The Political Consequences of Police Slowdowns

Arvind Krishnamurthy*

Elisa Wirsching†

Dvir Yogev‡

Abstract

Police increasingly resist reform by strategically reducing service quality to shape electoral outcomes, yet we lack systematic evidence on how this bureaucratic resistance moves reform attitudes among voters and elected officials. We argue that police wield two tools grounded in their day-to-day service provision: *fearmongering*, public warnings that tie a reform to future service decline, and *resistance shirking*, the quiet withdrawal of effort afterward. Because voters cannot easily distinguish policy failure from bureaucratic resistance, the mere prospect of pushback can deter voter and policymaker support for reform. We test this with two survey experiments, alongside observational evidence on the prevalence of police fearmongering and slowdowns. A 2×2 vignette experiment with 1,676 mass-public respondents varies whether a police union signals cooperation or fearmongering and whether post-reform service quality improves or worsens, and an experiment of more than 500 sitting local elected officials varies whether police threaten resistance to a qualified-immunity reform. In the mass experiment, fearmongering is independently costly to politicians regardless of service outcomes, and service quality strongly shapes support for both reforms and incumbents. In the elite experiment, a threatening police response substantially reduces officials' support for reform even though no service decline has occurred. Together, these results show that police possess strategic tools operating across the entire policy pipeline, creating anticipatory political costs for both voters and policymakers.

*Assistant Professor, Ohio State University, krishnamurthy.51@osu.edu

†Assistant Professor, The London School of Economics and Political Science, e.m.wirsching@lse.ac.uk

‡Post-doctoral Researcher, UC Berkeley, dyo@berkeley.edu

1 Introduction

Across dozens of American cities, elected officials who pursue police reforms have met a distinctive form of pushback that works through the everyday quality of policing services. The recent experience of San Francisco is a case in point. In 2019, voters elected Chesa Boudin, a career public defender, as district attorney on a platform of ending cash bail, reducing low-level incarceration, and holding officers accountable in criminal court. As he began implementing that agenda, officer-initiated activity fell sharply, and residents repeatedly complained to city officials and the media that police were no longer responding to crime (Kyriazis, Schechter and Yogev, 2023). On body-worn camera footage, officers reportedly told victims there was nothing they could do and reminded them, "Don't forget to vote in the upcoming recall election" (Heo and Wirsching, 2026). Boudin was recalled after eighteen months, and police activity and responsiveness markedly rebounded (Kyriazis, Schechter and Yogev, 2023). Whether the decline reflected deliberate sabotage, ordinary attrition, or the disruptions of the period is hard to establish from the outside, and voters watching public safety deteriorate could not easily tell a failed reform from a bureaucracy declining to do its job.

We argue that this ambiguity is a source of political power. Because bureaucrats are the actors who deliver public services, they have a distinctive, *indirect* channel of political influence. By changing the quality of the services voters experience and instigating fear of poor public safety, they can shape how voters judge the policies and politicians behind them. This channel operates alongside well-studied forms of power, such as endorsements, campaign contributions, and organized lobbying (Moe, 2006; Carreri, Teso and Yu, 2025; Gaudette, 2024). Police can deploy this influence in two ways. Before a reform is enacted, they can engage in *fearmongering*, warning the public that it will make communities less safe, suggesting that it may reduce proactive enforcement, and motivating them to vote reformers out of office. After a disfavored policy takes effect, they turn to *resistance shirking*, quietly withdrawing effort so that the resulting decline in service shapes how voters judge the officials

responsible. Both work through the same ambiguity: when service deteriorates, neither voters nor officials can tell whether the cause is a poorly designed policy or a bureaucracy withholding effort, and that uncertainty is what gives the threat its force.

This matters for democratic accountability. Elections are meant to discipline elected officials for the quality of governance they deliver. But if voters cannot distinguish policy failure from bureaucratic resistance, and if officials understand that they may be held responsible for outcomes they do not fully control, the mere prospect of police pushback can deter reform long before any reform is attempted. Although scholars increasingly recognize that bureaucrats can resist policies they oppose for political leverage (Wirsching, 2026; Heo and Wirsching, 2026), we still know little about how voters and officials actually interpret these signals and adjust their behavior. What are the political consequences when police threaten, or deliver, declining service in response to an unfavorable policy regime? Can police use the services they provide as a lever to move voters? And does the prospect of such resistance shape whether elected officials are willing to pursue reform in the first place?

These questions are difficult to answer with observational data. First, bureaucratic resistance is hard to observe, takes many different forms across departments, is only imperfectly captured by any single measure, and inherently difficult to disentangle from confounding factors (just as it is hard for voters). Second, anticipatory behavior by politicians complicates causal identification of electoral and behavioral consequences. Reform-minded politicians may avoid or even double down on certain policies precisely because they expect bureaucratic resistance, or they may design policies differently when resistance is anticipated (Heo and Wirsching, 2026).

We therefore turn to two experiments that can hold the underlying policy fixed and vary only the police signal. We first use a survey experiment on the mass public, which manipulates whether police are aligned with a reform and whether service quality improves or worsens after it. We find that fearmongering lowers voters' support for both the reform and the incumbent, whatever the service outcome, and realized service quality strongly

shapes both evaluations. Second, we field an experiment on local elected officials, which varies whether police meet a proposed reform with cooperation or with the threat of reduced service. A threatening response sharply lowers officials' own willingness to back the reform, even though no service decline has occurred. We situate these experiments within descriptive evidence on how commonly Americans encounter fearmongering and police slowdowns to establish that the behavior we study is real and widespread. Taken together, the studies suggest that police can shape the politics of reform both before and after reform is tried.

Our analyses make three contributions. First, we extend the study of bureaucratic politics. Research on police as political actors has focused on formal tools of influence, including union endorsements (Gaudette, 2024; Zoorob, 2019; Boudreau, MacKenzie and Simmons, 2019, 2022; Vaughn, Peyton and Huber, 2022), campaign contributions (Carreri, Teso and Yu, 2025), and collective bargaining (Anzia and Trounstine, 2024), all of which can reshape the electoral fortunes of elected principals (Moe, 2006). We turn instead to informal channels—fearmongering and resistance shirking—and to a different valence. Where existing work studies positive influence such as endorsements, we ask what happens when support turns to opposition and cooperation turns to active resistance.

Second, we advance the politics of policing. Work on how police behavior moves public opinion has concentrated on acute incidents of police violence and attitudes toward the police (Reny and Newman, 2021; Sances, 2023; Boudreau, MacKenzie and Simmons, 2019; Mullinix and Norris, 2019; Ang and Tebes, 2024; Burch, 2022). Instead, we shift from high-profile use-of-force to routine performance, such as 911 response times and crime rates,[1] and we ask whether elected principals, not just the police, are held accountable for that performance. The latter connects us to work on how service provision shapes principals' electoral fortunes (Martin and Raffler, 2021; Burnett and Kogan, 2017; Hopkins and Pettingill, 2018; Slough, 2024).

Third, we contribute to the policy feedback literature (Pierson, 1993; Campbell, 2004;

[1]See Arnold and Carnes (2012); Heo and Wirsching (2026); Hopkins and Pettingill (2018) for notable exceptions that consider retrospective evaluation after shifts in crime or other policing outcomes.

Hacker and Pierson, 2019; Weaver and Lerman, 2010; Soss and Schram, 2007; White, 2016; Walker, 2020). This work generally treats the policy experience as flowing directly from design to its targets. Joining recent research on the traceability of outcomes to government actors (Hamel, 2025; Jares and Malhotra, 2025), we show that politically motivated bureaucrats can strategically shape those experiences. Feedback effects may thus depend as much on bureaucratic behavior as on the policies themselves.

2 Theoretical Framework

Our argument begins from a familiar problem in democratic accountability: voters use public services to evaluate the quality of elected officials, but elected officials do not produce those services alone. Instead, politicians delegate implementation of public services to bureaucratic agents who possess expertise, discretion, and private information about their own effort (McCubbins and Schwartz, 1984; Lipsky, 2010). In policing, this delegation problem is especially acute. Patrol officers decide how quickly to respond to calls for service, how much proactive enforcement to undertake, and how much investigative effort to put into crime solving, among other decisions. A resident may observe that response times have grown longer or that crime feels less controlled, but she usually cannot observe whether that deterioration reflects a bad reform, resource constraints, ordinary administrative friction, or officers strategically withholding effort.

We argue that this information problem gives the police a distinctive source of political power. Standard principal-agent accounts emphasize the difficulty principals face in controlling agents with hidden information or hidden action (Moe, 1984; Kiewiet and McCubbins, 1991; Miller, 2005). But in democratic settings, because principals are elected, agents can also act politically against the principals who are supposed to control them (Moe, 2006). Police unions endorse candidates, lobby over work rules, contribute to campaigns, and bargain over contracts (Carreri, Teso and Yu, 2025; Gaudette, 2024; Anzia and Trounstine, 2024).

We, instead, focus on an indirect channel. Because voters infer the quality of a reform and its sponsors from outcomes officers help produce, a department that opposes a reform can quietly degrade service to turn voters against it, without any formal say over policy.

The key theoretical object is how voters update when service quality is jointly produced. Voters care about two things. They want to know whether a reform is good policy, such as whether a civilian oversight board, a budget cut, or a qualified-immunity reform improves or weakens public safety. And they want to know whether incumbent officials are competent, whether they chose wisely and managed implementation well. Service quality is an important signal about both. When public safety improves after a reform, voters should become more favorable toward the reform and toward the officials who adopted it. When service deteriorates, voters should become less favorable. This is the basic accountability logic linking bureaucratic performance to electoral evaluation (Ujhelyi, 2014; Martin and Raffler, 2021; Foarta, 2023; Slough, 2024).

But this signal becomes politically unstable when voters know that bureaucrats may be strategic. Police, like other bureaucrats, are organized actors with preferences over the policies that govern their work. Importantly, prior research on police supervision shows that officer effort varies with incentives, supervision, organizational culture, and solidarity among subordinates (Brehm and Gates, 1993, 1999; Mummolo, 2018). Recent work extends this insight to the political arena, showing how bureaucrats can exploit voter uncertainty by reducing service quality under reform-minded incumbents (Heo and Wirsching, 2026; Wirsching, 2026; Kyriazis, Schechter and Yogev, 2023). In this setting, the same bad outcome can support two interpretations. It may reveal that reform was harmful, or it may reveal that officers refused to faithfully implement it. Democratic accountability then depends on how voters allocate responsibility under this uncertainty.

We suggest that there are two tools that police use to exploit this challenge of democratic accountability: *fearmongering*, where police signal about the likely failure of some policy reform, and *resistance shirking*, where police reduce service quality in order to reshape public

views about politicians and policy quality. Below, we summarize the theoretical logic of these strategies and how they may affect voter updating.

2.1 Fearmongering as an Ex Ante Signal

Fearmongering is an ex ante signal that links a reform to future service deterioration and assigns political responsibility for that deterioration to elected officials. A police union might argue that elected officials decision to end qualified immunity will make officers hesitate before pursuing suspects, that reducing a police budget will leave 911 calls unanswered, or that a progressive prosecutor's policies will produce a crime wave. These are not mere theoretical possibilities. For example, the International Association of Chiefs of Police has warned that eliminating qualified immunity would have a "chilling effect" on officers' willingness to respond to 911 calls (International Association of Chiefs of Police, 2020); the Portland Police Union warned that budget cuts would mean "longer waits for crime victims" (Templeton, 2026); and the union representing New York City's police detectives predicted that reformer District Attorney Alvin Bragg's new charging policies would drive up crime and shootings (Newsday, 2022). These claims may be presented as predictions rather than threats, but they convey two pieces of information: police oppose the policy, and public safety may deteriorate if officials proceed.

Fearmongering can take multiple forms, but two archetypes are especially important for our argument. The first is *policy fearmongering*. In policy fearmongering, police claim that a reform will produce bad outcomes because of what the policy itself does to public safety. For example, a police union might argue that weakening qualified immunity will expose officers to lawsuits and make them hesitate in the field,[2] or that a budget cut will leave departments unable to answer 911 calls. Here, the threat of resistance shirking is often implicit; the message is framed as a prediction about policy quality or implementation, but does carry

[2]This example reflects the common public argument against qualified-immunity reform: that the doctrine protects officers from personal liability and that weakening it could make law enforcement work more difficult. See, for example, Bates (2021) on the claims that qualified immunity allows officers to do their jobs without fear of financial ruin from lawsuits.

some possible signal of forthcoming reduced effort. The second is *shirking fearmongering*. In shirking fearmongering police more explicitly signal that officers will reduce effort if the reform policy is adopted. The claim is not simply that this policy will fail, but that officers will pull back, slow down, stop engaging in proactive policing, and "*make* elected officials bear the blame for residents who fall through the cracks."[3] Both these forms of fearmongering differ from pure opposition because they do more than state that police dislike a reform. They forecast a deterioration in public safety and attach that deterioration to the officials who proceed.

We suggest that these classes of fearmongering can affect voters through two key mechanisms. First, fearmongering can reduce the perceived informativeness of future service outcomes. If voters know that officers may resist, a later decline in service is no longer a clean signal of policy failure. A pro-reform voter observing worse response times may reason that the policy never received a fair test. Conversely, a reform skeptic observing good outcomes may discount the apparent success as temporary, lucky, or unrelated to the policy because the implementation environment was noisy. In this sense, fearmongering can make voters rely more heavily on their prior views about the reform, a logic consistent with models in which bureaucratic resistance lowers the informational value of government performance (Heo and Wirsching, 2026).

Second, fearmongering can impose a direct political cost by making institutional conflict itself salient. Voters may not only ask "Whose fault was this specific outcome?" They may also ask "Will this city be governable if officials and police are at war, and police are not likely to work hard?" A warning that officers will hesitate, slow down, or scale back activity can generate fear of de-policing even before service has worsened. On this view, it is enough for voters to believe that conflict with police will make public safety harder to deliver. In San Francisco, for example, residents did not need to resolve whether declining police activity

[3]This kind of language is often used by police. For example NYPD Police Benevolent Association President Pat Lynch said that the city council, mayor's office, and state legislature had "handcuffed police officers and given the street back to the criminals" in the summer of 2020. https://www.foxnews.com/media/pat-lynch-nypd-de-blasio-crime-violence.amp

under Chesa Boudin reflected deliberate sabotage, ordinary staffing problems, or pandemic disruption for the issue to become politically damaging (Kyriazis, Schechter and Yogev, 2023; Knight, 2021; Swan, 2021; Pearson, 2023). The uncertainty itself was part of the political environment.[4]

These mechanisms generate our first set of expectations. Voters exposed to fearmongering should expect worse service provision and a greater likelihood of officer resistance. In turn, fearmongering is likely to reduce support for the reform and for incumbent officials directly, because it makes future implementation appear costly. At the same time, fearmongering should also alter how voters interpret later performance. If the reduced-informativeness mechanism dominates, voters should be more forgiving of poor outcomes when they have already been warned that police may resist. If the prospective-conflict mechanism dominates, fearmongering may instead be independently costly even when outcomes are good, because voters see the reform as likely to generate continuing conflict and portend poor police performance.

These arguments imply several empirical expectations about fearmongering. *H1 [expectations]*: voters exposed to fearmongering should expect police resistance to be more likely and future service quality to be worse. *H3a [reduced informativeness]*: if fearmongering primarily works by reducing the informativeness of later performance signals, then realized service quality should matter less for voter evaluations under fearmongering; both good and bad outcomes become less diagnostic of the reform's true quality. *H3b [prospective conflict]*: if fearmongering instead works by making institutional conflict and future de-policing salient, then it should directly reduce support for the reform and for incumbent officials, even when realized service quality is good. *H4a–H4d [prior amplification]*: because fearmongering makes outcomes harder to interpret, voters should rely more heavily on their prior views;

[4]Certainly, fearmongering can change expectations about implementation quality. If voters learn that the officers charged with carrying out a policy oppose it, they may expect worse service quality and a higher probability of shirking. This is a straightforward inference about agency loss: even a good reform may fail if the implementing bureaucracy is hostile to it. This component of fearmongering, however, merely reflects the effect of learning about police opposition to a policy. This is theoretically distinct from the institutional conflict or informativeness mechanisms that we suggest characterize fearmongering.

reform supporters should become more supportive and more forgiving of poor outcomes, while reform opponents should become less supportive and more skeptical of good outcomes.

2.2 Resistance Shirking and Ex Post Updating

The second strategy is *resistance shirking*: the ex post reduction of effort after a disfavored policy is adopted. Traditional shirking is a dyadic agency problem between supervisor and subordinate. Resistance shirking is triadic. Officers reduce effort in a way that affects voters' experience of public services, and voters may then punish the elected officials associated with the reform (Kyriazis, Schechter and Yogev, 2023; Heo and Wirsching, 2026; Wirsching, 2026). The point is not merely to avoid work, but to make the elected principal politically vulnerable through the principal's relationship with the public.

This strategy is attractive because many forms of police effort are hard to verify. Officers can write fewer tickets, initiate fewer stops, respond more slowly, or decline to engage in discretionary enforcement while offering plausible explanations for these changes: staffing shortages, legal exposure, low morale, or new procedural constraints. In Minneapolis, a former council member described officers arriving late to calls and telling constituents to ask their council member why help could not come sooner (Blumgart, 2020). In Atlanta, roughly 170 officers called out sick in the days after the Fulton County district attorney filed murder charges against the officer who killed Rayshard Brooks in June 2020, a coordinated "Blue Flu" that thinned downtown coverage even as the police union publicly disavowed it (The Atlanta Journal-Constitution, 2020). In Philadelphia, arrests declined during the city's protracted conflict with reform prosecutor Larry Krasner; police leadership attributed the drop to the pandemic and a national post-Floyd pullback, while Krasner's critics blamed his policies and Krasner denied any link to rising crime (The Philadelphia Inquirer, 2022). And in San Francisco, declining police activity under Chesa Boudin was read as deliberate resistance, even as staffing shortages and pandemic disruption offered competing explanations (Kyriazis, Schechter and Yogev, 2023). In each case, the same decline in service could sustain more

than one story. That contestation and ambiguity are what allow a service decline to become politically useful, leaving voters with the challenging task of apportioning blame for poor service quality.

For voters, first, realized service quality provides direct welfare information. Residents care about response times, crime victimization, and visible disorder regardless of who caused them. Even if a voter suspects that police are partly responsible for deterioration, she may still penalize elected officials for failing to maintain effective governance. While voters may be less able to learn about the policy quality, they do learn something about the elected officials' ability to maintain effective governance. That information may be relevant in their determination of a politician's quality.

Second, service quality can inform voters about policy quality. The complication is attribution. Voters with favorable priors toward reform should be more likely to attribute poor outcomes to police resistance and to treat police opposition as evidence that reform threatens entrenched interests. Voters with unfavorable priors should be more likely to attribute poor outcomes to the reform itself and to treat police warnings as vindicated. Thus, the same service decline may polarize interpretations even if it lowers average support (Heo and Wirsching, 2026).

The empirical implication here is that service quality should have strong main effects on attitudes toward both policies and politicians, but these effects may depend on prior beliefs and exposure to fearmongering. *H2a–H2b [performance updating]*: because voters use public services as signals of policy and politician quality, good service should increase support for the reform and the incumbent, while poor service should reduce support for both. *H5 [blame attribution]*: when service deteriorates, voters should differ in how they allocate responsibility; reform supporters should be more likely to blame police for poor outcomes, especially when fearmongering has made resistance salient, whereas reform opponents should be more likely to blame the reform itself. The attribution gap between supporters and opponents should therefore widen under fearmongering.

2.3 Politicians' Anticipatory Choices

Politicians observe this same environment before deciding whether to pursue reform. A local elected official considering qualified-immunity reform, civilian oversight, or a budget cut must weigh not only the policy's substantive merits but also the likelihood that police will resist and that voters will hold the official responsible for any deterioration in service. Fearmongering can therefore deter reform without any actual slowdown occurring, especially if there is a history of slowdowns within a jurisdiction or police force. If the threat is credible enough, the official may narrow the proposal, delay it, or abandon reform altogether. This logic could (partially) explain how, for example, 2020 saw unprecedented levels of activism and attitudinal shifts towards policing (Reny and Newman, 2021), yet was accompanied by no corresponding change to policing budgets even in cities with more BLM protests (Ebbinghaus, Bailey and Rubel, 2024).

This anticipatory channel is central because unrealized reforms are hard to observe in ordinary data. Election returns can reveal the consequences of a reform that happened, but they rarely reveal the proposals never introduced (or amended) because officials expected police opposition to be costly. Existing accounts of bureaucratic power emphasize how organized agents can help select their principals (Moe, 2006). Our argument adds that agents can also constrain sitting principals by shaping the expected electoral price of policy change. A policymaker may personally support a reform and still conclude that passing it would create an implementation fight that constituents would punish. Reports from local officials that police unions rely on intimidation or public safety threats illustrate how this logic can operate outside formal collective bargaining (Blumgart, 2020).

The implication for elite behavior is straightforward. When policymakers receive a fearmongering signal from police, they should become less willing to support reform than when they receive mere opposition without a threat of fearmongering. On our account, this is likely to occur because fearmongering leads politicians to update their beliefs about the probability of police shirking, and in turn politicians anticipate constituents will blame them if public

safety declines. Still, a fearmongering threat can reduce reform support even if it does not substantially change beliefs about whether officers are capable of resistance, because it makes the political cost of conflict salient as well.

The elite experiment tests the upstream implication of this argument. *H6 [elite deterrence]*: if officials anticipate that police resistance will impose political costs, then officials who receive a threatening frame should become less supportive of reform than those who receive a cooperative frame. *H7 [anticipated shirking]*: officials exposed to a threatening frame expect police to shirk more.

2.4 Officers' Strategic Logic

Finally, officers' choices are strategic but constrained. Fearmongering is relatively cheap and often deniable. It can be framed as professional expertise: officers warn that a policy will make residents less safe, that legal exposure will make them hesitate, or that staffing cuts will reduce capacity. Because the message is public, it can influence both voters and politicians before implementation. Its main risk is reputational. If voters view the message as self-interested intimidation, fearmongering may lower evaluations of police and strengthen support for reform.

Resistance shirking may be costlier. It may require coordination across officers, expose individual officers or departments to discipline, and harm the public services from which police also derive legitimacy. Organizational culture and solidarity can make such coordination easier (Brehm and Gates, 1993), while legal restrictions on strikes encourage more ambiguous forms of withdrawal. Shirking is most attractive when officers expect a service decline to be attributed at least partly to reforming politicians, when the relevant voters are persuadable, and when the costs of detection or sanction are low. It is less attractive when the public is strongly pro-reform, when media or administrative data can clearly document police responsibility, or when voters punish police for the decline.

The strategic logic across actors is therefore recursive. Police fearmongering changes

what voters expect from reform. Voters' expected updating changes what politicians are willing to attempt. Politicians' caution then changes the incentives officers face when deciding whether to threaten, cooperate, or resist. Our empirical design isolates key links in this chain: whether fearmongering changes voters' expectations and reform attitudes; whether service quality changes evaluations of policies and incumbents; whether fearmongering changes blame attribution under poor outcomes; and whether threatening police signals reduce reform support among the local officials who must decide whether to act in the first place.

Table 1: Summary of Hypotheses

Hypothesis	**Prediction**	**Mechanism**
Study 1 — Mass Public (fearmongering × realized service quality)		
H1	Fearmongering raises voters' expected likelihood of police resistance and lowers expected service quality.	Rational expectations about agency loss
H2a / H2b	Good service raises, and poor service lowers, support for the reform (H2a) and the incumbent (H2b).	Basic performance updating
H3a	Fearmongering attenuates the effect of realized service quality on support: both good and bad outcomes become less diagnostic.	Reduced signal informativeness
H3b	Fearmongering directly lowers support for the reform and the incumbent, and this cost persists even when realized service is good.	Prospective institutional conflict
H4a–H4d	Fearmongering amplifies priors: supporters grow more supportive (H4a) and more forgiving of poor outcomes (H4b); opponents grow less supportive (H4c) and more skeptical of good outcomes (H4d).	Greater reliance on priors
H5	Under poor service, supporters blame the police while opponents blame the reform; this gap widens under fearmongering.	Heterogeneous (directional) attribution
Study 2 — Elite Policymakers (threatening vs. cooperative frame)		
H6	Officials who receive a threatening frame are less supportive of reform than those who receive a cooperative frame.	Anticipatory deterrence
H7	Officials who receive a threatening frame expect police to shirk more.	Elite anticipated shirking

3 Fearmongering and Resistance Shirking in Practice

Before presenting our experimental results, we establish the real-world relevance of the phenomena we study. First, we draw on about 645,000 Facebook posts by US police departments and police union organizations in the 100 largest US cities between 2014 and 2025, provided by the Meta Content Library.[5] Second, we use an LLM-assisted news search to identify cases of reported police slowdowns and "blue flu" across the 100 largest US cities.[6] Finally, we present descriptive evidence from a survey of 1,800 Americans as well as our survey of elite policymakers on how commonly they encounter police slowdowns and how they interpret them.

As our theoretical argument suggests, voters rarely observe shirking in real time, since police organizations retain plausible deniability by attributing slower response, missed enforcement, or sick-outs to staffing shortages or operational pressures. Citizens may form general expectations about how often slowdowns occur, but in any given instance cannot tell with certainty whether reduced service reflects deliberate withdrawal or legitimate constraint. The observational evidence below therefore documents that fearmongering and slowdowns are real and recurring features of the policy landscape, but cannot pin down the exact prevalence of shirking or how voters and policymakers update when confronted with service lapses under uncertain attributability (Heo and Wirsching, 2026). This informational ambiguity is not a limitation of our evidence but the strategic environment we set out to study—and the one police organizations actively exploit.

3.1 Police Facebook Posts

We classify each post along four binary categories using an LLM (Gemini 2.5 Flash) with a structured prompt and worked examples, validated against human coding (Appendix A):

[5] Meta only allows downloading posts from pages with at least 15,000 followers or likes. This applies to 181 police departments and 19 unique unions among the top 100 US cities. See Appendix A for the classification procedure.

[6] See Appendix B for the search procedure.

policy opposition (opposes a named reform), *policy fearmongering* (predicts deteriorating public safety as a direct consequence of a specific policy or policymaker action), *blame* (attributes harms to named officials or political bodies), and *shirking fearmongering* (warns officers will reduce proactive policing or withdraw enforcement).

As Figure 1 shows, police organizations, particularly police unions, regularly use Facebook to oppose reform policies, predict safety harm from them, and assign blame to elected officials for crime outcomes. These three main categories appear in 8.6%, 5.7%, and 8.0% of union posts respectively, and peak above 20% around crucial points in the reform debate, such as the January 2020 New York bail-reform rollout and the mid-2020 George Floyd protests. Explicit shirking fearmongering is rarer and appears in about 0.7% of union posts. There is also substantial overlap between the categories: 68% of opposition posts use policy fearmongering and 38% explicitly attribute blame to policymakers.

Posts that oppose a specific reform or warn against its consequences typically pair their advocacy with electoral mobilization. Opposition to budget cuts, council resolutions, or ballot propositions, for example, is often paired with a call to vote or contact officials:

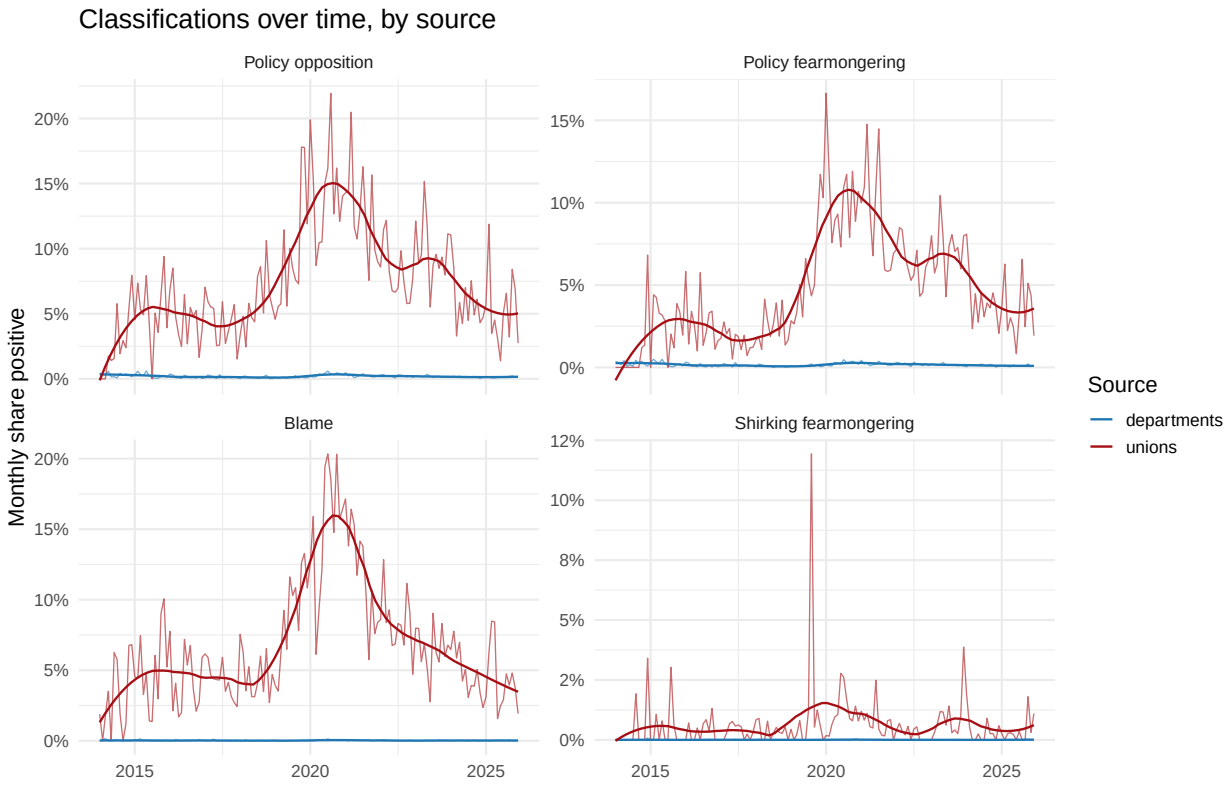
> "Early voting is upon us! ... Vote Against Prop B, it's a BAD Deal for Houston. If passed it will cause layoffs of Police Officers and Firefighters, causing a catastrophic impact to public safety in our community." (Houston Police Officers Union, October 2018).

> "No cash bail for a fatal hit and run. This is due to a COVID-19 directive in CA, but is part of criminal justice reform strategy. We don't want these problems here. The current Douglas County Board of Commissioners voted unanimously to support a no cash bail bill here in NE. Remember that at the November election." (Omaha Police Officers Association, May 2020).

Unlike pure opposition, *fearmongering* posts make a causal claim about future or ongoing danger to residents. The two most common framings either tie defunding decisions to slower response and unanswered 911 calls, or tie bail and parole policy to specific recidivism harms:

> "Underfunding and understaffing departments has real consequences. Those consequences can sometimes be deadly. The real victims of defunding police are the citizens. Citizens have the right to feel safe and when they need help they call

Figure 1: Monthly Distribution of Facebook Posts by Category



Note: Monthly share of posts classified as positive in each of four categories, by source. Thin lines show raw monthly shares; thick lines are LOESS smoothers (span = 0.3). $N = 645{,}540$ classified posts from 200 Pages (181 departments, 19 unions), January 2014 – December 2025. See Appendix A for the classification procedure.

> 911. What if they don't answer? What if they don't come to help?" (Dallas Police Association, September 2020).
>
> "Cops on the street are past the point of frustration, because we saw this new reality coming from a mile away. You can draw a direct line from Mayor de Blasio's anti-cop campaign in 2013 to the present disaster. Every action of our local and city government has aimed to shield criminals from consequences." (NYC Police Benevolent Association, January 2020).

Where fearmongering forecasts harm, posts coded as *blame* assign responsibility for that harm to specific officials rather than to vague systemic causes. The recurring targets are progressive prosecutors, city councils, and mayors, and the posts usually combine a current crime indicator with a named official, often implying causation to the reader:

> "Chairman Gregg Pemberton explains how Council Members Charles Allen and Phil Mendelson have created a crime crisis in the District by creating and passing anti police, anti victim, and pro criminal legislation." (DC Police Union, July 2023).

> "Our President Joe Gamaldi's comments this morning on DA Kim Ogg's continued sweetheart deals for violent, repeat offenders... 'He gets arrested for aggravated robbery, and the DA's office cuts him a sweetheart deal to give him deferred adjudication, which for those who don't know, means no jail time whatsoever."' (Houston Police Officers' Union, June 2019).

A rarer but distinctive subset of union posts moves beyond political argument and signals possible behavioral withdrawal, what we classify as *shirking fearmongering*. These typically take the form of roll-call briefings in which union trustees tell members to use procedural discretion to limit their exposure to discipline and litigation, often marked by the hashtags #TheJobIsDead and #NoConfidence. A second variant predicts that a specific accountability statute will produce a chilling effect on enforcement:

> "Bronx Trustee Betty Carradero and 47 Pct. delegates turned out the midnight and day tour roll calls with the PBA safety briefing. The message is simple: utilize the discretion afforded by current Department procedures to protect yourself from unwarranted discipline and legal liability. #TheJobIsDead" (NYC Police Benevolent Association, August 2019).

> "...a repeal or weakening of 50-a could affect the way police go about their jobs, given fears of retribution by disgruntled defendants and criminals... I believe it could have a chilling effect." (NYC Sergeants Benevolent Association, March 2019).

Notably, even in this most explicit category, posts do not claim deliberate withdrawal. Instead, they frame reduced enforcement as the inevitable consequence of the policy itself (e.g., a chilling effect or a defensive response to liability exposure), preserving the same plausible-deniability logic that operates in actual slowdown episodes.

This communication strategy also seems to attract viewership and engagement. Figure A2 shows that union pages reach a substantial audience (about 13,000 views per post on average) and that posts classified as policy opposition, fearmongering, or blame receive substantially more comments and shares (though not more likes) than other posts.

3.2 Shirking Cases in the Media

The Facebook record captures what police organizations say about reforms and politicians. To document cases where they have actually withdrawn effort, we assembled a hand-curated database of police slowdowns and "blue flu" incidents identified through journalistic, scholarly, and LLM-assisted media searches across the 100 largest US cities. This exercise reveals 46 documented cases between 2010 and 2023, of which 25 are supported by strong evidence such as multiple independent news sources, official acknowledgements, or scholarly documentation. Importantly, this collection is not a census but a lower bound on publicly documented episodes.[7]

The documented episodes span 22 states and 31 cities, with the largest cluster in 2020 (18 cases) and at least one episode in nearly every year of the window. Three triggers of slowdown incidents appear most often. The first is the prosecution of officers by progressive district attorneys, including Chesa Boudin (San Francisco), Kim Foxx (Chicago), Kim Gardner (St. Louis), Larry Krasner (Philadelphia), Paul Howard (Atlanta), Mike Schmidt (Portland), and José Garza (Austin). The second is budget cuts and "defund" resolutions adopted by city councils or mayors. The third is the public discipline of officers following high-profile use-of-force incidents. For example, in Atlanta, after DA Paul Howard charged Officer Garrett Rolfe with felony murder in the June 2020 shooting of Rayshard Brooks, officers staged a confirmed mass sick-out the same evening. Supporting evidence typically mentions quantitative administrative data on sick-call counts, arrests, or response times, acknowledgments from police chiefs or union presidents, and, in roughly a third of cases, peer-reviewed academic analyses.

3.3 Descriptive Survey Evidence

Finally, we asked respondents in our mass public survey and in our elite survey how commonly they perceive police to shirk and how they interpret police slowdowns. Using a separate

[7] We describe the construction and grading criteria in Appendix B.

sample of 1,800 respondents recruited via Verasight in September 2024, we find that 14% of survey participants experienced duty infractions by police, and more so among those who identify as Democrat or Independent, non-white and urban respondents (Figure A3).[8] Additionally, 79% of respondents disapprove of police slowdowns (Figure A4). The elite survey shows very similar figures. 18% of local officials report that police officers have resisted reform policies through shirking in their jurisdiction at least once, and 56% say that law enforcement is at least somewhat influential in local politics in their jurisdiction.

Taken together, the observational and descriptive evidence establishes that police slowdowns are a real phenomenon, that police organizations actively signal opposition to reform policies and warn of potential negative consequences, and that both voters and local officials encounter and disapprove of police shirking. Yet, we do not claim that police resistance or fearmongering is pervasive across departments, nor that police organizations are uniformly to blame for service shortfalls. Our interest is in a specific strategic behavior—the rhetorical and behavioral strategies some police organizations use against accountability reforms—and in how citizens and policymakers respond when they encounter it.

4 Study 1: Mass Public Survey Experiment

4.1 Study Population and Survey Platform

We recruited 1,676 participants through Prolific between April 1st and April 10th, 2025. The sample is quota-sampled to approximate the U.S. adult population on region, sex, age, race/ethnicity, and party identification.

[8]The question wording was "Have you ever experienced a police slowdown", where we give the following definition for "police slowdown" prior to the question: "Police slowdowns (or 'Blue Flu') can happen when police officers purposely take longer to respond to calls or don't enforce laws as strictly. They might do this to protest, usually when they feel they're being treated unfairly, when there's tension between them and the community, or when they're worried about being watched too closely. This is the definition of a 'police slowdown,' but not every delay is the result of intentional protest."

4.2 Survey Flow and Outcome Variables

Our experiment proceeds in several parts. First, respondents answer pre-treatment demographic and moderator questions, and we elicit pre-treatment beliefs about our key outcomes: support for the reform policy, perceived implications of the policy for public safety, and ratings of local institutions, including incumbent city council members and local police. Respondents are then randomized to the fearmongering treatment, after which we measure expectations about service quality and, for a 20% subsample, the likelihood of police resistance. Respondents then receive the service quality treatment, after which we measure post-treatment versions of all outcome measures plus blame attribution (for respondents in the poor service condition).

4.3 Experimental Conditions

4.3.1 Fearmongering Treatment

We vary whether respondents receive a cooperative police-union signal or a fearmongering signal. In the cooperative control condition, the news article states that the police union supports the reform and signals willingness to cooperate with city officials. In the fearmongering condition, the article states that the police union opposes the reform, warns that the city council will bear the blame for any decline in public safety, and notes that observers worry the union's opposition could lead to police slowdowns and on-the-job protests. In supplementary analyses, we use a union-opposition-only variant that omits explicit shirking language, allowing us to distinguish fearmongering from a pure endorsement cue.

4.3.2 Service Quality Treatment

We vary the quality of police services following implementation, showing either improved or worsened public safety outcomes. Respondents are randomly assigned to one of three service-quality measures: 911 response times, crime victimization rates, or traffic stops. They are

also randomly assigned to one of three policy reforms: cutting the police department's budget, creating a civilian complaint oversight board, or ending qualified immunity. We show below that these variations do not significantly alter our main findings.

4.4 Empirical Strategy

Our primary analyses estimate treatment effects using ANCOVA specifications in which the post-treatment outcome is the dependent variable and the corresponding pre-treatment measure is included as a control. For respondent i, let Y_i^{post} denote the post-treatment outcome and Y_i^{pre} the corresponding pre-treatment measure. For outcomes measured after both treatments are revealed, we estimate:

$$Y_i^{post} = \alpha + \beta_1 \text{PoorService}_i + \gamma Y_i^{pre} + \varepsilon_i \tag{1}$$

$$Y_i^{post} = \alpha + \beta_1 \text{Fearmongering}_i + \beta_2 \text{PoorService}_i + \beta_3 (\text{Fearmongering}_i \times \text{PoorService}_i) + \gamma Y_i^{pre} + \varepsilon_i \tag{2}$$

In equation (2), β_1 captures the effect of fearmongering under good service, β_2 captures the effect of poor service in the cooperative control condition, and β_3 tests whether fearmongering changes the informational effect of service quality. For outcomes measured before service quality is revealed, we estimate the corresponding pre-service model:

$$Y_i^{inter} = \alpha + \beta_1 \text{Fearmongering}_i + \gamma Y_i^{pre} + \varepsilon_i.$$

For heterogeneity analyses, we interact the treatment indicators with respondents' baseline support for the specific reform policy they were assigned to evaluate, classifying respondents as opponents, indifferent respondents, or supporters.

4.5 Descriptive Results: Pre-Treatment Attitudes

Before presenting experimental results, we describe the information environment and prior attitudes of our respondents. Nearly 78% of respondents report being at least somewhat informed about public safety in their town. The most common sources of information are local TV, friends and family, Facebook, and local newspapers. Public safety and crime ranks as the most important local-election issue in the sample, with 43% rating it "extremely important."

With respect to prior attitudes toward local institutions, 49.6% of respondents grade local police with a "D" or "F," compared with 38.8% for city council representatives and 44.3% for mayors or city managers. Respondents are ex-ante broadly supportive of civilian oversight boards and ending qualified immunity. 59.1% support or strongly support civilian oversight, and 58.7% support or strongly support ending qualified immunity. By contrast, respondents are more skeptical of police budget cuts, with 50.2% opposing or strongly opposing cuts to police funding. Respondents also expect civilian oversight boards and ending qualified immunity to improve public safety, while they expect police funding cuts to worsen public safety. These distributions are consistent with national opinion surveys (Sances, 2024) and validate the external plausibility of our hypothetical vignette design.

4.6 Experimental Results

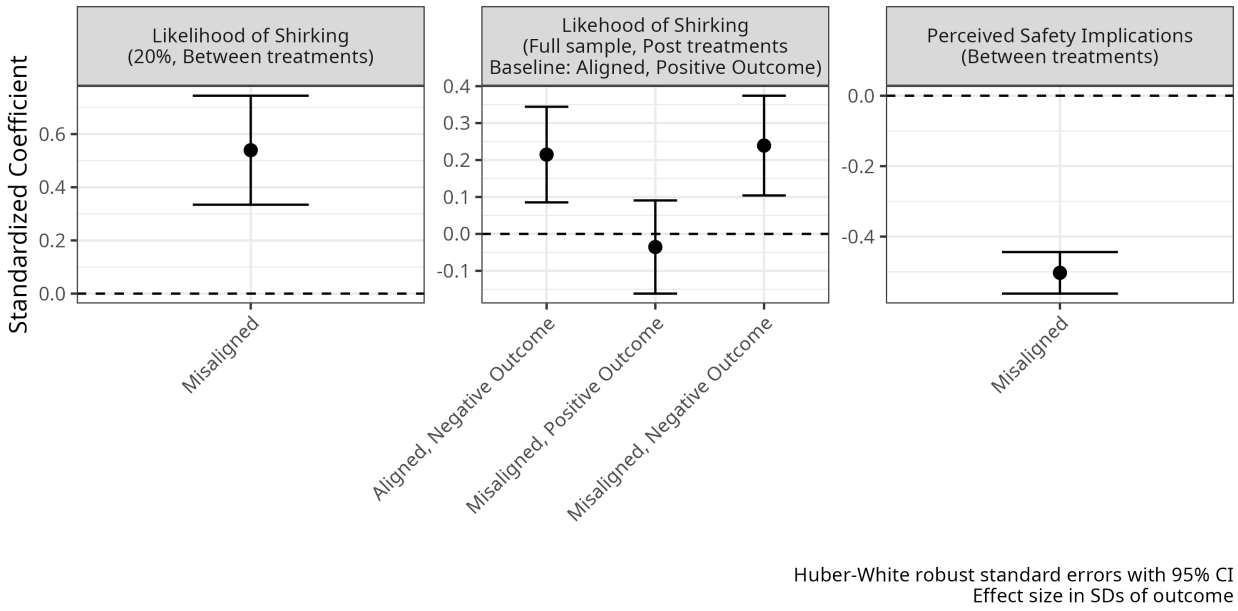
4.6.1 H1: Fearmongering Shifts Expectations

H_1 specifies that voters expect a higher likelihood of resistance and worse public safety implications when they receive a fearmongering signal. Columns 1 and 2 in Figure 2 presents results for two resistance measures captured at different points in the survey. The first is *anticipatory*. When asked between treatments, before any service outcome is shown, respondents in the fearmongering condition rate the likelihood of police resistance significantly higher than those in the cooperative control condition. This question was asked only to a

random 20% subsample to avoid priming resistance attributions for the full sample. The effect is substantively large, roughly 15% of the scale, supporting the hypothesis that fearmongering credibly raises expectations of shirking.

The last column in Figure 2 shows that fearmongering also shifts expected policy consequences. Respondents who receive the fearmongering signal are significantly less likely to expect the implemented policy to improve public safety. Both outcomes are measured before respondents observe actual service outcomes, which confirms that fearmongering shapes anticipatory beliefs.

Figure 2: Fearmongering Increases Expected Resistance and Reduces Expected Safety Benefits Before Service Outcomes Are Revealed (H1)



The second resistance measure is *ex post*. After both treatments are revealed, we ask how likely it is that police resistance actually occurred. These perceptions are driven primarily by observed service quality rather than by the earlier fearmongering signal. Once respondents observe a negative service outcome, they become more likely to perceive police resistance, regardless of whether they were previously warned about it; positive service outcomes, by contrast, reduce resistance attributions. Fearmongering thus shapes what voters *expect* be-

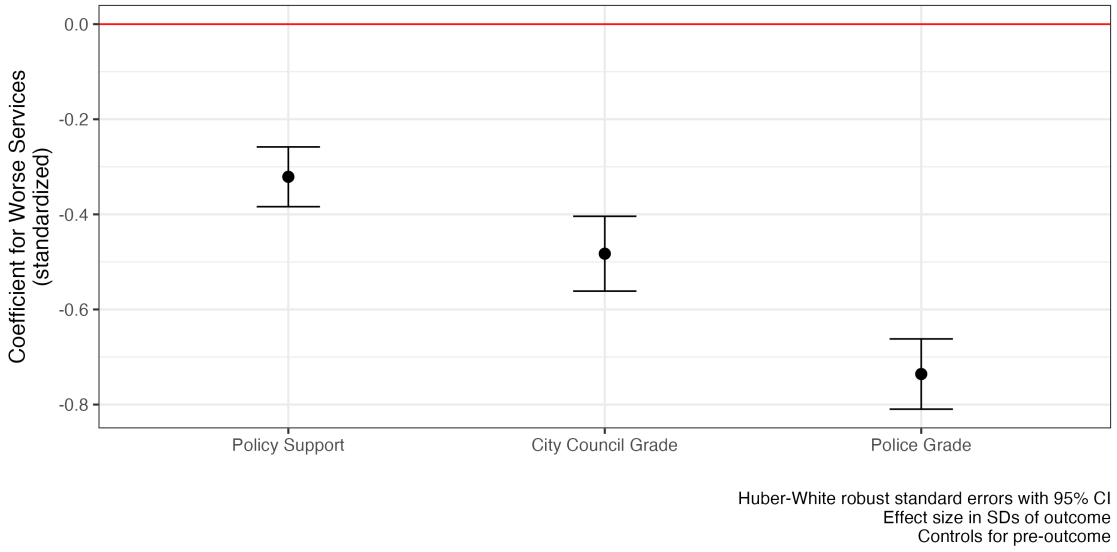
forehand, while realized service quality largely governs how they assess the possibility of shirking afterward.

4.6.2 H2: Service Quality Drives Updating on Policies and Politicians

H_2 predicts that voters update support for incumbent politicians and reform policies based on observed service quality. Figure 3 presents standardized OLS estimates for the effect of poor service relative to good service, controlling for the corresponding pre-treatment outcome. We find strong support for H_2. Poor service significantly reduces support for reform policies, lowers city council grades, and lowers police grades.

These effects are substantively large, ranging from roughly one-third to three-quarters of a standard deviation across outcomes. The substantial negative effect on police grades—not a pre-registered hypothesis—suggests that service quality outcomes have repercussions for all actors in the governance chain, not only the reform's political principals.

Figure 3: Poor Service Reduces Support for Policies, Politicians, and Police (H2)



4.6.3 H3: Bad Services Are Not Excused by Fearmongering

H_3 predicts that fearmongering weakens the effect of service quality on support, because knowing that shirking is possible should make voters more forgiving of poor outcomes. Figure 4 plots standardized outcomes for policy support, city council grade, and police grade by fearmongering condition and service quality condition.

We find limited evidence for H_3. The effect of service quality is somewhat smaller under fearmongering, but poor service is not meaningfully excused. Respondents in the fearmongering condition still evaluate policies and politicians negatively when service quality deteriorates. In other words, fearmongering does not lead voters to discount poor outcomes as merely the result of police resistance.

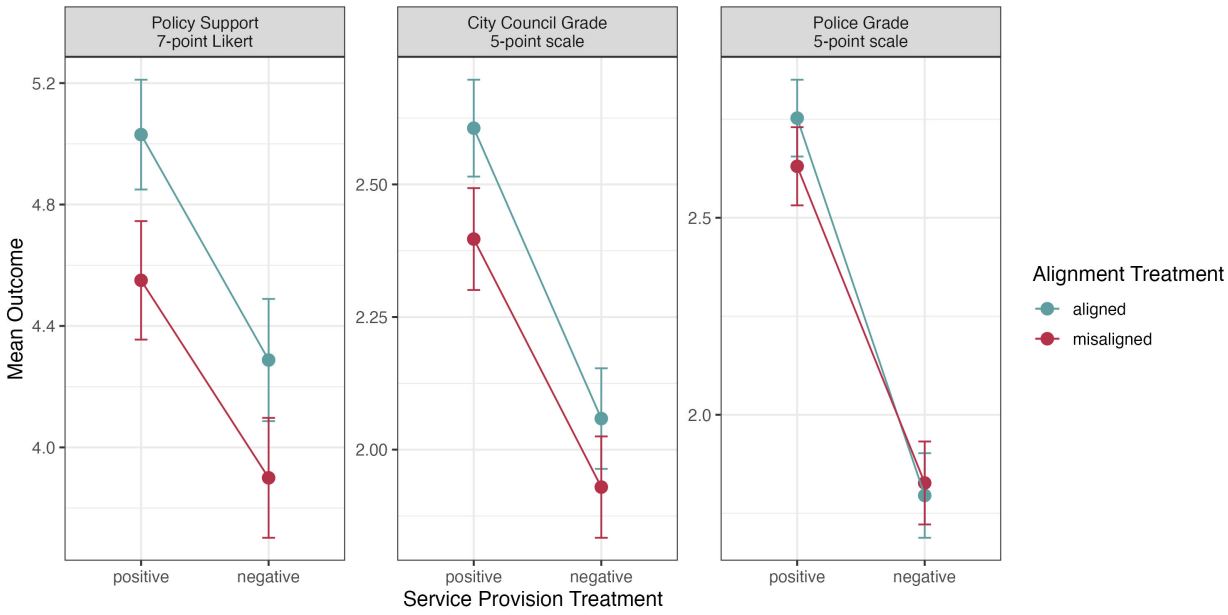
Instead of the interaction pattern predicted by H_3, we observe a broader downward shift: fearmongering appears independently costly across service-quality conditions. Respondents exposed to fearmongering evaluate reform policies and city council members less favorably under both good and poor service. The two experimental factors therefore appear closer to additive than interactive. Voters penalize politicians for institutional conflict itself, not only when that conflict results in poor service outcomes.

4.6.4 H4: Heterogeneous Updating by Prior Beliefs

H_4 predicts that fearmongering amplifies prior policy beliefs: supporters should become more supportive, while opponents should become less supportive, especially when service-quality signals conflict with those priors. Figure 5 shows the overall effect of fearmongering by baseline policy belief, pooling over service-quality conditions. Figure 6 shows treatment-combination effects by prior belief, relative to the cooperative-control + good-service condition.

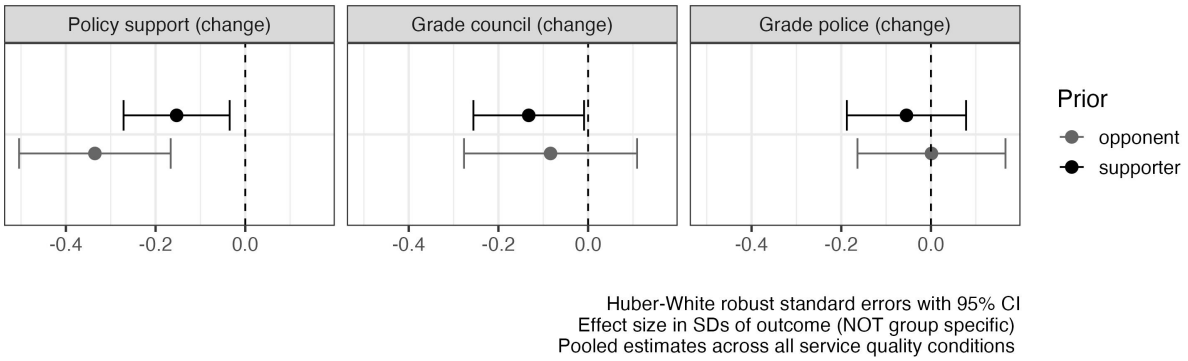
We find partial support for H_4. Heterogeneity is concentrated in policy support. For city council and police grades, supporters and opponents respond in broadly similar ways, with little evidence that prior policy beliefs condition evaluations of institutions. For policy

Figure 4: Bad Services Are Not Excused with Fearmongering (H3)



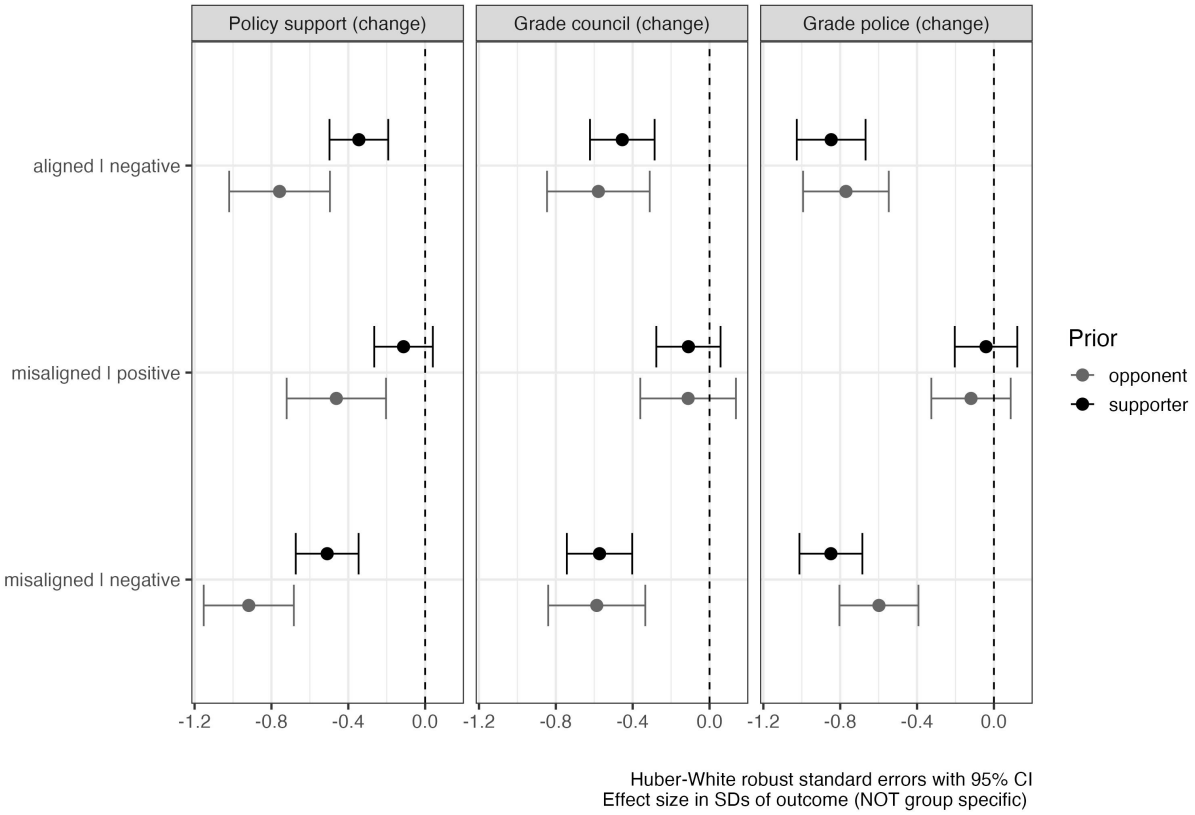
support, however, opponents respond more negatively to fearmongering than supporters. In the fearmongering/good-service condition, opponents remain less supportive even though service quality improves, while supporters are close to zero. This is consistent with the idea that opponents discount positive service signals when fearmongering makes resistance salient.

Figure 5: Fearmongering Effects by Prior Policy Beliefs (H4a, H4c)



We do not find the complementary pattern for supporters under poor service. Supporters do not become more forgiving of poor service when it occurs under fearmongering, i.e.,

Figure 6: Treatment-Combination Effects by Prior Policy Beliefs (H4b, H4d)



poor service still reduces support. Taken together, the heterogeneity evidence suggests partial support for H_4: fearmongering appears to make opponents more skeptical of positive outcomes, but it does not lead supporters to excuse poor outcomes.

4.6.5 H5: Blame Attribution

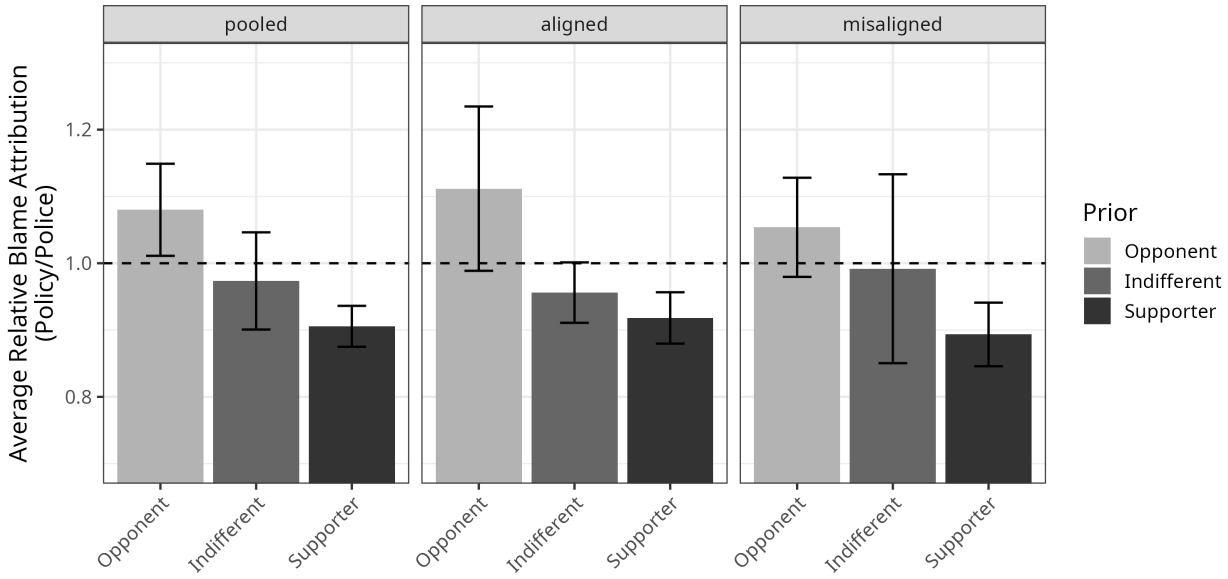
H_5 predicts that supporters blame police for poor service outcomes, opponents blame the policy, and this gap intensifies under fearmongering. Figure 7 shows relative blame attribution—policy blame divided by police blame—by prior support and fearmongering condition. Positive values indicate stronger blame attribution for the policy relative to the police.

We find mixed support for H_5. In line with our predictions, policy supporters are more likely to blame police than the policy when service quality is poor (an average of 3.58 vs. 3.08 on a 5-point likert scale, or an average 0.91 relative blame attribution). However, policy

opponents assign nearly equal levels of blame to police and the policy (an average of 3.95 vs. 3.94 and a relative blame attribution of 1.08), i.e., we do not find that opponents are much more likely to blame the policy itself. This suggests that the predicted asymmetry in blame attribution is largely present only on one side of the prior support distribution.

On the interaction with fearmongering, we do not find strong evidence that the supporter-opponent gap widens. In both the cooperative-control and fearmongering conditions, supporters assign relatively more blame to police than to the policy, while opponents assign roughly equal blame to police and the policy. The relative-blame pattern is therefore consistent with partial support for H_5, but not with the stronger prediction that fearmongering polarizes blame attribution by prior policy beliefs.

Figure 7: Relative Blame Attribution by Prior Support and Fearmongering Condition (H5)



4.7 Mechanisms: Open-Ended Responses

To illuminate the processes underlying our experimental results, we combine embedding regressions (Rodriguez, Spirling and Stewart, 2023) with sparse autoencoders (SAE) (Movva et al., 2025) to analyze respondents' open-ended explanations for their attitude changes.

Existing approaches to analyzing open-ended responses to randomized treatments either pre-specify dimensions before estimation (e.g., embedding regressions with target words (Rodriguez, Spirling and Stewart, 2023), hand-coded categories) or estimate unsupervised features in the corpus and then test the treatment effects on each feature separately (e.g., HypotheSAEs with LLM-annotated SAE features (Movva et al., 2025), STM with topic-level estimateEffect (Egami et al., 2022)). Both largely lose the multidimensional structure of how treatment affects text: the first by collapsing onto a small set of analyst-chosen target words, the second by atomizing the analysis into separate tests. We, instead, estimate the treatment effect as a single multidimensional object in embedding space, $\hat{\beta}$, test its overall significance with a permutation test, and project it onto the directions of an unsupervised SAE.This preserves the multidimensional structure throughout estimation and inference, while still delivering interpretable substantive conclusions.

Following the split-sample logic of Egami et al. (2022), we estimate the semantic shifts from the embedding regressions and the SAE on two different halfs of the data, so the feature directions and the effect they decompose are estimated from disjoint text. For the embedding regressions, we first obtain a $D = 1536$-dimensional embedding for the open ended responses e_i from a pre-trained large language model on the first half of the data (OpenAI's `text-embedding-3-small`)[9]. We asked respondents to expand on their reasons for the attitude change. The treatment effect on the average semantic content of responses is the difference of group means in embedding space. For the fearmongering treatment, for example, we estimate the average treatment effect as the difference in mean embeddings between the fearmongering and cooperative-control conditions,

$$\hat{\beta} \;=\; \bar{e}_{\text{fear}} - \bar{e}_{\text{con}}, \qquad \bar{e}_g = \frac{1}{n_g} \sum_{i:T_i=g} e_i. \tag{3}$$

[9]We apply only two filters before embedding. Responses that are missing or contain fewer than six non-whitespace characters are dropped, reducing the analysis sample to $N = 1{,}088$. The remaining responses are sent verbatim to the embedding model. We do not lowercase, strip punctuation, or remove stopwords, because the model is trained on raw text with its own subword tokenizer and expects unprocessed input.

Because the embedding function is fit on an external corpus and is invariant to our treatment assignments, this estimator inherits the identification properties of standard difference-in-means under randomization (Egami et al., 2022). We assess whether the treatment shifts text content overall by permuting the treatment vector 5,000 times and recomputing $\|\hat{\beta}\|^2$ on each draw. The squared norm is upward biased in finite samples, because each squared component picks up a variance term and is by definition positive. We therefore report a bias-corrected effect size following Green et al. (2024),

$$\|\hat{\beta}\|^2_{\text{cor}} \;=\; \sum_{k=1}^{D}\Big(\hat{\beta}_k^2 - \widehat{V}[\hat{\beta}_k]\Big)\,. \tag{4}$$

To interpret the substantive content of the treatment shift, we train a sparse autoencoder (SAE) on the document embeddings of the remaining 50% split of the responses, with $M = 8$ features and a TopK activation with $K = 2$, following Movva et al. (2025). We chose these parameters due to the short length of the open-ended responses, i.e., we target a small set of eight interpretable features and assume that each respondent at most mentions two of them. The SAE learns a set of unit-norm decoder directions $w_j \in \mathbb{R}^D$ that correspond to interpretable axes of the embedding space.[10] We then project our $\hat{\beta}$ onto each decoder direction, computing its share of the squared length of $\hat{\beta}$,

$$\alpha_j \;=\; w_j^\top \hat{\beta}, \qquad r_j \;=\; \frac{\alpha_j^2}{\|\hat{\beta}\|^2}, \tag{5}$$

with standard errors for α_j obtained via leave-one-out jackknife over the test responses.[11] The sign of scalar α_j indicates whether the treatment shifts responses toward (positive) or away from (negative) the concept that feature j captures. $r_j \in [0, 1]$, in turn, is the share of the treatment-induced semantic shift that runs along that feature's direction. To label each

[10] We use the decoder directions, since the decoder is approximately linear, and we can ask "how much does $\hat{\beta}$ overlap with the direction in embedding space that feature j contributes when active?"

[11] We do not propagate uncertainty in w_j. The global $\|\hat{\beta}\|^2$ permutation test does not require splitting—the embedding map is fixed externally and invariant to treatment—but we report it on the same held-out half for consistency.

w_j substantively, we feed the documents that most strongly activate the feature to a large language model (GPT-4.1) with a task-specific instruction and adopt the model's description of the common content as the feature's label (Movva et al., 2025).

Table 2 presents results for the interaction of fearmongering and service quality. The shift is largest for the comparison combining both fearmongering and poor service ($\|\hat{\beta}\|^2_{\text{cor}} = 0.0263$ on the held-out split). The global interaction test is not statistically significant ($p_{\text{perm}} = 0.386$), indicating that—similar to our experimental findings—we cannot formally reject an additive model of the two treatments. The cell-by-cell pattern, however, is substantively informative. The fearmongering shift under good service reduces a "positive police performance outcomes" framing in respondents' explanations (about 14% of the shift) while raising a "police accountability" framing (about 15%). Yet, the effect is small and does not reach significance ($p_{\text{perm}} = 0.121$).

A complementary pattern emerges when comparing the outcome effect among aligned respondents. The bad-service shift runs through the same two features: aligned respondents become less likely to invoke positive police performance outcomes ($\alpha = -0.075$, about 18% of the squared shift) and more likely to invoke police accountability and consequences ($\alpha = +0.078$, about 20%), both significant out of sample. For example, respondents in this cell write that "police officers should be held accountable" or question whether officers are doing their job. This suggests that, even without a separate experimental intervention, some respondents attribute poor service outcomes to police behavior and demand accountability.

When respondents *also* receive the fearmongering signal, the "positive police performance outcomes" framing drops further ($\alpha = -0.087$, about 21% of the shift) and the "police accountability" framing strengthens ($\alpha = +0.072$, about 14%), each significant out of sample. Together these two features account for roughly 35% of the combined semantic shift, consistent with fearmongering priming a shirking-attribution channel that respondents lean on more heavily once poor service gives them a concrete "proof."

Table 2: Embedding Regression and SAE Projection in the 2×2 Design — *Out-of-Sample* (50/50 split; directions from train, effects from test)

Feature	α_j (95% CI)	r_j
(1) Fearmongering effect, good service: good+misaligned vs. good+aligned $N = 125,\ 144$ (test); $\|\hat{\beta}\|^2_{\text{cor}} = 0.0024$; $p_{\text{perm}} = 0.121$		
Police accountability and shirking	0.045 [0.002, 0.088]	0.154
Positive police performance outcomes	−0.043 [−0.078, −0.008]	0.142
No opinion change	0.024 [−0.026, 0.073]	0.042
(2) Bad-outcome effect, aligned: bad+aligned vs. good+aligned $N = 134,\ 144$ (test); $\|\hat{\beta}\|^2_{\text{cor}} = 0.0206$; $p_{\text{perm}} < 0.001$		
Police accountability and shirking	0.078 [0.038, 0.119]	0.198
Positive police performance outcomes	−0.075 [−0.108, −0.043]	0.184
Unwavering support	−0.041 [−0.079, −0.003]	0.055
(3) Combined effect: bad+misaligned vs. good+aligned $N = 142,\ 144$ (test); $\|\hat{\beta}\|^2_{\text{cor}} = 0.0263$; $p_{\text{perm}} < 0.001$		
Positive police performance outcomes	−0.087 [−0.118, −0.055]	0.206
Police accountability and shirking	0.072 [0.031, 0.114]	0.143
Unwavering support	−0.036 [−0.072, 0.000]	0.035
(4) Interaction: (bad+mis − bad+ali) − (good+mis − good+ali) $N = 267,\ 278$ (test); $\|\hat{\beta}\|^2_{\text{cor}} = 0.0005$; $p_{\text{perm}} = 0.386$		

Note. Decoder directions w_j are estimated on a stratified 50% training split; each contrast's $\hat{\beta}$, its bias-corrected squared norm, permutation p-value (5,000 draws), and α_j (leave-one-out jackknife 95% CIs) are computed on the held-out 50% test split, so the directions and the projected effect share no responses. r_j is computed as a share of the *raw* squared norm $\sum_k \hat{\beta}_k^2$ (larger than the bias-corrected value shown) to ensure it is bounded in $[0, 1]$. Feature labels from GPT-4.1 on top-activating training responses. N gives (treatment, control) *test* cell sizes. Interaction features omitted (none significant).

4.8 Robustness

Primary results pool across all three policy reforms and all three service quality outcomes. Appendix figures show that there are no statistically significant differences in treatment effects across policy reforms or service quality measures. The "union signal only" condition (which signals opposition without mentioning shirking) produces weaker effects on perceived safety and policy support than our main fearmongering treatment, suggesting that the shirking component of the fearmongering message is doing independent work beyond simple endorsement signaling.

5 Study 2: Elite Policymaker Experiment

The mass experiment shows that voters use police service quality to evaluate both reform policies and the officials who adopt them, and that fearmongering is in and of itself politically costly. But our theory does not stop with voters. If elected officials anticipate that constituents will punish them for a decline in public safety—even when that decline may reflect police resistance rather than policy failure—then police can shape reform before any policy is implemented and before any actual shirking occurs (Heo and Wirsching, 2026). In that sense, Experiment 2 asks whether fearmongering has an *anticipatory deterrent* effect on the policymakers who decide whether reform will be attempted in the first place.

To examine this link in the democratic accountability chain, we partnered with CivicPulse[12] to embed an experiment in the 2026 Police Reform Policy Survey of local government officials, conducted between February and March of 2026. CivicPulse surveys have been used in a growing body of research on local policymakers, including studies of how officials respond to expert evidence, nationalized policy preferences, local news, and infrastructure investment (Lee, 2022; Lee, Landgrave and Bansak, 2023; Mullin and Hansen, 2023). The design holds fixed the underlying reform proposal and varies only the posture of local law enforcement: whether police leaders say they will cooperate with implementation or whether they warn that officers will scale back policing if the reform is adopted. This allows us to test whether a threatening police signal reduces officials' own support for reform and whether officials expect constituents to blame elected officials, rather than police, for any resulting decline in service.

5.1 Study Design

The CivicPulse 2026 Police Reform Policy Survey was administered to a national sample of local elected officials in early 2026. We embedded a qualified-immunity experiment within this survey. Qualified immunity is a legal doctrine that shields government officials, including

[12] https://www.civicpulse.org/

police officers, from civil liability in their individual capacities (Schwartz, 2023). We use this policy because qualified immunity reform makes our theory's mechanism especially concrete: the policy directly affects officers' legal exposure, and police organizations frequently argue that weakening immunity will make officers hesitant to engage in proactive policing.[13]

All respondents saw the same underlying policy: a proposal to make it *easier to sue individual officers* for misconduct, thereby weakening qualified immunity. We measured support for the policy before the treatment and again after it on a five-point scale from "strongly oppose" to "strongly support." This between-group pre/post design lets us estimate whether the police response changes officials' support relative to their own baseline position. Additionally, such a design has beneficial measurement properties, increasing precision without tradeoff in substantive treatment effects (Clifford, Sheagley and Piston, 2021; Jordan, Ollerenshaw and Trexler, 2026). Between the pre- and post-measures, respondents were randomized into one of two framing conditions describing how the local police union would respond to the reform.

Opposition without threat frame (control): Local policymakers read that the police union in their jurisdiction put out a statement saying officers opposed the policy change, and thought it was a bad idea, but would cooperate with the policy change, doing their best to maintain their efforts.

Threatening (fearmongering) frame (treatment): Local policymakers read that the police union in their jurisdiction put out a statement saying officers opposed the policy change and would likely scale back proactive policing in response to the policy, citing fears of personal legal liability, which could lead to fewer arrests and reduced public safety.

Randomization was at the individual level, so any difference in post-treatment responses is attributable to the union's posture, not the reform's content. Because in both conditions, the union puts forward a statement of opposition to the policy, we can attribute the differences in how politicians update their beliefs to the effect of fearmongering. The treatment is

[13] As an example, see the International Association for Police Chiefs statement opposing qualified immunity reform: https://www.iml.org/file.cfm?key=23108.

intentionally anticipatory, i.e., officials evaluate the reform after receiving a warning of future resistance, but before observing any actual deterioration in service.

We analyze two sets of outcomes to test the effects of fearmongering. First, we analyze reform support, which measures support for qualified immunity. Reform support is measured using a Likert scale both before and after the treatment. Second, we analyze perceived officer-resistance likelihood on a scale of (1–5), measuring how likely officers are to resist or retaliate. This item is not measured pre-treatment, as there is no inherent reason for politicians to express a likelihood of officer resistance when no reform is being proposed. Finally, though not a post-treatment outcome, we also measure blame attribution, which asks who politicians believe constituents would blame if proactive policing declined after a reform (elected officials who adopted the policy, the law-enforcement department/officers, both equally, neither, or "can't say").

The pre/post design supports an OLS estimator that conditions on the baseline attitude for precision, per the recommendations of Jordan, Ollerenshaw and Trexler (2026):

$$Y_i^{\text{post}} = \alpha + \beta\, D_i + \gamma\, Y_i^{\text{pre}} + \varepsilon_i, \tag{6}$$

where D_i indicates the threatening arm and β is the causal effect. We also report a difference-in-differences on the shift. Standard errors are HC2-robust (the appendix adds state-clustered (CR2) errors and covariate-adjusted specifications).

5.2 Sample and Representativeness

The survey was fielded January 29–March 10, 2026. CivicPulse drew the sample from a continuously maintained, phone-verified database of *elected officials*, including mayors, county commissioners, and city councilors in municipal, county, and township governments serving communities of 1,000 or more residents. Every respondent is therefore a sitting policymaker with local governing authority.

An important question is how our sample compares to the nation's elected officials. CivicPulse's sampling-frame statistics show that the respondent pool closely tracks that universe (Table A7): the median respondent's jurisdiction is modestly larger than the frame median, while college-education and Trump-vote shares are within a few points. The sample is therefore best understood as a cross-section of local police-policy authority rather than a cross-section of citizens. It accordingly skews toward small-town and exurban America and leans Republican (45% Republican or Republican-leaning vs. 33% Democratic or Democratic-leaning), reflecting the composition of local officeholding across the nation. There are more Republican local officials than Democratic local elected officials nationwide (CivicPulse and Carnegie Corporation of New York, 2024).

Of the 833 respondents, 420 were assigned to the cooperative arm and 413 to the threatening arm; the arms are balanced on observables and on baseline support (pre-treatment means 2.48 vs. 2.59). Experimental analyses use the $N = 550$ respondents with valid pre- *and* post-treatment answers (277 cooperative, 273 threatening). Item completion is symmetric across arms (≈66% each), and attritors and completers do not differ significantly on observables (all $p > 0.24$). Full balance and attrition tables are in the appendix (Section E.1).

5.3 Experimental Results

Table 3 presents the main treatment effects. Consistent with the anticipatory-deterrence logic, a fearmongering police response significantly reduces officials' support for qualified-immunity reform relative to mere policy opposition without fearmongering. The post-treatment level controlling for pre-treatment support decreases by 0.236 ($p < 0.001$) and the share of supporters decreases by 6.3 ($p = 0.02$) percentage points. These are substantively meaningful effects in a setting where officials are reacting to a short vignette about a single reform. Officials in the opposition without threat arm slightly *increase* support after the vignette (mean shift +0.12, $p = 0.04$), whereas officials in the threatening arm *decrease* support (mean shift −0.14, $p = 0.005$).

Table 3: Treatment Effects of Fearmongering

	Policy support		Resistance Expectations	
	Binary (0/1)	Likert (1-5)	Binary	Likert
Treatment (Fearmongering)	-0.06^{*}	-0.24^{***}	0.12	$0.08^{\dagger}$
	(0.03)	(0.07)	(0.10)	(0.04)
Control mean	0.29	2.60	2.62	0.48
Pre-outcome control	Yes	Yes	No	No
Num. obs.	550	550	542	542
Adj. R^2 (full model)	0.48	0.63	0.00	0.00

$^{\dagger}p < 0.1$;* $p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

On the five-point scale, the threat pushes the average official roughly a quarter of a point against the reform. Statistically, this translates to a Cohen's $d \approx -0.29$ (pooled SD = 0.88). Importantly, this difference is not merely driven by shifts in intensity of opposition or support for the policy reform. These shifts are also meaningful when translated into the real-world decision officials actually face: whether they would support the reform or not. For policymaking, a marginal change in the intensity of support matters less than whether an official moves across the support/opposition threshold. On that binary metric, support *rose* from 25.6% to 28.9% in the cooperative arm but *fell* from 27.5% to 23.8% in the threatening arm. The transition patterns show the same asymmetry. In the threatening arm, 21 officials who initially supported the reform now abandoned it (vs. 11 who newly joined), whereas the cooperative arm gained more supporters than it lost (22 vs. 13). This binary support effect survives randomization inference ($p = 0.0003$), state clustering ($p = 0.005$) and covariate adjustment correction. Politicians become less likely to support police reform when they see police fearmongering, relative to mere police opposition.

The threat also moved officials' beliefs about how officers would react, though the study is underpowered to certify the movement statistically. Mean perceived resistance rose from 2.62 in the cooperative arm to 2.74 in the threatening arm, a difference of +0.124 on the five-point scale, or roughly 0.11 SD. At the "likely to resist" threshold, the share of officials rating resistance somewhat, very, or extremely likely rose from 48.4% to 56.1%, an increase

of about eight percentage points. The shift is directionally exactly what the manipulation intends, but with $N \approx 542$ split across arms it is estimated somewhat imprecisely ($p = 0.06$). We read this as a real but modestly sized perceptual shift the experiment cannot pin down, not as evidence of no effect. Substantively, it implies that the drop in reform support is not driven primarily by a large change in beliefs about whether officers will resist. The threatening frame appears to do more than convey new information about the probability of shirking. It makes the political cost of conflict with police more salient, the upstream channel our theory identifies.

5.3.1 Mechanisms: Blame Attribution and Accountability Concerns

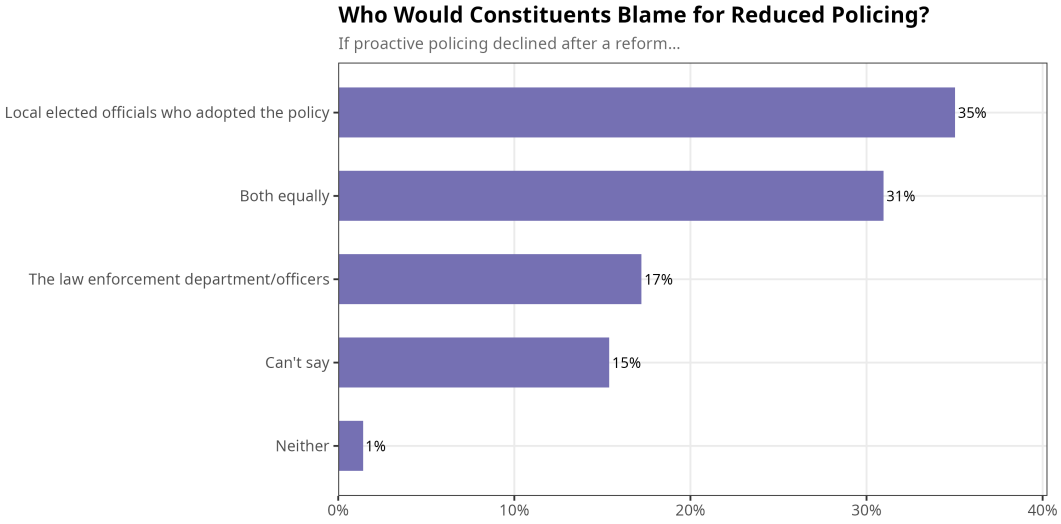
Importantly, our elite survey does find convincing descriptive evidence that politicians are concerned about being blamed for poor service quality, the exact proposition that undergirds our theoretical logic. Figure 8 shows how local officials believe constituents would assign blame if proactive policing declined following a policy reform. A plurality (35%) believe constituents would blame the local elected officials who adopted the policy. An additional 31% believe constituents would assign equal blame to both officials and police, meaning a majority (66%) of officials anticipate some blame accruing to policymakers. Only 17% believe constituents would primarily blame law enforcement.

The open-ended responses by officials point in the same direction, though they remain noisy at this sample size. A dictionary-free embedding regression finds no statistically significant difference in the language of the two arms (permutation $p = 0.15$ with GloVe, $p = 0.34$ with a neural encoder; Appendix E.8). With fewer than 130 substantive responses this null is unsurprising, and we read the following patterns as descriptive rather than confirmatory. Compared to the cooperative arm, officials in the threatening condition were somewhat more likely to discuss accountability (41.5% vs. 33.9%) and less likely to mention community relations (15.4% vs. 24.2%; Appendix E.7). The words most aligned with the estimated treatment direction lean toward reactive, emotional terms (absurd, ridiculous, blackmail,

wrongfully prosecuted) and away from constructive, institutional vocabulary (community, education, professional, implementation).

These findings link the elite experiment back to Study 1. In the mass experiment, respondents penalize reform-supporting politicians when service quality deteriorates. Local officials appear to anticipate exactly that accountability problem. If they believe constituents will hold them responsible for service deterioration, even when deterioration may be caused by police resistance, then a police threat can reduce reform support before any decline in service actually occurs.

Figure 8: Local Officials' Expectations of Constituent Blame Attribution (H7)



5.3.2 Who Moves? A Brief Look at Heterogeneity

The average effect masks meaningful heterogeneity, detailed in full in appendix E.1. The threat bites hardest among officials politically closest to law enforcement: it is driven by Republican officials (-0.33, $p = 0.006$; $d \approx -0.36$) and is essentially null among Democrats (-0.05, n.s.). We also test two specific moderators with formal treatment×moderator interactions. *(i) Police-performance grade.* Officials who grade their own department's 2025 performance highly do show a significant shift (-0.27, $p = 0.01$) while the lower-grading group does not reach significance (-0.21, n.s.)—but the *difference* between them is far

from significant (interaction $p = 0.77$). Officials with more *pro-police (anti-reform)* policy attitudes move more in point-estimate terms (-0.35, $p = 0.002$) than pro-reform officials (-0.20), consistent with the intuition that a police threat resonates most with pro-police officials—but again the interaction is not significant ($p = 0.39$). The pattern is suggestive but cannot statistically distinguish the moderated effects from the average effect.

6 Discussion and Conclusion

This paper examined how police resistance shapes political behavior, both before a reform through fearmongering and after it through service deterioration, among the US mass public and elite policymakers. Across two experiments, police emerge with strategic tools that operate across the entire policy pipeline. After implementation, the service quality voters observe strongly influences their evaluations of reform policies and incumbent politicians, and poor outcomes are not meaningfully excused even when voters know police resistance was possible. Before implementation, the credible prospect of resistance and pushback lowers mass support for reform and, as our elite experiment shows, the willingness of officials themselves to pursue it.

In both settings, the effect appears driven less by what people expect officers to do than by making the broader political cost of conflict with police salient. The possibility of shirking is a plausible part of the story. In the mass experiment, respondents facing fearmongering and poor service were more likely to attribute the outcome to officers scaling back effort, and among officials the threat moved perceived resistance in the expected direction. Yet, fearmongering does more than update beliefs about whether officers will resist or about how to read later outcomes. Among officials, the threat produces only a modest and statistically imprecise increase in perceived resistance. Among voters, fearmongering remains costly even when service turns out well, which a pure informational account of anticipated shirking would not predict. Hence, signals of police resistance act as a more general sign of police power and

influence in local politics, which reaches beyond the literal threat of slowdowns and work stoppages.

Yet, importantly, executing this power does not come without costs for police. When service deteriorates, voters lower their grades of the police substantially, and more so than for politicians. Among officials, 48% believe constituents would at least partly blame local law enforcement for worse services and reduced policing. Additionally, who bears the blame for service lapses depends on voters' prior perceptions of police. Hence, while the ambiguity of service provision provides police with plausible deniability and is an important source of political power, it also spreads blame across actors rather than shielding them entirely.

Several scope conditions structure these conclusions. Our outcomes are survey responses rather than observed legislative behavior or vote choice, and each experiment isolates a single policy area and a single bundled frame. The vignette design buys causal identification at some cost to external validity, and without randomizing actual shirking or finding a credible quasi-experiment that isolates partial effects, we cannot recover the total electoral impact. Hence, we focus on internal validity to identify partial effects rather than general equilibria or prevalence. Finally, the marginal effects we estimate do not imply that officers should always resist an unfavorable principal. In real elections voters weigh many issues, and concerns about long-term political relationships might add constraints for both policymakers and police.

Nonetheless, the results raise questions about democratic accountability in U.S. municipalities. If officials are punished for poor service whether it stems from policy failure or from police resistance, and if they anticipate that punishment, then police hold a strategic veto over reform that runs through the electoral mechanism itself rather than any formal institutional channel. The anticipatory costs we document may matter most where they are hardest to observe, in the reforms never proposed, the policies quietly abandoned, and the platforms moderated before any vote is cast.

Future work can extend this account in several directions. Better measurement of the mechanisms behind voter responses to fearmongering, especially the fear of de-policing, would

help separate informational from political channels. Careful observational data linking documented shirking episodes to the electoral fortunes of reform-minded officials would complement our experimental evidence. And because anticipatory deterrence appears to be a meaningful channel, future work should examine how officials condition their reform proposals on what they expect police to do.

References

Ang, Desmond and Jonathan Tebes. 2024. "Civic responses to police violence." *American Political Science Review* 118(2):972–987.

Anzia, Sarah F and Jessica Trounstine. 2024. "Civil Service Adoption in America: The Political Influence of City Employees." *American Political Science Review* pp. 1–17.

Arnold, R Douglas and Nicholas Carnes. 2012. "Holding mayors accountable: New York's executives from Koch to Bloomberg." *American Journal of Political Science* 56(4):949–963.

Bates, Josiah. 2021. "The Debate Over Qualified Immunity Is at the Heart of Police Reform. Here's What to Know." Time. https://time.com/6061624/what-is-qualified-immunity/.

Blumgart, Jake. 2020. "'They rely on you being intimidated': Local elected officials in the US describe how police unions bully them." CityMonitor, August 18 2020. https://citymonitor.ai/government/they-rely-you-being-intimidated-local-elected-officials-us-describe-how-police-unions

Boudreau, Cheryl, Scott A MacKenzie and Daniel J Simmons. 2019. "Police violence and public perceptions: an experimental study of how information and endorsements affect support for law enforcement." *The Journal of Politics* 81(3):1101–1110.

Boudreau, Cheryl, Scott A. MacKenzie and Daniel J. Simmons. 2022. "Police Violence and Public Opinion After George Floyd: How the Black Lives Matter Movement and Endorsements Affect Support for Reforms." *Political Research Quarterly* 75(2):497–511.

Brehm, John O and Scott Gates. 1999. *Working, shirking, and sabotage: Bureaucratic response to a democratic public*. University of Michigan Press.

Brehm, John and Scott Gates. 1993. "Donut shops and speed traps: Evaluating models of supervision on police behavior." *American Journal of Political Science* pp. 555–581.

Burch, Traci. 2022. "Not all black lives matter: officer-involved deaths and the role of victim characteristics in shaping political interest and voter turnout." *Perspectives on Politics* 20(4):1174–1190.

Burnett, Craig M and Vladimir Kogan. 2017. "The politics of potholes: Service quality and retrospective voting in local elections." *The Journal of Politics* 79(1):302–314.

Campbell, Andrea L. 2004. *How Policies Make Citizens: Senior Political Activism and the American Welfare State*. Princeton University Press.

Carreri, Maria, Edoardo Teso and Rui Yu. 2025. "The Influence of Police Unions in Local Elections. Evidence from U.S. Cities Spending and Performance." *Working Paper* .

CivicPulse and Carnegie Corporation of New York. 2024. "Local Government Navigates Negative Impact of Political Polarization Better than Federal Government According to New CivicPulse/Carnegie Survey." https://www.carnegie.org/news/articles/local-government-navigates-negative-impact-of-political-polarization-better-than-fede Survey of 1,412 local government leaders: 38% Republican, 29% Democrat, 33% Independent.

Clifford, Scott, Geoffrey Sheagley and Spencer Piston. 2021. "Increasing precision without

altering treatment effects: Repeated measures designs in survey experiments." *American Political Science Review* 115(3):1048–1065.

Ebbinghaus, Mathis, Nathan Bailey and Jacob Rubel. 2024. "The Effect of the 2020 Black Lives Matter Protests on Police Budgets: How "Defund the Police" Sparked Political Backlash." *Social Problems* p. spae004.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts and Brandon M. Stewart. 2022. "How to Make Causal Inferences Using Texts." *Science Advances* 8(42):eabg2652.

Foarta, Dana. 2023. "How Organizational Capacity Can Improve Electoral Accountability." *American Journal of Political Science* 67(3):776–789.

Gaudette, Jennifer. 2024. "Polarization in Police Union Politics." *American Journal of Political Science* 69(3):961–980.

Green, Breanna, William Hobbs, Sofia Avila, Pedro L. Rodriguez, Arthur Spirling and Brandon M. Stewart. 2024. "Measuring Distances in High Dimensional Spaces: Why Average Group Vector Comparisons Exhibit Bias, and What to Do About It." *Political Analysis* . Forthcoming.

Hacker, Jacob S. and Paul Pierson. 2019. "Policy Feedback in an Age of Polarization." *The ANNALS of the American Academy of Political and Social Science* 685(1):8–28.

Hamel, Brian T. 2025. "Traceability and Mass Policy Feedback Effects." *American Political Science Review* 119(2):778–793.

Heo, Kun and Elisa M. Wirsching. 2026. "Bureaucratic Resistance and Policy Inefficiency." *American Political Science Review* pp. 1–17. First View.

Hopkins, Daniel J and Lindsay M Pettingill. 2018. "Retrospective voting in big-city US mayoral elections." *Political Science Research and Methods* 6(4):697–714.

International Association of Chiefs of Police. 2020. "IACP Statement on Qualified Immunity." https://www.theiacp.org/sites/default/files/IACP%20Statement%20on%20Qualified%20Immunity.pdf. Accessed June 7, 2026.

Jares, Jake Alton and Neil Malhotra. 2025. "Policy impact and voter mobilization: Evidence from farmers' trade war experiences." *American Political Science Review* 119(2):847–869.

Jordan, Diana, Trent Ollerenshaw and Andrew Trexler. 2026. "New Evidence and Design Considerations for Repeated Measure Experiments in Survey Research." *American Political Science Review* .

Kiewiet, D Roderick and Mathew D McCubbins. 1991. *The logic of delegation*. University of Chicago Press.

Knight, Heather. 2021. "More S.F. residents share stories of police standing idly by as crimes unfold: 'They didn't want to be bothered'." San Francisco Chronicle. https://www.sfchronicle.com/sf/bayarea/heatherknight/article/sf-police-crime-16931399.php.

Kyriazis, Anna, Lauren Schechter and Dvir Yogev. 2023. The Day After the Recall: Policing and Prosecution in San Francisco. Technical report Working Paper. https://github.com/lrschechter/workingpapers/blob ....

Lee, Nathan. 2022. "Do Policy Makers Listen to Experts? Evidence from a National Survey of Local and State Policy Makers." *American Political Science Review* 116(2):677–688.

Lee, Nathan, Michelangelo Landgrave and Kirk Bansak. 2023. "Are Subnational Policymakers' Policy Preferences Nationalized? Evidence from Surveys of Township, Municipal, County, and State Officials." *Legislative Studies Quarterly* 48(2):441–454.

Lipsky, Michael. 2010. *Street-level bureaucracy: Dilemmas of the individual in public service*. Russell Sage Foundation.

Martin, Lucy and Pia J. Raffler. 2021. "Fault Lines: The Effects of Bureaucratic Power on Electoral Accountability." *American Journal of Political Science* 65(1):210–224.

McCubbins, Mathew D and Thomas Schwartz. 1984. "Congressional oversight overlooked: Police patrols versus fire alarms." *American Journal of Political Science* pp. 165–179.

Miller, Gary J. 2005. "The political evolution of principal-agent models." *Annu. Rev. Polit. Sci.* 8(1):203–225.

Moe, Terry M. 1984. "The new economics of organization." *American journal of political science* pp. 739–777.

Moe, Terry M. 2006. "Political Control and the Power of the Agent." *Journal of Law, Economics, and Organization* 22(1):1–29.

Movva, Rajiv, Kenny Peng, Nikhil Garg, Jon Kleinberg and Emma Pierson. 2025. Sparse Autoencoders for Hypothesis Generation. Vol. PLMR 267 of *Proceedings of the 42nd International Conference on Machine Learning*.

Mullin, Megan and Katy Hansen. 2023. "Local News and the Electoral Incentive to Invest in Infrastructure." *American Political Science Review* 117(3):1145–1150.

Mullinix, Kevin J. and Robert J. Norris. 2019. "Pulled-Over Rates, Causal Attributions, and Trust in Police." *Political Research Quarterly* 72(2):420–434.

Mummolo, Jonathan. 2018. "Modern police tactics, police-citizen interactions, and the prospects for reform." *The Journal of Politics* 80(1):1–15.

Newsday. 2022. "Manhattan District Attorney Alvin L. Bragg Jr. Crime-Priority Memo Spurs Criticism." https://www.nycpba.org/news-items/newsday/2022/manhattan-district-attorney-alvin-l-bragg-jr-crime-priority-memo-spurs-criticism/. Accessed June 7, 2026.

Pearson, Jake. 2023. "Police Resistance and Politics Undercut the Authority of Prosecutors Trying to Reform the Justice System." ProPublica, October 11, 2023. https://www.propublica.org/article/police-politicians-undermined-reform-prosecutors-chicago-philadelphia.

Pierson, Paul. 1993. "When Effect Becomes Cause: Policy Feedback and Political Change." *World Politics* 45(4):595–628.

Reny, Tyler T and Benjamin J Newman. 2021. "The opinion-mobilizing effect of social protest against police violence: Evidence from the 2020 George Floyd protests." *American Political Science Review* 115(4):1499–1507.

Rodriguez, Pedro L., Arthur Spirling and Brandon M. Stewart. 2023. "Embedding Regression: Models for Context-Specific Description and Inference." *American Political Science Review* 117(4):1255–1274.

Sances, Michael W. 2023. "Defund my police? The effect of George Floyd's murder on support for local police budgets." *The Journal of Politics* 85(3):1156–1160.

Sances, Michael W. 2024. "The Nationalized Politics of Police Reform." *The Forum* 22(1):71–96.
URL: *https://doi.org/10.1515/for-2024-2010*

Schwartz, Joanna. 2023. *Shielded: How the police became untouchable.* Penguin.

Slough, Tara. 2024. "Bureaucratic Quality and Electoral Accountability." *American Political Science Review* 118(4):1931–1950.

Soss, Joe and Sanford F Schram. 2007. "A public transformed? Welfare reform as policy feedback." *American political science review* 101(1):111–127.

Swan, Rachel. 2021. "San Francisco police just watch as burglary appears to unfold, suspects drive away, surveillance video shows." San

Francisco Chronicle. https://www.sfchronicle.com/bayarea/article/San-Francisco-police-only-watch-as-burglary-16647876.php.

Templeton, Amelia. 2026. "Portland Police, Fire Unions Push Back against Mayor's Proposed Budget Cuts: 'Dangerous and Reckless'." KGW. https://www.kgw.com/article/news/local/the-story/portland-police-fire-union-budget-cut-emergency-response-911/283-598595ec-1956-4d82-973d-1530c776f151. Accessed June 7, 2026.

The Atlanta Journal-Constitution. 2020. "170 Atlanta Police Officers Called Out Sick during 'Blue Flu' Protests, Records Show." https://www.ajc.com/news/crime--law/170-atlanta-police-officers-called-out-sick-during-blue-flu-protests-records-show/RIDIMWuApAM0Ytmdiwx01H/. Accessed June 7, 2026.

The Philadelphia Inquirer. 2022. "In the Move to Remove DA Larry Krasner from Office, a Fight over Gun Violence and Convictions." https://www.inquirer.com/news/da-larry-krasner-record-convictions-cases-guns-homicides-20221121.html. Accessed June 7, 2026.

Ujhelyi, Gergely. 2014. "Civil service reform." *Journal of Public Economics* 118:15–25.

Vaughn, Paige E, Kyle Peyton and Gregory A Huber. 2022. "Mass support for proposals to reshape policing depends on the implications for crime and safety." *Criminology & Public Policy* 21(1):125–146.

Walker, Hannah L. 2020. "Targeted: The mobilizing effect of perceptions of unfair policing practices." *The Journal of Politics* 82(1):119–134.

Weaver, Vesla M and Amy E Lerman. 2010. "Political consequences of the carceral state." *American Political Science Review* 104(4):817–833.

White, Ariel. 2016. "When threat mobilizes: Immigration enforcement and Latino voter turnout." *Political Behavior* 38:355–382.

Wirsching, Elisa Maria. 2026. "Political Power of Bureaucratic Agents: Evidence from Policing in New York City." *The Journal of Politics* . Forthcoming.

Zoorob, Michael. 2019. "Blue endorsements matter: How the fraternal order of police contributed to Donald Trump's victory." *PS: Political Science & Politics* 52(2):243–250.

Appendix: Supporting Information for *The Political Consequences of Police Slowdowns*

A Classification of Police Facebook Posts

To complement our experimental evidence with observational measures of bureaucratic fear-mongering in the real-world communications of police organizations, we assemble and classify a corpus of public Facebook posts from U.S. police departments and police unions. This appendix describes the data source and sample, the text preprocessing pipeline, the coding scheme, and the large language model (LLM) classification procedure used to produce the post-level labels analyzed in Section 3.

A.1 Data Source and Sample

We obtain posts from the Meta Content Library (MCL), Meta's research-access platform for public Facebook and Instagram content (the successor to CrowdTangle). MCL exposes posts from public Pages that exceed Meta's 15,000 follower threshold. From this universe, we hand-curated a roster of 200 Pages: 181 police department Pages and 19 police union Pages, covering departments and union locals in the 100 largest US cities. We download every public post made by these Pages between January 1, 2014 and December 31, 2025. The raw download contains 724,281 posts (664,074 from departments; 60,207 from unions).

A.2 Text Preprocessing

We apply a light, transparent preprocessing pipeline designed to remove non-substantive content without altering the political language of the posts. For each post, we (i) strip URLs and standardize HTML character entities (`&`, `>`, `<`, ` `); (ii) collapse runs of whitespace; and (iii) drop posts that, after cleaning, are shorter than 20 characters or whose lowercased text contains any of a small set of recurring non-political markers ("is hiring,"

"job opening," "apply now," "congratulations," "in memoriam"). These markers identify recruitment announcements, promotion notices, and standardized memorial posts that are common on department Pages but carry no political content. After preprocessing, 646,032 posts remain (594,565 department, 51,467 union), i.e.,89.2% of the raw corpus.

A.3 Coding Scheme

We classify each post along four binary categories that operationalize distinct components of policy opposition and bureaucratic resistance discourse:

1. **Policy opposition.** The post expresses opposition to a named reform or policy decision (e.g., defunding, civilian oversight, qualified immunity reform, bail reform).

2. **Policy fearmongering.** The post predicts deteriorating public safety as a direct consequence of a specific policy or policymaker action. This is the observational counterpart to the fearmongering treatment in Study 1.

3. **Blame.** The post attributes current or anticipated bad outcomes (e.g., crime, slow response times, low recruitment, declining morale) to political figures or named political bodies (e.g., mayor, city council, district attorney, named legislators).

4. **Shirking fearmongering.** The post warns that officers will reduce proactive policing, hesitate to intervene, or withdraw from enforcement activity, either in response to a named reform or via explicit union directives. This is the observational counterpart to the threatened-resistance signal in our experiments.

Because policy fearmongering is by construction a form of policy opposition, the coding rules require that any post coded as policy fearmongering also be coded as policy opposition. This priority rule is encoded directly in the classification prompt.

The coding prompt also included a fifth category—*ambient fearmongering*, intended to capture alarming rhetoric not tied to any named policy or political actor. Qualitative review

of a stratified sample of positive classifications revealed that this category had a substantially elevated false-positive rate, driven primarily by routine department crime-prevention bulletins and factual incident reports that the model coded as ambient fearmongering despite the prompt's explicit exclusions. The remaining four categories did not exhibit comparable systematic errors, and we therefore drop the ambient fearmongering category from analysis.

A.4 LLM Classification Procedure

We classify the corpus using Google's `gemini-2.5-flash` model accessed through the Gemini Batch API. The prompt comprises (i) a five-category coding instruction with definitions, (ii) explicit decision rules covering borderline cases (e.g., sharing a critical headline without counter-framing, rhetorical questions, working-conditions disputes, factual incident reports), and (iii) approximately 25 worked examples spanning department and union posts and all five categories. The full prompt is provided in the replication materials. The model is queried at temperature 0 and is required to respond with exactly five comma-separated 0/1 digits.

The corpus is processed in 130 batches of approximately 5,000 posts each. Of the 646,032 posts submitted, 645,540 (99.92%) received a parsed classification; the 492 misses (0.08%, all on the department side) are posts for which the Batch API did not return a parseable response. All returned responses conformed to the required `[01],[01],[01],[01],[01]` format; no further parsing repair was needed. Joint labels show high adherence to the priority rule: among posts coded as policy fearmongering, 97.8% are also coded as policy opposition, as required by construction.

To assess classifier validity, we draw an enriched validation sample stratified on the LLM prediction. For each of the four categories retained for analysis, we draw 40 LLM-predicted positives and 20 LLM-predicted negatives (10 per source type). A single human coder labeled each post in the sample. We use an enriched sample rather than a random sample given that accuracy measures are misleading with rare categories: a random sample would contain too few positives to produce meaningful precision and recall estimates.

For each category amd source type, precision is estimated directly from the predicted-positive stratum:

$$\widehat{\text{Precision}}_{c,s} \;=\; \frac{\text{TP}_{c,s}}{\text{TP}_{c,s} + \text{FP}_{c,s}}.$$

Recall cannot be read off the confusion matrix directly, because LLM-positives are over-sampled relative to LLM-negatives by design in the enriched sample. Using Bayes Rule, we instead combine the precision estimate with the false-negative rate $\widehat{r}_{\text{FN}|c,s} = \text{FN}_{c,s}/(\text{FN}_{c,s} + \text{TN}_{c,s})$ in the predicted-negative stratum, reweighted by the empirical share of LLM-positives in the full corpus $\widehat{p}_{\text{LLM}=1}|c, s$:

$$\widehat{\text{Recall}}_{c,s} \;=\; \frac{\widehat{\text{Precision}}_{c,s} \cdot \widehat{p}_{\text{LLM}=1}|c, s}{\widehat{\text{Precision}}_{c,s} \cdot \widehat{p}_{\text{LLM}=1}|c, s \;+\; \widehat{r}_{\text{FN}|c,s} \cdot (1 - \widehat{p}_{\text{LLM}=1}|c, s)}.$$

Two design details matter for the implementation of this. First, we report all metrics separately by source type (departments and unions), because the prevalence and rhetorical content of these categories differ substantively between the two. Additionally, within each source type, the 10 predicted-negatives drawn per category are a simple random sample from that source type's LLM-negative pool, so no reweighting across sources is required for $\widehat{r}_{\text{FN}|c,s}$, and $\widehat{p}_{\text{LLM}=1}$ in the recall formula is the empirical share of LLM-positives within source type.

Second, although a single post can be drawn into multiple categories' pools (e.g., a post that is LLM-positive on both *policy opposition* and *blame* may be drawn for both categories' positive pools), each post is deduplicated to a single row in the coded sample and labeled by the human coder on all four categories simultaneously. When computing metrics for category c, we restrict the predicted-positive and predicted-negative strata to posts that were drawn *directly* for c's pool, rather than to all coded posts where the LLM label matches. This ensures that the precision and false-negative-rate estimates are computed on simple random samples from the category-specific LLM-positive and LLM-negative pools, respectively.

Table A1 reports the per-(category, source type) confusion matrices on the enriched sample, and Table A2 reports the resulting precision and reweighted recall estimates with

Wilson 95% confidence intervals.

Three results are noteworthy. First, there are no false negatives observed in the LLM-negative stratum, which yields a recall point estimate of 1 across categories. Given the modest size of this stratum (10 posts per cell), this suggests lack of evidence for substantial recall loss rather than a guarantee of perfect recall. Second, precision is generally high for union posts (no less than 0.85 across categories), while the estimates are much lower, noisy, and largely underpowered for department posts due to the much lower prevalence of these categories in the department corpus. Third, the LLM identifies a non-trivial share of posts as positive for these categories, particularly among unions. For example, 8.6% of union posts are classified as expressing policy opposition, and 5.7% are classified as expressing policy fearmongering.

Table A1: Confusion Matrices on the Enriched Validation Sample, by Category and Source Type

Category	TP	FP	FN	TN	N_+	N_-
Departments						
Policy opposition	2	4	0	10	6	10
Policy fearmongering	5	6	0	10	11	10
Blame	1	0	0	10	1	10
Shirking fearmongering	1	0	0	10	1	10
Unions						
Policy opposition	33	1	0	10	34	10
Policy fearmongering	26	3	0	10	29	10
Blame	33	6	0	10	39	10
Shirking fearmongering	32	5	0	10	37	10

Table A2: Validation Metrics by Category and Source Type

Category	N_+	N_-	$\widehat{p}_{\text{LLM}=1}$	Precision [95% CI]	Recall [95% CI]	F1
Departments						
Policy opposition	6	10	0.002	0.33 [0.10, 0.70]	1.00 [1.00, 1.00]	0.50
Policy fearmongering	11	10	0.002	0.45 [0.21, 0.72]	1.00 [1.00, 1.00]	0.62
Blame	1	10	0.000	1.00 [0.21, 1.00]	1.00 [1.00, 1.00]	1.00
Shirking fearmongering	1	10	0.000	1.00 [0.21, 1.00]	1.00 [1.00, 1.00]	1.00
Unions						
Policy opposition	34	10	0.086	0.97 [0.85, 0.99]	1.00 [1.00, 1.00]	0.99
Policy fearmongering	29	10	0.057	0.90 [0.74, 0.96]	1.00 [1.00, 1.00]	0.95
Blame	39	10	0.080	0.85 [0.70, 0.93]	1.00 [1.00, 1.00]	0.92
Shirking fearmongering	37	10	0.007	0.86 [0.72, 0.94]	1.00 [1.00, 1.00]	0.93

A.5 Descriptive Analysis of Classified Posts

Figure A1: Share of Fearmongering Facebook Posts by Top 10 Pages

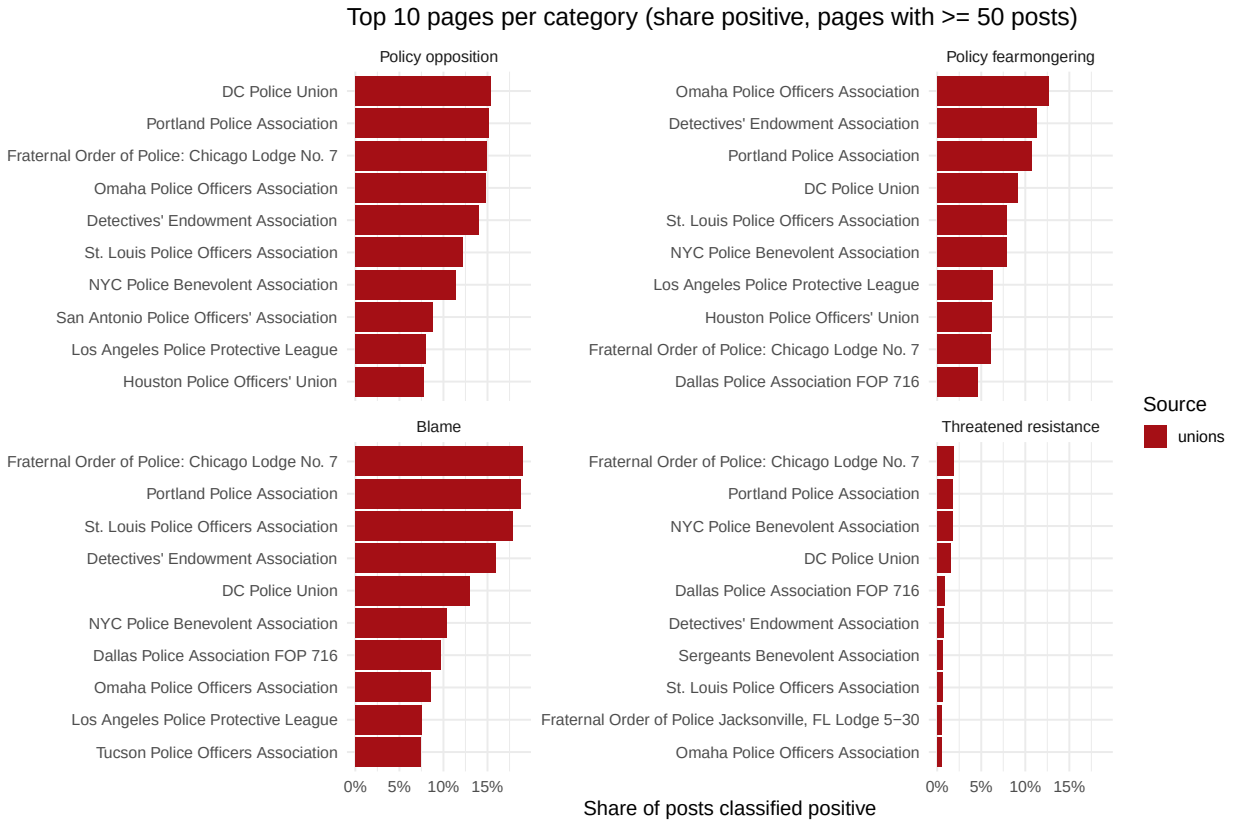
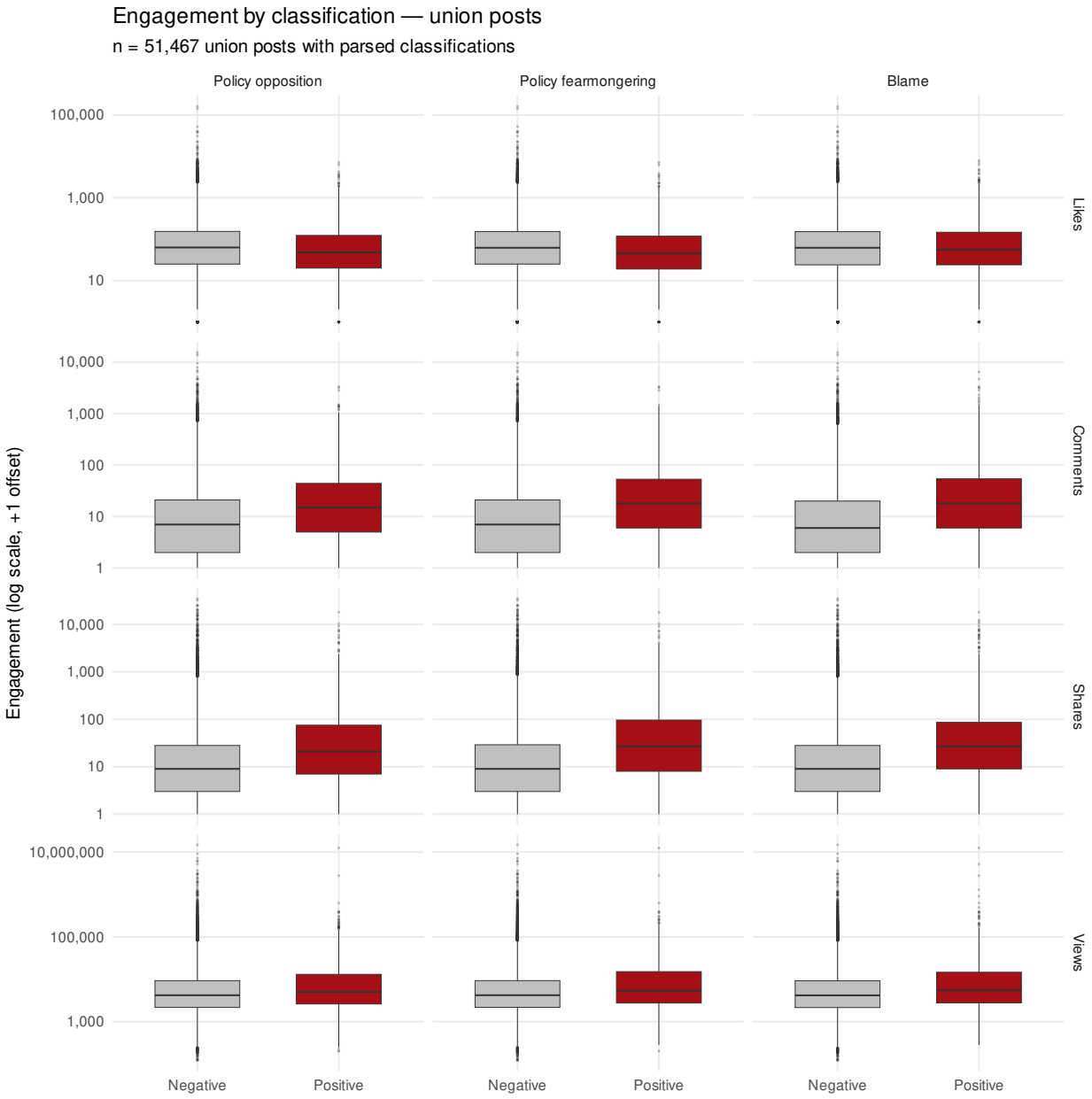


Figure A2: Engagement by Classification



B Identifying Police Slowdowns Across US Cities

To assess how far instances of police shirking extend across the US, we assembled a descriptive database of police slowdowns across the 100 largest U.S. cities since 2010. The database is not intended as a complete census. Coverage of these episodes is uneven, and many slowdowns are never reported or are reported only in local outlets. We therefore treat it as a lower bound on

publicly documented episodes, sufficient to show that coordinated slowdowns happen across cities, periods, and political contexts rather than being confined to a handful of well-known cases.

We identified candidate episodes from two sources. The first was existing academic and journalistic work on police work stoppages, blue-flu episodes, and de-policing. The second was a structured media search conducted with an LLM. We used Anthropic's Claude, accessed through the Research feature in claude.ai, to search news coverage city by city. For each city we issued an independent query, so that heavy coverage of prominent cases would not crowd out smaller ones, and we searched a common set of terms including "police slowdown," "blue flu," "sick-out," "work stoppage," "reduced enforcement," and "rising response times," combined with the city name and the window since 2010. The model formulated and executed iterative web searches, retrieved the underlying reporting, and returned candidate episodes together with the sources it relied on.

The model's role was confined to locating and summarizing coverage. It did not adjudicate whether an episode qualified, and we treated every candidate it surfaced as a lead to be verified rather than as a finding. We checked each candidate against its original sources and discarded those we could not corroborate. For episodes that survived this check, we recorded the city, the date or date range, the stated trigger, the actors involved, any electoral or policy target, and the supporting sources.

We graded the strength of the evidence in tiers. The strongest rating was reserved for episodes substantiated by quantitative administrative records such as response times or arrest and citation counts. Episodes supported only by news reporting or official acknowledgment received lower ratings. We also distinguished episodes in which a slowdown materialized from those in which one was only threatened, since threatened and realized actions carry different theoretical weight.

Two limitations follow from this design. First, recall is bounded by media coverage and by what the search surfaces, so the absence of a documented slowdown in a given city is

not evidence that none occurred. Second, language-model-assisted search is not perfectly reproducible. Results depend on the model version and on the web content available at the time of the search, and the underlying coverage itself changes as articles are revised or removed. For this reason we retain the sources for every coded episode, so that the database can be audited and extended independently of the search procedure that generated it.

B.1 Search Instructions Issued to the Language Model

The instruction was applied to each city independently; the city list and minor wording varied across batches.

```
Conduct a thorough, city-by-city search for documented cases of police
slowdowns, "blue flu," or police shirking in the following cities since
2010: [CITY LIST]. Search each city independently.

Look for incidents related to: budget cuts, police reform proposals,
officer discipline/prosecution, contract negotiations, controversial
policies.

EVIDENCE CLASSIFICATION STANDARDS:
- STRONGEST: Multiple sources + quantitative data (arrest/citation
  statistics, attendance records) + official confirmation
- STRONG: Multiple credible sources + clear documentation of operational
  changes
- DOCUMENTED: Credible reporting with specific details and dates
- THREATENED/UNCLEAR: Exclude these - only report materialized cases

FOR EACH CONFIRMED INCIDENT, COLLECT:
1. City and State
```

2. Year_Start and Year_End
3. Month_Start and Month_End
4. Reason_For_Shirk (what triggered it - e.g., specific policy, budget cut, officer prosecution)
5. Electoral_Target (who was being pressured - mayor, DA, city council, etc.)
6. Evidence_Type (what kind of shirking - sick-out, statistical decline in enforcement, service delays, etc.)
7. Evidence_Quality (STRONGEST/STRONG/DOCUMENTED following standards above)
8. Key_Details (specific numbers, quotes, timeline details)
9. Scholarly_Evidence (any academic studies documenting the case)
10. News_Sources (full URLs to all source articles)
11. Classification_notes (Strategic Resistance / Threatened Resistance / other relevant notes)

OUTPUT FORMAT:
Provide results in a structured table matching the exact column format of the example database provided, ready to be imported into Google Sheets. Include all source URLs for verification. If no confirmed cases are found for a city, explicitly state "No confirmed cases found" with a brief note about what was searched.

C Descriptive Survey (Pilot Study)

We conducted a pilot survey between September 6 and 12th, 2024 through the survey vendor Verasight. Data are weighted to match the July 2024 Current Population Survey on age,

race/ethnicity, sex, income, education, region, and metropolitan status, as well as partisanship benchmarks from the Pew Research Center NPORS surveys.

Figure A3: Shirking Experience Among Pilot Respondents

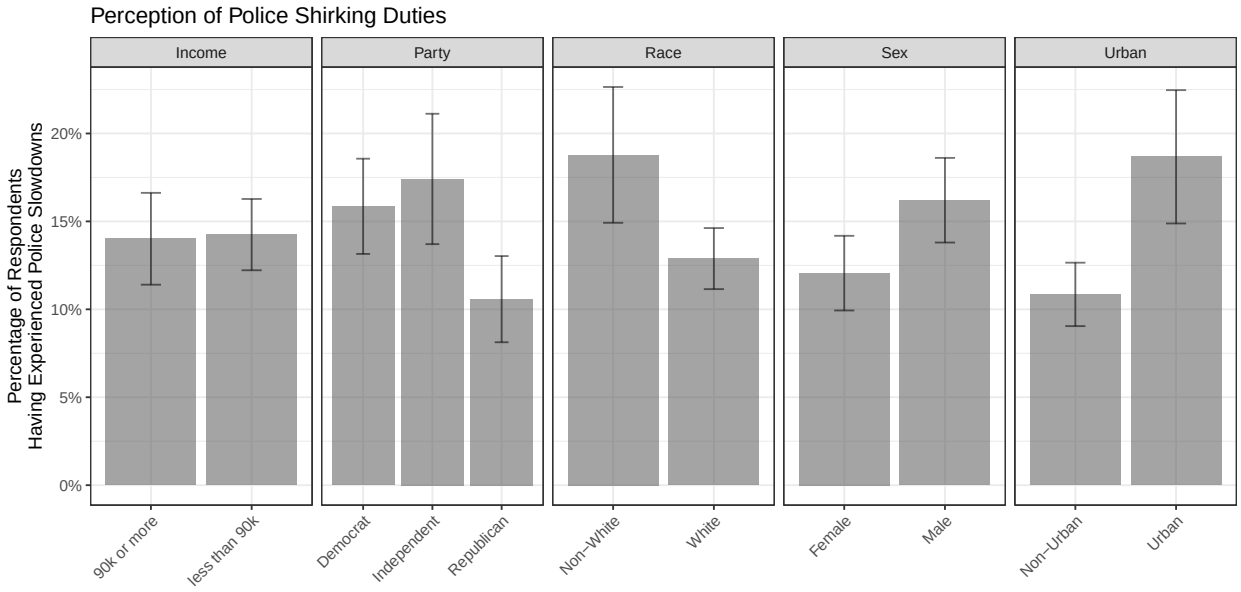
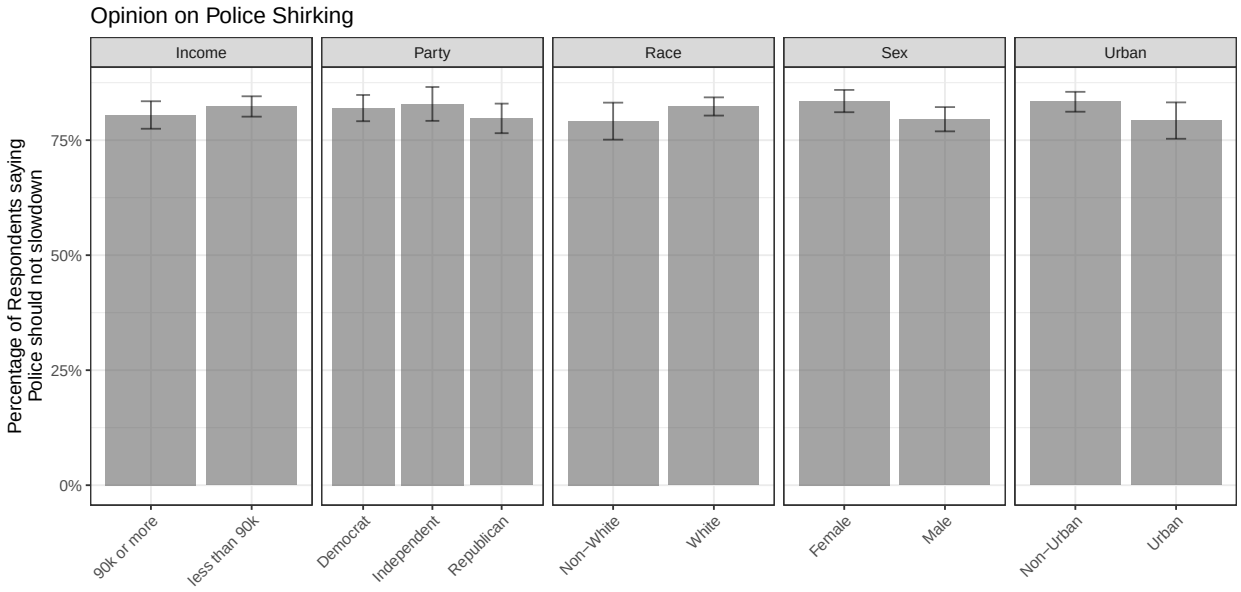


Figure A4: Opinion on Police Slowdowns



D Study 1

D.1 Mapping of Hypotheses to Outcomes

Table A3: Mapping of Hypotheses to Survey Measures and Empirical Models

H	**Concept**	**Survey Question**	**Variable Name**	**When Measured**	**Sample**
H1	Expected resistance	"How likely is it that police will reduce their effort?" (1–5)	`ExpectedResistance`	After fearmongering, before service	20% subsample
H1	Expected service quality	"How do you expect public safety to change?" (1–5)	Δ`PerceivedService`	After fearmongering, before service	Full sample
H2a	Policy support	"How much do you support [policy]?" (1–5)	Δ`PolicySupport`	Baseline & change scores	Full sample
H2b	Incumbent support	"How would you grade the City Council?" (A–F)	Δ`CouncilGrade`	Baseline & change scores	Full sample
H3	Reduced informativeness	Same as H2a/H2b	Interaction term	Change scores	Full sample
H4/H5	Prior heterogeneity	Same as H2a plus blame items	Subgroup analyses	Change scores	Full sample
H5	Blame attribution	"How likely is it that police response caused worse public safety?" (1–5)	`BlamePolice`; `BlamePolicy`	After both treatments	Poor service only

D.2 Descriptive Statistics

Table A4: Study 1 Descriptive Statistics

Variable	**N**	**Mean / Percent**	**SD**	**Median**
Duration (seconds)	1817	762.77	1755.22	562
Education (7-pt)	1794	4.31	1.35	5
Income (7-pt)	1786	3.38	1.64	3
Ideology (7-pt)	1789	3.80	–	4
Latino	1780	10.28%		0
White	1676	74.76%		1
Black	1676	13.48%		0
Democrat	1757	45.3%		0
Republican	1757	41.2%		0
Independent	1757	13.5%		0

Figure A5: Distribution of Grades for Local Institutions

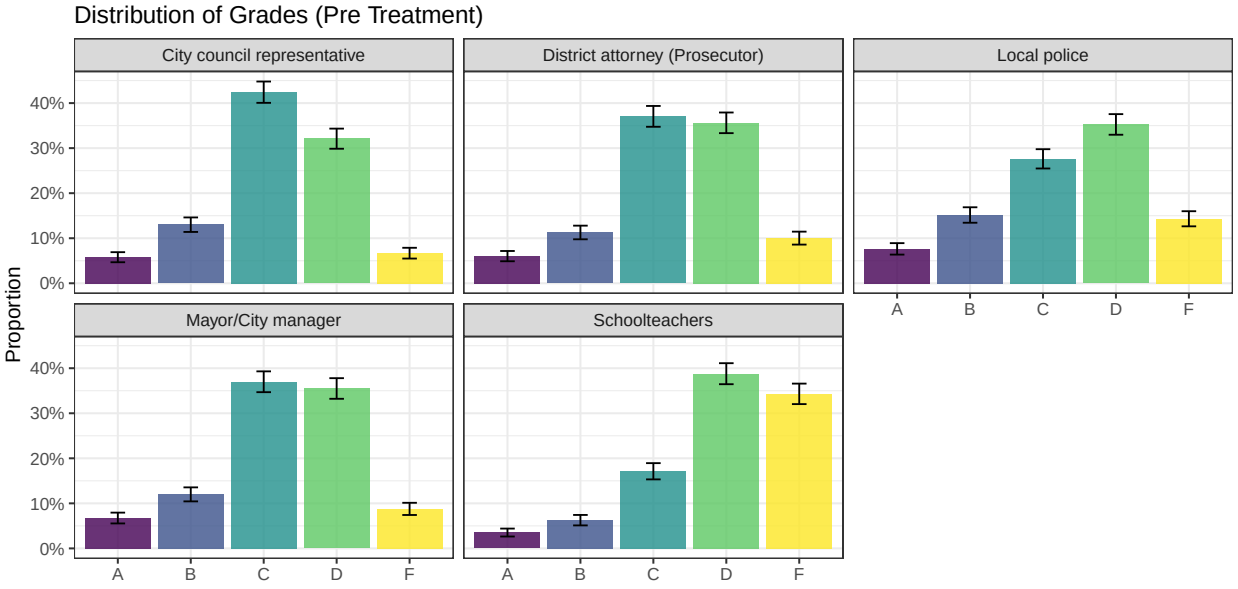
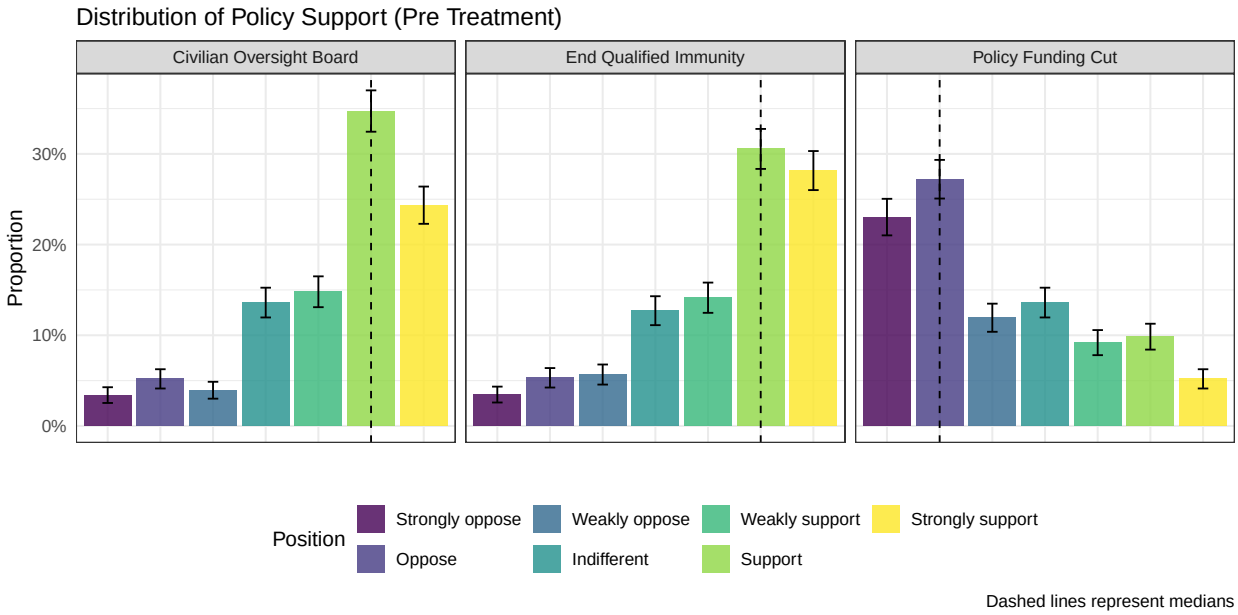


Figure A6: Distribution of Policy Positions



D.3 Additional Experimental Results

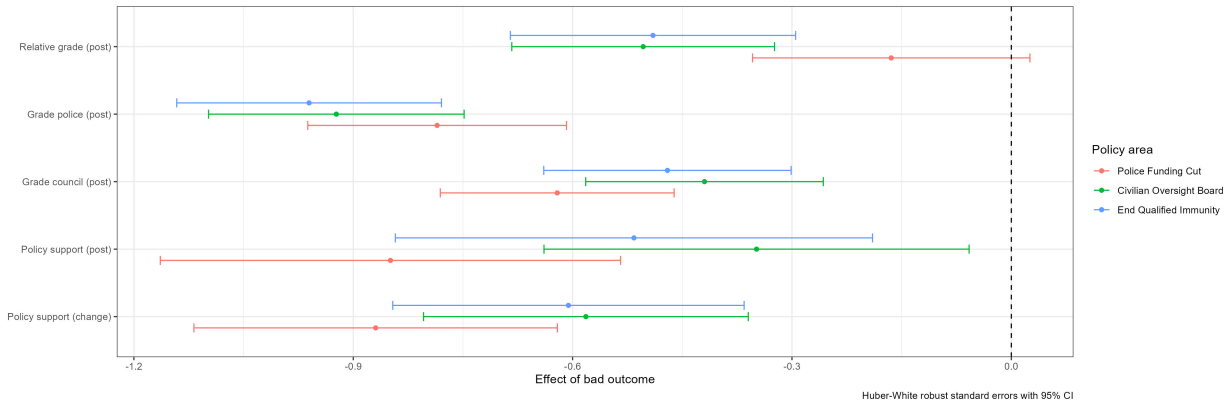


Figure A7: H1 Results by Policy Type

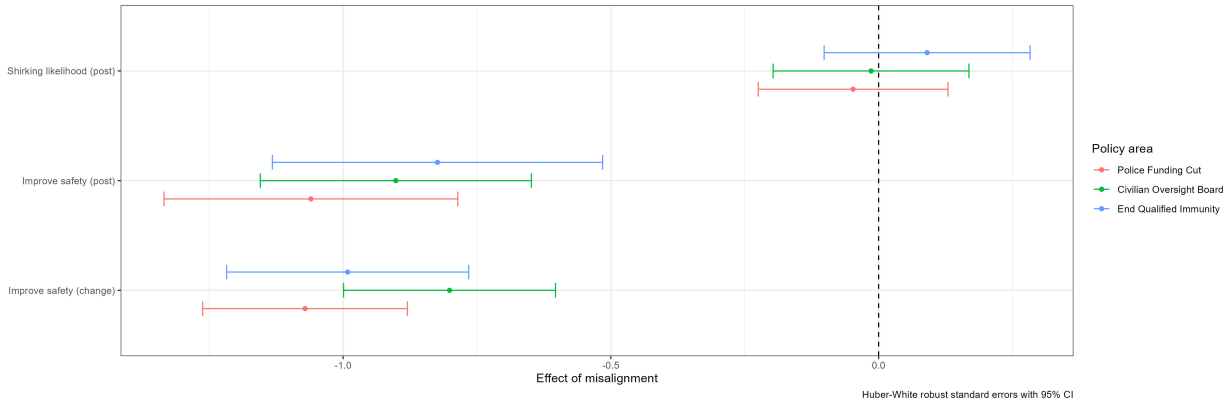


Figure A8: H2 Results by Policy Type

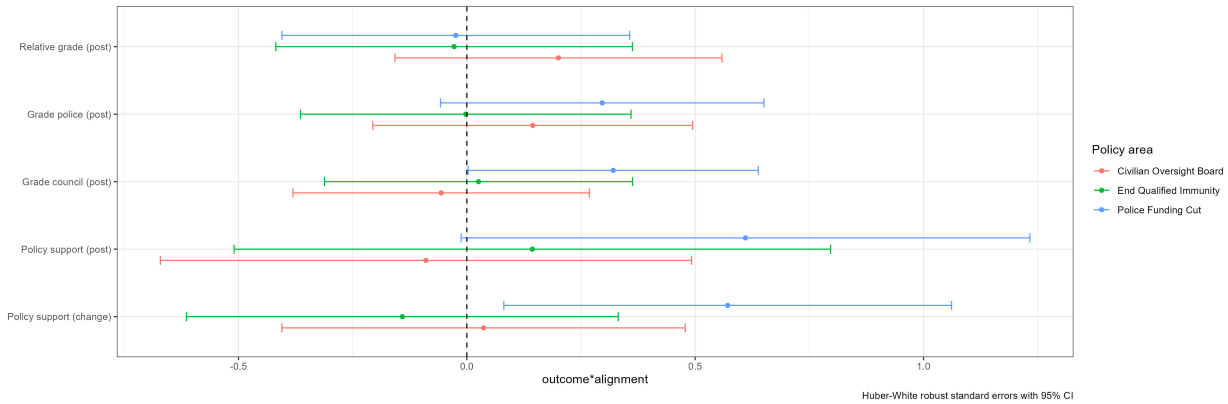


Figure A9: H3 Results by Policy Type

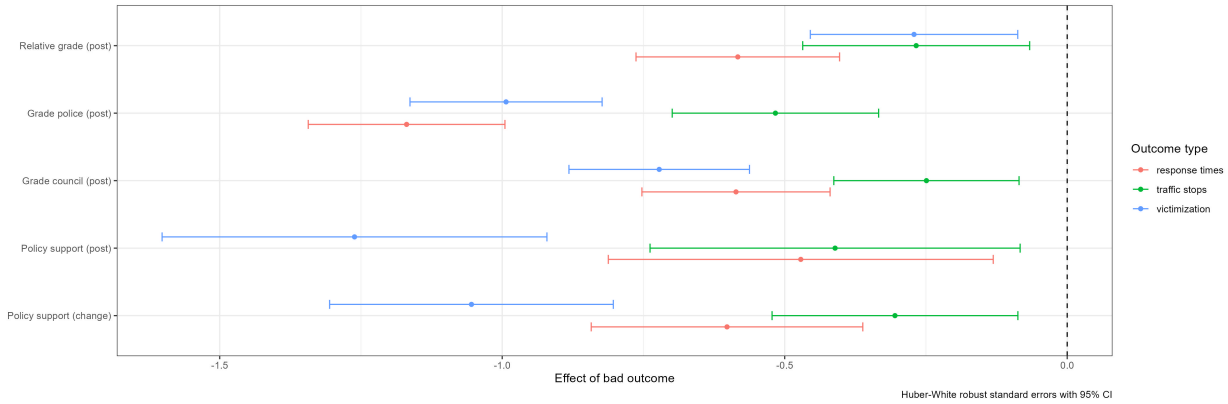


Figure A10: H1 Results by Outcome Type

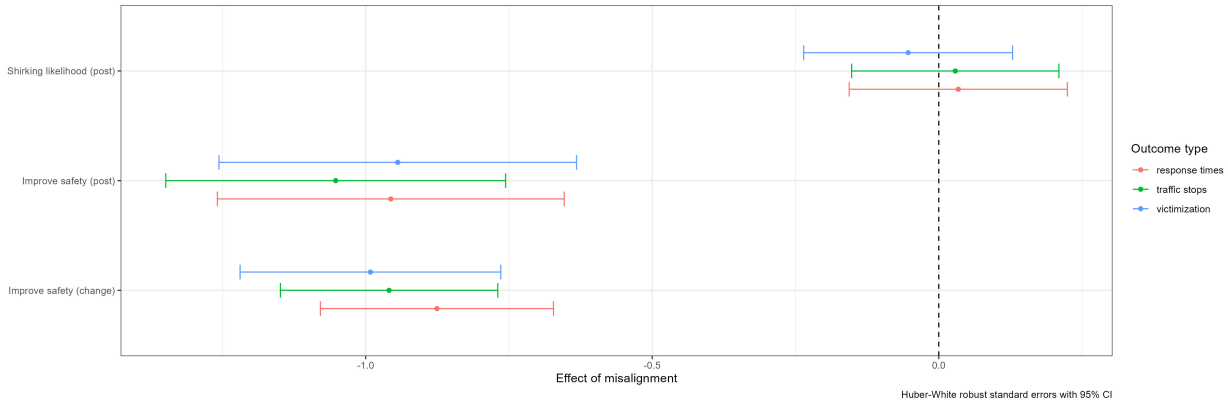


Figure A11: H2 Results by Outcome Type

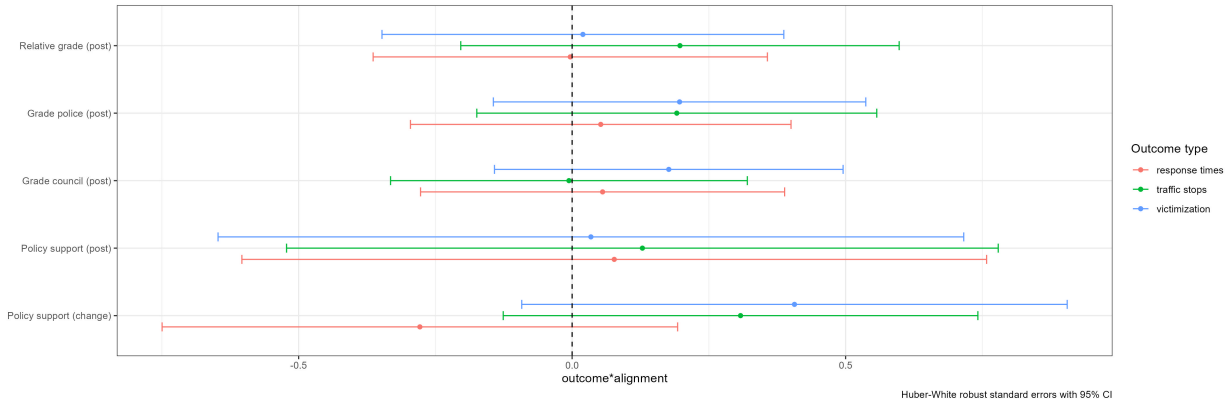


Figure A12: H3 Results by Outcome Type

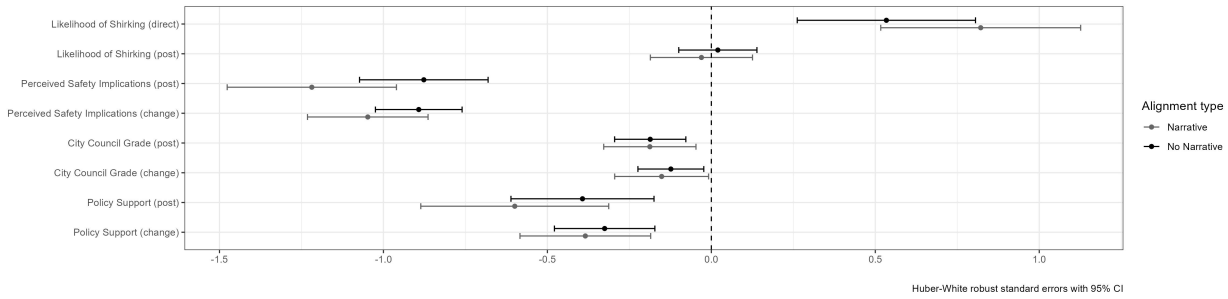


Figure A13: Effect Heterogeneity With and Without Added Narrative

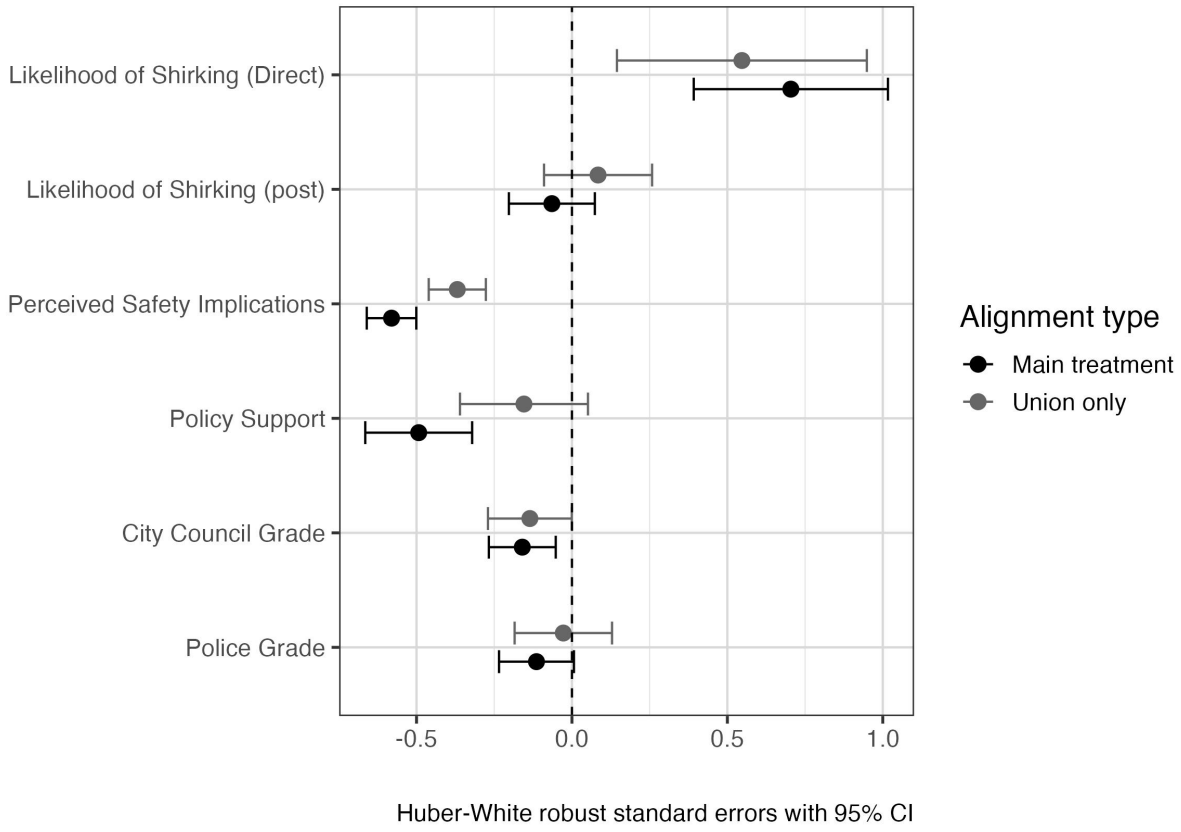


Figure A14: Effect Comparison: Main Fearmongering Treatment vs. Union Signal Only

D.4 Effects by Partisanship

Among Republican respondents, fearmongering is associated with a positive and statistically nonzero shift in police evaluations, suggesting Republicans may view misaligned police favorably as a check on the political establishment. Democrats and Liberals exhibit negative effects of fearmongering on police grades, though these are not statistically distinguishable from zero. For all subgroups, fearmongering produces a negative (if not always significant) shift in city council evaluations, suggesting that institutional conflict erodes perceived council effectiveness across partisan lines.

Figure A15: Effect of Fearmongering by Partisanship

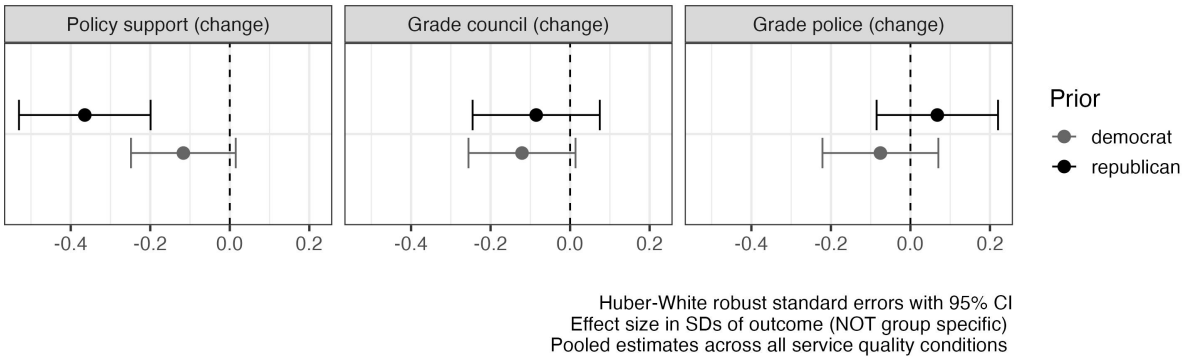
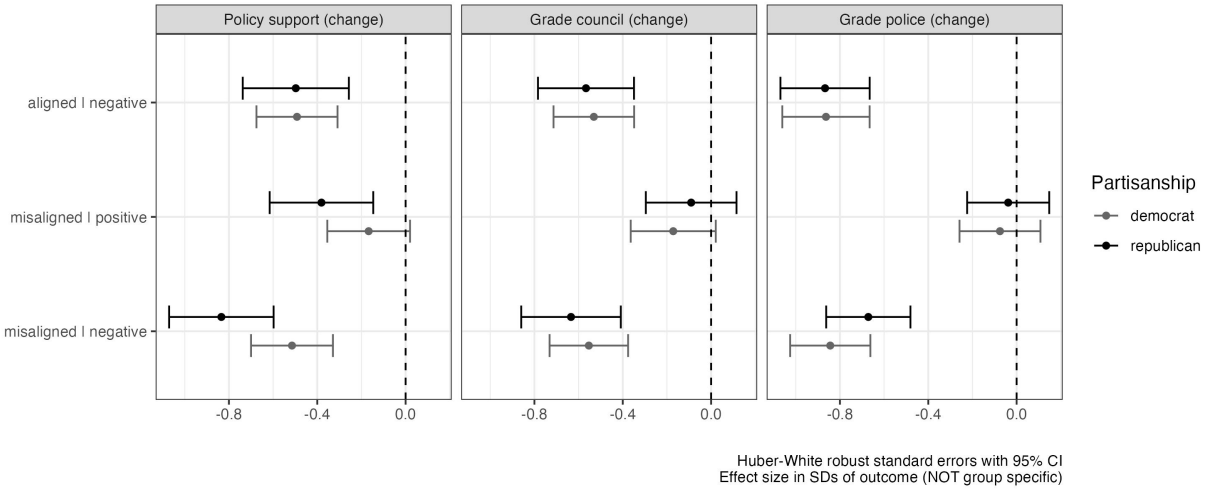


Figure A16: Conditional Treatment Effects by Partisanship



D.5 Embedding Regressions and SAE Projections

Table A5: Embedding Regression and SAE Projection: Fearmongering Treatment Effect on Open-Ended Responses

Panel A: Global treatment effect			
Bias-corrected squared norm $\|\hat{\beta}\|^2_{\text{cor}}$	0.0018		
Permutation p-value (5,000 draws)	0.007		
Panel B: SAE feature projections (top three by r_j)			
Feature	α_j	(95% CI)	r_j
Policy worked / showed improvements	−0.035	[−0.056, −0.014]	0.278
Police shirking / not performing duties	0.022	[−0.002, 0.046]	0.110
Uncertainty / inability to articulate	0.012	[−0.009, 0.032]	0.030

Note. $N = 1{,}088$ (533 misaligned, 555 aligned). 95% CIs for α_j from leave-one-out jackknife.

Table A6: Embedding Regression and SAE Projection: Service Quality Treatment Effect on Open-Ended Responses

Panel A: Global treatment effect			
Bias-corrected squared norm $\|\hat{\beta}\|^2_{\text{cor}}$	0.0165		
Permutation p-value (5,000 draws)	< 0.001		
Panel B: SAE feature projections (top five by r_j)			
Feature	α_j	(95% CI)	r_j
Policy worked / showed improvements	−0.087	[−0.107, −0.066]	0.391
Police shirking / not performing duties	0.057	[0.034, 0.081]	0.171
Uncertainty / inability to articulate	0.038	[0.018, 0.059]	0.077
Civilian review boards / accountability	0.030	[0.013, 0.047]	0.047
Police response times to 911 calls	0.017	[−0.003, 0.036]	0.014

Note. $N = 1{,}088$ (551 bad-service outcome, 537 good-service outcome). $\hat{\beta} = \bar{e}_{\text{bad}} - \bar{e}_{\text{good}}$, so positive α_j indicates a shift toward feature j under bad service. 95% CIs for α_j from leave-one-out jackknife.

E Study 2

E.1 CivicPulse Experiment Details

The CivicPulse 2026 Police Reform Survey was administered to a national sample of local government officials. The qualified immunity experiment randomly assigned officials to a cooperative or threatening frame describing how local police might respond to QI reform. The pre-post design measures QI support before and after exposure to the frame, using change scores as the primary outcome. Robust HC2 standard errors are used throughout.

The blame attribution question asks: "If proactive policing declined in your jurisdiction following a reform like removing qualified immunity, who do you think your constituents would most likely blame?" Response options: local elected officials who adopted the policy; the law enforcement department/officers; both equally; neither; can't say.

This appendix documents the full set of analyses underlying Study 2: the experimental design recap, randomization balance and attrition, the complete set of treatment-effect estimates and effect sizes, all robustness checks (randomization inference, state clustering, covariate adjustment, multiple-comparison correction), the mechanism/mediation analysis, the full heterogeneity battery (including the police-performance-grade and pro-/anti-police interaction tests), the shift typology and transition matrices, the open-ended theme analysis, and an index of all generated figures. Unless noted, treatment = threatening union response (vs. cooperative), the outcome scale runs 1 (strongly oppose) to 5 (strongly support), standard errors are HC2-robust, and the analytic sample is the $N = 550$ respondents with valid pre- and post-treatment answers.

E.2 Full Treatment-Effect Estimates and Effect Sizes

Perceived officer resistance, in detail. Mean perceived resistance is 2.62 (cooperative) vs. 2.74 (threatening); pooled SD = 1.13, so the +0.124 difference is ≈ 0.11 SD. On the "likely to resist" threshold (somewhat/very/ extremely likely), the share rises from **48.4%**

Table A7: Composition of the respondent sample ($N = 833$ local elected officials)

Dimension	Category	% of sample
Government type	Municipality	83.8
	County	16.2
Party (incl. leaners)	Republican / Rep.-leaning	45.2
	Democrat / Dem.-leaning	32.9
	Pure independent / other	21.9
Gender	Man	68.1
	Woman	30.1
	Prefer to self-describe	1.7
Jurisdiction size	Small (<2,308 residents)	16.6
(2023 ACS population)	Medium (2,308–8,732)	34.6
	Large (>8,732)	48.9
2024 Trump vote share	<55%	44.0
(jurisdiction)	55–70%	36.1
	>70%	19.9
Racial/ethnic minority	Yes	12.3
Police union present	Yes	53.3
	No / Don't know	46.7

Percentages among non-missing responses for each item.

Table A8: Covariate balance across arms (full randomized sample, $N = 833$)

Covariate	Cooperative ($n = 420$)	Threatening ($n = 413$)
% racial/ethnic minority	10.8	13.7
% large jurisdiction (>8,732)	48.6	49.2
% high Trump vote (>70%)	22.3	17.5
Pre-treatment support (1–5, mean)	2.48	2.59

Note: No across-arm difference is statistically significant, as expected under random assignment.

Table A9: Response completeness and attrition by arm

Arm	n (assigned)	% pre complete	% pre & post complete
Cooperative	420	68.1	66.0
Threatening	413	71.4	66.1

Note: Completion is symmetric across arms. An attritor-vs-completer comparison finds no significant differences on minority status, jurisdiction size, Trump vote, or baseline attitude (all $p > 0.24$); see `QI_attrition_covariate_table.csv`.

Table A10: Causal effect of a threatening (vs. cooperative) union response

Outcome / specification	Estimate	Robust SE	95% CI	p	p_{FDR}
Reform support					
Shift (post − pre), OLS	−0.258	0.074	[−0.405, −0.112]	0.0006	0.002
Binary support, shift	−0.069	0.030	[−0.127, −0.011]	0.020	0.029
Perceived officer-resistance likelihood					
Resistance (1–5)	+0.124	0.097	[−0.066, 0.315]	0.200	0.240

Analytic $N = 550$ (resistance: $N = 542$). p_{FDR} is the Benjamini–Hochberg-adjusted p-value across all primary outcomes. Binary support = 1 if "somewhat" or "strongly support."

Table A11: Pre/post means and within-arm shifts

Arm	n	Mean pre	Mean post	Mean shift	Within-arm p
Cooperative	277	2.48	2.60	+0.116	0.037
Threatening	273	2.59	2.44	−0.143	0.005

Note: The arms move in *opposite* directions: cooperation nudges support up, threat pushes it down.

to 56.1% (+7.7 points). The effect is directionally consistent with the manipulation but underpowered ($p = 0.20$).

Blame attribution. The blame item does not move with treatment (Table A15): the distribution is nearly identical across arms, a precisely estimated null.

Table A12: Primary treatment effects with BH-FDR correction

Outcome	Estimate	SE	95% CI	p	p_{FDR}
Shift (post − pre)	−0.258	0.074	[−0.405, −0.112]	0.0006	0.0025
Post support (ANCOVA)	−0.236	0.070	[−0.373, −0.099]	0.0007	0.0038
Binary support (shift)	−0.069	0.030	[−0.127, −0.011]	0.020	0.029
Binary support (ANCOVA)	−0.063	0.027	[−0.117, −0.010]	0.020	0.029
Perceived resistance	+0.124	0.097	[−0.066, 0.315]	0.200	0.240

Note: Source: `QI_key_outcomes_with_FDR.csv`. All three support outcomes survive FDR adjustment; resistance does not.

Table A13: Effect sizes (Cohen's d), pooled SD = 0.88

Quantity	Cohen's d
Cooperative arm (within-arm shift)	+0.131
Threatening arm (within-arm shift)	−0.162
Treatment effect (threat − coop)	−0.293

Table A14: Binary support rates and pre→post transitions, by arm

Arm	% support pre	% support post	Crossed *to* support	Crossed *from* support
Cooperative	25.6	28.9	22	13
Threatening	27.5	23.8	11	21

Note: The threatening arm loses nearly twice as many supporters as it gains; the cooperative arm does the reverse.

Table A15: Blame attribution by arm ("Who would constituents blame for reduced policing?")

Response	Cooperative (%)	Threatening (%)
Local elected officials who adopted the policy	34.9	35.2
Both equally	31.0	31.4
The law-enforcement department/officers	17.2	17.2
Can't say	16.5	13.8
Neither	0.4	2.3

Table A16: Robustness of the support effect across specifications

Outcome	Specification	Estimate	SE	p
Shift	HC2 robust (baseline)	−0.258	0.074	0.0006
Shift	Randomization inference (4,000 perms)	−0.258	—	0.0003
Shift	State-clustered (CR2, 47 clusters)	−0.264	0.086	0.005
Shift	Covariate-adjusted[a]	−0.262	—	—
Post	ANCOVA, HC2 (baseline)	−0.236	0.070	0.0007
Post	ANCOVA, state-clustered (CR2)	−0.229	0.077	0.007
Post	Covariate-adjusted[a]	−0.247	—	—
Post (DIM)[b]	Randomization inference	−0.156	—	0.185
Resist	Randomization inference	+0.124	—	0.214
Resist	State-clustered (CR2)	+0.120	0.113	0.297

Notes: [a] Adjusted for government type, party, jurisdiction population, union presence, and state fixed effects; point estimates are essentially unchanged. (HC2 SEs are unstable with sparse state dummies; CR2-clustered inference above is the preferred test.) [b] Unconditional difference-in-means on post support (no pre control) is noisier than the ANCOVA/shift specifications, which is why conditioning on the baseline matters.

E.3 Mechanism: Resistance and Mediation

Officials who perceive higher officer resistance shift more against the reform (resistance coefficient -0.113, $p = 0.002$ in a shift model controlling for treatment and pre). However, the treatment itself does not significantly raise perceived resistance, so resistance cannot carry the support effect. A formal causal-mediation analysis confirms this (Table A17).

Table A17: Causal mediation: does perceived resistance transmit the threat effect?

Quantity	Estimate	95% CI	p
ACME (indirect, via resistance)	-0.009	$[-0.028, 0.004]$	0.20
ADE (direct)	-0.238	$[-0.378, -0.088]$	< 0.001
Total effect	-0.247	$[-0.386, -0.094]$	< 0.001
Proportion mediated	0.029	$[-0.018, 0.150]$	0.20

Note: Quasi-Bayesian CIs, 1,000 simulations; mediator = perceived resistance (1–5), $N = 540$. Only $\sim$3% of the total effect runs through perceived resistance, and that share is not significant. The mechanism is direct/political rather than perceptual.

E.4 Subgroup estimates

E.5 Formal interaction tests: police grade and pro-/anti-police orientation

Subgroup estimates differ partly because of differing precision (subgroup sizes vary), so we test moderation directly with treatment$\times$moderator interaction models on the support shift. Table A19 reports the interaction coefficient (the differential effect) and its p-value.

High-police-grade jurisdictions do not shift significantly more than low-grade ones (interaction $p = 0.77/0.85$); the larger, significant subgroup estimate for the high-grade group reflects its larger sample, not a stronger effect. *More pro-police (anti-reform) officials* do show a larger point-estimate shift (-0.35 vs. -0.20), consistent with the hypothesis that a pro-police threat resonates most with pro-police officials, but the interaction is not significant ($p = 0.39/0.52$)—suggestive but underpowered.

Table A18: Treatment effect on the support shift, by subgroup

Moderator	Subgroup	Estimate	95% CI	p
Party	Republicans (incl. leaners)	−0.325	[−0.556, −0.095]	0.006
	Democrats (incl. leaners)	−0.051	[−0.280, 0.179]	0.66
	Pure independents	−0.369	[−0.922, 0.184]	0.19
Government	County	−0.452	[−0.770, −0.134]	0.006
	Municipality	−0.221	[−0.385, −0.058]	0.008
Prior QI attitude	Initially opposed (pre 1–2)	−0.285	[−0.476, −0.093]	0.004
	Initially neutral (pre 3)	−0.188	[−0.446, 0.070]	0.15
	Initially supportive (pre 4–5)	−0.182	[−0.478, 0.114]	0.23
Police union	Present	−0.186	[−0.385, 0.013]	0.067
	Absent	−0.200	[−0.452, 0.052]	0.12
	Don't know	−0.518	[−0.954, −0.082]	0.020
Jurisdiction size	Medium (2,308–8,732)	−0.280	[−0.479, −0.081]	0.006
	Large (>8,732)	−0.234	[−0.482, 0.014]	0.065
	Small (<2,308)	−0.258	[−0.680, 0.164]	0.23
Police-perf. grade	High 2025 performance	−0.271	[−0.486, −0.056]	0.014
	Low 2025 performance	−0.215	[−0.517, 0.088]	0.16
Policy reformism	Anti-reform (more pro-police)	−0.350	[−0.572, −0.128]	0.002
	Pro-reform	−0.202	[−0.460, 0.057]	0.13
Staffing	About right	−0.310	[−0.533, −0.087]	0.007
	Too few officers	−0.293	[−0.563, −0.023]	0.034

Table A19: Treatment × moderator interaction tests (outcome: support shift)

Moderator	Form	Interaction est.	SE	p	N
Police-performance grade	Median split (High vs. Low)	−0.056	0.188	0.77	387
Police-performance grade	Continuous (centered)	+0.028	0.153	0.85	387
Policy reformism	Median split (Pro- vs. Anti-reform)	+0.148	0.173	0.39	408
Policy reformism	Continuous (centered)	+0.092	0.142	0.52	408
Pre-treatment QI support	Continuous (centered)	+0.033	0.053	0.53	550

Note: For policy reformism, the baseline category is anti-reform (more pro-police); the positive interaction means pro-reform officials shift *less* (a smaller negative effect). For pre-treatment QI support, the positive interaction means initial supporters shift slightly less than opponents. **None of the interactions is statistically significant**: although point estimates run in the intuitive direction (more pro-police and more reform-skeptical officials move more), the data cannot distinguish these moderated effects from the average effect.

E.6 Shift typology

Decomposing movement into threshold-crossing (changing the support/oppose side) vs. intensity shifts (moving within a side) shows that in the threatening arm, threshold losses exceed gains (21 crossed from support, 11 to support), while "no change" dominates both arms (≈66–68%). See `plot_QI_shift_typology.png` and `plot_QI_shift_groups.png`.

E.7 Open-ended responses: keyword themes

A free-text follow-up (`Qualified_LOE`) was answered by $n = 127$ respondents (15% of the sample; 62 cooperative, 65 threatening arm). We analyze it in two ways: a transparent keyword/theme dictionary, and a data-driven embedding regression. The two are complementary— the dictionary fixes the concepts *a priori* and counts them; the embedding regression asks, with no dictionary, whether the *language as a whole* differs by arm.

Each response was coded for ten themes via a regex dictionary (`civicpulse_QI_analysis.R`). Table A20 reports overall and by-arm prevalence. The dominant themes were *accountability* (37.8%), *community trust/relations* (19.7%), and *public-safety concerns* (18.1%). Directionally the threatening arm invoked accountability, public-safety, "support reform regardless," and political-pressure themes somewhat more, while the cooperative arm more often raised community trust, officer morale, and officer-resistance themes. With cells this small (most themes < 15 mentions per arm) none of these arm differences is statistically reliable and none survives FDR correction; they are descriptive only.

E.8 Open-ended responses: embedding regression

E.9 Open-ended responses: embedding regression

As a dictionary-free check, we run an *embedding regression* (Rodriguez, Spirling and Stewart, 2023). For each open-ended response to the qualified-immunity prompt (`Qualified_LOE`),

Table A20: Open-ended theme prevalence, overall and by treatment arm ($n = 127$).

Theme	Overall %	Cooperative %	Threatening %
Accountability	37.8	33.9	41.5
Community trust/relations	19.7	24.2	15.4
Public-safety concerns	18.1	14.5	21.5
Support reform regardless	8.7	3.2	13.8
Oppose/concerned about reform	8.7	9.7	7.7
Officer morale/retention	6.3	9.7	3.1
Officers would resist/retaliate	6.3	9.7	3.1
Political pressure/influence	6.3	3.2	9.2
Didn't affect my view	1.6	3.2	0.0
Need more information	0.0	0.0	0.0

we embed the text and form the treatment–control difference vector

$$\hat{\beta} \;=\; \bar{e}_{\text{threatening}} - \bar{e}_{\text{cooperative}},$$

where $\bar{e}_g$ is the mean embedding in arm g. We test the bias-corrected squared norm $\|\hat{\beta}\|^2_{\text{cor}} = \sum_k (\hat{\beta}_k^2 - \widehat{V}[\hat{\beta}_k])$ by permutation (5,000 label reshuffles, refitting $\hat{\beta}$ each time), following the bias correction of Spirling et al. (2024). The omnibus test asks whether the two arms differ in language at all, with no *a priori* concepts.

Text preprocessing. The two embedding models require different preprocessing. GloVe represents individual words, so we tokenize each response with `quanteda`, lowercase it, and remove punctuation, numbers, symbols, and English stopwords (the Snowball and SMART lists), then match the remaining tokens to the GloVe 6B vocabulary (99% of tokens match). Following the à la carte (ALC) approach, we represent each document as the average of its matched word vectors left-multiplied by a pre-trained transformation matrix A (the Khodak et al. matrix, estimated on a large general corpus and distributed with `conText`). The QI corpus ($n \approx 119$) is far too small to estimate A itself, which is exactly the setting ALC is built for. By contrast, `text-embedding-3-small` embeds whole passages, so we pass the raw response text to the API without tokenization or stopword removal. In both cases we

keep responses with more than five non-whitespace characters and a non-missing treatment assignment, leaving $n = 119$ for the OpenAI encoder; the GloVe sample is $n = 118$ after dropping one response with no in-vocabulary tokens.

We run the test under two embedding models: GloVe 6B 300d with the ALC transform (via `conText`) and OpenAI `text-embedding-3-small` (1536d).

Table A21: Embedding-regression omnibus test, qualified-immunity open-ended responses.

Model	Dim	$\|\hat{\beta}\|^2_{\text{raw}}$	$\|\hat{\beta}\|^2_{\text{cor}}$	p_{perm}
GloVe 6B 300d (ALC transform)	300	0.061	0.010	0.153
`text-embedding-3-small`	1536	0.023	0.001	0.338

The omnibus test is null under both models ($p = 0.153$ for GloVe, $p = 0.338$ for the OpenAI encoder). We find no robust evidence that the open-ended language differs between the threatening and cooperative arms. This is consistent with the closed-ended picture: treatment moved the support *scale*, but among the small set of substantive open-ended responses it left no detectable whole-text semantic signature at this sample size.

Directional vocabulary (descriptive). As a further descriptive read, we identify the GloVe content words whose direction best aligns with $\hat{\beta}$. The threatening arm leans toward reactive, emotional terms (*absurd*, *ridiculous*, *unfounded*, *wrongfully*, *blackmail*, *nonsense*) and away from constructive, institutional vocabulary (*community*, *issues*, *support*, *professional*, *implementation*, *education*, *legislation*). The two arms' group-mean nearest neighbors are nearly identical, however, reinforcing the null: both arms draw on the same core vocabulary, and the small $\hat{\beta}$ shift is not driven by distinct lexical choices. We report these patterns only to characterize the direction of $\hat{\beta}$, not as evidence of an effect.

Interpretable (SAE) companion. For interpretation, we also project $\hat{\beta}$ onto the decoder directions of a TopK sparse autoencoder ($M = 8$ features, $K = 2$ active) trained unsupervised on the response embeddings, reporting $\alpha_j = w_j^\top \hat{\beta}$ with leave-one-out jackknife standard errors (script `embedding_regression_sae.py`). Because the global test is null,

these projections are descriptive only and should not be read as treatment effects. The most $\hat{\beta}$-aligned feature—"discusses following departmental policy or training as a condition for qualified-immunity protection"—reaches $\alpha = 0.055$ ($z = 2.24$, $p = 0.025$), accounting for about 13% of $\|\hat{\beta}\|^2$; no other feature is distinguishable from zero, and none survives once the null omnibus result is taken into account.