (2005). A political economy model of congressional careers. *American Economic Review 95*(1), 347–373.

DiLeo, Daniel (1997). Dynamic Representation in the United States: Effects of the Public Mood on Governors' Agendas. *State and Local Government Review 29*(2), 98–109.

Dincecco, Mark and Yuhua Wang (2018). Violent Conflict and Political Development over the Long Run: China versus Europe. *Annual Review of Political Science 21*, 341–358.

Dower, Paul Castañeda, Evgeny Finkel, Scott Gehlbach, and Steven Nafziger (2018). Collective Action and Representation in Autocracies: Evidence from Russia's Great Reforms. *American Political Science Review 112*(1), 125–147.

Dower, Paul C and Andrei Markevich (2019). The Stolypin Reform and Agricultural Productivity in Late Imperial Russia. *European Review of Economic History 23*(3), 241–267.

Dur, Robert AJ (2001). Why Do Policy Makers Stick to Inefficient Decisions? *Public Choice 107*(3-4), 221–234.

Eliaz, Kfir and Ran Spiegler (2020). A Model of Competing Narratives. *American Economic Review 110*(12), 3786–3816.

Ely, Jeffrey C and Juuso Välimäki (2003). Bad Reputation. *The Quarterly Journal of Economics 118*(3), 785–814.

Emmons, Terence LeRoy (1966). *The Russian Landed Gentry and the Peasant Emancipation of 1861*. University of California, Berkeley.

Evans, Robert John Weston and Hartmut Pogge Von Strandmann (2002). *The Revolutions in Europe, 1848-1849: From Reform to Reaction*. Oxford University Press on Demand.

Ferguson, Margaret Robertson (2003). Chief executive success in the legislative arena. *State Politics & Policy Quarterly 3*(2), 158–182.

Ferreira, Fernando and Joseph Gyourko (2009). Do political parties matter? Evidence from US cities. *The Quarterly journal of economics 124*(1), 399–422.

Field, Daniel (1976). *The End of Serfdom: Nobility and Bureaucracy in Russia, 1865-1861*. Harvard University Press.

Finkel, Evgeny, Scott Gehlbach, and Tricia D Olsen (2015). Does Reform Prevent Rebellion? Evi-

dence from Russia's Emancipation of the Serfs. *Comparative Political Studies 48*(8), 984–1019.

Folke, Olle, Johanna Rickne, Seiki Tanaka, and Yasuka Tateishi (2020). Sexual harassment of women leaders. *Daedalus 149*(1), 180–197.

Fu, Qiang and Ming Li (2014). Reputation-Concerned Policy Makers and Institutional Status Quo Bias. *Journal of Public Economics 110*, 15–25.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–574.

Gentzkow, Matthew, Jesse M Shapiro, and Matt Taddy (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica 87*(4), 1307–1340.

Gerschenkron, Alexander (1962). *Economic Backwardness in Historical Perspective (1962)*. The Belknap Press of Harvard University Press.

Goldgeier, James and Elizabeth N Saunders (2018). The unconstrained presidency: checks and balances eroded long before Trump. *Foreign Aff. 97*, 144.

Grossmann, Matt, Marty P Jordan, and Joshua McCrain (2021). The correlates of state policy and the structure of state panel data. *State Politics & Policy Quarterly 21*(4), 430–450.

Grumbach, Jacob M (2018). From backwaters to major policymakers: policy polarization in the states, 1970–2014. *Perspectives on Politics 16*(2), 416–435.

Hansen, Stephen, Michael McMahon, and Andrea Prat (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. *The Quarterly Journal of Economics 133*(2), 801–870.

Heidbreder, Brianne (2012). Agenda setting in the states: How politics and policy needs shape gubernatorial agendas. *Politics & Policy 40*(2), 296–319.

Hersch, Joni (2015). Sexual harassment in the workplace. *IZA world of labor*.

Hersch, Joni (2018). Valuing the risk of workplace sexual harassment. *Journal of Risk and Uncertainty 57*(2), 111–131.

Holmström, Bengt (1999). Managerial Incentive Problems: A Dynamic Perspective. *The Review of Economic Studies 66*(1), 169–182.

Hopkins, Daniel J, Eric Schickler, and David L Azizi (2022). From many divides, one? The polarization and nationalization of American state party platforms, 1918–2017. *Studies in American*

Political Development 36(1), 1–20.

Iaryczower, Matias, Gabriel Lopez-Moctezuma, and Adam Meirowitz (2024). Career concerns and the dynamics of electoral accountability. *American Journal of Political Science 68*(2), 696–713.

Ichihashi, Shota (2019). Limiting Sender's information in Bayesian persuasion. *Games and Economic Behavior 117*, 276–288.

Izzo, Federica (2024). With Friends Like These, Who Needs Enemies? *The Journal of Politics 86*(3), 835–849.

Izzo, Federica, Gregory J Martin, and Steven Callander (2021). Ideological Competition. *SocArXiv. February 19*.

Jacobson, Gary C (2006). The polls: polarized opinion in the states: partisan differences in approval ratings of governors, senators, and George W. Bush. *Presidential Studies Quarterly 36*(4), 732–757.

Jeffries, John W. (1996). A "Third New Deal"? Liberal Policy and the American State, 1937-1945. *Journal of Policy History 8*(4), 387–409.

Jensen, Jacob, Suresh Naidu, Ethan Kaplan, Laurence Wilse-Samson, David Gergen, Michael Zuckerman, and Arthur Spirling (2012). Political polarization and the dynamics of political language: Evidence from 130 years of partisan speech [with comments and discussion]. *Brookings Papers on Economic Activity*, 1–81.

Judd, Gleason (2017). Showing off: promise and peril in unilateral policymaking. *Quarterly Journal of Political Science 12*(2), 241–268.

Kamenica, Emir and Matthew Gentzkow (2011). Bayesian Persuasion. *American Economic Review 101*(6), 2590–2615.

Kartik, Navin, Francesco Squintani, Katrin Tinn, et al. (2015). Information Revelation and Pandering in Elections. *Columbia University, New York 36*, 8.

Khristoforov, IA (2007). The Fate of Reform: The Russian Peasantry in Government Policy and Public Opinion from the Late 1860s to the Early 1880s. *Russian Studies in History 46*(1), 24–42.

Khristoforov, Igor' (2009). Nineteenth-Century Russian Conservatism: Problems and Contradictions. *Russian Studies in History 48*(2), 56–77.

Khristoforov, Igor (2022). From Speransky to Stolypin: Peasant Reform and the Problem of Land Management. *Herald of the Russian Academy of Sciences 92*(3), S161–S173.

Khristoforov, Igor and Christopher Gilley (2016). Blurred Lines. *Cahiers du Monde Russe 57*(1), 31–54.

Kingston-Mann, Esther (1991). *Peasant Communes and Economic Innovation: A Preliminary Inquiry*, pp. 1–51. Princeton University Press.

Klarner, Carl (2013). State Partisan Balance Data, 1937 - 2011.

Knott, Stephen (2024). George H.W. Bush: Impact and Legacy. Accessed: 2024-10-28.

Kousser, Thad and Justin H Phillips (2012). *The power of American governors: Winning on budgets and losing on policy*. Cambridge University Press.

Kurian, George (2011). Constitutional Monarchy. *The Encyclopedia of Political Science*, 316–317.

Larimer, Christopher W. (2015). Title of Larimer's Article Here. In D. P. Redlawsk (Ed.), *The American Governor: Power, Constraint, and Leadership in The States*. Springer.

Lee, Frances Xu and Wing Suen (2020). Credibility of crime allegations. *American Economic Journal: Microeconomics 12*(1), 220–259.

Leigh, Andrew (2008). Estimating the impact of gubernatorial partisanship on policy settings and economic outcomes: A regression discontinuity approach. *European Journal of Political Economy 24*(1), 256–268.

Levy, Gilat and Ronny Razin (2021a). A Maximum Likelihood Approach to Combining Forecasts. *Theoretical Economics 16*(1), 49–71.

Levy, Gilat and Ronny Razin (2021b). Short-Term Political Memory and the Inevitability of Polarisation. *Unpublished*.

Leyden, Kevin M and Stephen A Borrelli (1995). The effect of state economic conditions on gubernatorial elections: Does unified government make a difference? *Political Research Quarterly 48*(2), 275–290.

Li, Yuan, Mario Gilli, et al. (2014). Accountability in Autocracies: The Role of Revolution Threat. Technical report, Stockholm School of Economics, Stockholm China Economic Research Institute.

Lipnowski, Elliot, Doron Ravid, and Denis Shishkin (2022). Persuasion via weak institutions. *Journal of Political Economy 130*(10), 2705–2730.

Lushkov, Gili (2019). *Governors' Rhetorical Attention to Education: Text Analysis, Historical Perspectives and Building a Gubernatorial Addresses Repository*. Plymouth State University.

Majumdar, Sumon and Sharun W Mukand (2004). Policy Gambles. *American Economic Review 94*(4), 1207–1222.

Manso, Gustavo (2011). Motivating Innovation. *The Journal of Finance 66*(5), 1823–1860.

Markevich, Andrei and Ekaterina Zhuravskaya (2018). The Economic Effects of the Abolition of Serfdom: Evidence from the Russian Empire. *American Economic Review 108*(4-5), 1074–1117.

Maskin, Eric and Jean Tirole (2004). The politician and the judge: Accountability in government. *American Economic Review 94*(4), 1034–1054.

McCarty, Nolan (2019). *Polarization: What Everyone Needs to Know*. Oxford University Press.

Mokyr, Joel (2016). *A Culture of Growth*. Princeton University Press.

Montiel Olea, José Luis, Pietro Ortoleva, Mallesh M Pai, and Andrea Prat (2022, 04). Competing Models. *The Quarterly Journal of Economics*. qjac015.

Moon, David (2014). *The Abolition of Serfdom in Russia: 1762-1907*. Routledge.

Mukand, Sharun W and Dani Rodrik (2005). In Search of the Holy Grail: Policy Convergence, Experimentation, and Economic Performance. *American Economic Review 95*(1), 374–383.

Nafziger, Steven (2010). Peasant Communes and Factor Markets in Late Nineteenth-Century Russia. *Explorations in Economic History 47*(4), 381–402.

Nafziger, Steven (2012). Serfdom, Emancipation, and Off-Farm Labour Mobility in Tsarist Russia. *Economic History of Developing Regions 27*(1), 1–37.

Nafziger, Steven (2016). Communal Property Rights and Land Redistributions in Late Tsarist Russia. *The Economic History Review 69*(3), 773–800.

National Conference of State Legislatures (2024). State Partisan Composition. Accessed: 2024-08-20.

National Governors Association (2018). Governors' Office Onboarding Guide: Legislative Relations. Accessed: 2024-08-16.

National Governors Association (2019, October). Overseeing State Government Operations in the Governor's Office. Technical report, National Governors Association.

National Governors Association (2024). Former Governors. Accessed: 2024-09-18.

Pallot, Judith (1999). *Land Reform in Russia, 1906-1917: Peasant Responses to Stolypin's Project of Rural Transformation*. Clarendon press.

Patty, John W and Ian R Turner (2021). Ex post review and expert policy making: When does

oversight reduce accountability? *The journal of politics 83*(1), 23–39.

Pereira, NGO (1980). Alexander II and the Decision to Emancipate the Russian Serfs, 1855-61. *Canadian Slavonic Papers 22*(1), 99–115.

Peterson, Andrew and Arthur Spirling (2018). Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. *Political Analysis 26*(1), 120–128.

Peterson, Erin (2024). Presidential Power Surges. Accessed: 2024-10-25.

Polunov, Aleksandr, Thomas C Owen, and Larissa Georgievna Zakharova (2015). *Russia in the Nineteenth Century: Autocracy, Reform, and Social Change, 1814-1914: Autocracy, Reform, and Social Change, 1814-1914*. Routledge.

Pomeranz, Kenneth (2021). *The Great Divergence*. Princeton University Press.

Prat, Andrea (2005). The Wrong Kind of Transparency. *American Economic Review 95*(3), 862–877.

Pravilova, Ekaterina (2014). *A Public Empire: Property and the Quest for the Common Good in Imperial Russia*. Princeton University Press.

Prendergast, Canice and Lars Stole (1996). Impetuous Youngsters and Jaded Old-Timers: Acquiring a Reputation for Learning. *Journal of Political Economy 104*(6), 1105–1134.

Pulejo, Massimo and Pablo Querubín (2021). Electoral concerns reduce restrictive measures during the COVID-19 pandemic. *Journal of public economics 198*, 104387.

Reynolds, Molly (2024). Improving Congressional Capacity to Address Problems and Oversee the Executive Branch. Accessed: 2024-10-25.

Riley, Russell (2024). Bill Clinton. Accessed: 2024-10-29.

Roland, Gerard (2002). The Political Economy of Transition. *Journal of Economic Perspectives 16*(1), 29–50.

Rosenthal, Jean-Laurent and R Bin Wong (2011). *Before and Beyond Divergence: The Politics of Economic Change in China and Europe*. Harvard University Press.

Scheppach, Raymond C. (2005, November). What Makes a Successful Governor? *Stateline*.

Schwartzstein, Joshua and Adi Sunderam (2021). Using Models to Persuade. *American Economic Review 111*(1), 276–323.

Shor, Boris and Nolan McCarty (2011). The ideological mapping of American legislatures. *American Political Science Review 105*(3), 530–551.

Siggelkow, Benjamin Florian, Jan Trockel, and Oliver Dieterle (2018). An inspection game of inter-

nal audit and the influence of whistle-blowing. *Journal of Business Economics 88*, 883–914.

Singer, Matthew M (2023). Dynamics of Gubernatorial Approval: Evidence from a New Database. *State Politics & Policy Quarterly 23*(3), 306–326.

Sperber, Jonathan (2005). *The Pre-Revolutionary Political Universe* (2 ed.)., pp. 56?108. New Approaches to European History. Cambridge University Press.

Spiegler, Ran (2016). Bayesian Networks and Boundedly Rational Expectations. *The Quarterly Journal of Economics 131*(3), 1243–1290.

Starr, Frederick S (2015). *Decentralization and Self-Government in Russia, 1830-1870*. Princeton University Press.

Stokey, Nancy L, Robert E Lucas, and Edward C Prescott (1989). *Recursive Methods in Economic Dynamics*. Harvard University Press.

Strong, Robert A. (2024). Jimmy Carter: Domestic Affairs. Accessed: 2024-10-28.

The American Presidency Project (2024). Presidential Statistics. Accessed: 2024-10-29.

Tilly, Charles (1989). State and Counterrevolution in France. *Social Research*, 71–97.

Tomasi, Arduino (2023). The Stakes and Informativeness Trade-Off: Electoral Incentives to Implement Programmatic Transfers. *The Journal of Politics 85*(2), 654–666.

Vasudevan, Hari S (1988). Peasant Land and Peasant Society in Late Imperial Russia. *The Historical Journal 31*(1), 207–222.

Wang, Shaoda and David Y Yang (2021). Policy Experimentation in China: The Political Economy of Policy Learning. Technical report, National Bureau of Economic Research.

Warner, Seth B (2023). Measuring executive ideology and its influence. *State Politics & Policy Quarterly 23*(1), 97–116.

Wcislo, Francis William (2014). *Reforming Rural Russia*. Princeton University Press.

Weinberg, Micah (2010). Measuring governors' political orientations using words as data. *State Politics & Policy Quarterly 10*(1), 96–109.

Williams, Stephen F (2013). *Liberal Reform in an Illiberal Regime: The Creation of Private Property in Russia, 1906-1915*. Number 545. Hoover Institution Press.

Wolak, Jennifer and Srinivas Parinandi (2022). Does the public hold governors accountable for policy outcomes? *Political Research Quarterly 75*(4), 1051–1064.

Wortman, Richard (1989). Property Rights, Populism, and Russian Political Culture. *Civil Rights*

in Imperial Russia, 13–32.

Yaney, George L (1964). The Concept of the Stolypin Land Reform. *Slavic Review 23*(2), 275–293.

Zenkovsky, Serge A (1961). The Emancipation of the Serfs in Retrospect. *Russian Review*, 280–293.

Zhu, John Y (2024). Better monitoring, Worse outcome? *The RAND Journal of Economics 55*(4), 550–572.